ROGÉRIO GUERRA BORIN

Detecção de Atividade Vocal empregando Máquinas de Boltzmann Restritas

São Paulo
2016

ROGÉRIO GUERRA BORIN

Detecção de Atividade Vocal empregando Máquinas de Boltzmann Restritas

Dissertação apresentada à Escola Politécnica da Universidade de São Paulo para obtenção do título de Mestre em Ciências

São Paulo

2016

ROGÉRIO GUERRA BORIN

Detecção de Atividade Vocal empregando Máquinas de Boltzmann Restritas

Dissertação apresentada à Escola Politécnica da Universidade de São Paulo para obtenção do título de Mestre em Ciências

Área de Concentração: Sistemas Eletrônicos

Orientador: Prof. Dr. Magno T. M. Silva

São Paulo
2016

Este exemplar foi revisado e corrigido em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, _____ de _____ de _____

Assinatura do autor: _____

Assinatura do orientador: _____

Catalogação-na-publicação

Borin, Rogério Guerra
Detecção de Atividade Vocal empregando Máquinas de Boltzmann Restritas / R. G. Borin -- versão corr. -- São Paulo, 2016.
155 p.

Dissertação (Mestrado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Sistemas Eletrônicos.

1.Inteligência artificial 2.Processamento de som 3.Processamento de sinais 4.Telefonia I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Sistemas Eletrônicos II.t.

AGRADECIMENTOS

Gostaria de agradecer ao meu orientador, Prof. Magno T. M. Silva, pela orientação, dedicação e apoio constante na realização deste trabalho. Agradeço ainda pela amizade e confiança durante todo o processo de mestrado.

Ao Prof. Miguel Arjona Ramírez e à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), sob os processos 2012/24789-0 e 2015/25512-0, que viabilizaram a utilização da base de dados AURORA-2 neste trabalho.

Ao Prof. Emilio Del Moral Hernandez, pela leitura detalhada do texto de qualificação, bem como pelo apoio e pelas sugestões dadas no exame de qualificação.

Aos Profs. Romis Attux e Ricardo Suyama, pela revisão meticulosa deste trabalho e pelas valiosas observações e sugestões oferecidas na defesa da dissertação. Ao Prof. Ricardo Suyama agradeço também pelas ideias dadas durante a qualificação.

À toda minha família, em especial aos meus pais Nelson e Natália, pelo amor incondicional e apoio constante; à minha irmã, Renata, meu cunhado, José Antônio, e minha sobrinha, Sarah, por estarem sempre ao meu lado.

Ao colega Flávio R. M. Pavan, pela disponibilização do modelo de referência em LaTeX sobre o qual o presente texto foi redigido.

À Escola Politécnica da Universidade de São Paulo, pela oportunidade que tive de participar do programa de mestrado.

A Deus, por ter me dado saúde e força para realizar este trabalho.

RESUMO

Neste trabalho, uma versão de RBM (*Restricted Boltzmann Machine*) tendo uma camada de classificação é adaptada a fim de permitir o seu uso com dados definidos num domínio contínuo. Essa adaptação dá origem a uma variante do modelo para o qual são desenvolvidas as regras de atualização de parâmetros dos treinamentos discriminativo, generativo e híbrido. A aplicação da variante como classificador no problema de VAD (*Voice Activity Detection*) é então investigada. Por meio de simulações envolvendo o *corpus* NOIZEUS e empregando como entradas do classificador tanto MFCCs (*Mel-Frequency Cepstral Coefficients*) quanto FBEs (*Filter-Bank Energies*), são obtidos resultados comparáveis aos de detectores considerados como estado da arte, com um menor custo computacional. A variante de RBM é comparada também com as SVMs (*Support Vector Machines*) lineares e com núcleo gaussiano. Com treinamento discriminativo, a RBM fornece desempenhos intermediários entre as duas versões de SVM, porém um custo computacional que é consideravelmente inferior aos de ambas. Adicionalmente, um conjunto de medidas do áudio que tiveram seu uso em VAD proposto recentemente são avaliadas com o emprego da RBM com treinamento discriminativo. Embora os resultados não sejam conclusivos, os desempenhos conseguidos indicam que essas medidas não são vantajosas quando comparadas com os tradicionais MFCCs.

Palavras-chave: Inteligência artificial. Processamento de som. Processamento de sinais. Telefonia.

ABSTRACT

In this work, a type of Restricted Boltzmann Machine (RBM) having a classification layer is adapted to allow its use with data defined in a continuous domain. Such adaptation gives rise to a variant of the model for which the parameter update rules are developed for the discriminative, generative and hybrid types of training. The application of the variant as a classifier to the Voice Activity Detection (VAD) problem is then investigated. By means of simulations involving the *corpus* NOIZEUS and employing Mel-Frequency Cepstral Coefficients (MFCCs) or Filter-Bank Energies (FBEs) as classifier inputs, results comparable to those of state-of-the-art detectors are achieved with a lower computational cost. The RBM variant is also compared to the linear and Gaussian kernel Support Vector Machines (SVMs). With the discriminative training, the RBM provides intermediate performances between the two SVM types, but a computational cost that is considerably lower than theirs. Additionally, a set of measures from the audio whose application in VAD has been recently proposed are evaluated by employing the RBM with discriminative training. Although the results are not conclusive, the performances obtained indicate that the measures are not advantageous when compared to the traditional MFCCs.

Keywords: Artificial intelligence. Sound processing. Signal processing. Telephony.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AFE	*Advanced Front-End*
AIS	*Annealed Importance Sampling*
AMDF	*Average Magnitude Difference Function*
AMR	*Adaptive Multirate*
ANN	*Artificial Neural Network* (Rede Neuronal Artificial)
BA	*Balanced Accuracy* (Acurácia Balanceada)
CD	*Contrastive Divergence*
CPP	*Cepstral Peak Prominence* (Proeminência dos Picos Cepstrais)
CRF	*Conditional Random Field* (Campo Aleatório Condicional)
DBN	*Deep Belief Network* (Rede de Crença Profunda)
DCT	*Discrete Cosine Transform* (Transformada Discreta do Cosseno)
DFT	*Discrete Fourier Transform* (Transformada de Fourier Discreta)
DRBM	*Discriminative Restricted Boltzmann Machine* (RBM Discriminativa)
FBE	*Filter-Bank Energy* (Energia de Banco de Filtros)
GBRBM	*Gaussian-Bernoulli Restricted Boltzmann Machine* (RBM Gauss-Bernoulli)
GMM	*Gaussian Mixture Model* (Modelo de Mistura Gaussiana)
HDRBM	*Hybrid Discriminative Restricted Boltzmann Machine* (RBM híbrida)
HMM	*Hidden Markov Model* (Modelo Oculto de Markov)
HPS	*Harmonic Product Spectrum* (Espectro do Produto de Harmônicas)
ITU	*International Telecommunication Union*
LTSD	*Long-Term Spectral Divergence*
LTSV	*Long-Term Signal Variability*
LRT	*Likelihood-Ratio Test* (Teste de Razão de Verossimilhança)
MCMC	*Markov Chain Monte Carlo* (Monte Carlo via cadeias de Markov)

MFCC	*Mel-Frequency Cepstral Coefficient* (Coeficiente mel-cepstral)
MLP	*Multi-Layer Perceptron* (Perceptron de Múltiplas Camadas)
MRF	*Markov Random Field* (Campo Aleatório de Markov)
PoE	*Product of Experts* (Produto de Especialistas)
PCD	*Persistent Contrastive Divergence*
PT	*Parallel Tempering*
RBM	*Restricted Boltzmann Machine* (Máquina de Boltzmann Restrita)
ROC	*Receiver Operating Characteristic* (Característica de Operação do Receptor)
SC	*Seleção de Características*
SD	*Seleção de Dados*
SGD	*Stochastic Gradient Descent* (Gradiente Descendente Estocástico)
SNR	*Signal-to-Noise Ratio* (Relação Sinal-Ruído)
SRH	*Summation of the Residual Harmonics* (Somatório das Harmônicas do Resíduo)
SM	*Seleção de Modelo*
SVM	*Support Vector Machine* (Máquina de Vetor de Suporte)
VAD	*Voice Activity Detection* (Detecção de Atividade Vocal)
VAD	*Voice Activity Detector* (Detector de Atividade Vocal)

SUMÁRIO

1 INTRODUÇÃO

Neste capítulo, o presente trabalho é contextualizado. Inicialmente, é fornecida uma breve perspectiva histórica incluindo a Máquina de Boltzmann Restrita (RBM – *Restricted Boltzmann Machine*), que é o principal objeto de estudo deste trabalho. Em seguida é sucintamente apresentado o problema que é aqui abordado: a detecção de atividade vocal empregando RBMs. Estabelecido, assim, o contexto do trabalho, são então declarados os seus objetivos com suas respectivas justificativas. Finalmente, são apresentadas as contribuições e a organização da dissertação, a metodologia empregada na pesquisa e uma descrição da notação usada no texto.

1.1 PERSPECTIVA HISTÓRICA

Nos últimos anos, um ramo da aprendizagem de máquina que vem ganhando importância é a chamada aprendizagem profunda (*Deep Learning*), cuja principal característica é a modelagem de dados em múltiplos níveis de abstração. Nessa modelagem, são tipicamente utilizadas redes neuronais artificiais (ANNs – *Artificial Neural Networks*) e, por isso, fala-se que a aprendizagem profunda seria o renascimento dessas redes. Durante a década de 1990, o interesse em ANNs diminuiu em favor de modelos mais simples e cujo treinamento era mais eficiente como, por exemplo, as Máquinas de Vetor de Suporte (SVMs – *Support Vector Machines*) e os Campos Aleatórios Condicionais (CRFs – *Conditional Random Fields*) (DENG; YU, 2014). Entretanto, Hinton, Osindero e Teh (2006) introduziram um método de pré-treinamento para o Perceptron de Múltiplas Camadas (MLP – *Multi-Layer Perceptron*), um tipo de ANN, baseado numa estrutura denominada Rede de Crença Profunda (DBN – *Deep Belief Network*). Com esse método, a dificuldade no treinamento foi aliviada consideravelmente, trazendo de volta o interesse da comunidade científica pelas ANNs. Na última década, ANNs com múltiplas camadas têm se destacado ao obter resultados de estado da arte em várias tarefas nas áreas de reconhecimento de fala e de imagem, visão computacional, processamento de linguagem natural, dentre outras.

No algoritmo proposto por Hinton, Osindero e Teh (2006), uma DBN é treinada camada por camada. Cada camada é de fato uma RBM, cujo treinamento pode ser realizado de forma eficiente usando-se, por exemplo, o algoritmo CD (*Contrastive Divergence*) (HINTON, 2002). Destaca-se, assim, que as RBMs são utilizadas efetivamente como blocos construtivos de estruturas maiores e que a eficiência do seu treinamento foi um fator de considerável importância para o ressurgimento das ANNs e, de certa forma, para o surgimento do ramo da aprendizagem profunda.

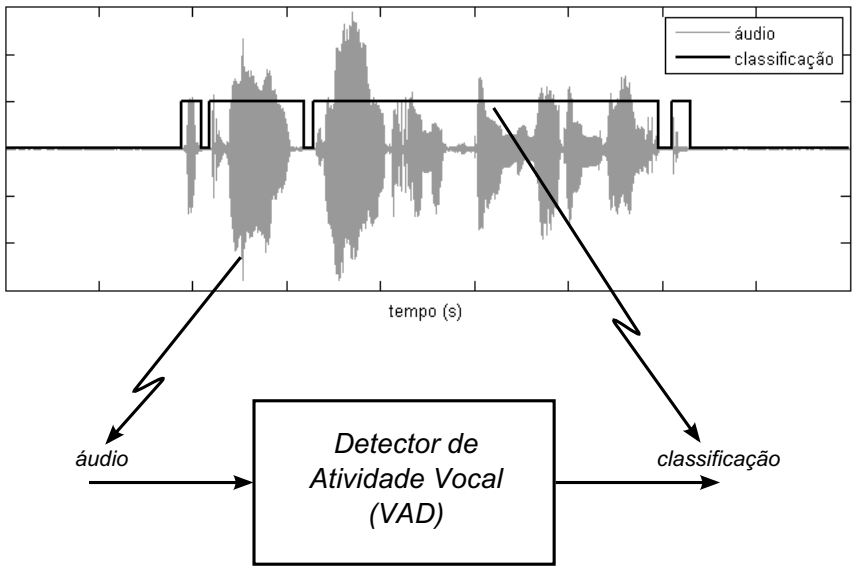
O treinamento de RBMs é normalmente feito de forma não supervisionada, isto é,

sem que os dados ou exemplos de treinamento tenham sido previamente rotulados. Com efeito, o modelo aprende a distribuição de probabilidade desses exemplos. Entretanto, Larochelle e Bengio (2008) mostraram que as RBMs poderiam ser usadas de forma bem sucedida como classificadores isolados (isto é, não fazendo parte de uma estrutura maior) sendo, para tanto, treinadas de forma supervisionada. Especificamente, os autores avaliaram RBMs obtidas com diferentes tipos de treinamento em duas tarefas: reconhecimento de dígitos escritos a mão e classificação de documentos. Em ambos os casos, as RBMs atingiram desempenhos comparáveis aos de modelos mais complexos.

1.2 DETECÇÃO DE ATIVIDADE VOCAL

A detecção de atividade vocal (VAD – *Voice Activity Detection*) consiste em se identificar em um áudio, os trechos contendo voz humana. A Figura 1 ilustra o trabalho realizado por um VAD. Na entrada do detector, tem-se as amostras do áudio, enquanto, na saída, o detector comumente produz um valor binário indicando a ausência ou presença de voz, como mostrado na figura. Dentre outros usos, esses detectores são de considerável importância na indústria de telefonia, pois permitem uma significativa redução na largura de banda utilizada por comunicações de voz. Vale dizer que, apesar da ideia aparentemente simples, a diversidade dos ruídos possivelmente presentes no áudio unida à ampla variabilidade da voz humana tornam a detecção de presença de voz em meio ao ruído uma tarefa difícil. De fato, muito embora o problema de VAD tenha sido estudado por muitos anos, considera-se que os detectores de estado da arte atuais ainda não produzem resultados satisfatórios, especialmente em relações sinal-ruído baixas (ZOU et al., 2014).

Figura 1 – Detector de atividade vocal.



Fonte: produção do próprio autor.

Como será visto na Seção 2.4, na abordagem de aprendizagem de máquina, o áudio é tipicamente dividido em curtos segmentos, dos quais são extraídas medidas ou características. Estas fornecem uma representação sucinta das amostras no segmento. As características alimentam então um classificador que fornece a saída do VAD para cada segmento.

1.3 OBJETIVOS E JUSTIFICATIVA

Os principais objetivos do presente trabalho são enumerados a seguir:

1. **Adaptação do modelo a entradas contínuas**: Larochelle e Bengio (2008) apresentaram uma versão de RBM cujo treinamento seria feito de forma supervisionada. Analisando-se a literatura posterior, notou-se que a versão descrita pelos autores não foi efetivamente explorada. Uma dificuldade com essa versão de RBM é que os dados de entrada a serem classificados devem ser binários. Tendo em vista os bons resultados obtidos pelos autores nas tarefas de classificação de documentos de texto e no reconhecimento de imagens de dígitos escritos a mão, veio a ideia de adaptar o modelo por eles apresentado para permitir entradas definidas num domínio contínuo (podendo assumir valores reais) e, assim, obter uma nova versão (ou variante) de RBM de uso mais geral.

2. **Aplicação do modelo proposto ao problema de VAD empregando medidas clássicas**: A fim de avaliar o modelo proposto em um caso real, escolheu-se aplicá-lo a um problema reconhecidamente difícil, como é o caso da detecção de atividade vocal. Numa primeira abordagem, são consideradas medidas tradicionalmente utilizadas para representação do áudio (ambas descritas na Seção 2.4): os coeficientes mel-cepstrais (MFCCs – *Mel-frequency cepstral coefficients*) e um subproduto do cálculo dos mesmos, as energias de banco de filtros (FBEs – *Filter-bank energies*). Deseja-se responder às seguintes perguntas:

   a) Como um classificador baseado no modelo proposto lida com medidas não (ou pouco) correlacionadas, como é o caso dos MFCCs, e com medidas fortemente correlacionadas, como as FBEs?

   b) Como um VAD baseado no modelo proposto se compara com outros VADs conhecidos na literatura?

   c) Como um VAD baseado no modelo proposto se compara com um baseado em SVM (outro modelo empregado em aprendizagem de máquina)?

3. **Aplicação do modelo proposto ao problema de VAD empregando medidas propostas recentemente**: Drugman et al. (2015) propuseram o uso de um conjunto de medidas para representação do áudio que, segundo os autores, produziram um

VAD de estado da arte. Deseja-se aqui empregar as referidas medidas com o modelo proposto para eventualmente confirmar o potencial das mesmas na aplicação de VAD.

1.4 CONTRIBUIÇÕES DA DISSERTAÇÃO

As principais contribuições do presente trabalho são:

1. **Introdução de uma variante de RBM**.
   Uma versão de RBM tendo uma camada de classificação e capaz de lidar com dados de entrada binários foi adaptada, neste trabalho, a fim de permitir o seu uso com dados definidos num domínio contínuo. Para essa nova variante de RBM, foram desenvolvidas as regras de atualização para o seu treinamento e as equações que permitem a realização de inferência, amostragem e classificação com o modelo, as quais são apresentadas nesta dissertação.
2. **Avaliação da variante no problema de VAD**.
   A nova variante foi aplicada ao problema de VAD e comparada a detectores adotados pela indústria e a outros conhecidos na literatura, incluindo três detectores que são considerados como estado da arte. A variante foi comparada também com as SVMs lineares e com núcleo gaussiano. Foi constatado que a variante proposta tem vantagens em relação aos VADs de estado da arte e também às SVMs.

Como resultado dessas contribuições, foi publicado o seguinte artigo:

- BORIN, Rogério G.; SILVA, Magno Teófilo Madeira da. Detecção de atividade vocal com o uso de máquinas de Boltzmann restritas discriminativas. In: **XXXIV Simpósio Brasileiro de Telecomunicações e Processamento de Sinais**, 2016, Santarém. Anais do SBrT'2016, 2016. p. 75-79.

Outras contribuições, que podem ser consideradas secundárias, são enumeradas a seguir:

1. **Comparação da RBM Gauss-Bernoulli com GMM**.
   A RBM Gauss-Bernoulli (uma das variantes de RBM apresentadas neste texto) foi comparada com o modelo denominado *Gaussian Mixture Model* (GMM). Foi mostrado que esse tipo de RBM pode ser visto como uma mistura de gaussianas independentes.

2. **Interpretação para algoritmos de treinamento**.
   Neste trabalho, é dada um interpretação visual para o processo conhecido como *geração de partículas negativas*, que faz parte dos principais algoritmos de treinamento

generativo de RBM. Acredita-se que tal interpretação auxilie no entendimento das similaridades e diferenças entre os algoritmos.

3. **Apresentação de regras de atualização**.
   Na literatura tomada como referência para o presente trabalho, as regras de atualização de parâmetros dos diferentes algoritmos de treinamento são apresentadas em termos de esperanças de variáveis aleatórias (calculadas nas distribuições dos dados de treinamento e do modelo). Embora essa forma de apresentação seja bastante sucinta, ela não permite uma implementação direta dos algoritmos. Neste texto, todas as regras de atualização são fornecidas numa forma diretamente implementável.

4. **Determinação das distribuições das medidas recentes**.
   Foram determinadas, sobre o *corpus* NOIZEUS (HU; LOIZOU, 2008), as distribuições empíricas das medidas propostas recentemente, nas condições de ausência e de presença de voz. Para cada medida, foram também obtidas as distâncias de Bhattacharyya entre as distribuições nas duas condições.

5. **Derivação das principais equações das RBMs**.
   Nos apêndices desta dissertação, as principais equações de duas versões de RBM são derivadas bem como todas as regras de atualização e equações relativas à variante proposta.

1.5 ORGANIZAÇÃO DA DISSERTAÇÃO

Em linhas gerais, o conteúdo da dissertação está dividido em três partes. A primeira delas fornece o arcabouço teórico associado ao projeto de pesquisa e é composta pelos Capítulos 2 e 3. Destes, o primeiro é baseado predominantemente na literatura existente, enquanto que o segundo concentra as contribuições teóricas do presente trabalho. A segunda parte da dissertação é dada no Capítulo 4 e tem foco majoritariamente experimental. A terceira parte, no Capítulo 5, contém as conclusões gerais do trabalho. A seguir os capítulos mencionados são melhor descritos.

O Capítulo 2 apresenta e discute em detalhes três versões de RBM – a RBM original e duas variantes. De fato, a versão original serve de base para as referidas variantes, as quais embasam uma nova variante introduzida como parte desta dissertação. No mesmo capítulo, são detalhados o problema de detecção de atividade vocal e a técnica de aprendizagem de máquina pela qual o problema é abordado neste trabalho. Resultados de simulações são também mostrados a fim de ilustrar alguns tópicos da teoria exposta no texto.

No Capítulo 3, é apresentada a versão de RBM efetivamente proposta nesta dissertação. Para essa versão, são desenvolvidas as regras de atualização de parâmetros que permitem o treinamento do modelo e as equações que possibilitam seu uso como

classificador. Por meio de simulações, a variante proposta é analisada num problema clássico de classificação.

No Capítulo 4, o modelo proposto é avaliado na tarefa de VAD mediante uma variedade de experimentos computacionais. Os detectores baseados no modelo são comparados com outros existentes na literatura e também com outro mecanismo reconhecido de aprendizagem de máquina (SVM). Adicionalmente, a variante proposta é empregada na avaliação de certas medidas extraídas do áudio cujo uso em VAD foi proposto recentemente.

Por fim, no Capítulo 5, são apresentadas as conclusões do trabalho e as propostas para trabalhos futuros.

1.6 METODOLOGIA

A metodologia empregada no trabalho consistiu de levantamentos e estudos bibliográficos, desenvolvimentos teóricos, implementações de programas e simulações computacionais. Os algoritmos avaliados foram implementados e executados no MATLAB versão 7.11.

1.7 NOTAÇÃO

Neste texto, escalares são representados por letras maiúsculas ou minúsculas em itálico como, por exemplo, Z, n_h e $\epsilon(\mathbf{x})$. Vetores coluna são indicados por letras minúsculas em negrito, como $\mathbf{b}$ e $\boldsymbol{\theta}$. Matrizes são indicadas por letras maiúsculas em negrito, assim como vetores aleatórios. Em especial, o estado de um vetor aleatório representado por uma certa letra maiúscula, por exemplo, $\mathbf{X}$, é indicado pela letra minúscula correspondente, $\mathbf{x}$, neste caso.

Toda função representando uma distribuição de probabilidade (seja seu domínio discreto, contínuo ou misto), é indicada pela letra P, possivelmente com um índice subscrito, como $P_g(\mathbf{v})$. Mais ainda, os parâmetros da função indicam a quais vetores aleatórios (ou variáveis aleatórias) a mesma se refere. Assim, $P_g(\mathbf{v})$ está associada ao vetor aleatório $\mathbf{V}$. As notações $P(\mathbf{x}|\mathbf{y})$ e $P(\mathbf{x}|\mathbf{Y}{=}\mathbf{y})$ representam a distribuição condicional de um certo vetor aleatório $\mathbf{X}$ conhecido o fato de que $\mathbf{Y}{=}\mathbf{y}$. A segunda notação é usada quando houver ambiguidade sobre o vetor aleatório sendo referenciado. Notações análogas são usadas com variáveis aleatórias (escalares).

O operador esperança matemática é denotado por $\mathbb{E}[\cdot]$ e o operador variância por $\mathbf{VAR}[\cdot]$. A esperança calculada sobre uma distribuição $P(\cdot)$ específica é indicada como $\mathbb{E}_{P(\cdot)}[\cdot]$. O caso especial da esperança condicional é mais facilmente explicado com um exemplo: $\mathbb{E}_{P(\cdot|\mathbf{y})}[\cdot|\overline{\mathbf{y}}]$ representa a esperança da expressão à esquerda da barra entre colchetes calculada sobre a distribuição $P(\cdot|\mathbf{Y}{=}\overline{\mathbf{y}})$.

2 REFERENCIAL TEÓRICO

Neste capítulo, é fornecido o alicerce teórico sobre o qual o projeto de pesquisa é desenvolvido. Na Seção 2.1, é introduzida a RBM Bernoulli-Bernoulli, que é a base para duas variantes de RBMs – a RBM Gauss-Bernoulli e a RBM com uma camada de classificação, as quais são discutidas nas Seções 2.2 e 2.3, respectivamente. Essas duas variantes são a fundação para uma nova variante proposta no presente trabalho e que será apresentada no Capítulo 3. A pesquisa envolve o uso dessa nova variante na detecção de atividade vocal, aplicação que é discutida na Seção 2.4.

2.1 RBM BERNOULLI-BERNOULLI

Uma RBM é uma rede neuronal estocástica capaz de reproduzir uma distribuição de probabilidade. Essa estrutura foi introduzida por Smolensky (1986) com o nome de *Harmonium*. Entretanto, a comunidade científica passou a dar maior atenção ao seu estudo somente anos mais tarde quando Hinton (2002) apresentou um algoritmo rápido para seu treinamento. De acordo com Desjardins et al. (2010), motivos para essa atenção incluem a capacidade da estrutura de representar distribuições discretas arbitrariamente complexas (ROUX; BENGIO, 2008), a facilidade para se realizar inferências condicionais entre as variáveis do modelo, e o fato de essas estruturas terem sido aplicadas de forma bem sucedida como blocos construtivos de arquiteturas com múltiplas camadas, como as DBNs (HINTON; OSINDERO; TEH, 2006).

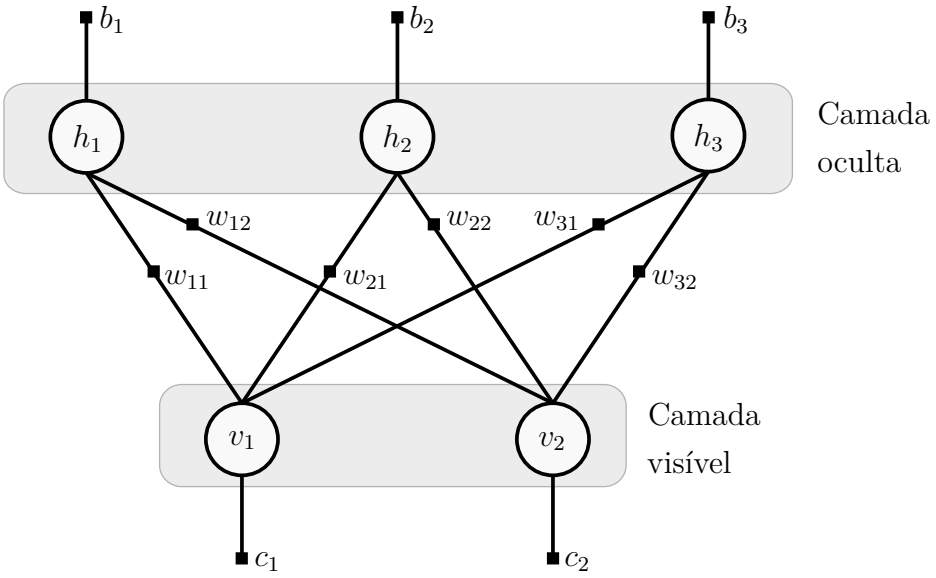
Nesta seção, a RBM na sua forma original, denominada RBM Bernoulli-Bernoulli, é descrita em detalhes. Na Seção 2.1.1 é apresentada a estrutura do modelo e na 2.1.2, a sua descrição matemática. Em seguida, a Seção 2.1.3 aborda os diferentes casos de uso do modelo, onde estão incluídas as tarefas de amostragem e de treinamento de RBMs. A Seção 2.1.4 relaciona a RBM Bernoulli-Bernoulli com alguns modelos conhecidos na literatura. Como será visto nessa seção, existe uma interessante relação entre as RBMs e as redes neuronais clássicas. Por fim, a Seção 2.1.5 discute brevemente a notável flexibilidade do modelo.

2.1.1 Estrutura

Uma RBM é formada por um conjunto de unidades visíveis e outro de unidades ocultas (SMOLENSKY, 1986; FREUND; HAUSSLER, 1994; HINTON, 2002). A Figura 2 fornece uma representação gráfica de uma RBM com duas unidades visíveis e três ocultas, em que os círculos representam as unidades e os pequenos quadrados, os parâmetros do modelo. Na figura, há n_h=3 unidades ocultas, denotadas por h_j, j=$1, 2, 3$ e n_v=2 unidades

visíveis denotadas por v_i, i=1, 2. As linhas ligando parâmetros a unidades indicam que esses estão associados. Assim sendo, o parâmetro c_i está associado à unidade v_i e é denominado viés dessa unidade visível. Similarmente, b_j é o viés da unidade oculta h_j. Cada parâmetro w_{ji} está simultaneamente associado às unidades h_j e v_i, sendo denominado peso da conexão entre as mesmas.

Figura 2 – RBM com duas unidades vísiveis (n_v=2) e três unidades ocultas (n_h=3).



Fonte: modificado de Larochelle (2013).

Algumas características estruturais do modelo podem ser salientadas. Em primeiro lugar, nota-se que as unidades são separadas em camadas, de modo que o modelo pode ser representado por um grafo bipartido. Em outras palavas, não existem ligações entre unidades pertencentes à mesma camada, o que justifica a palavra "restrita" no nome Máquina de Boltzmann Restrita. Em segundo lugar, destaca-se que o modelo é simétrico, muito embora sejam dados nomes diferentes às unidades das duas camadas. Vale, portanto, dizer que as unidades visíveis são assim chamadas pelo fato de serem, por definição, aquelas que interagem diretamente com o meio externo, em contraste com as ocultas, cujas interações com o meio são indiretas e probabilísticas, uma nomenclatura oriunda dos Modelos Ocultos de Markov (HMMs – *Hidden Markov Models*).

2.1.2 Descrição matemática

As unidades de uma RBM são formalmente modeladas como variáveis aleatórias com distribuição de Bernoulli (PAPOULIS; PILLAI, 2002). No presente texto, essas variáveis serão representadas pelos símbolos V_i e H_j, sendo o primeiro associado à unidade v_i e o segundo à h_j. Além de exprimirem os nomes das unidades, os símbolos v_i e h_j serão usados para representar valores específicos ou *estados* assumidos pelas variáveis V_i e H_j,

respectivamente. Para concisão de notação, o conjunto de variáveis aleatórias associadas às unidades visíveis será expresso pelo vetor aleatório

$$\mathbf{V} = [V_1 \dots V_{n_v}]^T \tag{2.1}$$

e o conjunto de estados dessas variáveis, pelo vetor

$$\mathbf{v} = [v_1 \dots v_{n_v}]^T, \tag{2.2}$$

em que o sobrescrito T indica transposição. Similarmente, para as unidades ocultas, tem-se

$$\mathbf{H} = [H_1 \dots H_{n_h}]^T \tag{2.3}$$

e

$$\mathbf{h} = [h_1 \dots h_{n_h}]^T. \tag{2.4}$$

Definem-se também a matriz $n_h \times n_v$ de pesos de conexão

$$\mathbf{W} = (w_{ji}), \tag{2.5}$$

com $j \in \{1, \dots, n_h\}$, $i \in \{1, \dots, n_v\}$ e os vetores de vieses

$$\mathbf{b} = [b_1 \dots b_{n_h}]^T \tag{2.6}$$

e

$$\mathbf{c} = [c_1 \dots c_{n_v}]^T. \tag{2.7}$$

Feitas essas definições, os principais conceitos relativos às RBMs serão estabelecidos a seguir.

2.1.2.1 Energia global

As RBMs fazem parte de uma classe de modelos probabilísticos que são baseados em energia. Nesses modelos, atribui-se um escalar, chamado de energia (total ou global), a cada estado do sistema. Para RBMs, tal estado é especificado pelo par $(\mathbf{v}, \mathbf{h})$ e é mapeado em um escalar por meio de uma *função de energia*, definida como (DESJARDINS et al., 2010)

$$E(\mathbf{v}, \mathbf{h}) = -\sum_{j=1}^{n_h} b_j h_j - \sum_{i=1}^{n_v} c_i v_i - \sum_{i=1}^{n_v} \sum_{j=1}^{n_h} h_j w_{ji} v_i. \tag{2.8}$$

Essa definição tem suas origens nas *redes de Hopfield* (HINTON; SEJNOWSKI, 1986) e foi chamada de energia por analogia a sistemas físicos. Hopfield (1982) mostrou que, se as unidades são simetricamente conectadas (o que é sempre verdade nas RBMs) e são atualizadas uma por vez, cada atualização reduz (ou não aumenta) o valor de uma função perda que, quando particularizada para o caso das RBMs, fornece a expressão para $E(\mathbf{v}, \mathbf{h})$. Consequentemente, reiteradas atualizações das unidades encontrariam, garantidamente, um mínimo de energia, já que a atualização de cada unidade é realizada para minimizar a energia global do sistema.

2.1.2.2 Distribuição de probabilidade

Com base na sua energia, $E(\mathbf{v}, \mathbf{h})$, um dado estado do sistema $(\mathbf{v}, \mathbf{h})$ tem a si atribuída uma certa probabilidade, dada pela função massa de probabilidade conjunta

$$P(\mathbf{v}, \mathbf{h}) = \frac{\exp(-E(\mathbf{v}, \mathbf{h}))}{Z}, \tag{2.9}$$

em que Z, chamada função de partição, é uma constante que garante que o somatório de $P(\mathbf{v}, \mathbf{h})$ sobre seu domínio seja unitário. Recordando que as variáveis que compõem os vetores $\mathbf{V}$ e $\mathbf{H}$ são do tipo Bernoulli, nota-se que o domínio de $P(\mathbf{v}, \mathbf{h})$ é discreto e dado explicitamente por $\{0, 1\}^{n_v} \times \{0, 1\}^{n_h}$ (i.e., todas as combinações binárias possíveis de valores das variáveis visíveis e ocultas). Dessa forma, tem-se que

$$Z = \sum_{\mathbf{v} \in \{0,1\}^{n_v}} \sum_{\mathbf{h} \in \{0,1\}^{n_h}} \exp(-E(\mathbf{v}, \mathbf{h})). \tag{2.10}$$

Destaca-se que $P(\mathbf{v}, \mathbf{h})$ tem a forma de uma distribuição de Boltzmann (LANDAU; LIFSHITZ, 1980), justificando novamente o nome das RBMs.

Como ficará evidente nas próximas seções, todas as características das RBMs podem ser deduzidas a partir da expressão de $P(\mathbf{v}, \mathbf{h})$ apresentada na Equação (2.9). Pode-se notar, entretanto, que $P(\mathbf{v}, \mathbf{h})$ depende exclusivamente de $E(\mathbf{v}, \mathbf{h})$, de modo que todas as informações sobre o modelo devem estar presentes na sua função de energia. De fato, examinando-se a expressão de $E(\mathbf{v}, \mathbf{h})$ na Equação (2.8), percebe-se que uma RBM fica totalmente especificada pelos seus vieses e pesos de conexão, isto é, pelo conjunto de parâmetros $\{\mathbf{W}, \mathbf{b}, \mathbf{c}\}$.

A partir das Equações (2.8) e (2.9), é interessante observar como os parâmetros do modelo atuam sobre sua função de energia global e, consequentemente, sobre sua distribuição de probabilidade. Por exemplo, analisando-se o primeiro termo do primeiro somatório em (2.8), percebe-se que:

a) Para h_1=0, o termo $b_1 h_1$ não contribui para a medida de energia global.

b) Para h_1=1, o termo $b_1 h_1$ contribui com $-b_1$ para a energia global. Assim, um aumento no viés b_1 causa uma diminuição na energia global e, conforme (2.9), produz um aumento na medida de probabilidade relativa ao estado conjunto do sistema.

Sendo assim, quanto maior o viés de uma unidade, maior será a probabilidade de que a variável aleatória correspondente assuma o valor um (nesse caso, a variável H_1). A mesma conclusão vale para os vieses das demais unidades ocultas e também para aqueles das unidades visíveis. Uma análise similar para os pesos de conexão permite concluir que estes controlam o acoplamento entre variáveis visíveis e ocultas. Por exemplo, incrementando-se o peso w_{12}, obtém-se um aumento na probabilidade de que as variáveis H_1 e V_2 assumam simultaneamente o valor um.

2.1.2.3 Energia livre

As variáveis visíveis do modelo são aquelas observáveis externamente, de modo que seria normalmente de interesse conhecer a distribuição de probabilidades das mesmas independentemente das variáveis ocultas. Com esse fim, pode-se marginalizar $\mathbf{h}$ na Equação (2.9) e, assim, determinar a função massa de probabilidade marginal

$$P(\mathbf{v}) = \sum_{\mathbf{h}\in\{0,1\}^{n_h}} \frac{\exp(-E(\mathbf{v},\mathbf{h}))}{Z}. \tag{2.11}$$

Recordando-se que na Equação (2.9) fora feito um mapeamento entre a energia global, $E(\mathbf{v},\mathbf{h})$, e sua probabilidade, $P(\mathbf{v},\mathbf{h})$, pode-se usar do mesmo expediente para definir a *função de energia livre*, $\mathcal{F}(\mathbf{v})$, de modo a satisfazer (BENGIO, 2009)

$$P(\mathbf{v}) = \frac{\exp(-\mathcal{F}(\mathbf{v}))}{Z}, \tag{2.12}$$

igualdade que associa uma probabilidade a cada nível de *energia livre* do sistema. Explicitamente, tem-se de (2.11) e (2.12) que a energia livre atribuída ao estado $\mathbf{v}$ é dada por

$$\mathcal{F}(\mathbf{v}) = -\ln\left[\sum_{\mathbf{h}\in\{0,1\}^{n_h}} \exp(-E(\mathbf{v},\mathbf{h}))\right]. \tag{2.13}$$

Na forma apresentada na Equação (2.13), a determinação da energia livre para um dado $\mathbf{v}$ necessitaria da avaliação de um somatório com 2^{n_h} termos. Como mostrado no Apêndice A, $\mathcal{F}(\mathbf{v})$ pode ser escrita como

$$\mathcal{F}(\mathbf{v}) = -\sum_{i=1}^{n_v} c_i v_i - \sum_{j=1}^{n_h} \ln\left(1 + \exp\left(b_j + \sum_{i=1}^{n_v} w_{ji} v_i\right)\right), \tag{2.14}$$

cuja avaliação é mais eficiente computacionalmente, pois tem uma complexidade linear em n_h.

Anteriormente citou-se que a energia global fora assim denominada por analogia a sistemas físicos. O mesmo ocorre com a energia livre. Em especial, ressalta-se que ambas têm a si associadas probabilidades segundo a distribuição de Boltzmann. Essa distribuição aparece no estudo de mecânica estatística dos gases em equilíbrio térmico e tem a forma

$$P_T(\mathbf{x}) = \frac{\exp(-\epsilon(\mathbf{x})/kT)}{Q(T)}, \tag{2.15}$$

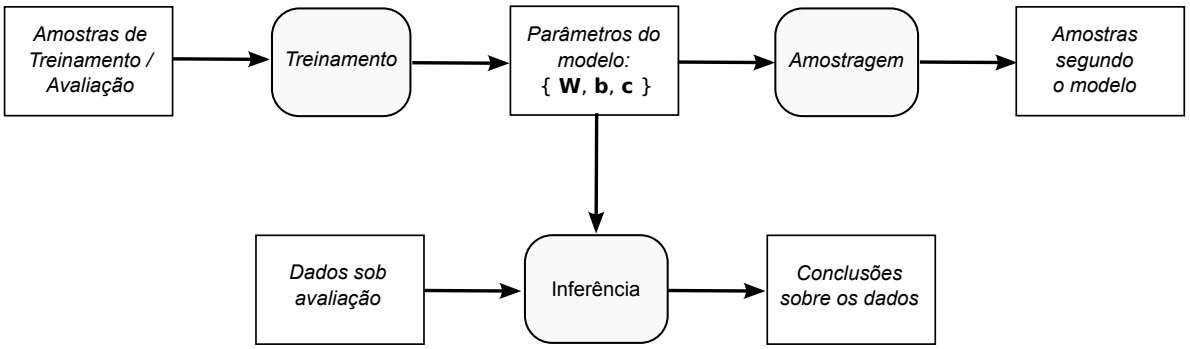
em que $\epsilon(\mathbf{x})$ é a energia associada ao estado $\mathbf{x}$, k é a constante de Boltzmann, T é a temperatura absoluta do sistema e $Q(T)$ é a função de partição que garante que o somatório de $P_T(\mathbf{x})$ seja unitário. Com base na expressão de $P_T(\mathbf{x})$, deve-se notar alguns fatos gerais sobre distribuições dessa forma (tais como $P(\mathbf{v},\mathbf{h})$ e $P(\mathbf{v})$). Em primeiro lugar, fixada uma temperatura, tem-se que estados com maior energia tem menores probabilidades. Em segundo lugar, um aumento de temperatura faz com que a expressão dentro da função

exponencial da Equação (2.15) varra uma faixa menor de valores. Sendo assim, com T tendendo a infinito seria obtida uma distribuição uniforme. Esses fatos são explorados na amostragem e no treinamento de RBMs, como será visto mais à frente.

2.1.3 Tarefas associadas ao modelo

Existem diferentes casos de uso associados ao modelo das RBMs. Um deles seria determinar o conjunto de parâmetros $\{\mathbf{W}, \mathbf{b}, \mathbf{c}\}$ de modo a se obter uma distribuição de probabilidade (ou, equivalentemente, energia) especificada, tarefa essa conhecida como *treinamento*. Em contrapartida, numa tarefa chamada de *amostragem*, o objetivo seria obter amostras de uma RBM cujos parâmetros tenham sido especificados. Um terceiro caso, ainda, consistiria em avaliar hipóteses com base no modelo, num denominado processo de *inferência*. A Figura 3 fornece uma visão geral das tarefas mencionadas. Nas seções seguintes, cada uma delas será abordada.

Figura 3 – Tarefas associadas ao modelo das RBMs.



Fonte: produção do próprio autor.

2.1.3.1 Inferência

Inferência estatística consiste em tirar conclusões gerais a respeito de uma população baseado no exame de apenas uma parte do todo (CASELLA; BERGER, 2002). No caso das RBMs, os dados extraídos da população são representados pelos estados das variáveis visíveis e o processo de inferência é realizado em duas etapas:

1. Os parâmetros do modelo são obtidos num treinamento;
2. De posse dos parâmetros, as distribuições relevantes são determinadas e utilizadas na obtenção das conclusões de interesse.

Um exemplo simples é o de obter a esperança dos dados da população: conhecendo-se $\{\mathbf{W}, \mathbf{b}, \mathbf{c}\}$, a distribuição $P(\mathbf{v})$ fica determinada e, daí, obtém-se a esperança $\mathbb{E}[\mathbf{V}] =$

$\sum_{\mathbf{v}} \mathbf{v}P(\mathbf{v})$. Deve-se perceber, entretanto, que essa simplicidade do ponto de vista teórico não necessariamente implica uma baixa complexidade na computação da expressão. No cálculo da esperança, é necessário avaliar $P(\mathbf{v})$ conforme Equação (2.11), expressão em que aparece a função de partição, Z. A determinação exata dessa constante, todavia, é considerada intratável por envolver um somatório com um número de termos que cresce exponencialmente com a quantidade de unidades. A ineficiência na determinação de Z pode ser mitigada pelo uso de um método de amostragem por importância, AIS (*Annealed Importance Sampling*), conforme mostrado por Salakhutdinov e Murray (2008). Na prática, porém, procura-se evitar a necessidade de uma determinação explícita de Z.

Rosales e Sclaroff (2006) discutiram o uso de modelos generativos (modelos com a capacidade de gerar dados de uma distribuição, como é o caso das RBMs) para a estimação de poses articuladas, tais como do corpo humano ou da mão humana, a partir de uma única imagem. No uso de RBMs nesse tipo de aplicação, as variáveis visíveis representariam a imagem e as ocultas, a pose que se deseja estimar. Esse exemplo demonstra um caso típico em que surge a necessidade de se determinar a distribuição das variáveis ocultas dadas as visíveis, $P(\mathbf{h}|\mathbf{v})$. A estrutura bipartida do modelo das RBMs faz com que suas distribuições condicionais $P(\mathbf{h}|\mathbf{v})$ e $P(\mathbf{v}|\mathbf{h})$ tenham expressões cuja avaliação é particularmente eficiente. Esse fato é explorado não apenas na aplicação do modelo em problemas de inferência, mas também nos processos de treinamento e amostragem. Pode-se dizer que essa eficiência foi um fator importante na popularização das RBMs (DESJARDINS et al., 2010).

Devido à sua importância no contexto das RBMs, as distribuições condicionais $P(\mathbf{v}|\mathbf{h})$ e $P(\mathbf{h}|\mathbf{v})$ serão apresentadas aqui. Conforme demonstrado no Apêndice A, fixados os estados das variáveis visíveis, as variáveis ocultas tornam-se independentes, ou seja,

$$P(\mathbf{h}|\mathbf{v}) = \prod_{j=1}^{n_h} P(h_j|\mathbf{v}). \tag{2.16}$$

Na mesma condição, cada H_j tem distribuição de Bernoulli com probabilidade de ativação (probabilidade da variável binária assumir o valor 1), denotada por $P(h_j{=}1|\mathbf{v})$, dada por

$$P(h_j{=}1|\mathbf{v}) = \varphi\left(b_j + \sum_{i=1}^{n_v} w_{ji} v_i\right), \tag{2.17}$$

em que $\varphi(x) = 1/(1+e^{-x})$ é a chamada função sigmóide. Considerando-se a simetria do modelo, resultados análogos aos indicados anteriormente são obtidos para as variáveis visíveis quando fixados os estados das ocultas, a saber,

$$P(\mathbf{v}|\mathbf{h}) = \prod_{i=1}^{n_v} P(v_i|\mathbf{h}), \tag{2.18}$$

$$P(v_i{=}1|\mathbf{h}) = \varphi\left(c_i + \sum_{j=1}^{n_h} w_{ji} h_j\right). \tag{2.19}$$

Deve-se destacar que as eqs. (2.16) a (2.19) formam o elo de ligação entre as RBMs e as redes neuronais clássicas, como será explicado na Seção 2.1.4.

2.1.3.2 Amostragem

No processo de amostragem, objetiva-se produzir amostras segundo um modelo cujos parâmetros sejam conhecidos. Como já mencionado, a interação das RBMs com o meio externo é realizada pelas unidades visíveis. Assim, nesses modelos, o principal interesse é obter amostras segundo a distribuição marginal das variáveis visíveis, $P(\mathbf{v})$. Com esse objetivo, utiliza-se um método de Monte Carlo via cadeias de Markov (MCMC - *Markov Chain Monte Carlo*) conhecido como *amostragem de Gibbs* (GEMAN; GEMAN, 1984; GELFAND; SMITH, 1990). Na literatura, um sistema que produza amostras utilizando tal método é comumente chamado de *amostrador de Gibbs*. A amostragem de Gibbs é especialmente apropriada para situações em que amostrar a partir das distribuições condicionais do modelo seja simples, como ocorre com as RBMs.

Como será mostrado na Seção 2.1.3.3, a geração de amostras das variáveis visíveis do modelo é uma tarefa importante também durante o treinamento, na medida em que tais amostras são usadas na estimação de estatísticas cujo cálculo seria normalmente intratável. Cumpre dizer ainda que diferentes algoritmos de treinamento utilizam a amostragem de Gibbs de maneiras distintas. As particularidades de cada algoritmo com relação à amostragem serão apresentadas posteriormente usando a base introduzida aqui.

Seguindo Haykin (2009), na apresentação do amostrador de Gibbs será considerado um vetor aleatório

$$\mathbf{X} = [X_1 \dots X_K]^T, \tag{2.20}$$

para o qual são consideradas conhecidas as distribuições condicionais de variáveis individuais dadas as variáveis restantes, $P(x_k|X_1{=}x_1, \dots, X_{k-1}{=}x_{k-1}, X_{k+1}{=}x_{k+1}, \dots, X_K{=}x_K)$, para $k = 1, \dots, K$. Em um passo do algoritmo, comumente chamado de passo de Gibbs, são produzidas, em sequência, K amostras, sendo uma para cada distribuição condicional. Explicitamente, o n-ésimo passo é dividido nos seguintes K sub-passos de amostragem:

- Produz-se $x_1^{(n)}$ segundo $P(x_1|X_2{=}x_2^{(n-1)}, X_3{=}x_3^{(n-1)}, \dots, X_K{=}x_K^{(n-1)})$;
- Produz-se $x_2^{(n)}$ segundo $P(x_2|X_1{=}x_1^{(n)}, X_3{=}x_3^{(n-1)}, \dots, X_K{=}x_K^{(n-1)})$;

$\vdots$

- Produz-se $x_k^{(n)}$ segundo $P(x_k|X_1{=}x_1^{(n)}, \dots, X_{k-1}{=}x_{k-1}^{(n)}, X_{k+1}{=}x_{k+1}^{(n-1)}, \dots, X_K{=}x_K^{(n-1)})$;

$\vdots$

- Produz-se $x_K^{(n)}$ segundo $P(x_K|X_1{=}x_1^{(n)}, X_2{=}x_2^{(n)}, \dots, X_{K-1}{=}x_{K-1}^{(n)})$.

O índice entre parênteses sobrescrito aos nomes das amostras indica o passo de Gibbs em que as mesmas seriam produzidas. Nota-se que na execução do primeiro passo (n=1) seria necessário o conhecimento dos valores $x_2^{(0)}, \ldots, x_K^{(0)}$, os quais são escolhidos arbitrariamente. Os dados de saída de um passo, ou seja, o conjunto de valores $x_1^{(n)}, \ldots, x_K^{(n)}$, podem ser vistos como realizações de variáveis aleatórias, aqui indicadas como $X_1^{(n)}, \ldots, X_K^{(n)}$. Destaca-se que a sequência dessas variáveis criadas nos vários passos do algoritmo formam uma cadeia de Markov. Sob condições bastante gerais, valem os seguintes teoremas a respeito da amostragem de Gibbs (GEMAN; GEMAN, 1984):

1. **Teorema da convergência.** A distribuição da variável aleatória $X_k^{(n)}$ converge para a distribuição verdadeira de X_k, $k = 1, \ldots, K$, quando n tende a infinito.

2. **Teorema da taxa de convergência.** A distribuição cumulativa conjunta das variáveis aleatórias $X_1^{(n)}, \ldots, X_K^{(n)}$ converge para a distribuição cumulativa conjunta verdadeira de $X_1, \ldots, X_K$ a uma taxa geométrica em n.

3. **Teorema da ergodicidade.** Para qualquer função mensurável g das variáveis aleatórias $X_1^{(n)}, \ldots, X_K^{(n)}$ cuja esperança exista, tem-se

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} g(X_1^{(i)}, \ldots, X_K^{(i)}) \to \mathbb{E}\left[g(X_1, \ldots, X_K)\right] \tag{2.21}$$

   com probabilidade 1, sendo que $\mathbb{E}\left[\cdot\right]$ representa o operador esperança matemática.

Esses resultados garantem a convergência do algoritmo de amostragem, bem como a usabilidade das amostras por ele produzidas para a determinação de estatísticas de interesse.

A aplicação do amostrador de Gibbs às RBMs apresenta um fator simplificador: as variáveis ocultas são conjuntamente independentes dados os estados das visíveis (Seção 2.1.3.1 e Apêndice A). Esse fato permite que todas as variáveis ocultas sejam amostradas simultaneamente segundo a distribuição $P(\mathbf{h}|\mathbf{v})$. Conclusão semelhante vale para as variáveis visíveis. Dessa forma, para as RBMs, um passo de amostragem de Gibbs pode ser escrito com apenas dois sub-passos:

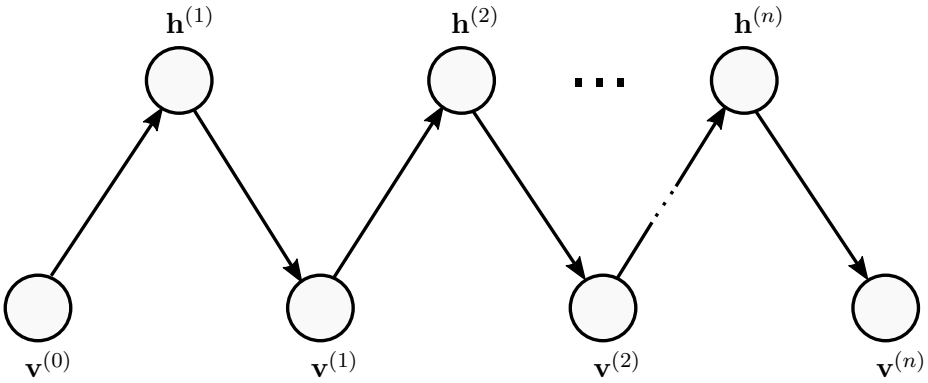
- Produz-se $\mathbf{h}^{(n)}$ segundo $P(\mathbf{h}|\mathbf{V} = \mathbf{v}^{(n-1)})$;
- Produz-se $\mathbf{v}^{(n)}$ segundo $P(\mathbf{v}|\mathbf{H} = \mathbf{h}^{(n)})$.

Vê-se que agora o resultado de cada sub-passo é um vetor de amostras. A produção de $\mathbf{h}^{(n)}$ consiste no simples sorteio de amostras independentes com distribuição de Bernoulli cada qual tendo uma probabilidade de sucesso igual a $P(h_j{=}1|\mathbf{V} = \mathbf{v}^{(n-1)})$, para $j = 1, \ldots, n_h$. Como mencionado, essa independência entre amostras permite que os sorteios das mesmas

seja feito em paralelo, algo frequentemente utilizado em implementações reais do algoritmo. A produção de $\mathbf{v}^{(n)}$, por sua vez, é feita de forma análoga.

A Figura 4 apresenta uma forma de se visualizar a sequência ou cadeia de amostragens definida pelos passos e sub-passos do amostrador de Gibbs. Na figura, os círculos representam os vetores de amostras e as setas, as amostragens propriamente ditas. Por exemplo, a seta de $\mathbf{v}^{(0)}$ para $\mathbf{h}^{(1)}$ indica a amostragem de $\mathbf{h}^{(1)}$ segundo a distribuição $P(\mathbf{h}|\mathbf{V} = \mathbf{v}^{(0)})$, como seria o primeiro sub-passo do amostrador de Gibbs. Nota-se também que a sequência de conexões $\mathbf{v}^{(0)} \rightarrow \mathbf{h}^{(1)} \rightarrow \mathbf{v}^{(1)}$ representa o primeiro passo completo do amostrador. Esse tipo de diagrama será útil na explanação dos algoritmos de treinamento.

Figura 4 – Representação gráfica da cadeia de amostragem de Gibbs.



Fonte: produção do próprio autor (inspirado em Cho, Ilin e Raiko (2011)).

2.1.3.3 Treinamento

No processo de treinamento, deseja-se determinar os parâmetros do modelo de forma a se obter uma distribuição de probabilidade especificada. Na prática, tal distribuição é fornecida implicitamente por meio de um conjunto de amostras de treinamento, as quais teriam sido produzidas segundo a distribuição que se deseja obter (ou aproximar). Como no treinamento das RBMs as amostras não requerem uma classificação prévia, a aprendizagem é dita não supervisionada. Isso também ocorre em outros modelos generativos, como GMMs e DBNs. No treinamento de modelos discriminativos como, por exemplo, SVMs e redes neuronais clássicas, as amostras de treinamento devem ter sido previamente classificadas, ou seja, a aprendizagem é supervisionada. Dada a imensa disponibilidade de dados não classificados, essa distinção representa uma importante vantagem prática dos modelos generativos frente aos discriminativos.

Os principais algoritmos usados no treinamento de RBMs são baseados no método do gradiente descendente estocástico (SGD – *Stochastic Gradient Descent*). Nesse método, o gradiente de uma função perda é determinado usando-se uma única amostra de treinamento

e os parâmetros do modelo são ajustados no sentido oposto ao do vetor gradiente, isto é, no sentido de minimizar a perda. Esse processo é repetido até que algum critério de parada seja atingido.

Visando o uso do SGD no treinamento de RBMs, são apresentados a seguir a função perda e o seu gradiente, os quais são compartilhados pelos algoritmos de treinamento usuais que serão introduzidos mais à frente.

Função perda e seu gradiente

No treinamento de RBMs, considera-se uma amostra $\mathbf{v}^{(t)}$ extraída da distribuição empírica dos dados de treinamento, e objetiva-se determinar os parâmetros do modelo que minimizem a função de log-verossimilhança negativa, definida como

$$\mathcal{L}(\boldsymbol{\theta}; \mathbf{v}^{(t)}) = -\ln P(\mathbf{v}^{(t)}), \tag{2.22}$$

sendo que $\boldsymbol{\theta}$ é um vetor em que cada componente consiste em um dos parâmetros do modelo. $\mathcal{L}(\boldsymbol{\theta}; \mathbf{v}^{(t)})$ representa a função perda (ou custo) a ser minimizada via SGD. Em razão do sinal negativo na Equação (2.22), percebe-se que a minimização de $\mathcal{L}(\boldsymbol{\theta}; \mathbf{v}^{(t)})$ é equivalente à maximização da log-verossimilhança da amostra de treinamento, um objetivo de otimização que é comumente empregado na estimação de parâmetros de modelos estatísticos.

Diferenciado-se (2.22) com relação ao vetor de parâmetros ($\boldsymbol{\theta}$), chega-se ao gradiente da função perda que, idealmente, seria usado no SGD. Tal gradiente pode ser escrito de duas formas particularmente interessantes, ambas demonstradas no Apêndice A:

1. Em termos da energia livre, chega-se a

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta}; \mathbf{v}^{(t)})}{\partial \boldsymbol{\theta}} = \frac{\partial \mathcal{F}(\mathbf{v}^{(t)})}{\partial \boldsymbol{\theta}} - \mathbb{E}_{P(\mathbf{v})}\left[\frac{\partial \mathcal{F}(\mathbf{V})}{\partial \boldsymbol{\theta}}\right], \tag{2.23}$$

   sendo que a notação $\mathbb{E}_{\mathcal{P}}[\cdot]$ representa a esperança da expressão entre colchetes sobre a distribuição $\mathcal{P}$ que estiver indicada. Vê-se que o gradiente da função perda é composto por dois termos, ambos envolvendo $\partial\mathcal{F}(\cdot)/\partial\boldsymbol{\theta}$ (gradiente da função de energia livre). O primeiro termo, denominado *fase positiva*, depende da amostra de treinamento ($\mathbf{v}^{(t)}$) e, por isso, pode-se interpretar que essa parcela de $\partial\mathcal{L}(\cdot)/\partial\boldsymbol{\theta}$ é responsável por aumentar a probabilidade de se observar $\mathbf{v}^{(t)}$ segundo o modelo sendo treinado (diminuindo a energia livre relativa a essa amostra). Em contrapartida, o segundo termo, chamado de *fase negativa*, corresponde à esperança da expressão $\partial\mathcal{F}(\cdot)/\partial\boldsymbol{\theta}$ calculada sobre toda a distribuição das variáveis visíveis. Interpreta-se, assim, que a fase negativa é responsável por diminuir a probabilidade das amostras geradas pelo modelo (HINTON; SEJNOWSKI, 1986).

2. Em função da energia global, tem-se

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta}; \mathbf{v}^{(t)})}{\partial \boldsymbol{\theta}} = \mathbb{E}_{P(\mathbf{h}|\mathbf{v})}\left[\frac{\partial E(\mathbf{v}^{(t)}, \mathbf{H})}{\partial \boldsymbol{\theta}}\bigg|\mathbf{v}^{(t)}\right] - \mathbb{E}_{P(\mathbf{v},\mathbf{h})}\left[\frac{\partial E(\mathbf{V}, \mathbf{H})}{\partial \boldsymbol{\theta}}\right]. \tag{2.24}$$

Novamente, o primeiro e o segundo termos correspondem às fases positiva e negativa, respectivamente. Com o gradiente da função perda escrito nessa forma, pode-se perceber que, na fase positiva, fixam-se as variáveis visíveis no estado $\mathbf{v}^{(t)}$ e, então, realiza-se a amostragem das variáveis ocultas do modelo. Já na fase negativa, toda a distribuição conjunta do modelo é amostrada (BENGIO, 2009).

A Equação (2.23) é especialmente conveniente para os desenvolvimentos que seguem. Como será mostrado em breve, o cálculo da fase positiva segundo essa equação é eficiente. A determinação da fase negativa, todavia, é intratável. Os algoritmos de treinamento devem, portanto, lidar de alguma maneira com essa intratabilidade.

Algoritmo CD

Hinton (2002) introduziu o modelo denominado Produto de Especialistas (PoE – *Product of Experts*), formado pelo produto de um conjunto de distribuições. No mesmo trabalho, o autor propôs o algoritmo CD (*Contrastive Divergence*) visando o treinamento de PoEs. Hinton também apontou que o algoritmo seria aplicável naturalmente ao treinamento de RBMs, visto que uma RBM pode ser entendida como sendo um PoE com um especialista por unidade oculta.

Inicia-se a derivação do algoritmo CD obtendo-se uma expressão tratável para o gradiente da função perda. Partindo da Equação (2.23) fazem-se duas aproximações (BENGIO, 2009). A primeira delas consiste em estimar a esperança no segundo termo do lado direito daquela equação usando-se uma única amostra, ou seja, assume-se

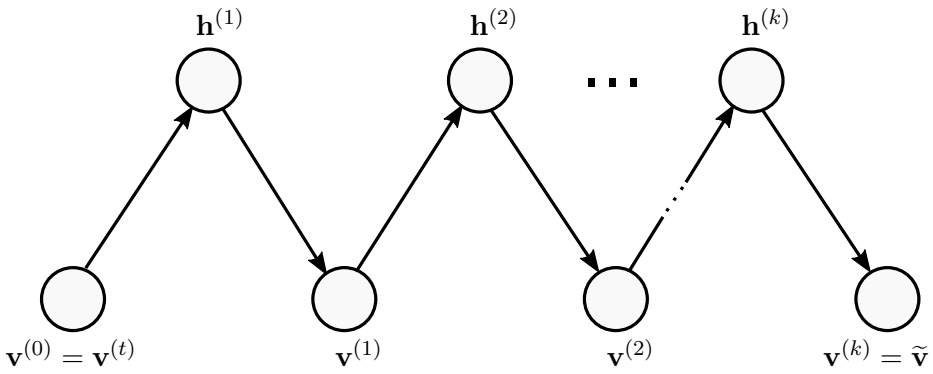
$$\mathbb{E}_{P(\mathbf{v})}\left[\frac{\partial \mathcal{F}(\mathbf{V})}{\partial \boldsymbol{\theta}}\right] \approx \frac{\partial \mathcal{F}(\tilde{\mathbf{v}})}{\partial \boldsymbol{\theta}}, \tag{2.25}$$

sendo $\tilde{\mathbf{v}}$ uma amostra representativa da distribuição marginal das variáveis visíveis. Essa amostra poderia, como visto anteriormente, ser obtida via amostragem de Gibbs, o que envolveria um número potencialmente elevado de passos de modo a garantir sua convergência. A segunda aproximação lida com essa possibilidade definindo que na amostragem de Gibbs sejam usados k passos e que a cadeia de amostragem seja iniciada na amostra de treinamento, $\mathbf{v}^{(t)}$. Nessas condições, denota-se o algoritmo por CD-k. A Figura 5 apresenta o esquema de amostragem usado pelo CD-k usando a representação gráfica introduzida na Seção 2.1.3.2.

De (2.23) e (2.25), tem-se agora a versão aproximada do gradiente da função perda

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta}; \mathbf{v}^{(t)})}{\partial \boldsymbol{\theta}} \approx \frac{\partial \mathcal{F}(\mathbf{v}^{(t)})}{\partial \boldsymbol{\theta}} - \frac{\partial \mathcal{F}(\tilde{\mathbf{v}})}{\partial \boldsymbol{\theta}}. \tag{2.26}$$

Figura 5 – Esquema de amostragem do CD-k.



Fonte: produção do próprio autor.

Nessa expressão, nota-se que as fases positiva e negativa consistem na avaliação da mesma função ($\partial\mathcal{F}(\cdot)/\partial\boldsymbol{\theta}$) usando-se diferentes amostras ($\mathbf{v}^{(t)}$ e $\widetilde{\mathbf{v}}$). Uma nomenclatura habitualmente usada na literatura chama as amostras utilizadas para o cálculo da fase positiva de *partículas positivas*, como é o caso de $\mathbf{v}^{(t)}$. Aquelas usadas no cálculo da fase negativa, como $\widetilde{\mathbf{v}}$, são denominadas *partículas negativas*.

Segundo o método do gradiente descendente estocástico, a regra de atualização de parâmetros pode ser escrita como

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \lambda \frac{\partial \mathcal{L}(\boldsymbol{\theta}; \mathbf{v}^{(t)})}{\partial \boldsymbol{\theta}}, \tag{2.27}$$

de modo que o uso de (2.26) nessa última expressão produz

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \lambda \left(\frac{\partial \mathcal{F}(\mathbf{v}^{(t)})}{\partial \boldsymbol{\theta}} - \frac{\partial \mathcal{F}(\widetilde{\mathbf{v}})}{\partial \boldsymbol{\theta}} \right), \tag{2.28}$$

em que λ é uma constante positiva denominada *taxa de aprendizagem*. No Apêndice A, essa expressão é ainda desenvolvida para cada um dos parâmetros do modelo chegando-se, assim, ao conjunto de regras de atualização a ser usado no algoritmo de treinamento:

$$b_j \leftarrow b_j + \lambda\Big[P(h_j{=}1|\mathbf{v}^{(t)}) - P(h_j{=}1|\widetilde{\mathbf{v}})\Big], \tag{2.29}$$

$$c_i \leftarrow c_i + \lambda\Big[v_i^{(t)} - \widetilde{v}_i\Big], \tag{2.30}$$

$$w_{ji} \leftarrow w_{ji} + \lambda\Big[P(h_j{=}1|\mathbf{v}^{(t)})\, v_i^{(t)} - P(h_j{=}1|\widetilde{\mathbf{v}})\, \widetilde{v}_i\Big], \tag{2.31}$$

sendo $v_i^{(t)}$ e $\widetilde{v}_i$, as i-ésimas componentes do vetores $\mathbf{v}^{(t)}$ e $\widetilde{\mathbf{v}}$, respectivamente. A função que fornece a probabilidade condicional de ativação das unidades ocultas para um dado vetor visível, $P(h_j{=}1|\mathbf{v})$, foi apresentada na Equação (2.17). Observa-se, naquela equação, que a referida probabilidade é calculada simplesmente aplicando-se a função sigmóide a uma combinação linear de valores. Assim, conhecida a partícula $\widetilde{\mathbf{v}}$, a avaliação dessas regras de atualização pode ser considerada eficiente.

Tabela 1 – Iteração do algoritmo CD-k

$\mathbf{v}^{(t)}$ $\leftarrow$ *Amostra (partícula positiva) obtida do conjunto de treinamento*;
$\tilde{\mathbf{v}}$ $\leftarrow$ *Amostra (partícula negativa) obtida através de k passos de Gibbs iniciando a cadeia de amostragem na amostra de treinamento,* $\mathbf{v}^{(t)}$;
Uso das regras de atualização de parâmetros:
forall $j \in \{1, \ldots, n_h\}$ **do**
  $b_j \leftarrow b_j + \lambda\Big[P(h_j{=}1|\mathbf{v}^{(t)}) - P(h_j{=}1|\tilde{\mathbf{v}})\Big]$;
end
forall $i \in \{1, \ldots, n_v\}$ **do**
  $c_i \leftarrow c_i + \lambda\Big[v_i^{(t)} - \tilde{v}_i\Big]$;
end
forall $(i, j) \in \{1, \ldots, n_v\} \times \{1, \ldots, n_h\}$ **do**
  $w_{ji} \leftarrow w_{ji} + \lambda\Big[P(h_j{=}1|\mathbf{v}^{(t)})\, v_i^{(t)} - P(h_j{=}1|\tilde{\mathbf{v}})\, \tilde{v}_i\Big]$;
end

Na Tabela 1, é fornecido o pseudocódigo relativo a uma iteração do algoritmo CD-k. Com isso, deve ficar claro o uso das regras de atualização. Na prática, essas iterações seriam executadas em vários ciclos (épocas) de treinamento, sendo que, em cada ciclo, todas as amostras de treinamento seriam apresentadas ao algoritmo. Como citado anteriormente, no presente texto, considera-se o cálculo do gradiente com o uso de uma única amostra de treinamento. A adaptação das regras de atualização para a utilização de múltiplas amostras no cálculo do gradiente é uma prática comum, mas, por ser trivial, não será mostrada aqui.

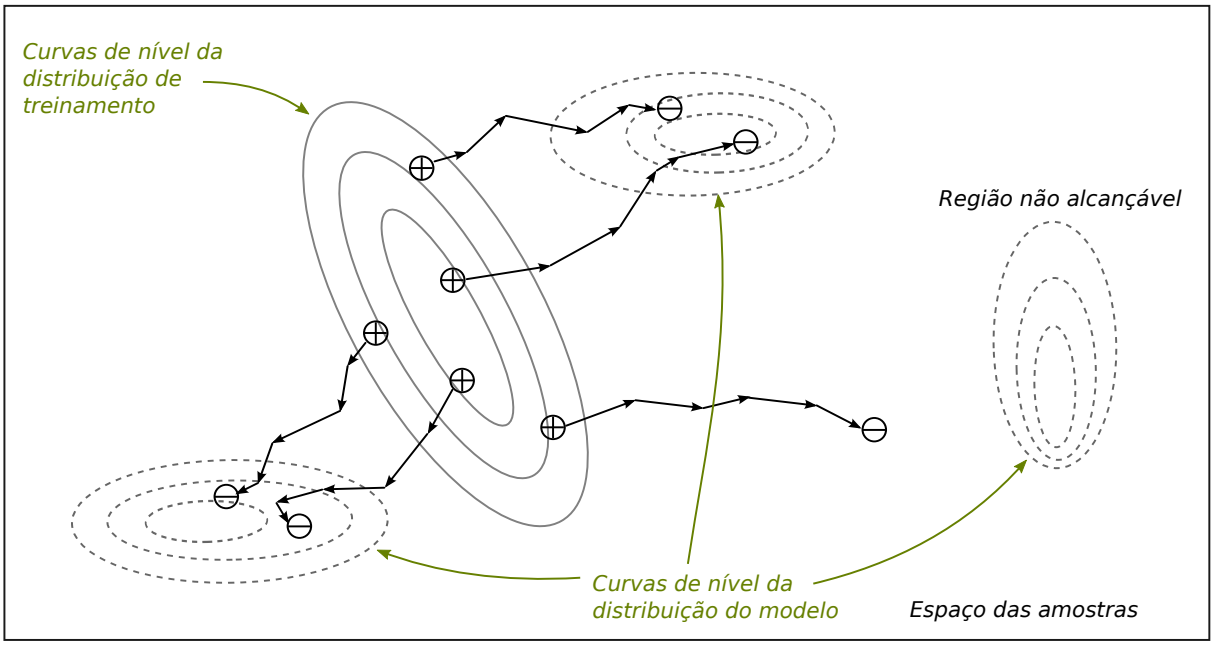
Por fim, com base no entendimento dado anteriormente para as fases do gradiente, pode-se chegar a uma interpretação também para o trabalho realizado pelo algoritmo CD-k. Na Figura 6, partículas positivas e negativas foram consideradas, para efeitos de ilustração, como pontos num espaço contínuo de duas dimensões e indicadas pelos símbolos $\oplus$ e $\ominus$, respectivamente. No mesmo espaço, foram projetadas as curvas de nível de duas distribuições:

1. A distribuição empírica dos dados de treinamento, da qual partículas positivas são retiradas (elipses em linhas contínuas);

2. A distribuição das variáveis visíveis segundo o modelo sendo treinado (elipses em linhas tracejadas).

As sequências de setas ligando partículas positivas às negativas representam possíveis caminhos produzidos pelos passos de Gibbs (conforme CD-k). Nota-se que a amostragem de Gibbs produz, normalmente, partículas negativas em regiões que, segundo o modelo, têm maior probabilidade. Deve-se perceber, entretanto, que devido à limitação no número de passos de amostragem, certas regiões podem não ser alcançadas (como aquela indicada à

Figura 6 – Geração de partículas negativas conforme CD-k. As sequências de setas ligando partículas positivas às negativas representam possíveis caminhos produzidos pelos passos de Gibbs.



Fonte: produção do próprio autor.

direita na figura). Recorda-se agora que a atualização dos parâmetros do modelo (via SGD) produz, pela interpretação das fases do gradiente, os efeitos de aumentar a probabilidade das partículas positivas e, simultaneamente, diminuir a probabilidade das negativas. Assim, pode-se esperar que uma sequência de atualizações retire aquelas massas de probabilidade das regiões indesejadas movendo-as para os entornos das partículas positivas. Essa é uma característica desejável para o algoritmo, na medida em que amostras de treinamento teriam probabilidades relativamente elevadas segundo um modelo treinado. Em contrapartida, regiões externas a um, por assim dizer, raio de k passos de Gibbs das partículas positivas não teriam suas probabilidades diminuídas. O efeito seria a presença de modos espúrios (isto é, inexistentes no conjunto de treinamento) na distribuição do modelo treinado. Esse é considerado como um ponto fraco do CD-k (ao menos para valores pequenos de k) e que motivou o desenvolvimento de outros algoritmos.

A fim de discutir algumas dificuldades práticas no treinamento e ilustrar a operação do algoritmo CD, são mostrados aqui os resultados de uma simulação em que esse algoritmo foi empregado no treinamento de uma RBM tendo n_v=3 unidades visíveis e n_h=2 unidades ocultas. Em particular, usou-se k=1 passo de Gibbs para obtenção das partículas negativas (algoritmo CD-1) e uma taxa de aprendizagem λ=0,1. Na Figura 7, são mostradas, em um único gráfico de barras, três distribuições de probabilidade. A primeira delas, correspondendo à primeira coluna em cada grupo de barras (em azul), representa a distribuição dos dados de treinamento. Percebe-se que apenas duas combinações de estados

das variáveis visíveis estão presentes nesses dados, ou seja,

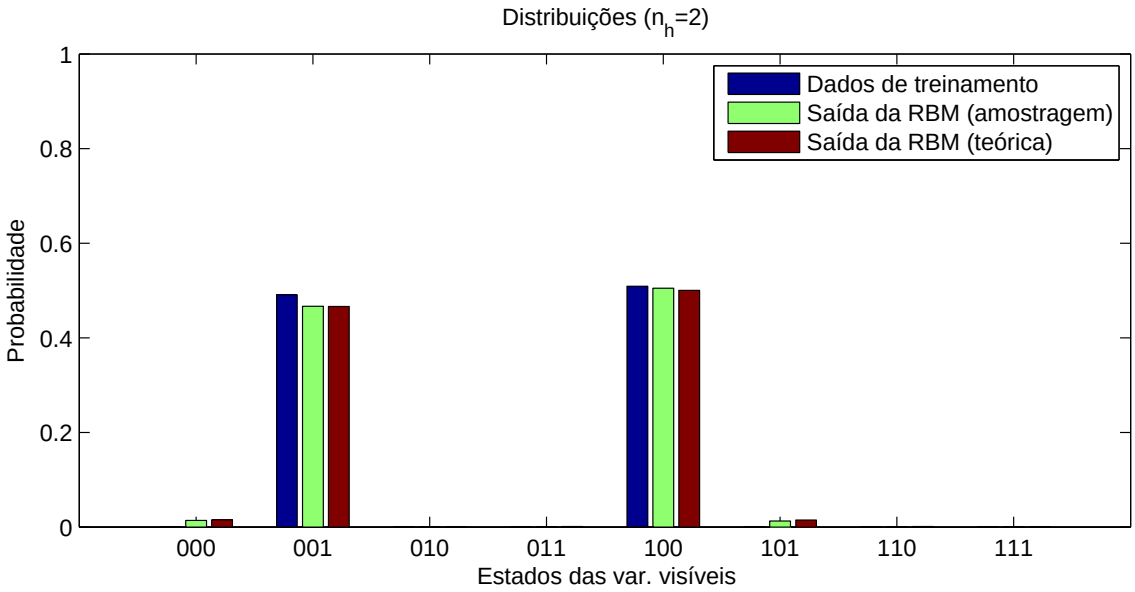
$$(v_1, v_2, v_3) \in \{(0,0,1), (1,0,0)\}$$

e esses estados são equiprováveis. Na segunda coluna de cada grupo de barras (em verde), tem-se a distribuição dos dados produzidos na amostragem do modelo treinado. Comparando essas duas distribuições entre si, é possível notar que a RBM é capaz de capturar razoavelmente bem as características estatísticas dos dados de treinamento. Mais ainda, o processo de amostragem permite que dados com características similares àqueles do treinamento sejam gerados pelo modelo resultante, fato que justifica a classificação das RBMs como modelos generativos. Ainda com relação à Figura 7, a terceira distribuição (em vermelho) expressa a distribuição teórica do modelo. Verifica-se que essa se assemelha bastante à distribuição dos dados obtidos com a amostragem de Gibbs. Vale dizer que a distribuição teórica pode ser obtida avaliando-se explicitamente a expressão

$$P(\mathbf{v}) = \sum_{\mathbf{h}^*} P(\mathbf{v}, \mathbf{h}^*).$$

Destaca-se que a avaliação dessa expressão é intratável e, por isso, só é realizada na prática quando o número de unidades do modelo é pequeno, como no presente caso.

Figura 7 – Distribuições de probabilidade dos dados de treinamento e do modelo resultante para RBM Bernoulli-Bernoulli treinada com o algoritmo CD-1.



Fonte: produção do próprio autor.

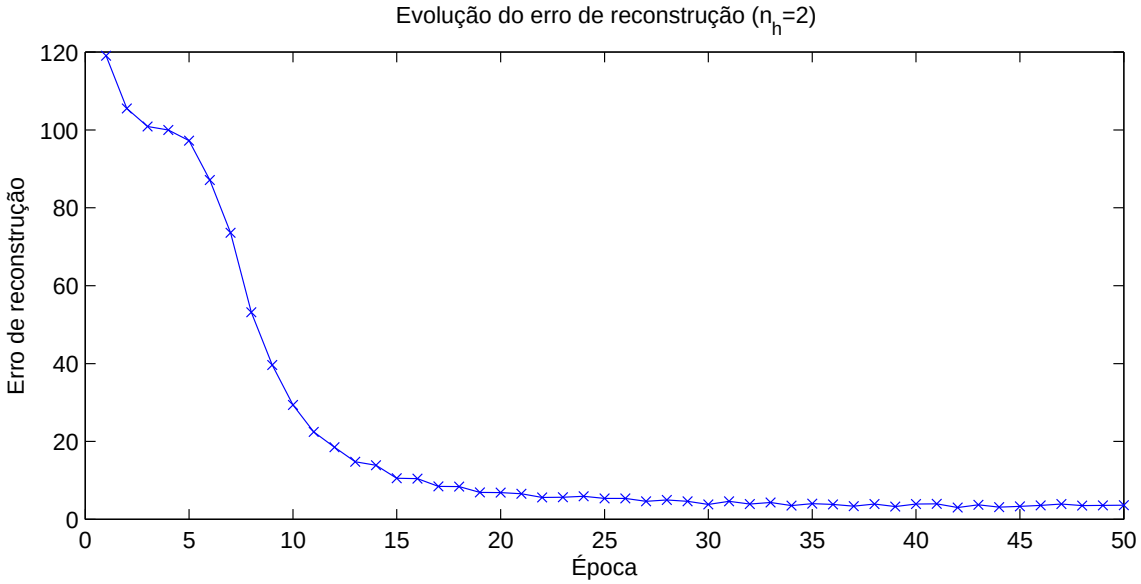
Como explicado anteriormente, no algoritmo CD (assim como nos outros a serem apresentados) busca-se maximizar a verossimilhança dos dados de treinamento. Entretanto, como a avaliação de $P(\mathbf{v})$ é intratável, o mesmo ocorre com a função perda definida em (2.22). Surge, portanto, a questão de como se acompanhar a evolução do treinamento,

dada a dificuldade em se avaliar a função perda que estaria sendo minimizada. A técnica mais comumente empregada para esse acompanhamento consiste no uso de uma medida denominada erro de reconstrução, definida como

$$\varepsilon_r = \left|\left|\tilde{\mathbf{v}} - \mathbf{v}^{(t)}\right|\right|^2, \tag{2.32}$$

em que a norma utilizada é a Euclidiana. Nessa equação, considerou-se o erro relativo a uma única amostra de treinamento. Comumente, a soma dessa medida sobre todas amostras do conjunto de treinamento é visualizada ao longo das épocas de treinamento, como na Figura 8.

Figura 8 – Evolução do erro de reconstrução para o treinamento da Figura 7.



Fonte: produção do próprio autor.

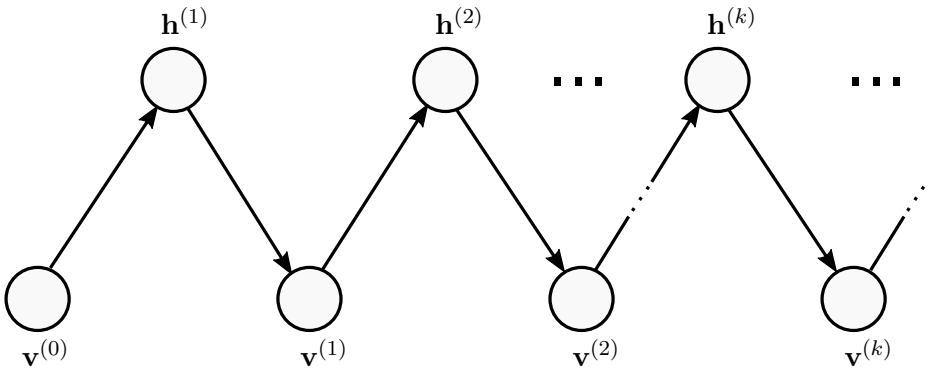
Devido ao seu uso frequente, vale aqui uma digressão a respeito do erro de reconstrução. Embora essa medida seja usada com diversos algoritmos de treinamento, ela é particularmente útil e mais facilmente interpretável no contexto do algoritmo CD-k. Com já explicado, segundo o algoritmo CD, a partícula negativa, $\tilde{\mathbf{v}}$, é obtida executando-se k passos de Gibbs a partir da amostra de treinamento $\mathbf{v}^{(t)}$. A amostra $\tilde{\mathbf{v}}$ pode ser interpretada como uma reconstrução de $\mathbf{v}^{(t)}$ segundo o raciocínio que será exposto a seguir. No primeiro passo de Gibbs (conforme Figura 5), que consiste na sequência de amostragens $\mathbf{v}^{(t)} \rightarrow \mathbf{h}^{(1)} \rightarrow \mathbf{v}^{(1)}$, $\mathbf{h}^{(1)}$ pode ser vista como uma representação interna (latente) da amostra $\mathbf{v}^{(t)}$, uma vez que a primeira tem suas características estatísticas determinadas pela segunda, conforme $P(\mathbf{h}|\mathbf{v}^{(t)})$ (Equação (2.16)). Já a amostra $\mathbf{v}^{(1)}$ pode ser considerada como uma primeira reconstrução de $\mathbf{v}^{(t)}$ a partir de $\mathbf{h}^{(1)}$ por dois motivos: em primeiro lugar, porque $\mathbf{h}^{(1)}$ define a distribuição de $\mathbf{v}^{(1)}$, conforme $P(\mathbf{v}|\mathbf{h}^{(1)})$ (Equação (2.18)); e, em segundo lugar, pela observação de que na proporção em que o modelo melhor captura as características

da distribuição de treinamento, as amostras por ele produzidas deveriam também ser mais parecidas com as de treinamento, ao menos em termos médios. Empregando-se esse raciocínio para os passos de Gibbs posteriores, chega-se à já mencionada conclusão sobre $\tilde{\mathbf{v}}$. Vale destacar que o erro de reconstrução não está diretamente relacionado com o objetivo de treinamento. Tipicamente, entretanto, aumentos recorrentes no erro de reconstrução durante o treinamento indicam a divergência do algoritmo – daí a utilidade da medida.

Algoritmo PCD

O algoritmo PCD (*Persistent Contrastive Divergence*) foi proposto por Tieleman (2008). A técnica utilizada no algoritmo, entretanto, teve sua efetividade demonstrada anteriormente por Neal (1992) para o caso de Máquinas de Boltzmann (não restritas). A ideia básica no PCD é remover a limitação, como ocorre com o CD-k, no número de passos de Gibbs usados para se obter as partículas negativas. Para isso, ao invés de executar uma curta cadeia de amostragem de Gibbs em cada iteração do algoritmo de treinamento, mantém-se uma única cadeia de amostragem ao longo de todas iterações (ou seja, durante todo o treinamento) e retiram-se amostras dessa cadeia a cada passo de Gibbs executado. Após um número de iterações, espera-se a produção de amostras bastante representativas do modelo. Deve-se notar que essa conclusão assume que a amostragem seja feita num modelo com parâmetros constantes, o que não é verdade durante o treinamento. Assim, no PCD supõe-se que as mudanças na distribuição devidas às atualizações de parâmetros ocorram lentamente, de modo que a cadeia de amostragem possa acompanhá-las. A Figura 9 ilustra a cadeia de amostragem do PCD, da qual cada uma das amostras $\mathbf{v}^{(k)}, k = 1, 2, \ldots,$ seria tomada como partícula negativa ($\mathbf{v}^{(0)}$ seria escolhida aleatoriamente).

Figura 9 – Esquema de amostragem do PCD.



Fonte: produção do próprio autor.

Na Tabela 2, é fornecido o pseudocódigo do PCD. A única diferença em relação ao CD-k está na produção da partícula negativa.

Tabela 2 – Iteração do algoritmo PCD

$\mathbf{v}^{(t)}$ $\leftarrow$ *Amostra (partícula positiva) obtida do conjunto de treinamento*;
$\widetilde{\mathbf{v}}$ $\leftarrow$ *Amostra (partícula negativa) obtida através de um passo de Gibbs continuando a cadeia de amostragem a partir do estado anterior,* $\widetilde{\mathbf{v}}$;
Uso das regras de atualização de parâmetros:
forall $j \in \{1, \ldots, n_h\}$ **do**
  $b_j \leftarrow b_j + \lambda\Big[P(h_j{=}1|\mathbf{v}^{(t)}) - P(h_j{=}1|\widetilde{\mathbf{v}})\Big]$;
end
forall $i \in \{1, \ldots, n_v\}$ **do**
  $c_i \leftarrow c_i + \lambda\Big[v_i^{(t)} - \widetilde{v}_i\Big]$;
end
forall $(i, j) \in \{1, \ldots, n_v\} \times \{1, \ldots, n_h\}$ **do**
  $w_{ji} \leftarrow w_{ji} + \lambda\Big[P(h_j{=}1|\mathbf{v}^{(t)})\, v_i^{(t)} - P(h_j{=}1|\widetilde{\mathbf{v}})\, \widetilde{v}_i\Big]$;
end

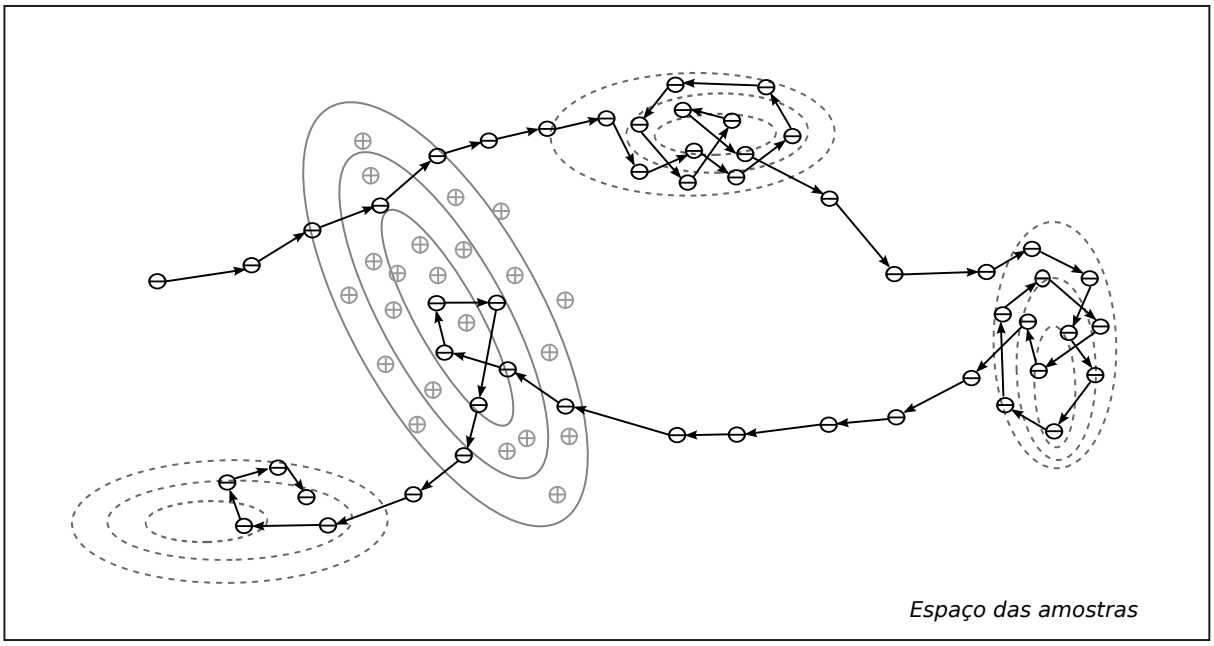
Na Figura 10, ilustra-se (com a mesma representação usada na Figura 6 para o CD-k) o efeito do esquema de amostragem do PCD na geração de partículas negativas. Percebe-se que, no PCD, existe uma única cadeia de amostragem da qual são retiradas as partículas negativas e que as mesmas não têm relação direta com as positivas. A continuidade da cadeia de amostragem permite que modos da distribuição do modelo não alcançáveis no CD-k sejam agora atingidos (e, consequentemente, possam ter suas probabilidades diminuídas na atualização de parâmetros). Outro efeito visível, uma característica da amostragem de Gibbs, é que há uma tendência para a geração sequencial de partículas negativas dentro de uma região de concentração de massas de probabilidade segundo o modelo. Esse efeito pode ser considerado inconveniente na medida em que ele torna mais lenta (ou até mesmo impede) a exploração, por parte das partículas negativas, de outros modos que possam estar presentes na distribuição do modelo. Como será visto à frente, o algoritmo chamado PT busca remediar essa situação.

Na Figura 11, é mostrado o resultado de uma simulação análoga àquela mostrada para o CD, empregando agora o algoritmo PCD, n_v=3, n_h=2, λ=0,01 (frequentemente o PCD opera melhor com taxas de aprendizagem menores do que as usadas com o CD), e uma distribuição de treinamento na forma de uma rampa (de probabilidades). Verifica-se novamente que a RBM resultante captura e reproduz as características dos dados de treinamento.

Algoritmo PT

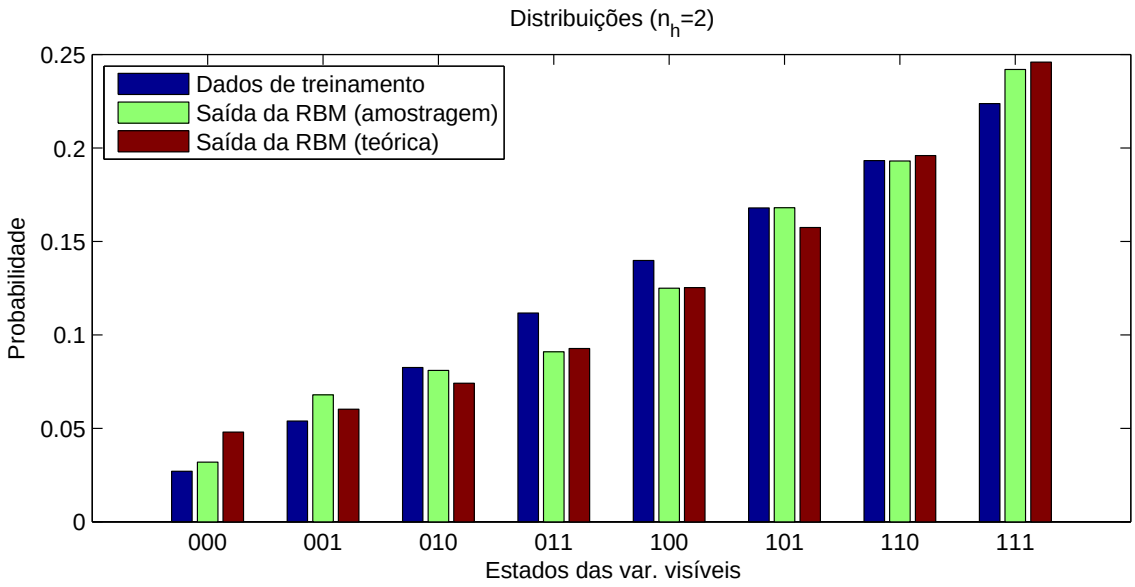
O uso do método conhecido como PT (*Parallel Tempering*) para treinamento de RBMs foi proposto por Desjardins et al. (2010). Segundo Earl e Deem (2005), as origens da ideia propriamente dita podem ser rastreadas ao trabalho de Swendsen e Wang (1987)

Figura 10 – Geração de partículas negativas conforme PCD.



Fonte: produção do próprio autor.

Figura 11 – Distribuições de probabilidade dos dados de treinamento e do modelo resultante para RBM Bernoulli-Bernoulli treinada com o algoritmo PCD.



Fonte: produção do próprio autor.

em simulações de mecânica estatística. No trabalho, fora introduzido um método de Monte Carlo em que réplicas de um sistema de interesse eram simuladas a diferentes temperaturas com eventuais trocas de informações de estado entre essas réplicas. Ao longo do tempo, o mesmo método passou a ser usado em muitas áreas incluindo química, biologia, engenharia e ciência dos materiais.

Nas RBMs, a utilização do PT teria o objetivo de evitar que o processo de amostragem responsável pela geração de partículas negativas ficasse preso em regiões de maior probabilidade do modelo. A analogia das RBMs a sistemas físicos torna bastante natural a formulação de uma versão do modelo ao qual se associe uma temperatura. Especificamente, a Equação (2.15) sugere a definição da função massa de probabilidade conjunta (adaptado de (DESJARDINS et al., 2010))

$$P_{T_m}(\mathbf{v}, \mathbf{h}) = \frac{\exp(-E(\mathbf{v}, \mathbf{h})/T_m)}{Z(T_m)}, \tag{2.33}$$

que representa uma RBM a uma temperatura T_m. Como já mencionado com relação a distribuições de Boltzmann em geral, o efeito da elevação da temperatura é o de suavizar a distribuição tornando-a, no limite ($T_m \to \infty$), uniforme.

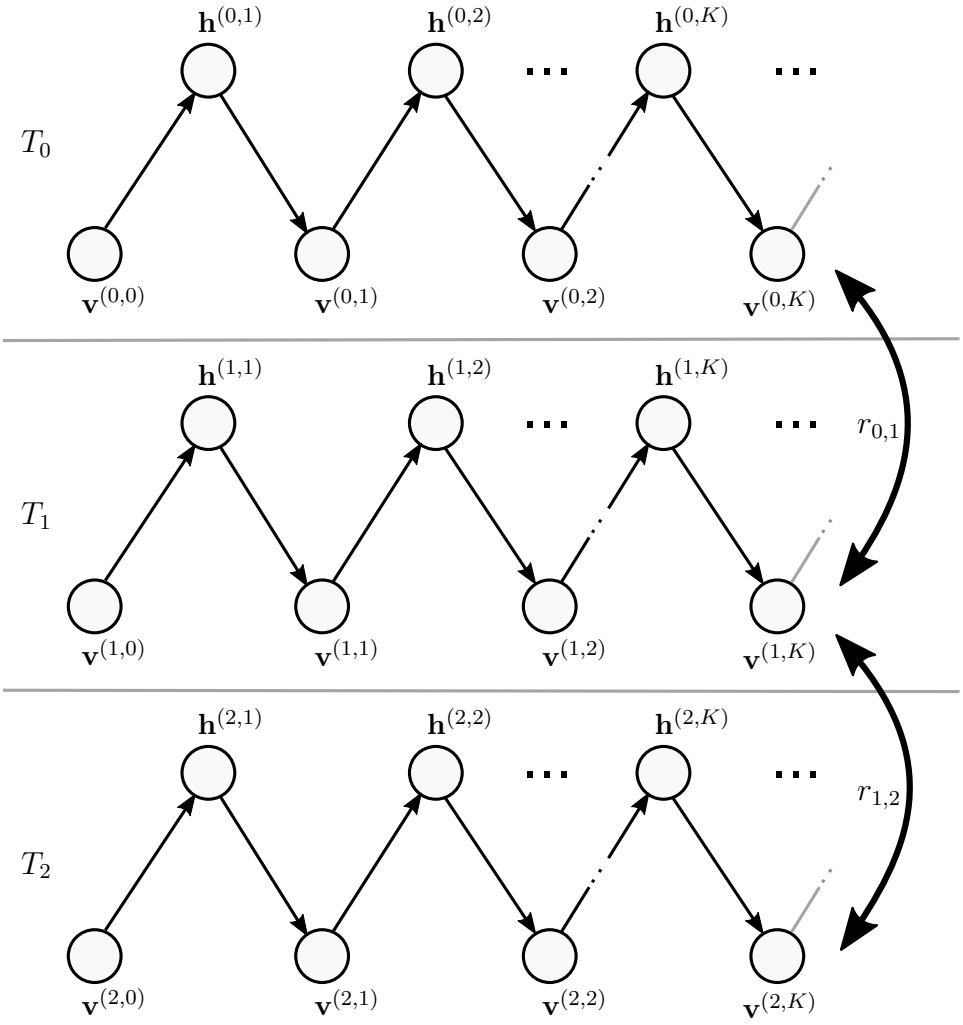
A Figura 12 ilustra o esquema de amostragem segundo o PT para o caso em que duas réplicas do sistema são simuladas. A amostragem do sistema (RBM) original é representada pela cadeia de Gibbs sendo executada à temperatura $T_0 = 1$. Similarmente ao PCD, somente dessa cadeia seriam extraídas partículas negativas $(\mathbf{v}^{(0,1)}, \mathbf{v}^{(0,2)}, \ldots)$ para serem utilizadas na atualização de parâmetros. As réplicas estão representadas pelas cadeias executadas às temperaturas T_1 e T_2. Em geral, assume-se que $T_k < T_{k+1}$, para $k = 0, 1, \ldots, M$, em que M representa o número de réplicas. Em temperaturas mais elevadas, o problema devido à concentração de massas de probabilidade é dirimido, mas a distribuição produzida não é aquela em que se tem interesse. Surge, portanto, a questão de como se empregar as amostras produzidas nessas cadeias. No PT, faz-se a troca de partículas entre cadeias adjacentes (de índices m e $m+1$) segundo a probabilidade (regra de Metropolis)

$$r_{m,m+1} = \min\left[\frac{P_{T_m}\left(\mathbf{v}^{(m+1,k)}\right)P_{T_{m+1}}\left(\mathbf{v}^{(m,k)}\right)}{P_{T_m}\left(\mathbf{v}^{(m,k)}\right)P_{T_{m+1}}\left(\mathbf{v}^{(m+1,k)}\right)}, 1\right], \tag{2.34}$$

em que k indica o número do passo de Gibbs e os índices entre parênteses sobrescritos às amostras indicam a cadeia e o passo de Gibbs (respectivamente) das quais as amostras se originam. É interessante interpretar a Equação (2.34). Na divisão envolvendo probabilidades, pode-se entender o denominador como sendo a probabilidade de se observar simultaneamente duas partículas segundo as distribuições em que cada uma delas foi produzida. Em contrapartida, no numerador mede-se a probabilidade de se observar as mesmas partículas se cada uma delas tivesse sido produzida pela outra distribuição. Assim, pode-se dizer que a regra em questão favorece a realização das trocas se essas resultarem em partículas melhor explicadas pelas novas distribuições. Na Figura 12, as potenciais trocas de partículas entre cadeias foram indicadas pelas setas bidirecionais no passo de Gibbs de índice K.

Exceto pela mudança no passo de geração de partículas negativas, o pseudocódigo do algoritmo PT é idêntico ao do PCD e, por isso, não será repetido aqui. A geração das

Figura 12 – Esquema de amostragem do PT.



Fonte: produção do próprio autor (inspirado em Cho, Ilin e Raiko (2011)).

partículas segundo o esquema de amostragem do PT é detalhada na Tabela 3, em que as potenciais trocas de partículas são consideradas a cada K passos de amostragem, isto é, periodicamente, como feito comumente na prática.

Na Figura 13, é ilustrado (utilizando novamente a representação usada na Figura 6 no caso do CD-k) o efeito do esquema de amostragem do PT na geração de partículas negativas. Na figura, duas cadeias são executadas em paralelo. A primeira delas, identificada pela cor azul, representa o sistema original, isto é, a RBM à temperatura T_0=1. A segunda cadeia, indicada pela cor vermelha, representa uma réplica do sistema à uma temperatura $T_1 > T_0$. Pode-se perceber que as partículas negativas são extraídas sempre da cadeia à temperatura T_0 e que, como no PCD, essas partículas não têm relação direta com as amostras de treinamento (partículas positivas). Tentou-se também mostrar o fato de que a cadeia a uma maior temperatura teria um caminho mais livre no espaço de amostras. A

Tabela 3 – Geração de partícula negativa segundo PT (k-ésimo passo de Gibbs).

Passos de Gibbs das cadeias individuais:
forall $m \in \{0, \ldots, M\}$ **do**
  $\mathbf{v}^{(m,k)} \leftarrow$ *Amostra obtida através de um passo de Gibbs continuando a cadeia de amostragem a partir de* $\mathbf{v}^{(m,k-1)}$ *em uma RBM à temperatura* T_m*;*
end

Troca periódica de partículas entre cadeias:
if k *é múltiplo de* K **then**
  $m \leftarrow M - 1$;
  while $m \geq 0$ **do**
    Trocar partículas $\mathbf{v}^{(m,k)} \leftrightarrow \mathbf{v}^{(m+1,k)}$ *com probabilidade* $r_{m,m+1}$ (*Equação* (2.34));
    $m \leftarrow m - 1$;
  end
end

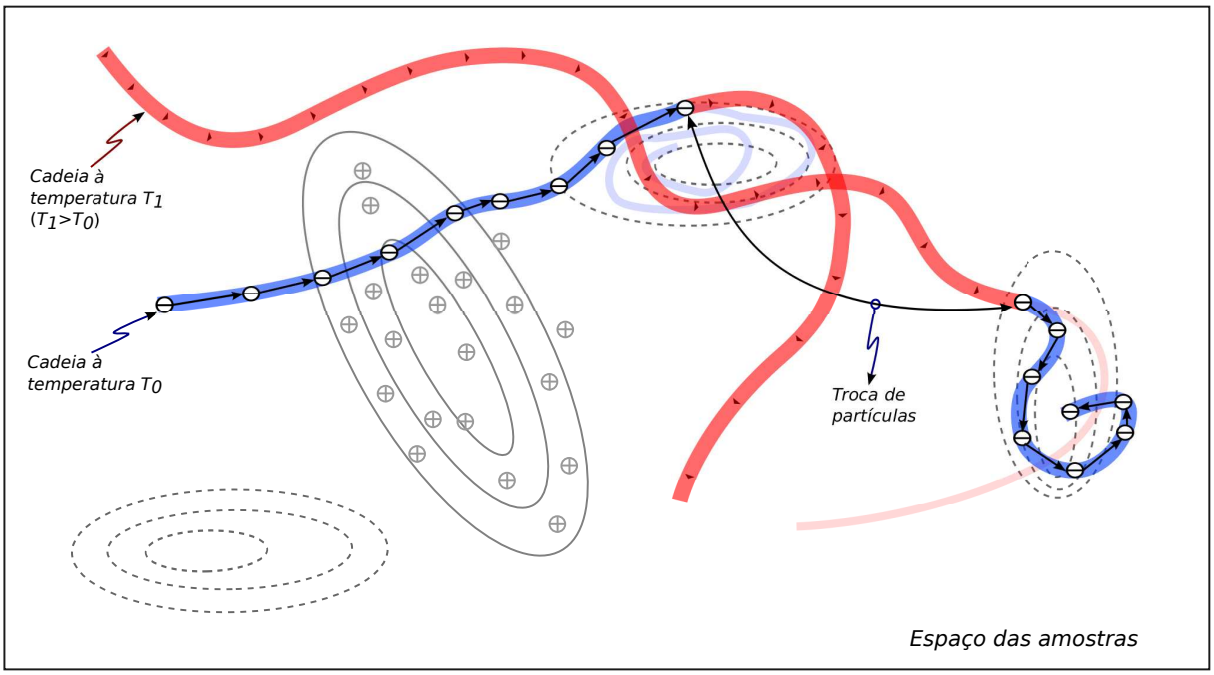
Retornar $\mathbf{v}^{(0,k)}$ *como resultado da amostragem*;

Informações adicionais:
- *Assume-se que* k*, o número do passo de Gibbs atual, seja fornecido*;
- *Assume-se que o conjunto de estados das cadeias,* $\{\mathbf{v}^{(0,k)}, \mathbf{v}^{(1,k)}, \ldots, \mathbf{v}^{(M,k)}\}$*, seja mantido entre execuções desse algoritmo*;
- *Considera-se que o os estados iniciais das cadeias,* $\mathbf{v}^{(0,0)}, \mathbf{v}^{(1,0)}, \ldots, \mathbf{v}^{(M,0)}$*, sejam iniciados com valores aleatórios antes da primeira execução desse algoritmo*;
- *Com a substituição* $\{\mathbf{W}, \mathbf{b}, \mathbf{c}\} \rightarrow \{(1/T_m)\mathbf{W}, (1/T_m)\mathbf{b}, (1/T_m)\mathbf{c}\}$*, chega-se a uma RBM à temperatura* T_m *a partir de uma à temperatura ambiente.*

troca de partículas identificada na figura demonstra como o esquema de amostragem do PT permite que a cadeia à temperatura T_0 escape de uma região de alta probabilidade segundo o modelo. No exemplo, não fosse a troca de partículas, essa cadeia teria seguido o caminho em azul claro, como ocorreria no PCD. É interessante perceber também que a regra de Metropolis desfavorece trocas em que uma cadeia esteja em região de probabilidade alta enquanto a outra esteja numa região de probabilidade muito baixa. Portanto, a troca identificada na figura representa uma situação comum na qual as cadeias saltam entre modos da distribuição ou, em outras palavras, entre regiões representativas da distribuição do modelo, uma característica bastante desejável.

Para ilustrar as vantagens do PT em relação ao PCD, é mostrado aqui um experimento extraído de Desjardins et al. (2010) em que uma RBM tendo n_h=500 unidades ocultas é treinada com ambos os algoritmos (PCD e PT). Os dados de treinamento são imagens de dígitos escritos à mão tendo 28×28 pixels de tamanho. Deve-se notar que, nesse caso, o vetor de estado das variáveis visíveis, $\mathbf{v}$, tem dimensão n_v=28^2 e representa uma imagem completa. Após o treinamento, amostras são produzidas a partir dos modelos obtidos empregando, para isso, o mesmo processo de geração de partículas negativas dos respectivos algoritmos. Na Figura 14, são apresentadas amostras produzidas pelo processo

Figura 13 – Geração de partículas negativas conforme PT.



Fonte: produção do próprio autor (inspirado em Earl e Deem (2005)).

descrito, sendo que imagens consecutivas (ao longo das linhas) são separadas entre si por 50 passos de Gibbs. Do lado esquerdo da figura, tem-se os resultados para o PCD. Observando-se as amostras produzidas por esse algoritmo, percebe-se que o mesmo tem uma propensão a produzir imagens parecidas ou amostras que tendem a vir de um mesmo modo da distribuição representada pelo modelo. Já para o PT, do lado direito da figura, as imagens consecutivas não se repetem em grande quantidade. Na literatura, costuma-se dizer que o esquema de amostragem do PT provê uma melhor taxa de mistura (*mixing rate*) do que a do PCD, o que, por motivos já explicados, é visto como uma vantagem do PT sobre o PCD.

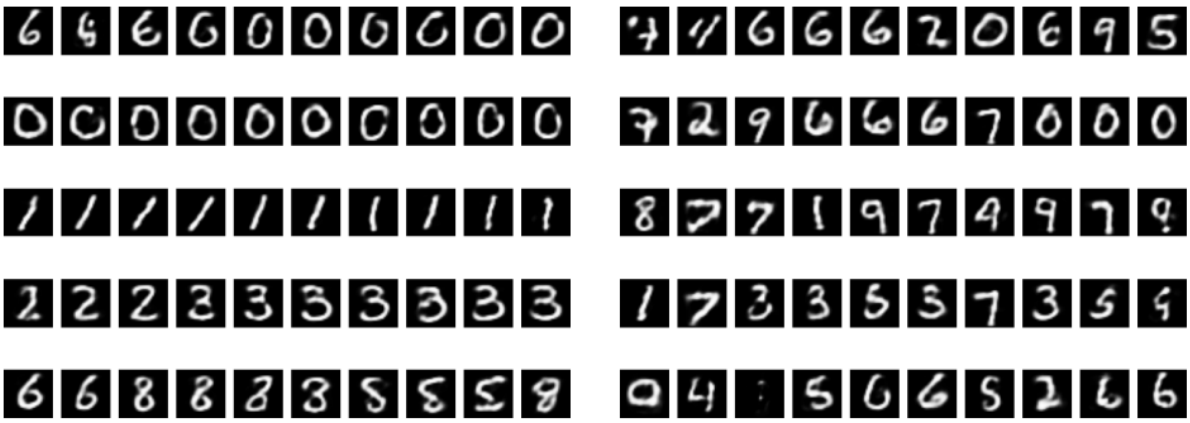
2.1.4 Relações com outros modelos

Nesta seção, as RBMs são relacionadas com alguns modelos existentes na literatura.

Máquinas de Boltzmann e Produtos de Especialistas

Como mencionado na Seção 2.1.1, as RBMs são um caso particular de Máquinas de Boltzmann (ACKLEY; HINTON; SEJNOWSKI, 1985) em que não existem conexões entre unidades de uma mesma camada. A ausência de tais conexões, por sua vez, leva à independência entre as variáveis ocultas dado o estado das visíveis, conforme Equação (2.16). Consequentemente, uma RBM pode ser vista como um *Produto de Especialistas* (HINTON, 2002) com um especialista por unidade oculta. De fato, Hinton (2002) aponta que as RBMs

Figura 14 – Amostras produzidas pelos algoritmos PCD (esquerda) e PT (direita).



Fonte: extraído de Desjardins et al. (2010).

são a intersecção entre as Máquinas de Boltzmann e os Produtos de Especialistas, duas classes bem diferentes de modelos.

Campos Aleatórios de Markov

Uma Máquina de Boltzmann é considerada também como um modelo gráfico probabilístico, visto que o modelo pode ser expresso por um grafo em que as conexões representam as dependências condicionais entre variáveis aleatórias. Mais ainda, como as conexões entre as variáveis são não direcionadas, as Máquinas de Boltzmann (e, portanto, as RBMs) pertencem a um conjunto específico de modelos gráficos probabilísticos conhecidos como Campos Aleatórios de Markov (MRF – *Markov Random Field*) (FISCHER; IGEL, 2014).

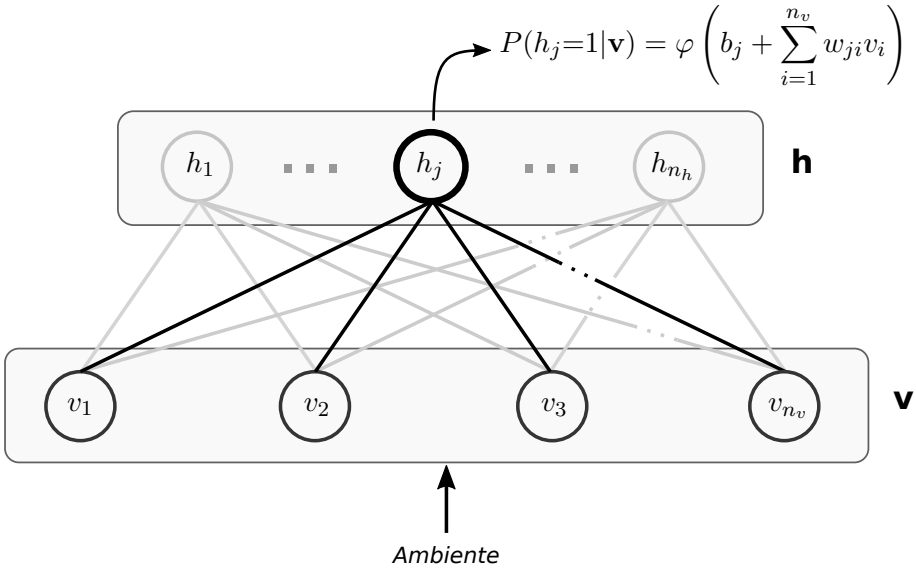
Redes Neuronais Artificiais

As RBMs podem ainda ser comparadas com as redes neuronais clássicas, que representam modelos determinísticos. Na Figura 15, é mostrada uma RBM em que uma unidade, h_j, está destacada juntamente com as conexões da mesma com as unidades visíveis. Considerando agora as unidades como variáveis aleatórias, tem-se que, fixado o estado da camada visível, $\mathbf{v}$, as variáveis ocultas são independentes entre si e, portanto, podem ser analisadas individualmente. Na mesma condição (fixado $\mathbf{v}$), a probabilidade de ativação da variável H_j é dada pela Equação (2.17) e é indicada na figura. Para comparação entre os modelos, na Figura 16 é apresentado um neurônio clássico tendo a função sigmóide, $\varphi(\cdot)$, como função de ativação. A saída do neurônio é dada por

$$y_j = \varphi\left(b_j + \sum_{i=1}^{m} w_{ji} x_i\right), \tag{2.35}$$

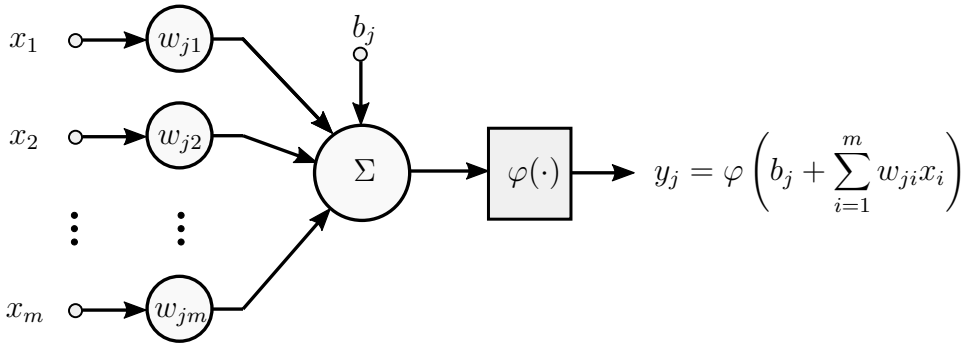
também indicada na figura. Comparando-se as Figuras 15 e 16, percebe-se a similaridade das estruturas e nota-se que as expressões para $P(h_j{=}1|\mathbf{v})$ e y_j são análogas. Entretanto, no neurônio clássico, y_j representa efetivamente a saída (valor real) do neurônio, enquanto que, na RBM, $P(h_j{=}1|\mathbf{v})$ expressa a probabilidade de que a variável H_j assuma o valor 1. Assim, cada unidade oculta da RBM pode ser vista como uma versão estocástica de neurônio. É importante lembrar que, diferentemente do neurônio clássico, as RBMs têm um grafo não direcionado, de modo que a mesma conclusão seria obtida com relação às unidades visíveis se as unidades ocultas fossem consideradas como suas entradas. Com essa interpretação, pode-se dizer que uma RBM é uma estrutura formada por neurônios estocásticos (PAPA et al., 2015).

Figura 15 – RBM com uma unidade oculta isolada.



Fonte: produção do próprio autor.

Figura 16 – Neurônio clássico.



Fonte: Haykin (2009).

No processo conhecido como pré-treinamento de redes neuronais profundas, as

DNNs, os parâmetros de modelos construídos pelo empilhamento de RBMs são transportados para um rede neuronal *feed-forward* (HINTON et al., 2012). Essa técnica é justificada pela analogia que acaba de ser descrita entre as unidades das RBMs e os neurônios.

2.1.5 Variantes do modelo

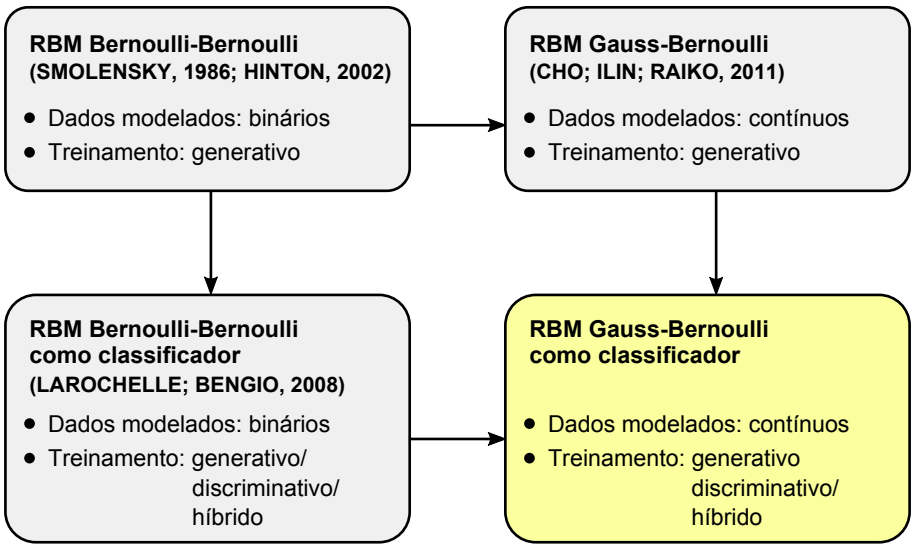
Na sua forma original, as unidades visíveis de uma RBM são binárias e, consequentemente, os dados por ela modelados também são binários. Felizmente, modificando-se a definição da função de energia global, pode-se obter variantes do modelo com a capacidade de representar distribuições contínuas, ampliando, assim, o espectro de aplicações para esse tipo de modelo. Uma formulação para a energia global baseada em estatísticas suficientes atribui distribuições da família exponencial às unidades visíveis e ocultas (WELLING; ROSEN-ZVI; HINTON, 2004). Outras opções incluem o emprego de unidades ocultas binárias em conjunto com unidades visíveis com distribuição do tipo exponencial truncada (BENGIO et al., 2007) ou então Gaussiana (CHO; ILIN; RAIKO, 2011). Existem ainda extensões para uso com dados categóricos (SALAKHUTDINOV; MNIH; HINTON, 2007) e ordinais (PHUNG; VENKATESH et al., 2009). Como será visto, a facilidade para se obter diferentes variantes de RBMs assinalam a considerável flexibilidade desse modelo.

Na Figura 17, são indicadas as versões de RBM discutidas neste texto. A RBM Bernoulli-Bernoulli introduzida na atual seção representa a versão original do modelo. A RBM Gauss-Bernoulli (Seção 2.2) modifica a Bernoulli-Bernoulli no sentido de permitir a modelagem de dados contínuos, enquanto que a versão indicada aqui como "RBM Bernoulli-Bernoulli como classificador" (Seção 2.3) adiciona ao modelo original uma camada de classificação e novas formas de treinamento. Por fim, a variante "RBM Gauss-Bernoulli como classificador" é proposta no presente trabalho e une as características das duas variantes mencionadas. Vale reiterar que todas essas versões do modelo são obtidas modificando-se a definição da função de energia global.

2.2 RBM GAUSS-BERNOULLI

Nesta seção, é introduzida a variante denominada RBM *Gauss-Bernoulli* ou, abreviadamente, GBRBM (*Gaussian-Bernoulli Restricted Boltzmann Machine*) cujas unidades visíveis são Gaussianas e as ocultas, binárias. Considera-se aqui que os vários conceitos introduzidos com as RBMs Bernoulli-Bernoulli sejam conhecidos, de modo que o foco na presente seção são as diferenças decorrentes da mudança na definição da energia global que leva às GBRBMs. A organização dos tópicos a seguir é similar à da Seção 2.1. Nas Seções 2.2.1 a 2.2.3, são vistas a estrutura, a descrição matemática e as tarefas associadas ao modelo, nessa ordem. A Seção 2.2.4 ilustra algumas características do modelo por meio de simulações computacionais. Ao final, a Seção 2.2.5 compara a GBRBM com outros modelos.

Figura 17 – Modelos descritos no documento.



Fonte: produção do próprio autor.

2.2.1 Estrutura

Do ponto de vista estrutural, as GBRBMs são idênticas às RBMs originais (Figura 2): o modelo é formado por um conjunto de unidades ocultas e outro de unidades visíveis separados nas respectivas camadas. As conexões entre camadas opostas, como em toda Máquina de Boltzmann, são não direcionadas. Em razão de uma nova formulação para a energia global, surge um novo conjunto de parâmetros associados às unidades visíveis, como será visto a seguir.

2.2.2 Descrição matemática

No presente texto, considera-se a formulação para a GBRBM introduzida por Cho, Ilin e Raiko (2011), na qual a função de energia global é definida como

$$E_g(\mathbf{v}, \mathbf{h}) = -\sum_{j=1}^{n_h} b_j h_j + \sum_{i=1}^{n_v} \frac{(v_i - c_i)^2}{2\sigma_i^2} - \sum_{i=1}^{n_v}\sum_{j=1}^{n_h} h_j w_{ji} \frac{v_i}{\sigma_i^2}. \tag{2.36}$$

Nessa expressão, todos os símbolos comuns à função de energia global das RBMs (Equação (2.8)) mantêm os seus significados. Surge agora um novo conjunto de parâmetros, as variâncias das unidades visíveis, referenciadas por σ_i^2, com $i = 1, \ldots, n_v$. As definições dadas nas eqs. (2.1) a (2.7) são reutilizadas aqui. Adicionalmente, define-se o vetor de variâncias como

$$\boldsymbol{\sigma}^2 = [\sigma_1^2 \ldots \sigma_{n_v}^2]^T, \tag{2.37}$$

de modo que o modelo fica totalmente especificado pelo conjunto de parâmetros $\{\mathbf{W}, \mathbf{b}, \mathbf{c}, \boldsymbol{\sigma}^2\}$.

Como em todas versões de RBMs, a probabilidade associada a um dado estado das variáveis, $(\mathbf{v}, \mathbf{h})$, decai exponencialmente com a energia global, de modo que a distribuição

conjunta do modelo é definida como

$$P_g(\mathbf{v}, \mathbf{h}) = \frac{\exp(-E_g(\mathbf{v}, \mathbf{h}))}{Z_g}. \tag{2.38}$$

Nota-se que essa definição tem a mesma forma daquela fornecida na Equação (2.9) para as RBMs Bernoulli-Bernoulli. Entretanto, $P_g(\mathbf{v}, \mathbf{h})$ tem um domínio misto, contínuo para os estados das variáveis visíveis e discreto para as ocultas, dado explicitamente por $\mathbb{R}^{n_v} \times \{0, 1\}^{n_h}$. A constante no denominador (função de partição) pode, portanto, ser escrita como

$$Z_g = \int_{\mathbb{R}^{n_v}} \sum_{\mathbf{h} \in \{0,1\}^{n_h}} \exp(-E_g(\mathbf{v}, \mathbf{h}))\, \mathbf{dv}. \tag{2.39}$$

A partir de (2.36) e (2.38), todos os resultados antes obtidos para RBMs podem ser derivados para o caso das GBRBMs. As demonstrações são apresentadas no Apêndice B.

2.2.3 Tarefas associadas ao modelo

Como será mostrado nos tópicos seguintes, as tarefas de inferência (condicional entre variáveis do modelo), amostragem e treinamento apresentadas para as RBMs originais podem ser facilmente adaptadas para as GBRBMs. Vale dizer que as várias discussões e interpretações antes feitas para as RBMs continuam válidas para as GBRBMs e, por isso, não são repetidas aqui.

2.2.3.1 Inferência

Para realizar inferências condicionais entre variáveis visíveis e ocultas, deve-se conhecer as expressões de $P_g(\mathbf{h}|\mathbf{v})$ e $P_g(\mathbf{v}|\mathbf{h})$. Nas GBRBMs, também vale a independência condicional das variáveis de uma camada dado o estado da camada oposta, ou seja,

$$P_g(\mathbf{h}|\mathbf{v}) = \prod_{j=1}^{n_h} P_g(h_j|\mathbf{v}) \tag{2.40}$$

e

$$P_g(\mathbf{v}|\mathbf{h}) = \prod_{i=1}^{n_v} P_g(v_i|\mathbf{h}). \tag{2.41}$$

Fixando-se $\mathbf{v}$, cada variável oculta tem distribuição de Bernoulli (como nas RBMs) com probabilidade de ativação (ou sucesso)

$$P_g(h_j{=}1|\mathbf{v}) = \varphi\left(b_j + \sum_{i=1}^{n_v} w_{ji} \frac{v_i}{\sigma_i^2}\right). \tag{2.42}$$

Além disso, fixado $\mathbf{h}$, cada uma das variáveis visíveis tem distribuição Gaussiana dada por

$$P_g(v_i|\mathbf{h}) = \mathcal{N}\left(v_i \middle| c_i + \sum_{j=1}^{n_h} h_j w_{ji}, \sigma_i^2\right), \tag{2.43}$$

em que a notação $\mathcal{N}(\cdot|\mu, \sigma^2)$ representa a função de densidade de probabilidade Gaussiana com média μ e variância σ^2.

2.2.3.2 Amostragem

O simples uso das eqs. (2.40) a (2.43) no amostrador de Gibbs permite que a amostragem seja realizada nas GBRBMs.

2.2.3.3 Treinamento

Como demonstrado no Apêndice B, as regras de atualização para os vieses e pesos de conexão de uma GBRBM são as seguintes:

$$b_j \leftarrow b_j + \lambda\Big[P_g(h_j{=}1|\mathbf{v}^{(t)}) - P_g(h_j{=}1|\widetilde{\mathbf{v}})\Big], \tag{2.44}$$

$$c_i \leftarrow c_i + \frac{\lambda}{\sigma_i^2}\left[v_i^{(t)} - \widetilde{v}_i\right], \tag{2.45}$$

$$w_{ji} \leftarrow w_{ji} + \frac{\lambda}{\sigma_i^2}\left[P_g(h_j{=}1|\mathbf{v}^{(t)})\, v_i^{(t)} - P_g(h_j{=}1|\widetilde{\mathbf{v}})\, \widetilde{v}_i\right]. \tag{2.46}$$

Na atualização das variâncias, utiliza-se um artifício com o objetivo de assegurar que esses parâmetros sejam sempre positivos, algo que não é naturalmente garantido pelo SGD. Para todo $i = 1, \ldots, n_v$, define-se um parâmetro auxiliar, $z_i = \ln(\sigma_i^2)$, o qual é efetivamente atualizado via SGD. Da definição de z_i, tem-se que $\sigma_i^2 = e^{z_i}$, de modo que, para todo valor (real) de z_i, as variâncias serão positivas, como desejado. A atualização dos parâmetros σ_i^2 é realizada pela execução em sequência das seguintes regras:

$$z_i \leftarrow \ln(\sigma_i^2), \tag{2.47}$$

$$\begin{aligned} z_i \leftarrow z_i + \frac{\lambda}{\sigma_i^2}\Bigg[& \left(\frac{(v_i^{(t)} - c_i)^2}{2} - P_g(h_j{=}1|\mathbf{v}^{(t)})\, w_{ji} v_i^{(t)}\right) \\ & - \left(\frac{(\widetilde{v}_i - c_i)^2}{2} - P_g(h_j{=}1|\widetilde{\mathbf{v}})\, w_{ji}\widetilde{v}_i\right)\Bigg], \end{aligned} \tag{2.48}$$

$$\sigma_i^2 \leftarrow \exp(z_i). \tag{2.49}$$

Embora as regras de atualização para as variâncias sejam fornecidas aqui, deve-se dizer que, na prática, procura-se evitar o ajuste desses parâmetros fazendo-se, para isso, uma normalização prévia dos dados de treinamento. Essa restrição é normalmente aplicada porque a atualização das variâncias frequentemente faz com que o treinamento divirja, ou seja, se afaste recorrentemente do seu objetivo de minimizar a função perda. Em todas as regras fornecidas, $v_i^{(t)}$ e $\widetilde{v}_i$ representam as i-ésimas componentes do vetores $\mathbf{v}^{(t)}$ e $\widetilde{\mathbf{v}}$, respectivamente, sendo que $\mathbf{v}^{(t)}$ é a amostra de treinamento e $\widetilde{\mathbf{v}}$, a partícula negativa; $P_g(h_j{=}1|\mathbf{v})$ é dada na Equação (2.42). Com as regras de atualização apresentadas, é possível utilizar qualquer um dos algoritmos de treinamento detalhados na Seção 2.1.3.3 (para as RBMs Bernoulli-Bernoulli). Para tanto, vale lembrar que os algoritmos estudados (CD-k, PCD e PT) se diferenciam pela técnica usada na produção de partículas negativas, ou seja, pela forma como realizam a amostragem, a qual já foi descrita para as GBRBMs. O

treinamento e a amostragem do modelo são ilustrados na Seção 2.2.4. Como será visto, as GBRBMs conseguem capturar as características probabilísticas dos dados de treinamento, inclusive dependências entre as variáveis que os representam.

2.2.4 Experimentos

Os resultados de algumas simulações envolvendo GBRBMs serão mostrados a seguir. Na Seção 2.2.4.1, o modelo é treinado num caso simples, em que os dados de treinamento são independentes entre si. Aproveita-se o exemplo para apontar algumas dificuldades que são encontradas na prática no treinamento de GBRBMs. Ilustra-se também o uso do erro de reconstrução (introduzido na pág. 36 para as RBMs originais) para acompanhamento do treinamento do modelo. Na Seção 2.2.4.2, é apresentado um experimento que mostra a capacidade das GBRBMs de modelar dependências entre suas variáveis visíveis.

2.2.4.1 RBM Gauss-Bernoulli reproduzindo uma distribuição

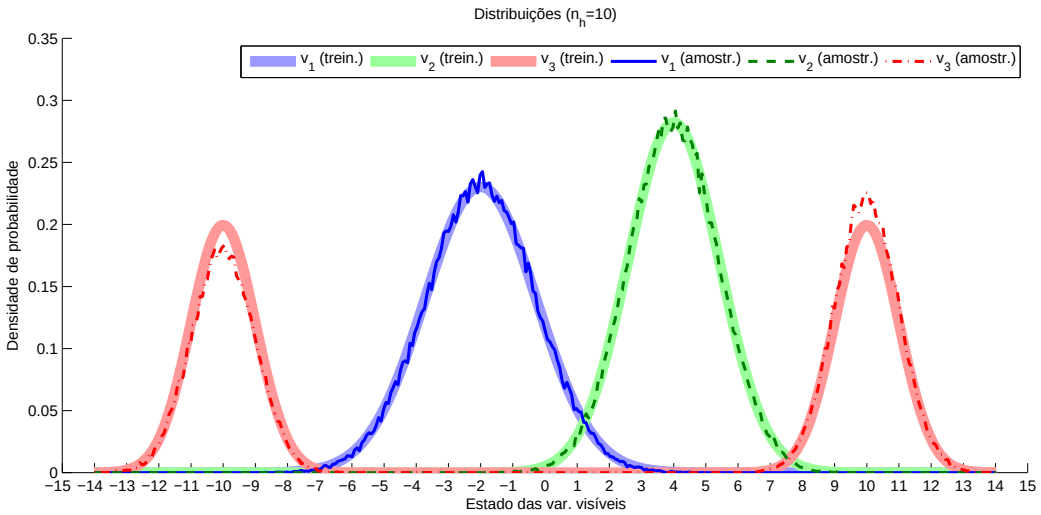
No primeiro experimento, uma GBRBM com n_v=3 unidades visíveis e n_h=10 unidades ocultas é treinada com dados gerados a partir de distribuições conjuntamente independentes com as seguintes características:

1. Unidade v_1: dados gerados a partir de uma distribuição gaussiana com média -2 e variância 3.
2. Unidade v_2: dados gerados a partir de uma distribuição gaussiana com média 4 e variância 2.
3. Unidade v_3: dados gerados em igual proporção a partir de duas distribuições gaussianas com variância unitária e médias -10 e 10, respectivamente.

Na Figura 18, são mostradas, sobrepostas a distribuições de treinamento teóricas, as densidades de probabilidade empíricas resultantes da amostragem do modelo, o qual foi treinado por meio do algoritmo CD-5 e amostrado com 80 passos de Gibbs para geração de cada amostra. É interessante notar que as distribuições associadas às unidades v_1 e v_2 são gaussianas simples (unimodais), enquanto que para a unidade v_3 os dados formam uma mistura de gaussianas. Nessa simulação, ambos os tipos de distribuição foram aproximados com uma boa precisão. Esse resultado chama a atenção para um fato notável: muito embora as distribuições das variáveis visíveis sejam condicionalmente gaussianas (conforme Equação (2.43)), verifica-se que a distribuição marginal do modelo, $P(\mathbf{v})$ é mais geral, podendo representar uma mistura de gaussianas.

No treinamento da Figura 18, foram feitos os ajustes dos parâmetros de variância do modelo (σ_i^2), que costumam causar problemas de divergência no treinamento. Para diminuir as chances dessa ocorrência, fez-se uso da seguinte técnica: no início do treinamento, os

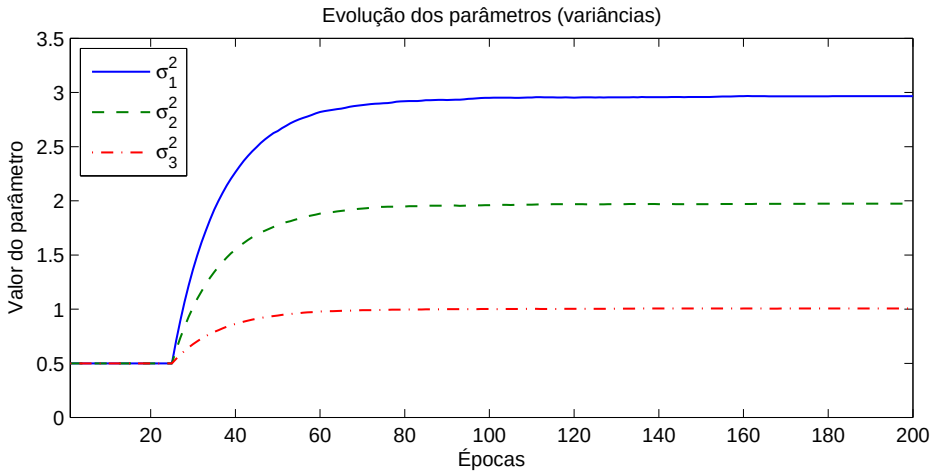
Figura 18 – Distribuições dos dados produzidos por GBRBM treinada com o CD-5.



Fonte: produção do próprio autor.

parâmetros de variâncias não são ajustados (neste caso, nas primeiras 25 épocas de treinamento), permitindo que o modelo se ajuste às médias dos dados. Após essa fase inicial, o ajuste das variâncias é liberado. A evolução desses parâmetros ao longo das épocas de treinamento conforme a técnica descrita é apresentada na Figura 19. Nessa figura, pode-se perceber que o parâmetro σ_3^2 converge para a variância individual das distribuições que formam a mistura de gaussianas referentes à unidade v_3.

Figura 19 – Evolução dos parâmetros de variância no treinamento da Figura 18.



Fonte: produção do próprio autor.

Na Figura 20, é mostrada a evolução do erro de reconstrução (ainda para o treinamento da Figura 18), em que se verifica que o erro citado aumenta com o ajuste da variâncias, muito embora o modelo esteja, ao longo das épocas, se tornando mais representativo dos dados sendo modelados. De fato, nesse caso de teste simples, em que os

vetores de treinamento são conhecidamente independentes, o valor esperado para o erro de reconstrução pode ser estimado com base na expressão

$$\overline{\varepsilon_r} = \mathbb{E}\left[\left|\left|\widetilde{\mathbf{V}} - \mathbf{V}^{(t)}\right|\right|^2\right], \tag{2.50}$$

que corresponde à *esperança* do erro de reconstrução dado na Equação (2.32), considerando as amostras de treinamento ($\mathbf{v}^{(t)}$) e partículas negativas ($\widetilde{\mathbf{v}}$) como vetores aleatórios ($\mathbf{V}^{(t)}$ e $\widetilde{\mathbf{V}}$, respectivamente). Expandindo os vetores nas suas componentes, tem-se

$$\overline{\varepsilon_r} = \mathbb{E}\left[\sum_{i=1}^{n_v}(\widetilde{V}_i - V_i^{(t)})^2\right] = \sum_{i=1}^{n_v}\mathbb{E}\left[(\widetilde{V}_i - V_i^{(t)})^2\right]. \tag{2.51}$$

Considerando agora que as médias individuais das variáveis visíveis já tenham sido descobertas pelo modelo, então $\mathbb{E}[(\widetilde{V}_i - V_i^{(t)})] = 0$ e os termos do somatório na equação anterior se igualam às variâncias de $(\widetilde{V}_i - V_i^{(t)})$, ou seja,

$$\overline{\varepsilon_r} = \sum_{i=1}^{n_v}\mathbf{VAR}\left[\widetilde{V}_i - V_i^{(t)}\right]. \tag{2.52}$$

Finalmente, assumindo que as amostras de treinamento e as partículas negativas sejam independentes (o que é garantido com um número suficientemente grande de passos de Gibbs), chega-se a

$$\overline{\varepsilon_r} = \sum_{i=1}^{n_v}\mathbf{VAR}\left[\widetilde{V}_i\right] + \mathbf{VAR}\left[V_i^{(t)}\right]. \tag{2.53}$$

O valor mostrado no eixo vertical da Figura 20 corresponde, na realidade, à soma dos erros de reconstrução médio de 200 lotes de amostras (com cada lote contendo 100 amostras). Com a observação de que os parâmetros de variâncias do modelo estavam fixados no valor 0,5 (conforme Figura 19) antes do ajuste desses parâmetros ser liberado, estima-se o erro de reconstrução como

$$\overline{\varepsilon_{r1}} = 200[(3 + 0,5) + (2 + 0,5) + (1 + 0,5)] = 1500. \tag{2.54}$$

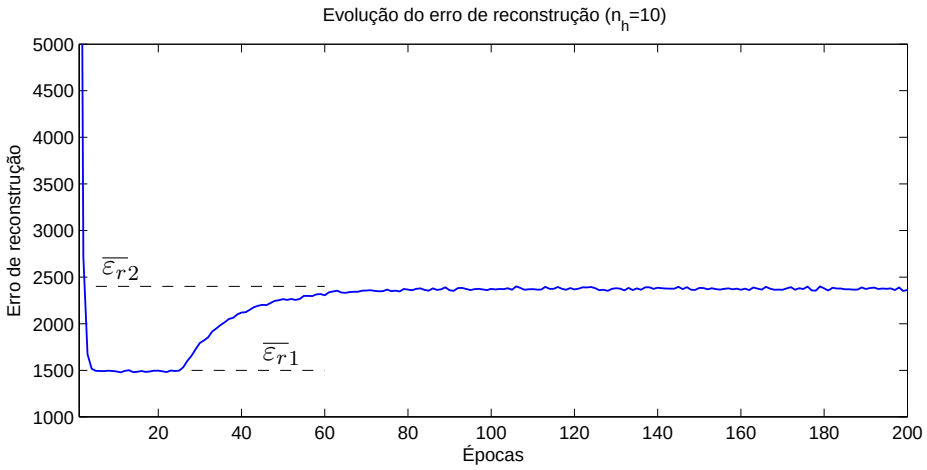
Já após a liberação dos ajustes das variâncias e assumindo que o modelo convirja para os valores das variâncias dos dados, tem-se a segunda estimativa

$$\overline{\varepsilon_{r2}} = 200[(3 + 3) + (2 + 2) + (1 + 1)] = 2400. \tag{2.55}$$

Percebe-se que os valores $\overline{\varepsilon_{r1}}$ e $\overline{\varepsilon_{r2}}$ são condizentes com os patamares constatados para o erro de reconstrução na Figura 20, indicando que as hipóteses simplificadoras aqui assumidas são razoáveis.

Pode-se verificar o efeito negativo do ajuste das variâncias por meio de um caso de teste relativamente mais difícil para o modelo. Tornando mais próximas as médias das gaussianas que formam a mistura na entrada v_3 do caso de teste anterior, a dificuldade em se distinguir a presença de mais de um modo nos dados de treinamento também

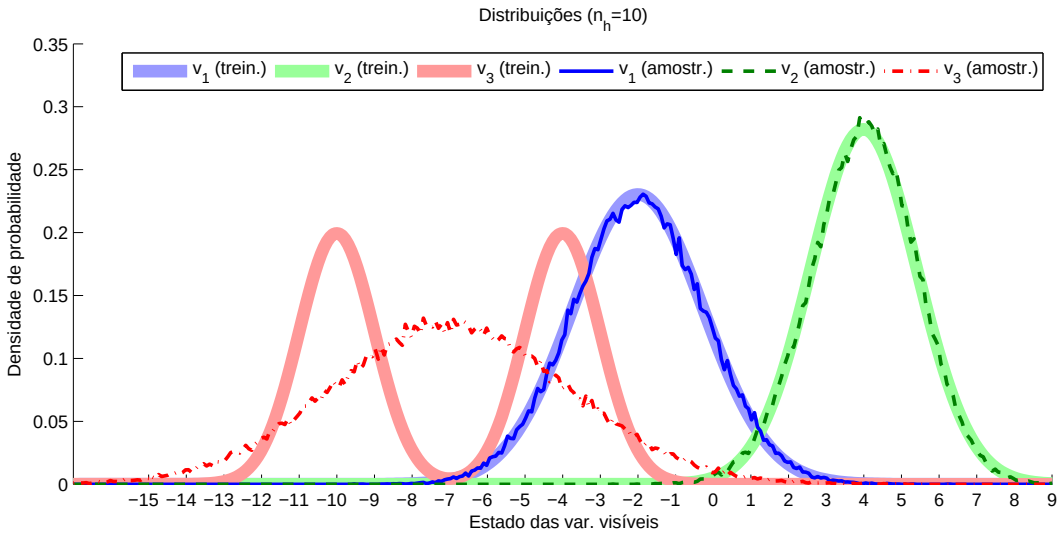
Figura 20 – Evolução do erro de reconstrução no treinamento da Figura 18.



Fonte: produção do próprio autor.

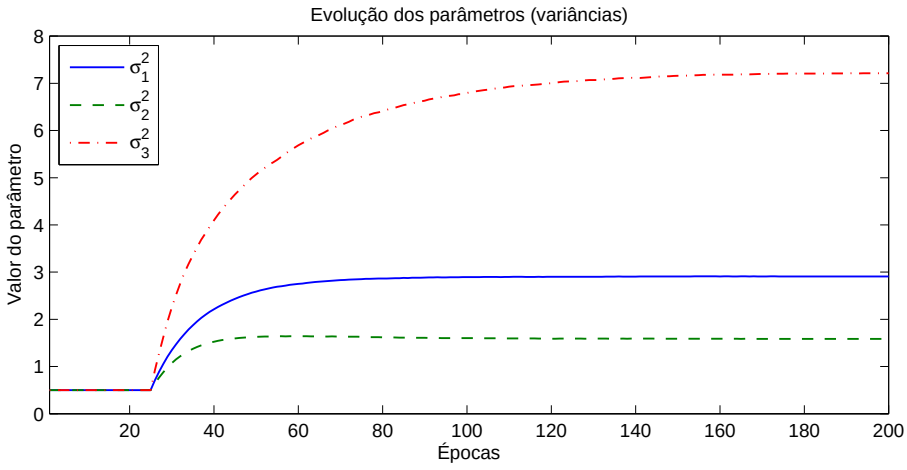
aumenta. Esse fato, aliado ao ajuste das variâncias, pode fazer com que o modelo termine por representar a mistura como uma única gaussiana com uma variância englobando os dois modos presentes nas distribuições, como ilustrado na Figura 21. A evolução das variâncias correspondente ao mesmo caso é fornecida na Figura 22, na qual se percebe que o parâmetro σ_3^2 tende a um valor aproximadamente 7 vezes maior que o das variâncias das gaussianas individuais da mistura. Repetindo-se o treinamento com os mesmos dados desse caso de teste, mas sem o ajuste das variâncias, chega-se a um modelo cujas distribuições individuais das variáveis visíveis são mostradas na Figura 23. Visivelmente, o modelo assim treinado representa melhor as características da distribuição dos dados.

Figura 21 – Distribuições dos dados produzidos por GBRBM treinada com o CD-5 (aproximação menos precisa).



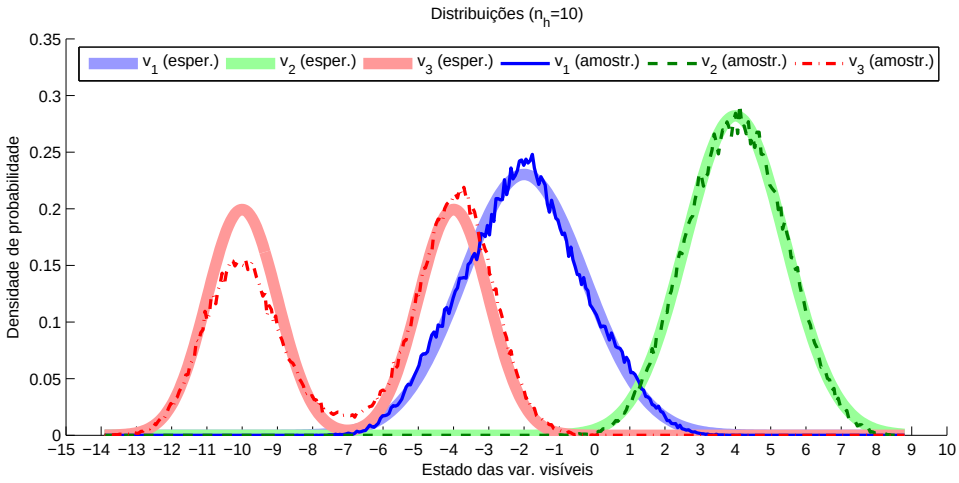
Fonte: produção do próprio autor.

Figura 22 – Evolução dos parâmetros de variância no treinamento da Figura 21.



Fonte: produção do próprio autor.

Figura 23 – Distribuições dos dados produzidos por GBRBM treinada com o CD-5 (sem ajuste de variâncias).



Fonte: produção do próprio autor.

2.2.4.2 RBM Gauss-Bernoulli reproduzindo dependências entre variáveis

O objetivo do experimento apresentado na presente seção é o de verificar o comportamento do modelo quando os dados apresentam dependências entre si. Para isso, os dados de treinamento são produzidos a partir de uma distribuição conjuntamente gaussiana

tendo uma matriz de covariância definida manualmente como:

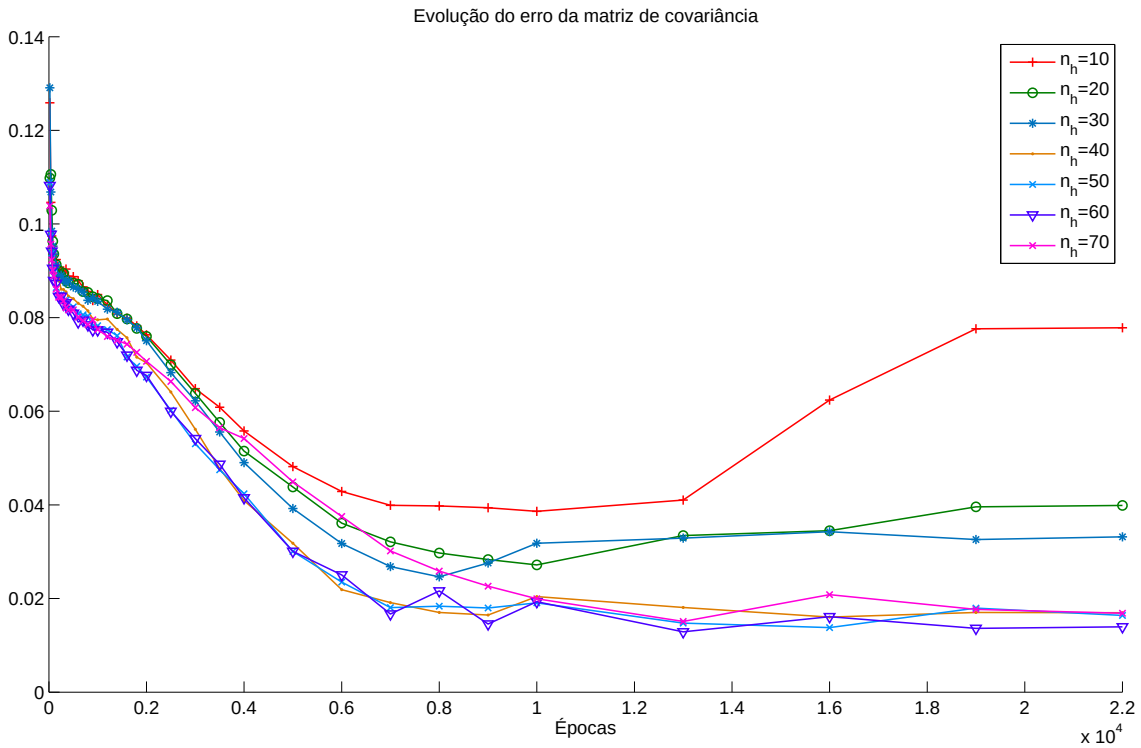
$$\mathbf{C}_0 = \begin{pmatrix} 1{,}00 & 0 & 0{,}10 & 0 & 0 & 0 & 0{,}30 & 0 & 0 & 0 \\ 0 & 1{,}00 & 0 & 0{,}10 & 0 & 0 & 0 & 0{,}30 & 0 & 0 \\ 0{,}10 & 0 & 1{,}00 & 0 & 0{,}10 & 0 & 0 & 0 & 0{,}40 & 0 \\ 0 & 0{,}10 & 0 & 2{,}00 & 0 & 0{,}10 & 0 & 0 & 0 & 0{,}40 \\ 0 & 0 & 0{,}10 & 0 & 2{,}00 & 0 & 0{,}20 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0{,}10 & 0 & 2{,}00 & 0 & 0{,}20 & 0 & 0 \\ 0{,}30 & 0 & 0 & 0 & 0{,}20 & 0 & 2{,}00 & 0 & 0{,}20 & 0 \\ 0 & 0{,}30 & 0 & 0 & 0 & 0{,}20 & 0 & 1{,}00 & 0 & 0{,}20 \\ 0 & 0 & 0{,}40 & 0 & 0 & 0 & 0{,}20 & 0 & 1{,}00 & 0 \\ 0 & 0 & 0 & 0{,}40 & 0 & 0 & 0 & 0{,}20 & 0 & 1{,}00 \end{pmatrix}.$$

Amostras produzidas de modo a apresentar essa matriz de covariância são usadas no treinamento de uma GBRBM com n_v=10 unidades visíveis e diferentes quantidades de unidades ocultas. Durante o treinamento, a GBRBM é amostrada e a matriz de covariância $\mathbf{C}_1$ dos dados por ela produzidos é determinada. O aqui chamado *erro da matriz de covariância* é definido como a norma de Frobenius da diferença $(\mathbf{C}_1 - \mathbf{C}_0)$ (MEYER, 2000). A Figura 24 mostra a evolução desse erro nos diversos treinamentos. Na figura, verifica-se que, conforme o treinamento progride, o erro da matriz de covariância tende a diminuir por um número relativamente alto de épocas até que se estabiliza em um certo nível, exceto no caso de n_h=10, em que houve um aparente início de divergência do algoritmo. Nota-se ainda que, aumentando-se o número de unidades ocultas, chega-se também a menores níveis de erro, efeito constatado até se atingir n_h=40. Um aumento acima dessa quantidade não parece benéfico. Para n_h=60, chegou-se, ao final da última época de treinamento, à matriz

$$\mathbf{C}_1 = \begin{pmatrix} 1{,}00 & 0{,}00 & 0{,}09 & 0{,}00 & 0{,}02 & 0{,}00 & 0{,}27 & 0{,}00 & 0{,}01 & 0{,}00 \\ 0{,}00 & 0{,}99 & -0{,}01 & 0{,}09 & -0{,}01 & 0{,}01 & -0{,}01 & 0{,}28 & 0{,}00 & 0{,}02 \\ 0{,}09 & -0{,}01 & 0{,}98 & 0{,}00 & 0{,}09 & -0{,}01 & 0{,}03 & -0{,}01 & 0{,}38 & 0{,}00 \\ 0{,}00 & 0{,}09 & 0{,}00 & 2{,}00 & -0{,}02 & 0{,}09 & 0{,}00 & 0{,}02 & 0{,}03 & 0{,}39 \\ 0{,}02 & -0{,}01 & 0{,}09 & -0{,}02 & 2{,}00 & -0{,}01 & 0{,}17 & -0{,}01 & 0{,}02 & 0{,}00 \\ 0{,}00 & 0{,}01 & -0{,}01 & 0{,}09 & -0{,}01 & 1{,}99 & 0{,}00 & 0{,}19 & -0{,}01 & 0{,}02 \\ 0{,}27 & -0{,}01 & 0{,}03 & 0{,}00 & 0{,}17 & 0{,}00 & 1{,}99 & 0{,}00 & 0{,}19 & 0{,}01 \\ 0{,}00 & 0{,}28 & -0{,}01 & 0{,}02 & -0{,}01 & 0{,}19 & 0{,}00 & 1{,}00 & 0{,}01 & 0{,}20 \\ 0{,}01 & 0{,}00 & 0{,}38 & 0{,}03 & 0{,}02 & -0{,}01 & 0{,}19 & 0{,}01 & 0{,}99 & 0{,}02 \\ 0{,}00 & 0{,}02 & 0{,}00 & 0{,}39 & 0{,}00 & 0{,}02 & 0{,}01 & 0{,}20 & 0{,}02 & 1{,}00 \end{pmatrix},$$

que pode ser confrontada com $\mathbf{C}_0$. Verifica-se que, não apenas os elementos da diagonal principal de $\mathbf{C}_1$ (variâncias individuais das variáveis visíveis) se parecem com aqueles de $\mathbf{C}_0$, mas que o mesmo acontece com os elementos fora da diagonal principal. Esse fato mostra que as GBRBMs têm a capacidade de modelar dependências entre variáveis.

Figura 24 – Evolução do erro da matriz de covariância ao longo das épocas de treinamento.



Fonte: produção do próprio autor.

2.2.5 Relações com outros modelos

Conceitualmente, as mesmas relações que as RBMs (Bernoulli-Bernoulli) têm com outros modelos (conforme apresentado na Seção 2.1.4) são válidas também para as GBRBMs. Com a mudança da formulação da energia global, entretanto, a função de ativação do neurônio clássico análogo a uma unidade visível da GBRBM tem agora a forma de uma função gaussiana, conforme Equação (2.43).

A capacidade da GBRBM de reproduzir distribuições formadas por misturas de gaussianas leva à interessante questão sobre a relação desse modelo com aquele que é denominado Modelo de Mistura Gaussiana (GMM – *Gaussian Mixture Model*). Neste, uma distribuição é escrita na forma (BISHOP, 2006)

$$P_{\text{gmm}}(\mathbf{x}) = \sum_{k=1}^{K_{\text{gmm}}} \pi_k \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right), \tag{2.56}$$

sendo que π_k é o coeficiente de mistura, $\boldsymbol{\mu}_k$ é o vetor média e $\boldsymbol{\Sigma}_k$ é a matriz de covariância, cada qual correspondente à k-ésima componente da mistura. K_{gmm} define o número de gaussianas que compõem a mistura. A notação $\mathcal{N}(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ é similar àquela usada anteriormente neste texto para representar uma distribuição gaussiana, porém aqui deve ficar claro que os parâmetros $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$ são, respectivamente, o vetor média e a matriz de covariância da distribuição, tratando-se portanto, de uma distribuição multivariada. A fim

de relacionar GMMs com GBRBMs, pode-se escrever

$$P_g(\mathbf{v}) = \sum_{\mathbf{h}^* \in \{0,1\}^{n_h}} P_g(\mathbf{v}, \mathbf{h}^*) = \sum_{\mathbf{h}^* \in \{0,1\}^{n_h}} P_g(\mathbf{h}^*) P_g(\mathbf{v}|\mathbf{h}^*)$$

e notar que $P_g(\mathbf{v}|\mathbf{h})$ corresponde, conforme Equações (2.41) e (2.43), a uma distribuição conjuntamente gaussiana tendo média

$$\boldsymbol{\mu}_g(\mathbf{h}) = \begin{bmatrix} c_1 + \sum_{j=1}^{n_h} h_j w_{j1} \\ \vdots \\ c_{n_v} + \sum_{j=1}^{n_h} h_j w_{jn_v} \end{bmatrix} \tag{2.57}$$

e matriz de covariância $\boldsymbol{\Sigma}_g = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_{n_v}^2)$ (matriz diagonal formada pelas variâncias das variáveis visíveis), de modo que

$$P_g(\mathbf{v}) = \sum_{\mathbf{h}^* \in \{0,1\}^{n_h}} P_g(\mathbf{h}^*) \mathcal{N}\left(\mathbf{v} \middle| \boldsymbol{\mu}_g(\mathbf{h}^*), \boldsymbol{\Sigma}_g\right). \tag{2.58}$$

A comparação dessa última equação com (2.56) sugere que a distribuição das variáveis visíveis da GBRBM, $P_g(\mathbf{v})$, pode ser vista como uma mistura de gaussianas (KENNY, 2011). Nota-se também que:

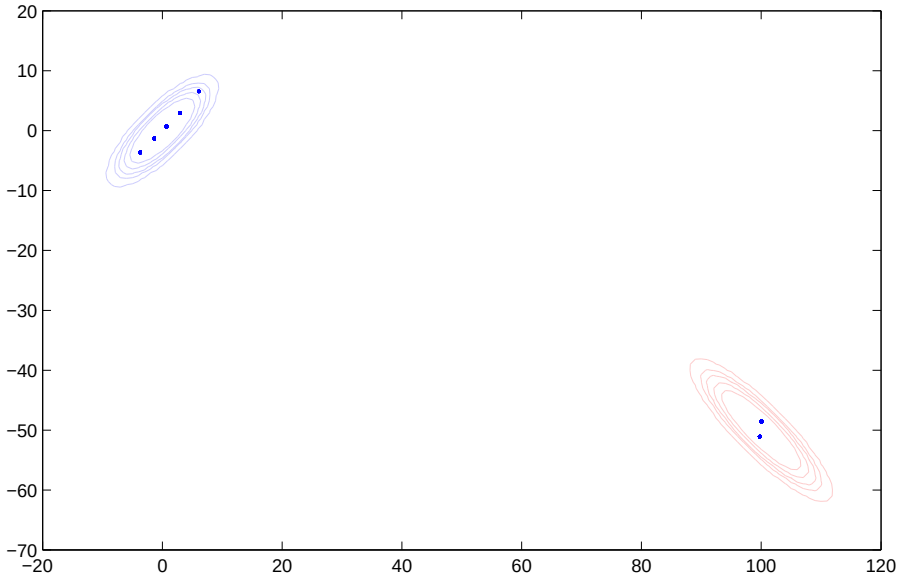
a) cada gaussiana da mistura é independente (a matriz de covariância é diagonal);
b) a média de cada componente da mistura é definida pelo estado da camada oculta, segundo Equação (2.57); e
c) o coeficiente (ou peso) da componente está associado à probabilidade do respectivo estado da camada oculta.

A partir do item a), pode-se concluir que eventuais dependências entre variáveis visíveis devem necessariamente ser representadas no modelo por meio das variáveis ocultas. De fato, essa necessidade vem da ausência de conexões entre variáveis visíveis, uma característica comum entre as RBMs e variantes (como a GBRBM). Mais ainda, a Equação (2.58) permite que se perceba como as dependências entre variáveis visíveis ficam estabelecidas numa GBRBM. Basicamente, o estado da camada oculta define simultaneamente as médias de todas as variáveis gaussianas que formam a camada visível do modelo, gerando, portanto, um acoplamento entre elas.

A Figura 25 ilustra como uma GBRBM pode aproximar uma mistura de duas gaussianas formadas por variáveis dependentes entre si. Vale notar de antemão que uma mistura assim seria representada de forma direta por um GMM com duas componentes tendo matrizes de covariância não diagonais. Uma GBRBM com n_h=30 e $\boldsymbol{\Sigma}_g$=$\mathbf{I}$ foi treinada com os dados da mistura cujas curvas de nível estão representadas por elipses na Figura 25. Já os pontos na mesma figura indicam as médias das gaussianas produzidas pelo

modelo resultante do treinamento. Esses pontos foram obtidos por meio de simulações, como segue: primeiramente, foram extraídas amostras das variáveis ocultas da GBRBM via amostragem de Gibbs e, em seguida, essas amostras foram mapeadas, segundo a Equação (2.57), nos referidos pontos. Em torno deles poderiam ainda ser imaginadas circunferências correspondentes às curvas de nível das componentes individuais da mistura representada pela GBRBM. Com o exemplo mostrado aqui, deve ficar claro que a mistura de algumas distribuições gaussianas dependentes seria representado numa GBRBM por um número potencialmente grande de gaussianas independentes.

Figura 25 – Aproximação de uma mistura de duas gaussianas dependentes for várias gaussianas independentes (GBRBM).



Fonte: produção do próprio autor.

2.3 RBM BERNOULLI-BERNOULLI COMO CLASSIFICADOR

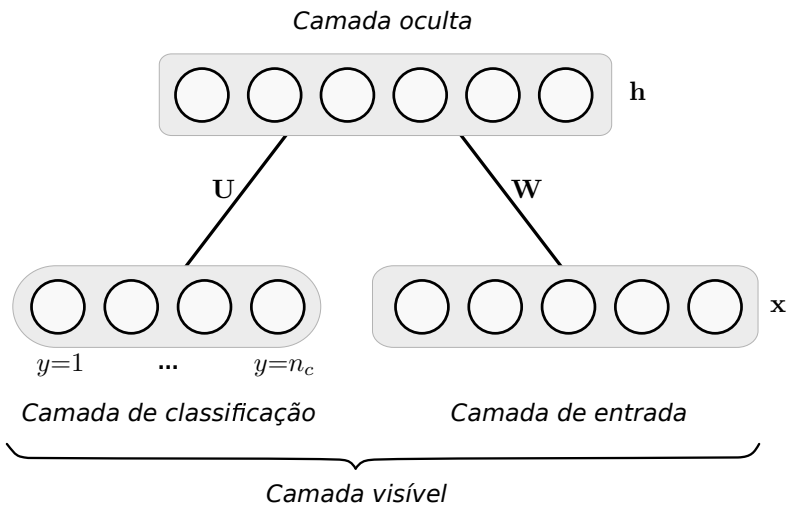
A variante de RBM apresentada nesta seção adiciona ao modelo original (Bernoulli-Bernoulli) a capacidade de classificação. A formulação para a energia global e as diferentes equações daí resultantes foram extraídas de Larochelle e Bengio (2008) com pequenas adaptações de notação. Visando novamente facilitar a comparação entre variantes de RBMs, os tópicos à frente seguem a organização da Seção 2.1. Inicialmente, a estrutura e a descrição matemática da variante são introduzidas nas Seções 2.3.1 e 2.3.2, respectivamente. Na Seção 2.3.3, tem-se a apresentação das tarefas associadas à presente versão de RBM. Em relação às versões vistas anteriormente, surgem aqui a tarefa de classificação e novos tipos de treinamento, os quais são introduzidos na mesma seção. Por fim, a Seção 2.3.4 compara a atual variante com uma rede neuronal artificial. Os tópicos a seguir são apresentados de forma sucinta. Em particular, detalhes relativos ao treinamento e ao uso do modelo

como classificador são adiados para o Capítulo 3, no qual a variante proposta no presente trabalho é apresentada.

2.3.1 Estrutura

Na Figura 26, é mostrada a estrutura de uma RBM adaptada para permitir o seu uso direto em problemas de classificação. A notação utilizada na figura é mais sucinta do que aquela usada para introduzir as RBMs originais. Em particular, aqui os vieses não são mostrados e o conjunto de conexões entre as unidades de um par de camadas são representadas por uma única linha interligando essas mesmas camadas. O rótulo acima da linha indica o nome da matriz de pesos de conexão correspondente. Como indicado na figura, a camada visível é de fato composta pelas camadas de entrada e de classificação. As unidades da camada de classificação são binárias (assumindo apenas os valores 0 ou 1) e considera-se que, em qualquer momento, somente uma delas possa estar ativa (i.e., assuma o valor 1). Para uma dada entrada, a classe a ela associada é definida pelo índice da unidade ativa na camada de classificação. Vale antecipar que, apesar do nome, a camada de entrada pode operar efetivamente tanto como uma entrada quanto como uma saída do modelo.

Figura 26 – RBM com uma camada de classificação.



Fonte: extraído de Larochelle e Bengio (2008).

2.3.2 Descrição matemática

Na formulação matemática, as unidades da camada de entrada serão representadas pelo vetor aleatório

$$\mathbf{X} = [X_1 \dots X_{n_d}]^T \tag{2.59}$$

e os estados dessas variáveis, pelo vetor

$$\mathbf{x} = [x_1 \dots x_{n_d}]^T, \tag{2.60}$$

sendo n_d o número de unidades da camada de entrada. Dada a restrição antes mencionada de que apenas uma das unidades da camada de classificação pode estar ativa em qualquer momento, o conjunto dessas unidades é modelado por uma única variável aleatória, Y, a qual pode assumir valores no conjunto $\{1, 2, \dots, n_c\}$, sendo n_c a quantidade de classes existentes. Um estado específico da variável Y é indicado pelo escalar y. Para as variáveis ocultas, as definições são iguais àquelas da RBM original, dadas nas Equações (2.3) e (2.4).

A RBM contendo uma camada de classificação é obtida definindo-se a sua energia global como (LAROCHELLE; BENGIO, 2008)

$$\begin{aligned} E_c(y, \mathbf{x}, \mathbf{h}) = &- \sum_{j=1}^{n_h} b_j h_j - \sum_{i=1}^{n_d} c_i x_i - \sum_{i=1}^{n_d} \sum_{j=1}^{n_h} h_j w_{ji} x_i \\ &- \sum_{k=1}^{n_c} d_k \delta_{k,y} - \sum_{k=1}^{n_c} \sum_{j=1}^{n_h} h_j u_{jk} \delta_{k,y}, \end{aligned} \tag{2.61}$$

em que a notação $\delta_{r,s}$ representa o delta de Kronecker e w_{ji}, b_j, c_i, d_k e u_{jk}, com i=1, ..., n_d, j=1, ..., n_h, k=1, ..., n_c, constituem os parâmetros do modelo. Os parâmetros d_k são os vieses da camada de classificação e os u_{jk} são os pesos de conexão entre as camadas oculta e de classificação. Definem-se a matriz $n_h \times n_d$ de pesos de conexão entre as camadas oculta e de entrada

$$\mathbf{W} = (w_{ji}) \tag{2.62}$$

e a matriz $n_h \times n_c$ de pesos de conexão entre as camadas oculta e de classificação

$$\mathbf{U} = (u_{jk}). \tag{2.63}$$

Similarmente, são definidos os vetores de vieses

$$\mathbf{b} = [b_1 \dots b_{n_h}]^T, \tag{2.64}$$

$$\mathbf{c} = [c_1 \dots c_{n_d}]^T \tag{2.65}$$

e

$$\mathbf{d} = [d_1 \dots d_{n_c}]^T. \tag{2.66}$$

Vale dizer que $\mathbf{W}$, $\mathbf{b}$ e $\mathbf{c}$ têm os mesmos significados dos símbolos de mesmos nomes antes definidos para a RBM Bernoulli-Bernoulli (Equações (2.5) a (2.7) da página 23) se a aqui chamada camada de entrada for considerada como a camada visível do modelo original. Com as definições anteriores, nota-se que a RBM com a camada de classificação fica totalmente especificada pelo conjunto de parâmetros $\{\mathbf{W}, \mathbf{b}, \mathbf{c}, \mathbf{d}, \mathbf{U}\}$.

É interessante comparar a definição da energia global na Equação (2.61) com aquela na Equação (2.8) para as RBMs originais. Pode-se perceber que a primeira linha à direita da igualdade na Equação (2.61) torna-se idêntica à expressão de $E(\mathbf{v}, \mathbf{h})$ se x_i for substituído por v_i em $E_c(y, \mathbf{x}, \mathbf{h})$. Conclui-se que os termos na primeira linha da Equação (2.61) representam o modelo original. Logo, a segunda linha da mesma equação responde pela camada de classificação propriamente dita.

A distribuição conjunta do modelo é definida como

$$P_c(y, \mathbf{x}, \mathbf{h}) = \frac{\exp(-E_c(y, \mathbf{x}, \mathbf{h}))}{Z_c} \tag{2.67}$$

e tem um domínio dado por $\{1, 2, \ldots, n_c\} \times \{0, 1\}^{n_d} \times \{0, 1\}^{n_h}$. A constante no denominador garante que $P_c(y, \mathbf{x}, \mathbf{h})$ tenha soma unitária no seu domínio.

2.3.3 Tarefas associadas ao modelo

As RBMs com camada de classificação admitem os mesmos casos de uso (ou tarefas) introduzidos na Seção 2.1.3: inferência, amostragem e treinamento. Além desses, surge o caso de uso de classificação, no qual busca-se determinar a classe à qual pertence uma dada amostra de entrada. A seguir, os casos de uso mencionados são brevemente discutidos.

2.3.3.1 Inferência

Como apontado por Larochelle e Bengio (2008), pode-se mostrar que

$$P_c(\mathbf{x}|\mathbf{h}) = \prod_{i=1}^{n_d} P_c(x_i|\mathbf{h}) \tag{2.68}$$

e

$$P_c(\mathbf{h}|y, \mathbf{x}) = \prod_{j=1}^{n_h} P_c(h_j|y, \mathbf{x}), \tag{2.69}$$

ou seja, conhecidos os estados das variáveis em uma camada, as variáveis na camada oposta são independentes. Além disso, as variáveis de entrada e ocultas são binárias (Bernoulli) com probabilidades de ativação dadas, respectivamente, por

$$P_c(x_i{=}1|\mathbf{h}) = \varphi\left(c_i + \sum_{j=1}^{n_h} h_j w_{ji}\right) \tag{2.70}$$

e

$$P_c(h_j{=}1|y, \mathbf{x}) = \varphi\left(b_j + u_{jy} + \sum_{i=1}^{n_d} w_{ji} x_i\right), \tag{2.71}$$

sendo $\varphi(\cdot)$ a função sigmóide já introduzida. Por fim, conhecido $\mathbf{h}$, a distribuição de probabilidade das classes é dada por

$$P_c(y|\mathbf{h}) = \frac{\exp\left(d_y + \sum\limits_{j=1}^{n_h} h_j u_{jy}\right)}{\sum\limits_{k=1}^{n_c} \exp\left(d_k + \sum\limits_{j=1}^{n_h} h_j u_{jk}\right)}. \tag{2.72}$$

As Equações (2.68) a (2.72) permitem a realização de inferências condicionais entre as camadas oculta e visível do modelo. Vale lembrar que tais tipos de inferência são úteis na realização da amostragem de Gibbs do modelo.

2.3.3.2 Amostragem

A amostragem da RBM com camada de classificação pode ser realizada de forma semelhante à amostragem nas RBMs originais. Para isso, basta notar que uma amostra de estado visível na atual variante é representada pelo par $(y, \mathbf{x})$. Dessa forma, observando a Figura 4, o primeiro passo de Gibbs seria composto pela sequência de amostras $(y^{(0)}, \mathbf{x}^{(0)}) \rightarrow \mathbf{h}^{(1)} \rightarrow (y^{(1)}, \mathbf{x}^{(1)})$, sendo que $\mathbf{h}^{(1)}$ seria produzida segundo a distribuição $P_c(\mathbf{h}|y^{(0)}, \mathbf{x}^{(0)})$ e as componentes do par $(y^{(1)}, \mathbf{x}^{(1)})$, segundo as distribuições $P_c(y|\mathbf{h}^{(1)})$ e $P_c(\mathbf{x}|\mathbf{h}^{(1)})$, respectivamente.

2.3.3.3 Treinamento

O treinamento do presente modelo pode ser feito de forma análoga ao das RBMs originais, descrito na Seção 2.1.3.3. Nesse caso, dada uma amostra de treinamento, $(y^{(t)}, \mathbf{x}^{(t)})$, busca-se minimizar a função perda generativa,

$$\mathcal{L}_{gen}(\boldsymbol{\theta}; y^{(t)}, \mathbf{x}^{(t)}) = -\ln P_c(y^{(t)}, \mathbf{x}^{(t)}), \tag{2.73}$$

com relação a $\boldsymbol{\theta}$ (tido como um vetor contendo todos os parâmetros do modelo). Essa função perda é dita generativa porque a sua minimização equivale a se maximizar a verossimilhança de o modelo gerar a amostra de entrada $\mathbf{x}^{(t)}$ juntamente com sua respectiva classificação, $y^{(t)}$, fato evidenciado pelo uso da distribuição conjunta de classes e entradas, $P_c(y, \mathbf{x})$, na Equação (2.73). O treinamento visando minimizar $\mathcal{L}_{gen}(\cdot)$ é denominado *treinamento generativo*.

Uma segunda possibilidade que surge em razão da presença da camada de classificação é treinar o modelo especificamente para classificação. Isso é feito minimizando-se a função perda dita discriminativa,

$$\mathcal{L}_{disc}(\boldsymbol{\theta}; y^{(t)}, \mathbf{x}^{(t)}) = -\ln P_c(y^{(t)}|\mathbf{x}^{(t)}), \tag{2.74}$$

com relação a $\boldsymbol{\theta}$, que é uma função definida em termos da distribuição condicional $P_c(y|\mathbf{x})$ (Equação (2.76)). A minimização de $\mathcal{L}_{disc}(\cdot)$ é o objetivo do denominado *treinamento discriminativo* e o modelo resultante desse tipo de treinamento recebe o nome de RBM Discriminativa ou DRBM (*Discriminative RBM*).

Por fim, pode-se ainda combinar as duas funções perda anteriores, o que dá origem à chamada função perda híbrida,

$$\mathcal{L}_{hyb}(\boldsymbol{\theta}; y^{(t)}, \mathbf{x}^{(t)}) = \mathcal{L}_{disc}(\boldsymbol{\theta}; y^{(t)}, \mathbf{x}^{(t)}) + \alpha \mathcal{L}_{gen}(\boldsymbol{\theta}; y^{(t)}, \mathbf{x}^{(t)}), \tag{2.75}$$

em que α é o peso dado ao critério generativo. Quando objetiva-se minimizar $\mathcal{L}_{hyb}(\cdot)$, tem-se o chamado treinamento híbrido e o modelo resultante é denominado RBM discriminativa híbrida ou HDRBM (*Hybrid Discriminative RBM*), indicando que o seu treinamento é majoritariamente discriminativo.

Cada uma das funções perda apresentadas ($\mathcal{L}_{gen}(\cdot)$, $\mathcal{L}_{disc}(\cdot)$ e $\mathcal{L}_{hyb}(\cdot)$) dá origem a diferentes regras de atualização de parâmetros. Essas regras não serão fornecidas aqui, pois são de pouco interesse para esta dissertação. Entretanto, os três tipos de treinamento introduzidos na seção atual serão desenvolvidos para o modelo proposto no Capítulo 3. Lá, as regras de atualização de parâmetros serão apresentadas e os tipos de treinamento, discutidos com mais detalhes.

2.3.3.4 Classificação

Na tarefa de classificação, os parâmetros do modelo são conhecidos, tendo sido obtidos num treinamento prévio, e é então fornecida uma amostra de entrada cuja classe deseja-se estimar com o uso do modelo. Na RBM com a camada de classificação, esse problema pode ser visto como um tipo de inferência realizada por meio da distribuição condicional $P_c(y|\mathbf{x})$, a qual pode ser escrita como

$$P_c(y|\mathbf{x}) = \frac{\exp\left[d_y + \sum\limits_{j=1}^{n_h} \zeta\Big(b_j + u_{jy} + \sum\limits_{i=1}^{n_d} w_{ji}x_i\Big)\right]}{\sum_{y^*=1}^{n_c} \exp\left[d_{y^*} + \sum\limits_{j=1}^{n_h} \zeta\Big(b_j + u_{jy^*} + \sum\limits_{i=1}^{n_d} w_{ji}x_i\Big)\right]}, \tag{2.76}$$

sendo $\zeta(z) = \ln(1 + e^z)$, função conhecida como *softplus*. Como apontado por Larochelle e Bengio (2008), $P_c(y|\mathbf{x})$ pode ser avaliada num tempo $\mathcal{O}(n_h n_d + n_h n_c)$. Normalmente, a classificação de uma amostra $\mathbf{x}^{(c)}$ é realizada escolhendo-se a classe $y^{(c)}$ que maximize o valor da função $P_c(y^{(c)}|\mathbf{x}^{(c)})$. Vale dizer que $P_c(y|\mathbf{x})$ seria usada na classificação independentemente do tipo de treinamento (generativo, discriminativo ou híbrido) anteriormente utilizado para se determinar os parâmetros do modelo.

2.3.4 Relações com outros modelos

Conforme apontado por Larochelle e Bengio (2008), existe uma similaridade entre RBMs discriminativas e redes neuronais. Especificamente, o cálculo de $P_c(y|\mathbf{x})$ pode ser implementado por uma rede neuronal tendo uma camada oculta com função de ativação *softplus* e uma camada de saída do tipo *softmax* (BISHOP, 2001).

2.4 DETECÇÃO DE ATIVIDADE VOCAL

A detecção de atividade vocal (VAD – *Voice Activity Detection*), introduzida na Seção 1.2, consiste em se identificar em um áudio, os trechos contendo voz. Devido

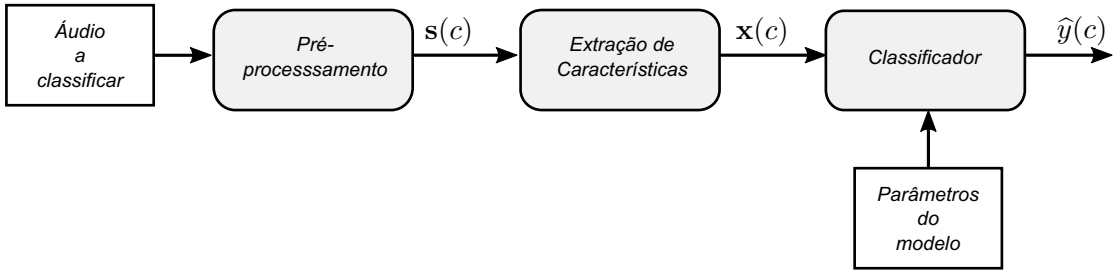
à sua importância prática, esse assunto que tem recebido a atenção da comunidade científica por muitos anos (RAMIREZ et al., 2004; RABINER; SAMBUR, 1977; JUNQUA; REAVES; MAK, 1991; TUCKER, 1992; HAIGH; MASON, 1993; HOYT; WECHSLER, 1994; RENEVEY; DRYGAJLO, 2001; NEMER; GOUBRAN; MAHMOUD, 2001; LI; SWAMY; AHMAD, 2005; SOHN; KIM; SUNG, 1999; GAZOR; ZHANG, 2003; CHANG; KIM; MITRA, 2006). Em sistemas de telefonia, detectores de atividade vocal permitem uma significativa redução na largura de banda usada para comunicações de voz. VADs são também utilizados em sistemas de redução de ruído, nos quais se faz a estimação do espectro do ruído durante períodos de ausência de voz (RAMIREZ et al., 2004).

Técnicas tradicionais de detecção de atividade vocal incluem medidas baseadas em energia, parâmetros de codificação preditiva linear (LPC – *Linear Predictive Coding*) (RABINER; SAMBUR, 1977), taxa de cruzamento por zero (JUNQUA; REAVES; MAK, 1991), medidas de periodicidade (TUCKER, 1992), características cepstrais (HAIGH; MASON, 1993), configuração de formantes (HOYT; WECHSLER, 1994) e entropia espectral (RENEVEY; DRYGAJLO, 2001). Destacam-se também as técnicas envolvendo estatísticas de ordem superior (NEMER; GOUBRAN; MAHMOUD, 2001; LI; SWAMY; AHMAD, 2005) e modelos estatísticos avançados (SOHN; KIM; SUNG, 1999; GAZOR; ZHANG, 2003; CHANG; KIM; MITRA, 2006). Essas técnicas buscam estabelecer medidas que salientem as diferenças entre as condições de presença e de ausência de voz. Tais medidas são então usadas em regras de decisão definidas empiricamente ou segundo algum critério objetivo como, por exemplo, satisfazer uma condição definida estatisticamente.

Recentemente, técnicas envolvendo aprendizagem de máquina vem se popularizando em muitas tarefas, inclusive na detecção de atividade vocal. Em uma das abordagens, utilizam-se medidas obtidas pelas já citadas técnicas tradicionais como entradas do classificador, o qual é treinado de modo a se obter a classificação desejada. Nesse caso, apenas o encargo da tomada de decisão do detector é passado para o classificador. Em outra abordagem, o mecanismo de aprendizagem é alimentado com medidas que buscam representar o som propriamente dito, como seria o caso dos coeficientes mel-cepstrais (ZOU et al., 2014). Assim, o classificador torna-se responsável não apenas pela decisão final do detector, mas também pela descoberta das características importantes para discriminar a presença ou ausência de voz. Independentemente da abordagem utilizada, vale notar que o uso de mecanismos de aprendizagem torna direta a adição de novas informações a um detector, o que enseja o emprego desses mecanismos.

A seguir, a Seção 2.4.1 detalha a realização de um detector baseado em aprendizagem de máquina, na forma como foi utilizado neste trabalho. As várias medidas que tiveram seu uso investigado como entradas do classificador são apresentadas à parte na Seção 2.4.2.

Figura 27 – VAD empregando um classificador baseado em aprendizagem de máquina.



Fonte: produção do próprio autor.

2.4.1 Detector baseado em aprendizagem de máquina

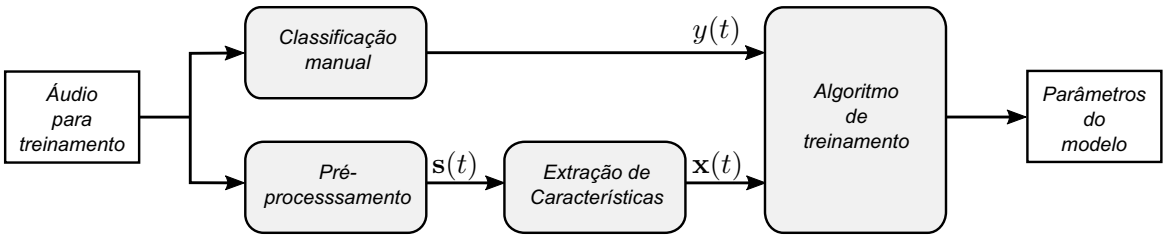
A detecção de atividade vocal pode ser vista como um problema de classificação binária, em que se atribui um rótulo (classe), indicando presença ou ausência de voz, a um dado segmento de áudio. Na Figura 27, é mostrado um diagrama de blocos de um VAD, no qual o classificador, bloco responsável pela decisão final (rotulação) do detector, seria um mecanismo previamente treinado para essa função. Percebe-se, na figura, que o áudio, que poderia ser uma sequência de amostras previamente adquiridas ou mesmo um fluxo de amostras obtidas em tempo real, passa primeiramente por um bloco de pré-processamento. Considera-se aqui que esse bloco seja responsável por dividir as amostras de entrada em segmentos (quadros), podendo, inclusive, haver sobreposição entre segmentos sucessivos (i.e., parte das amostras de um segmento serem compartilhadas com um segmento vizinho). O mesmo bloco poderia realizar outras tarefas, como filtrar o sinal antes da sua segmentação, por exemplo. Em seguida ao pré-processamento, um segmento de índice c, indicado pelo vetor $\mathbf{s}(c)$, passa ao bloco de extração de características que, por sua vez, produz o vetor de características, $\mathbf{x}(c)$. Este serve de entrada para o classificador, que finalmente fornece a estimativa $\widehat{y}(c)$ indicativa da presença ou ausência de voz em $\mathbf{s}(c)$.

Vale aqui fornecer algumas justificativas para a presença do bloco de extração de características no diagrama da Figura 27, visto que seria possível imaginar um cenário em que o classificador fosse alimentado diretamente com os segmentos de amostras de áudio. Um primeiro motivo para a extração de características vem do fato que, frequentemente, os dados originais contêm redundâncias. Eliminando-se essas redundâncias, chega-se a vetores com menores dimensões que as originais, o que contribui para um melhor uso de recursos computacionais no classificador. Constata-se também que entradas com maiores dimensões comumente levam à necessidade de classificadores com um maior número de parâmetros, os quais são mais propensos à situação de *overfitting*, já explicada. Por fim, muitas vezes os vetores de características (medidas) são projetados de modo a salientar informações importantes para a tarefa abordada, facilitando, portanto, o trabalho do classificador.

O uso do classificador na Figura 27 requer o conhecimento dos seus parâmetros, os quais seriam obtidos previamente num processo de treinamento. A Figura 28 mostra o fluxo de dados durante o processo citado. Nota-se que, como na classificação, o áudio de treinamento é pré-processado e dividido em segmentos. Um segmento de índice t, $\mathbf{s}(t)$, tem suas características extraídas e dá origem ao vetor $\mathbf{x}(t)$, que é uma amostra de entrada de treinamento. Além disso, segmentos do áudio (com as mesmas dimensões daqueles produzidos pelo bloco de pré-processamento) são classificados manualmente, produzindo a classe $y(t)$ à qual se deseja atribuir a amostra de entrada. Dessa forma, têm-se os pares $(y(t), \mathbf{x}(t))$ a serem empregados no treinamento supervisionado de um classificador.

Neste trabalho, RBMs são avaliadas no papel do classificador da Figura 27, com o uso de diferentes tipos de treinamento e de vetores de características. As medidas que costumam compor esses vetores são discutidas a seguir.

Figura 28 – Fluxo de dados no treinamento supervisionado.



Fonte: produção do próprio autor.

2.4.2 Medidas/características do áudio

Classificadores são normalmente alimentados com medidas ou características extraídas do sinal de interesse, as quais compõem vetores de características. Mais ainda, as referidas medidas são frequentemente *projetadas* para facilitarem a tarefa de classificação a ser realizada. Em aplicações de processamento de voz, os vetores de características são às vezes construídos a partir de parâmetros de modelos que tentam representar a forma como os sons são produzidos ou percebidos pelos sistemas vocal e auditivo humanos, respectivamente. Assim, na codificação preditiva linear (LPC –*Linear Predictive Coding*), usada na análise e síntese de voz (por exemplo, em telefonia), os efeitos do trato vocal são representados pelos coeficientes de um filtro linear tendo somente pólos (*all-pole filter*). Outros exemplos incluem os coeficientes cepstrais de predição linear (LPCC – *Linear Prediction Cepstral Coefficients*), que capturam informações sobre a propriedade de ressonância no trato vocal, e os coeficientes mel-cepstrais (MFCCs), que representam os sons com filtros que têm um papel similar ao da cóclea no sistema auditivo humano. Também costumam fazer parte dos vetores de características, medidas como a taxa de cruzamento

por zero, a energia de quadro e estatísticas de ordem superior dos resíduos de predição linear.

Como será visto no Capítulo 4, os MFCCs foram as principais medidas empregadas nos experimentos. O principal motivo para a escolha dos mesmos é a sua popularidade na literatura. De fato, ao longo dos anos, esses coeficientes têm sido usados como vetores de características em sistemas de reconhecimento de voz (ZWEIG; RUSSELL, 1998; MARTIN; DAMNATI; MAUUARY, 2001; MUDA; BEGAM; ELAMVAZUTHI, 2010; ITTICHAICHAREON; SUKSRI; YINGTHAWORNSUK, 2012) e de locutor (HASAN; JAMIL; RAHMAN, 2004; NAKAGAWA; ASAKAWA; WANG, 2007; ZULFIQAR; MUHAMMAD; ENRIQUEZ, 2009), bem como em detectores de atividade vocal (WU; CAO, 2005; ZOU et al., 2014). Em todas essas aplicações, os MFCCs produziram bons resultados. Além disso, Zou et al. (2014) compararam MFCCs, LPCs e LPCCs e concluíram que os MFCCs seriam mais separáveis do que os demais segundo uma medida de separabilidade de classes.

Além dos MFCCs, algumas medidas propostas recentemente também foram avaliadas experimentalmente. Especificamente, Sadjadi e Hansen (2013) propuseram um detector de atividade vocal empregando medidas sensíveis à presença no áudio de sons vocalizados (i.e., envolvendo a movimentação das cordas vocais). Com elas, os autores obtiveram ótimos resultados. As referidas medidas são: harmonicidade, clareza, ganho de predição e periodicidade. Drugman et al. (2015) combinaram essas quatro medidas, os MFCCs e outras medidas relacionadas com a fonte sonora: a chamada Proeminência dos Picos Cepstrais (CPP – *Cepstral Peak Prominence*) e o Somatório das Harmônicas do Resíduo (SRH – *Summation of the Residual Harmonics*) produzindo um VAD de estado da arte, de acordo com os autores. Daí a motivação para se avaliar as várias medidas citadas com o classificador baseado em RBM.

Mais a frente, serão descritas as medidas efetivamente usadas no experimentos do Capítulo 4. Nas descrições, considera-se que cada uma das medidas será extraída de um segmento do áudio, o qual é representado pelo vetor $\mathbf{s}(m)$ em que m indica o índice do segmento (quadro). Os elementos do vetor $\mathbf{s}(m)$ são denotados aqui por $s(m,n)$ e estão relacionados com as amostras do áudio segundo

$$s(m,n) = \begin{cases} a(mN_s + n), & \text{para } n = 0, \ldots, N_f - 1 \\ 0, & \text{caso contrário} \end{cases} \tag{2.77}$$

sendo que $a(n) \in \mathbb{R}$ é a n-ésima amostra de áudio, N_s é o deslocamento (em amostras) entre segmentos e N_f é o número de amostras num segmento. A taxa de amostragem do áudio, quando relevante, será indicada for f_s. Além disso, será útil nas descrições de algumas medidas (harmonicidade, clareza e ganho de predição) a autocorrelação do

segmento $\mathbf{s}(m)$ para um atraso $\tau \in \mathbb{Z}$, dada por (SADJADI; HANSEN, 2013)

$$r_{ss}(m,\tau) = \frac{\sum_{n=0}^{N_f-1} s(m,n)w_1(n)s(m,n+\tau)w_1(n+\tau)}{\sum_{n=0}^{N_f-1} w_1(n)w_1(n+\tau)}, \tag{2.78}$$

em que $w_1(n)$ representa uma janela de Hanning (OPPENHEIM; SCHAFER, 2010) de tamanho N_f.

2.4.2.1 Energia de quadro

A energia de quadro (FE – *Frame Energy*) é definida como a média quadrática das amostras no segmento, ou seja,

$$\mathrm{FE}(m) = \frac{1}{N_f} \sum_{n=0}^{N_f-1} s^2(m,n). \tag{2.79}$$

2.4.2.2 MFCCs e FBEs

A Figura 29 ilustra a sequência de operações para se obter os coeficientes mel-cepstrais (MFCCs) e as energias de banco de filtros (FBEs – *Filter-bank energies*) na forma como os mesmos foram calculados nos experimentos deste trabalho. Inicialmente, aplica-se uma janela de Hamming (OPPENHEIM; SCHAFER, 2010) de tamanho N_f (denotada por $w_2(n)$) ao segmento, obtendo-se

$$\widetilde{s}(m,n) = s(m,n)w_2(n). \tag{2.80}$$

Este é convertido para o domínio da frequência via transformada de Fourier discreta (DFT – *Discrete Fourier Transform*) conforme

$$\widetilde{S}(m,k) = \sum_{n=0}^{N_{DFT}-1} \widetilde{s}(m,n)e^{-j(2\pi/N_{DFT})nk}, \tag{2.81}$$

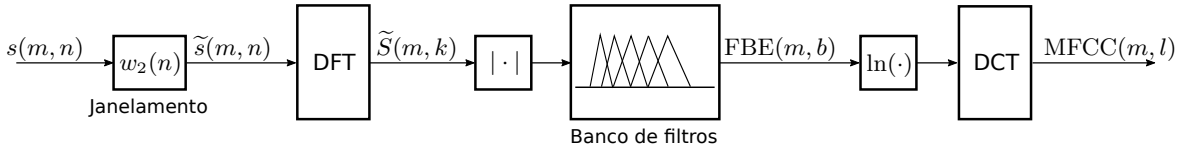
sendo que N_{DFT} indica o número de pontos da DFT. O espectro de magnitude $|\widetilde{S}(m,k)|$ é então passado por um banco de filtros triangulares espaçados linearmente na escala *mel*, produzindo as FBEs:

$$\mathrm{FBE}(m,b) = \sum_{k=0}^{N_{DFT}/2} |\widetilde{S}(m,k)|\,\Lambda_b(k), \tag{2.82}$$

em que $\Lambda_b(\cdot)$ representa o filtro triangular de índice b. Nos experimentos, foram usados $N_b = 23$ filtros na faixa de frequências de 300 Hz a 2700 Hz, os quais são apresentados na Figura 30. Conversões da frequência em Hertz para mels e vice-versa são feitas, respectivamente, segundo as fórmulas

$$M(f_{Hz}) = 1127\,\ln(1 + f_{Hz}/700) \tag{2.83}$$

Figura 29 – Cálculo dos MFCCs e FBEs.



Fonte: produção do próprio autor.

e

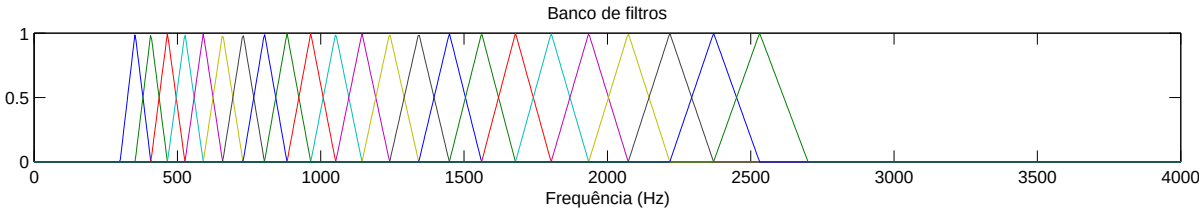
$$M^{-1}(f_{mel}) = 700\,[\exp(f_{mel}/1127) - 1]\,. \tag{2.84}$$

O MFCC de índice l é finalmente calculado como (YOUNG et al., 2002)

$$\mathrm{MFCC}(m,l) = \sqrt{\frac{2}{N_b}} \sum_{b=0}^{N_b-1} \ln(\mathrm{FBE}(m,b)) \cos\left(\frac{\pi l}{N_b}(b - \tfrac{1}{2})\right), \tag{2.85}$$

o que corresponde ao cálculo de uma transformada discreta do cosseno (DCT – *Discrete Cosine Transform*). Como os primeiros coeficientes da DCT estão associados a frequências mais baixas, os MFCCs com menores índices representam a envoltória do espectro do sinal, a qual se deseja capturar. Por isso e também por motivos históricos, são utilizados somente os 13 primeiros MFCCs (MFCC(m,l), $l = 0, \ldots 12$) na prática.

Figura 30 – Banco de filtros espaçados linearmente na escala mel.



Fonte: produção do próprio autor.

É um fato conhecido na literatura que sistemas de reconhecimento de voz podem ter seu desempenho melhorado informando-se ao classificador as derivadas temporais das medidas (além delas próprias) (YOUNG et al., 2002). As derivadas de primeira e segunda ordem tem uso comum com os MFCCs e são denotadas por Δ-MFCCs e $\Delta\Delta$-MFCCs, respectivamente. Neste trabalho, a derivada de primeira ordem de uma medida $\theta(m)$ qualquer associada ao segmento m é calculada como (YOUNG et al., 2002)

$$\Delta\theta(m) = \frac{\sum_{r=1}^{2} [\theta(m+r) - \theta(m-r)]\, r}{2 \sum_{r=1}^{2} r^2}. \tag{2.86}$$

Em (2.86), Δ pode ser visto como um operador de cálculo de derivada. As derivadas de segunda ordem são obtidas com duas aplicações consecutivas desse operador.

2.4.2.3 Harmonicidade

A harmonicidade (*harmonicity*) (relativa ao segmento de índice m) é definida como (SADJADI; HANSEN, 2013)

$$h(m) = \frac{r_{ss}(m, \tau_{max})}{r_{ss}(m, 0) - r_{ss}(m, \tau_{max})}, \tag{2.87}$$

sendo que

$$\tau_{max} = \underset{\tau \in \Phi_h}{\arg\max}\, r_{ss}(m, \tau). \tag{2.88}$$

Φ_h representa o conjunto de valores de τ para os quais o período do tom da voz (frequência principal percebida) é considerado plausível (2 ms a 16 ms, segundo Sadjadi e Hansen (2013)). Assim,

$$\Phi_h = \{\tau : \tau \in \mathbb{Z}, 2 \text{ ms} \leq \tau / f_s \leq 16 \text{ ms}\}. \tag{2.89}$$

A autocorrelação de um sinal periódico é também periódico com o mesmo período e o seu máximo assume valores próximos à autocorrelação com atraso zero. Assim, para segmentos vocalizados, $h(m)$ mostra picos bem definidos.

2.4.2.4 Clareza

A definição exata da medida denominada clareza (*clarity*) é fornecida em (SADJADI; HANSEN, 2013). Entretanto, os autores do mesmo artigo propõem o uso de uma aproximação para a clareza do segmento m, dada por

$$c(m) = 1 - \frac{\tilde{D}(m, \tau_{min})}{\tilde{D}(m, \tau_{max})}, \tag{2.90}$$

em que

$$\tau_{max} = \underset{\tau \in \Phi_c}{\arg\max}\, \tilde{D}(m, \tau) \tag{2.91}$$

e

$$\tau_{min} = \underset{\tau \in \Phi_c}{\arg\min}\, \tilde{D}(m, \tau). \tag{2.92}$$

Φ_c tem a mesma definição de Φ_h dada na Equação (2.89). $\tilde{D}(m, \tau)$ é uma aproximação para a chamada *average magnitude difference function* (AMDF) e é escrita como

$$\tilde{D}(m, \tau) = 0{,}8\, \sqrt{2\, [r_{ss}(m, 0) - r_{ss}(m, \tau)]}. \tag{2.93}$$

Segundo os autores mencionados, a clareza exibe valores elevados para segmentos parecidos com voz, enquanto se mantém em valores baixos para sons de fundo.

2.4.2.5 Ganho de predição

O ganho de predição (*prediction gain*) é definido como a razão entre a energia do sinal (segmento) e a energia do resíduo de predição linear (SADJADI; HANSEN, 2013), ou seja,

$$G_p(m) = \log_{10}\left(r_{ss}(m,0)/\epsilon^{(p)}\right). \tag{2.94}$$

Na equação, $\epsilon^{(p)}$ é o erro de predição linear de ordem p, extraído do último passo da recursão de Levinson-Durbin. A ideia para a definição do ganho de predição vem do entendimento que, em segmentos contendo voz, as amostras são mais correlacionadas e, portanto, mais fáceis de serem previstas, o que se reflete num menor valor de $\epsilon^{(p)}$. Com isso, $G_p(m)$ atinge valores maiores em segmentos vocalizados.

2.4.2.6 Periodicidade

A periodicidade (*periodicity*) é obtida a partir do espectro do produto de harmônicas (HPS – *Harmonic Product Spectrum*), o qual pode ser escrito no domínio log-espectral como

$$\mathrm{HPS}(m,k) = \sum_{l=1}^{N_{hps}} \log_{10} |\widetilde{S}(m,lk)|. \tag{2.95}$$

$\widetilde{S}(m,lk)$, a DFT de um segmento após janelamento, foi definida na Equação (2.81). Percebe-se que $\mathrm{HPS}(m,k)$ consiste na soma de várias (N_{hps}) cópias do log do espectro comprimidas no eixo das frequências. Com essa compressão, espera-se que harmônicas coincidam com a frequência da correspondente fundamental, o que justifica a definição da periodicidade como

$$P_{hps}(m) = \max_{k \in \Phi_{hps}} \mathrm{HPS}(m,k). \tag{2.96}$$

Φ_{hps} representa o conjunto de valores de k para os quais a frequência principal (tom) da voz é considerada plausível (62,5 Hz a 500 Hz, segundo Sadjadi e Hansen (2013)), isto é,

$$\Phi_{hps} = \{k : k \in \mathbb{Z}, 62{,}5 \text{ Hz} \leq (kf_s)/N_{DFT} \leq 500 \text{ Hz}\}. \tag{2.97}$$

2.4.2.7 SRH_1 e SRH_2

O cálculo do Somatório das Harmônicas do Resíduo (SRH – *Summation of the Residual Harmonics*) é baseado na análise do resíduo de predição linear autorregressiva. Seja $e(m,n)$ a amostra de índice n do resíduo de predição para o segmento m, e $\tilde{e}(m,n)$, o resultado da aplicação da janela de Hanning à $e(m,n)$. Seja também $\tilde{E}(m,k)$ a DFT de $\tilde{e}(m,n)$, calculada como

$$\tilde{E}(m,k) = \sum_{n=0}^{N_{DFT}-1} \tilde{e}(m,n) e^{-j(2\pi/N_{DFT})nk}. \tag{2.98}$$

O SRH a uma frequência k (ciclos por N_{DFT} amostras) é escrito como (DRUGMAN; ALWAN, 2011)

$$\mathrm{SRH}(m,k) = |\tilde{E}(m,k)| + \sum_{l=2}^{N_{srh}} \left[|\tilde{E}(m,lk)| - |\tilde{E}(m,(l-\tfrac{1}{2})k)| \right], \tag{2.99}$$

sendo N_{srh} o número de harmônicas consideradas. A primeira medida (escalar) derivada de (2.99) é indicada neste texto como

$$\mathrm{SRH}_1(m) = \underset{k \in \Phi_{srh}}{\arg\max}\, \mathrm{SRH}(m,k), \tag{2.100}$$

sendo que Φ_{srh} representa o conjunto de valores de k para os quais a frequência principal (tom) da voz é considerada plausível (80 Hz a 350 Hz)[1], isto é,

$$\Phi_{srh} = \{k : k \in \mathbb{Z}, 80 \text{ Hz} \leq (kf_s)/N_{DFT} \leq 350 \text{ Hz}\}. \tag{2.101}$$

Uma versão modificada de (2.100), difere desta somente por uma normalização (DRUGMAN et al., 2015) e é denotada aqui como

$$\mathrm{SRH}_2(m) = \frac{\mathrm{SRH}_1(m)}{\sqrt{\sum_{k=0}^{N_{DFT}/2} |\tilde{E}(m,k)|^2}}. \tag{2.102}$$

2.4.2.8 CPP

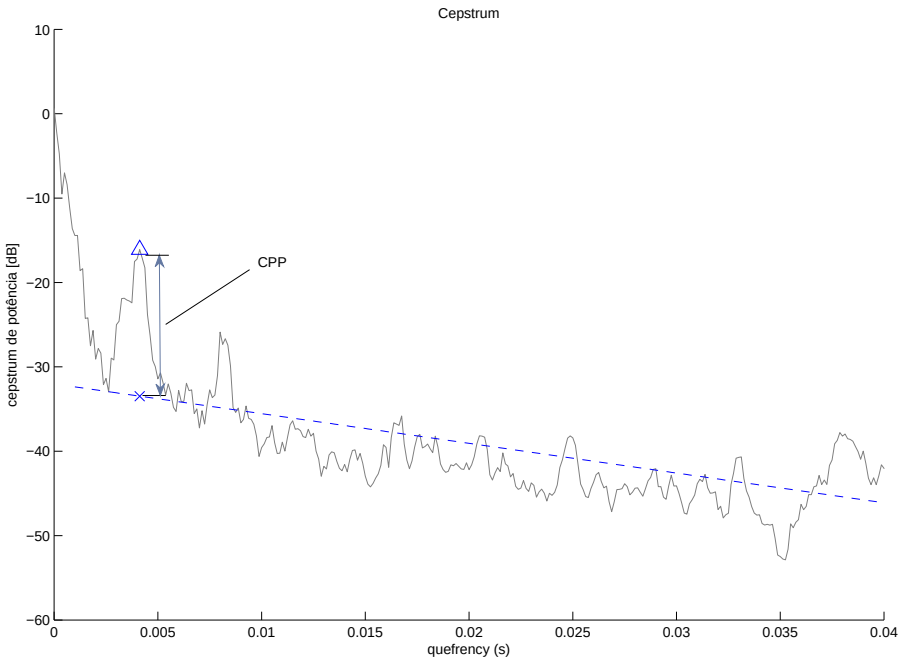
A medida CPP (HILLENBRAND; CLEVELAND; ERICKSON, 1994) é determinada a partir do *cepstrum* de potência do sinal (áudio). O *cepstrum* de potência de um sinal $\tilde{s}(t)$ de tempo contínuo é dado por (NORTON; KARCZUB, 2003)

$$C_{p\tilde{s}\tilde{s}}(\tau) = \left| \mathcal{F}^{-1} \left\{ \log_{10} \left| \mathcal{F}\left\{\tilde{s}(t)\right\} \right|^2 \right\} \right|^2, \tag{2.103}$$

sendo que $\mathcal{F}\{\cdot\}$ e $\mathcal{F}^{-1}\{\cdot\}$ representam as transformadas de Fourier direta e inversa, respectivamente. No cálculo da CPP, $\tilde{s}(t)$ corresponderia a um segmento do sinal após aplicação de uma janela de Hanning. A conversão de (2.103) para domínios discretos (no tempo e em frequência) é direta e será omitida aqui. A Figura 31 ilustra a determinação da medida. Basicamente, ela corresponde à diferença de amplitude (em dB) entre o pico cepstral e o valor correspondente da linha de regressão diretamente abaixo do pico. O valor em decibéis na figura (eixo vertical) é calculado como $10 \log_{10} C_{p\tilde{s}\tilde{s}}(\tau)$. A linha de regressão é determinada por mínimos quadrados utilizando todos os valores do *cepstrum* para tempos (*quefrency*) maiores que 1 milissegundo.

[1] Valores determinados a partir do código liberado pelos autores (Projeto COVAREP).

Figura 31 – Determinação da medida CPP a partir do *cepstrum* de potência.



Fonte: produção do próprio autor.

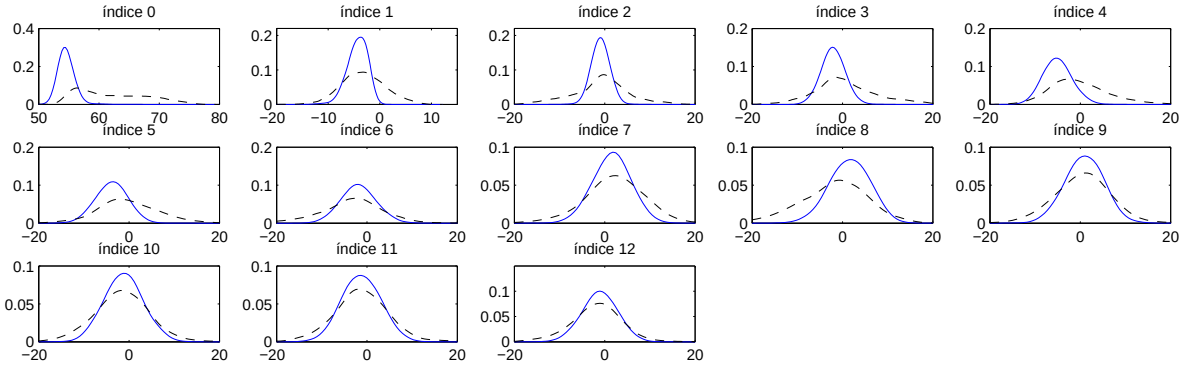
2.4.2.9 Experimentos

Nesta seção, são brevemente apresentados alguns resultados de experimentos envolvendo as diversas medidas vistas anteriormente. Nos experimentos, as medidas foram extraídas de áudios reais obtidos do *corpus* (base de dados) chamado NOIZEUS (HU; LOIZOU, 2008), o qual será melhor detalhado no Capítulo 4. Os arquivos de áudio que compõem o *corpus* foram rotulados manualmente para este trabalho, de modo que a informação sobre a ausência ou presença de voz em cada trecho de áudio é considerada conhecida.

No primeiro experimento, o áudio em cada arquivo foi dividido em segmentos de 25 ms de duração com 10 ms de deslocamento entre segmentos, a partir dos quais foram extraídos os MFCCs. De posse dessas medidas e das correspondentes classificações esperadas para cada segmento (da rotulação manual), foram determinadas as distribuições empíricas $\hat{P}(\theta|\mathcal{H}_0)$ e $\hat{P}(\theta|\mathcal{H}_1)$, em que θ representa, no caso atual, algum dos MFCCs. $\mathcal{H}_0$ e $\mathcal{H}_1$ denotam as hipóteses de ausência e de presença de voz, respectivamente. Os resultados para o áudio com SNR de 20 dB são mostrados na Figura 32. Pode-se perceber pela figura que as distribuições dos vários coeficientes mel-cepstrais podem ser razoavelmente aproximadas por distribuições gaussianas, tanto na hipótese $\mathcal{H}_0$ (linha contínua na figura) quanto na $\mathcal{H}_1$ (linha tracejada). Apenas o MFCC de índice 0 sob $\mathcal{H}_1$ é que parece ter sua distribuição melhor aproximada por uma mistura de gaussianas. De fato, esses resultados para os MFCCs serviram de motivação para o desenvolvimento da variante de RBM sendo

proposta nesta dissertação e detalhada no Capítulo 3.

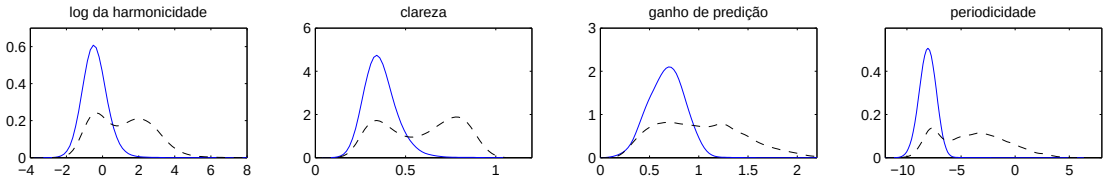
Figura 32 – Densidade de probabilidade de cada MFCC na ausência (linha contínua) e na presença de voz (linha tracejada).



Fonte: produção do próprio autor.

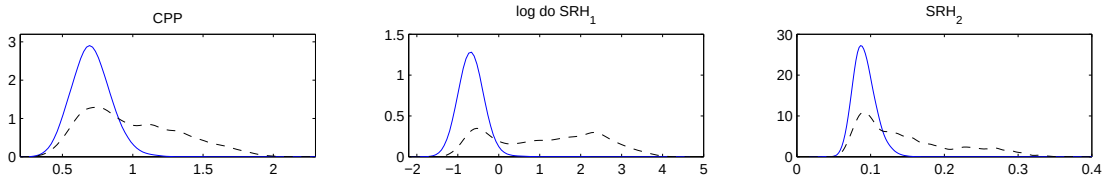
O experimento anterior foi repetido para as medidas propostas por Sadjadi e Hansen (2013) e para aquelas cujo uso em VAD foi sugerido por Drugman et al. (2015), sendo que suas distribuições são apresentadas nas Figuras 33 e 34, respectivamente. As medidas harmonicidade, e SRH_1 foram transformadas a fim de fazer com que as distribuições resultantes ficassem mais parecidas com gaussianas ou misturas de gaussianas. A simples aplicação da função log (nesse caso, na base e) às medidas mencionadas produz o efeito desejado, como pode ser verificado pelas figuras.

Figura 33 – Densidade de probabilidade de cada medida proposta por Sadjadi e Hansen (2013), na ausência (linha contínua) e na presença de voz (linha tracejada).



Fonte: produção do próprio autor.

Figura 34 – Densidade de probabilidade de cada medida proposta por Drugman et al. (2015), na ausência (linha contínua) e na presença de voz (linha tracejada).



Fonte: produção do próprio autor.

Observando-se as distribuições dos MFCCs e das medidas propostas recentemente, tem-se a impressão de que essas últimas teriam um potencial para produzir bons resultados

Tabela 4 – Distância de Bhattacharyya oferecida pelo uso de cada medida a diferentes SNRs (distância medida entre as distribuições obtidas nas condições de ausência e de presença de voz).

Medida	**∞ dB**	**20 dB**	**15 dB**	**10 dB**	**5 dB**	**0 dB**	**-5 dB**
MFCC$(\cdot, 0)$	**1,857**	**0,602**	**0,421**	**0,293**	**0,170**	**0,098**	**0,032**
MFCC$(\cdot, 1)$	0,373	0,158	0,102	0,070	0,037	0,012	0,002
MFCC$(\cdot, 2)$	0,241	0,203	0,167	0,132	0,084	0,046	**0,019**
MFCC$(\cdot, 3)$	0,229	0,180	0,140	0,102	0,060	0,027	0,007
MFCC$(\cdot, 4)$	0,143	0,177	0,136	0,100	0,067	0,030	0,013
MFCC$(\cdot, 5)$	0,133	0,134	0,106	0,075	0,037	0,023	0,004
MFCC$(\cdot, 6)$	0,101	0,075	0,054	0,040	0,024	0,011	0,004
MFCC$(\cdot, 7)$	0,087	0,047	0,035	0,025	0,016	0,005	0,002
MFCC$(\cdot, 8)$	0,118	0,086	0,056	0,040	0,028	0,009	0,003
MFCC$(\cdot, 9)$	0,049	0,031	0,022	0,015	0,006	0,005	0,001
MFCC$(\cdot, 10)$	0,052	0,031	0,023	0,017	0,007	0,002	0,002
MFCC$(\cdot, 11)$	0,049	0,028	0,022	0,015	0,008	0,003	0,001
MFCC$(\cdot, 12)$	0,061	0,031	0,023	0,020	0,009	0,003	0,002
$\ln(\mathrm{FE}(\cdot))$	**2,018**	**0,636**	**0,448**	**0,317**	**0,181**	**0,102**	**0,033**
$\ln(h(\cdot))$	0,418	0,268	0,201	0,134	0,068	0,014	0,002
$c(\cdot)$	0,565	0,283	0,224	0,141	0,073	0,017	0,001
$G_p(\cdot)$	**0,888**	0,256	0,163	0,088	0,037	0,003	0,002
$P_{hps}(\cdot)$	**1,822**	**0,582**	**0,475**	**0,344**	**0,207**	**0,082**	**0,023**
CPP$(\cdot)$	0,396	0,229	0,190	0,136	0,097	0,046	0,017
$\ln(\mathrm{SRH}_1(\cdot))$	**1,755**	**0,566**	**0,480**	**0,328**	**0,214**	**0,094**	**0,026**
$\mathrm{SRH}_2(\cdot)$	0,411	**0,308**	**0,253**	**0,173**	**0,107**	**0,050**	0,018

na aplicação de VAD, visto que a sobreposição entre suas distribuições sob as duas hipóteses parece menor do que para os MFCCs. Para confirmar essa impressão, foram calculadas, para cada medida, as distâncias de Bhattacharyya entre suas distribuições nas duas hipóteses. A distância de Bhattacharyya entre duas distribuições $P_1(x)$ e $P_2(x)$ quaisquer, definidas sobre um mesmo domínio, é dada por (KAILATH, 1967)

$$D_B(P_1, P_2) = -\ln\left(\int \sqrt{P_1(x)P_2(x)}dx\right). \tag{2.104}$$

Na Tabela 4, são fornecidos os resultados referentes a $D_B(\hat{P}(\theta|\mathcal{H}_0), \hat{P}(\theta|\mathcal{H}_1))$, sendo θ a medida indicada na primeira coluna. Os resultados estão separados por SNR do áudio (colunas da tabela) e, em cada coluna, estão destacadas em negrito as 5 medidas que exibiram os maiores valores de distância. Pode-se notar que algumas das medidas propostas recentemente (em especial, $P_{hps}(\cdot)$ e $\ln(\mathrm{SRH}_1)$) ficaram entre as 5 mais discriminativas (segundo a distância de Bhattacharyya) em todas SNRs. Entretanto, o mesmo fato é constatado com relação a algumas medidas tradicionais: o MFCC de índice 0 e o logaritmo da energia de quadro ($\ln(\mathrm{FE}(\cdot))$). Essa última foi incluída na avaliação por ser comumente utilizada em combinação com os MFCCs em sistemas de reconhecimento de fala (ZHENG; ZHANG, 2000).

Conclusões

Com relação às distribuições apresentadas, alguns pontos podem ser destacados:

1. Os MFCCs apresentam naturalmente distribuições bem aproximadas por gaussianas ou misturas de gaussianas;
2. As medidas recentes apresentam ou podem ser trivialmente transformadas de modo a apresentarem distribuições do mesmo tipo.

Assim, enquanto o Item 1 ensejou o desenvolvimento da variante de RBM proposta neste trabalho, o Item 2 indica que essa variante é facilmente aplicável às medidas recentes.

Considerando as distâncias de Bhattacharyya determinadas nos experimentos, pode-se concluir que algumas das medidas recentes fornecem valores maiores de distância do que a maioria dos MFCCs. Entretanto, o MFCC de índice 0 e o $\ln(\mathrm{FE}(\cdot))$ se mostram comparáveis ou até superiores às medidas recentes mais discriminativas ($P_{hps}(\cdot)$ e $\ln(\mathrm{SRH}_1)$), na maioria das SNRs. Além disso, as distâncias de Bhattacharyya avaliam as medidas individualmente. Não é possível inferir, a partir dos resultados obtidos, quais combinações de medidas (caso existam) poderiam ser mais poderosas no sentido de discriminar as hipóteses de interesse ($\mathcal{H}_0$ e $\mathcal{H}_1$). Os MFCCs são reconhecidamente pouco correlacionados entre si, de modo que a combinação de um maior número desses coeficientes deve também fornecer mais informações a um classificador. Já as medidas recentes não foram construídas com o objetivo de serem pouco correlacionadas entre si. Em vista desses fatores e da dificuldade em se avaliar combinações de medidas de forma independente da aplicação, definiu-se um dos objetivos do presente trabalho: avaliar combinações das várias medidas na aplicação de VAD, empregando a variante proposta de RBM.

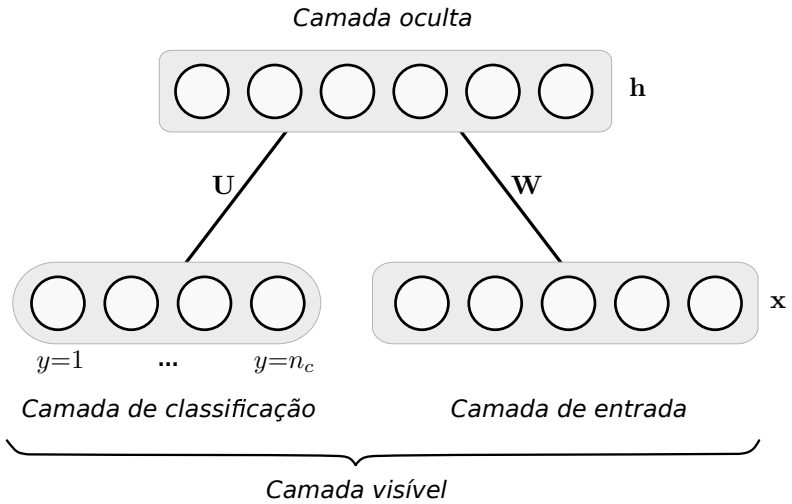
3 RBM GAUSS-BERNOULLI COMO CLASSIFICADOR

Neste capítulo, é proposta uma versão de RBM que possibilita seu uso em tarefas de classificação quando os dados de entrada são definidos num domínio contínuo. Os tópicos do presente capítulo seguem aproximadamente a organização da Seção 2.1, na qual as RBMs Bernoulli-Bernoulli foram introduzidas. Na Seção 3.1, é apresentada a estrutura do modelo, seguida, na Seção 3.2, pela sua formulação matemática. Na Seção 3.3, são detalhadas as diferentes tarefas associadas à variante proposta. Nessa seção, são apresentados os diferentes tipos de treinamento admitidos pelo modelo. Para cada tipo de treinamento, são inclusive fornecidas as regras de atualização de parâmetros necessárias à sua realização. Na Seção 3.4, a operação da variante é ilustrada por meio de experimentos computacionais. Por fim, a Seção 3.5 fornece as conclusões deste capítulo.

3.1 ESTRUTURA

A estrutura da variante é apresentada na Figura 35 e é idêntica à estrutura da RBM Bernoulli-Bernoulli com camada de classificação (Figura 26). Basicamente, existe uma camada oculta conectada às camadas de entrada e de classificação. Essas duas últimas formam a camada visível do modelo. As unidades da camada de classificação são binárias e considera-se que apenas uma delas esteja ativa (assuma o valor 1) a qualquer momento. Nesta versão de RBM, como será visto à frente, as unidades da camada oculta são binárias enquanto as de entrada podem assumir valores reais.

Figura 35 – Estrutura da variante proposta: estrutura idêntica à da RBM Bernoulli-Bernoulli com camada de classificação.



Fonte: extraído de Larochelle e Bengio (2008).

3.2 DESCRIÇÃO MATEMÁTICA

Assim como nas demais versões de RBM, as unidades são expressas matematicamente por variáveis aleatórias. Em particular, as unidades de entrada são representadas pelo vetor aleatório $\mathbf{X}$, de dimensão n_d, e um estado específico de $\mathbf{X}$, pelo vetor $\mathbf{x}$, conforme as definições nas Equações (2.59) e (2.60) da página 60. Similarmente, as unidades ocultas são descritas pelo vetor aleatório $\mathbf{H}$, de dimensão n_h, e um estado específico do mesmo, pelo vetor $\mathbf{h}$, conforme Equações (2.3) e (2.4). Finalmente, a unidade ativa na camada de classificação é indicada pela variável aleatória Y, a qual pode assumir valores no conjunto $\{1, 2, \ldots, n_c\}$, sendo n_c o número de classes (ou rótulos) nas quais uma entrada pode ser classificada. Um estado particular de Y é indicado pelo escalar y. Nota-se que essas definições são idênticas àquelas empregadas na RBM Bernoulli-Bernoulli com uma camada de classificação, descrita na Seção 2.3.

Com explicado na Seção 2.1.5, diferentes definições para a função de energia global levam a versões distintas do modelo. Na variante proposta, a função mencionada é definida como

$$\begin{aligned} E_{gc}(y, \mathbf{x}, \mathbf{h}) = &- \sum_{j=1}^{n_h} b_j h_j + \sum_{i=1}^{n_d} \frac{(x_i - c_i)^2}{2\sigma_i^2} - \sum_{i=1}^{n_d} \sum_{j=1}^{n_h} h_j w_{ji} \frac{x_i}{\sigma_i^2} \\ &- \sum_{k=1}^{n_c} d_k \delta_{k,y} - \sum_{k=1}^{n_c} \sum_{j=1}^{n_h} h_j u_{jk} \delta_{k,y}, \end{aligned} \tag{3.1}$$

sendo que a notação $\delta_{r,s}$ representa o delta de Kronecker e w_{ji}, b_j, c_i, σ_i^2, d_k, u_{jk}, com i=1, $\ldots$, n_d, j=1, $\ldots$, n_h, k=1, $\ldots$, n_c, constituem os parâmetros do modelo. As definições dadas nas Equações (2.62) a (2.66) da página 61 são reutilizadas aqui e as denominadas variâncias das variáveis de entrada são agrupadas no vetor

$$\boldsymbol{\sigma}^2 = [\sigma_1^2 \ldots \sigma_{n_d}^2]^T. \tag{3.2}$$

Se a camada de entrada for considerada como a camada visível da RBM Gauss-Bernoulli, pode-se ver que a definição na Equação (3.2) é igual àquela dada na Equação (2.37). Com as definições anteriores, o presente modelo fica totalmente especificado pelo conjunto de parâmetros $\{\mathbf{W}, \mathbf{b}, \mathbf{c}, \boldsymbol{\sigma}^2, \mathbf{d}, \mathbf{U}\}$.

Vale dizer que a forma da função de energia global expressa na Equação (3.1) foi inspirada nas variantes descritas nas Seções 2.2 e 2.3. Comparando-se as Equações (2.36) e (3.1), torna-se imediato notar que a expressão à direita da igualdade na primeira linha da Equação (3.1) corresponde à RBM Gauss-Bernoulli, enquanto que os termos na segunda linha da mesma equação refletem a adição da camada de classificação ao modelo, como feito na Equação (2.61) para incluir a mesma camada na RBM Bernoulli-Bernoulli.

A partir da energia global, define-se a distribuição de probabilidade conjunta das

variáveis do modelo como

$$P_{gc}(y, \mathbf{x}, \mathbf{h}) = \frac{\exp(-E_{gc}(y, \mathbf{x}, \mathbf{h}))}{Z_{gc}}. \tag{3.3}$$

Essa função tem domínio misto, contínuo para os estados das variáveis de entrada e discreto para os demais, sendo escrito explicitamente como $\{1, 2, \ldots, n_c\} \times \mathbb{R}^{n_d} \times \{0, 1\}^{n_h}$. A constante no denominador garante que $P_{gc}(y, \mathbf{x}, \mathbf{h})$ tenha soma unitária no seu domínio, ou seja,

$$Z_{gc} = \sum_{y \in \{1,2,\ldots,n_c\}} \sum_{\mathbf{h} \in \{0,1\}^{n_h}} \int_{\mathbb{R}^{n_d}} \exp(-E_{gc}(y, \mathbf{x}, \mathbf{h}))\, \mathbf{dx}. \tag{3.4}$$

Com base na Equação (3.3), pode-se determinar todas as distribuições necessárias à realização das diferentes tarefas relativas ao modelo, como será visto na próxima seção.

3.3 TAREFAS ASSOCIADAS AO MODELO

A variante proposta permite exatamente as mesmas tarefas suportadas pela RBM Bernoulli-Bernoulli com camada de classificação. Essas tarefas são detalhadas nas seções a seguir. Inicialmente, nas Seções 3.3.1 e 3.3.2, são abordadas a inferência e a amostragem do modelo, respectivamente. Os três tipos de treinamento introduzidos na Seção 2.3.3.3 são então discutidos individualmente nas Seções 3.3.3 a 3.3.5. A tarefa de classificação é finalmente examinada na Seção 3.3.6. No Apêndice C, são fornecidas as demonstrações das várias equações que serão mostradas aqui.

3.3.1 Inferência

Para as RBMs em geral, tem-se como principal interesse a realização de inferência entre as variáveis do modelo e, em particular, as inferências que permitam a realização da amostragem de Gibbs. Para isso, deve-se conhecer algumas das distribuições condicionais do modelo: $P_{gc}(\mathbf{x}|\mathbf{h})$, $P_{gc}(y|\mathbf{h})$ e $P_{gc}(\mathbf{h}|y, \mathbf{x})$. Assim, a partir da Equação (3.3) pode-se demonstrar que as variáveis de entrada são independentes entre si dado o estado das ocultas, ou seja,

$$P_{gc}(\mathbf{x}|\mathbf{h}) = \prod_{i=1}^{n_d} P_{gc}(x_i|\mathbf{h}) \tag{3.5}$$

e têm distribuição individualmente Gaussiana dada por

$$P_{gc}(x_i|\mathbf{h}) = \mathcal{N}\left(x_i \middle| c_i + \sum_{j=1}^{n_h} h_j w_{ji}, \sigma_i^2\right), \tag{3.6}$$

em que a notação $\mathcal{N}(\cdot|\mu, \sigma^2)$ representa a função de densidade de probabilidade Gaussiana com média μ e variância σ^2. Pode-se perceber que essa última equação tem forma idêntica àquela da Equação (2.43) relativa às RBMs Gauss-Bernoulli. Conhecido o estado das

variáveis ocultas, pode-se também inferir valores para a classe, y, segundo

$$P_{gc}(y|\mathbf{h}) = \frac{\exp\left(d_y + \sum\limits_{j=1}^{n_h} h_j u_{jy}\right)}{\sum\limits_{k=1}^{n_c} \exp\left(d_k + \sum\limits_{j=1}^{n_h} h_j u_{jk}\right)}. \tag{3.7}$$

Deve-se notar que as Equações (3.6) e (3.7) fornecem probabilidades associadas às variáveis visíveis do modelo. No sentido oposto, ou seja, dada uma amostra de entrada, $\mathbf{x}$, e sua classe, y, infere-se o valor das variáveis ocultas por meio de $P_{gc}(\mathbf{h}|y, \mathbf{x})$. Demonstra-se que as variáveis ocultas são independentes entre si dado o estado das visíveis, $(y, \mathbf{x})$, isto é,

$$P_{gc}(\mathbf{h}|y, \mathbf{x}) = \prod_{j=1}^{n_h} P_{gc}(h_j|y, \mathbf{x}) \tag{3.8}$$

e têm individualmente distribuição de Bernoulli com probabilidade de sucesso

$$P_{gc}(h_j{=}1|y, \mathbf{x}) = \varphi\left(b_j + u_{jy} + \sum_{i=1}^{n_d} w_{ji} \frac{x_i}{\sigma_i^2}\right), \tag{3.9}$$

sendo $\varphi(\cdot)$ a função sigmóide já introduzida. Vale antecipar que, dada uma amostra de entrada, pode-se inferir a sua classe empregando-se a distribuição condicional $P_{gc}(y|\mathbf{x})$. Esse caso particular de inferência é descrito no tópico sobre classificação mais à frente.

3.3.2 Amostragem

Como ocorre com as RBMs Bernoulli-Bernoulli, deseja-se gerar, na amostragem, amostras consistentes com a distribuição das variáveis visíveis do modelo. Na variante proposta, uma amostra da camada visível pode ser representada pelo par $(y, \mathbf{x})$, de modo que $P_{gc}(y, \mathbf{x})$ é a distribuição de interesse. A fim de produzir amostras segundo essa distribuição, usa-se a amostragem de Gibbs, cuja adaptação para essa versão de RBM é imediata. A Figura 36 mostra o resultado dessa adaptação usando a notação introduzida na Seção 2.1.3.2. Conforme a figura, o primeiro passo de Gibbs produz a sequência de amostras $(y^{(0)}, \mathbf{x}^{(0)}) \rightarrow \mathbf{h}^{(1)} \rightarrow (y^{(1)}, \mathbf{x}^{(1)})$, sendo que $\mathbf{h}^{(1)}$ é produzida segundo a distribuição $P_{gc}(\mathbf{h}|y^{(0)}, \mathbf{x}^{(0)})$ e as componentes do par $(y^{(1)}, \mathbf{x}^{(1)})$, segundo as distribuições $P_{gc}(y|\mathbf{h}^{(1)})$ e $P_{gc}(\mathbf{x}|\mathbf{h}^{(1)})$, respectivamente. Passos de amostragem seguintes seguem o mesmo processo.

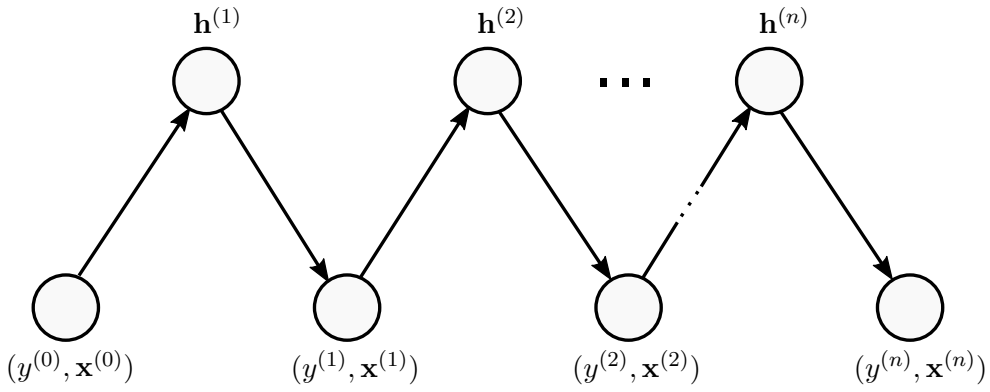
3.3.3 Treinamento generativo

De forma análoga à apresentada na Seção 2.3.3.3, a função perda generativa associada a uma dada amostra de treinamento, $(y^{(t)}, \mathbf{x}^{(t)})$, é definida como

$$\mathcal{L}_{gen}(\boldsymbol{\theta}; y^{(t)}, \mathbf{x}^{(t)}) = -\ln P_{gc}(y^{(t)}, \mathbf{x}^{(t)}), \tag{3.10}$$

na qual a função de distribuição de probabilidade conjunta das variáveis de entrada e da classe, $P_{gc}(y, \mathbf{x})$, é avaliada na amostra de treinamento. Nessa equação, $\boldsymbol{\theta}$ representa um

Figura 36 – Amostragem de Gibbs na variante proposta.



Fonte: produção do próprio autor.

vetor contendo todos os parâmetros do modelo. No treinamento generativo, objetiva-se minimizar $\mathcal{L}_{gen}(\boldsymbol{\theta}; y^{(t)}, \mathbf{x}^{(t)})$ com relação a $\boldsymbol{\theta}$, o que é equivalente a maximizar a verossimilhança de se observar simultaneamente $y^{(t)}$ e $\mathbf{x}^{(t)}$ segundo a distribuição de probabilidade das variáveis visíveis do modelo. Mais informalmente, pode-se dizer que o ajuste dos parâmetros visa maximizar a probabilidade de o modelo gerar a amostra $(y^{(t)}, \mathbf{x}^{(t)})$, fato que justifica a denominação de treinamento *generativo*.

Assim como nas RBMs Bernoulli-Bernoulli, a minimização de $\mathcal{L}_{gen}(\boldsymbol{\theta}; y^{(t)}, \mathbf{x}^{(t)})$ é normalmente obtida empregando-se o SGD. Para o seu uso, faz-se necessário conhecer o gradiente da função perda com relação a $\boldsymbol{\theta}$. Como demonstrado no Apêndice C, esse gradiente é dado por

$$\begin{aligned}\frac{\partial \mathcal{L}_{gen}(\boldsymbol{\theta}; y^{(t)}, \mathbf{x}^{(t)})}{\partial \boldsymbol{\theta}} &= \mathbb{E}_{P_{gc}(\mathbf{h}|y,\mathbf{x})}\left[\frac{\partial E_{gc}(y^{(t)}, \mathbf{x}^{(t)}, \mathbf{H})}{\partial \boldsymbol{\theta}}\middle| y^{(t)}, \mathbf{x}^{(t)}\right] \\ &- \mathbb{E}_{P_{gc}(y,\mathbf{x},\mathbf{h})}\left[\frac{\partial E_{gc}(Y, \mathbf{X}, \mathbf{H})}{\partial \boldsymbol{\theta}}\right].\end{aligned} \tag{3.11}$$

A esperança no segundo termo à direita da igualdade é intratável, como nas RBMs originais. Os algoritmos CD-k, PCD e PT (descritos na Seção 2.1.3.3), que podem ser facilmente adaptados para a variante proposta, lidam com a citada intratabilidade fazendo a aproximação

$$\mathbb{E}_{P_{gc}(y,\mathbf{x},\mathbf{h})}\left[\frac{\partial E_{gc}(Y, \mathbf{X}, \mathbf{H})}{\partial \boldsymbol{\theta}}\right] \approx \mathbb{E}_{P_{gc}(\mathbf{h}|y,\mathbf{x})}\left[\frac{\partial E_{gc}(\widetilde{y}, \widetilde{\mathbf{x}}, \mathbf{H})}{\partial \boldsymbol{\theta}}\middle| \widetilde{y}, \widetilde{\mathbf{x}}\right], \tag{3.12}$$

com o par $(\widetilde{y}, \widetilde{\mathbf{x}})$ obtido por meio da amostragem de Gibbs. Com o uso dessa aproximação na Equação (3.11), chega-se a uma expressão tratável para o gradiente da função perda,

cuja aplicação ao SGD leva às seguintes regras de atualização de parâmetros:

$$b_j \leftarrow b_j + \lambda_{gen}\Big[P_{gc}(h_j{=}1|y^{(t)}, \mathbf{x}^{(t)}) - P_{gc}(h_j{=}1|\widetilde{y}, \widetilde{\mathbf{x}})\Big], \tag{3.13}$$

$$c_i \leftarrow c_i + \lambda_{gen}\Big[x_i^{(t)} - \widetilde{x}_i\Big](\sigma_i^2)^{-1}, \tag{3.14}$$

$$d_k \leftarrow d_k + \lambda_{gen}\Big[\delta_{k,y^{(t)}} - \delta_{k,\widetilde{y}}\Big], \tag{3.15}$$

$$w_{ji} \leftarrow w_{ji} + \lambda_{gen}\Big[P_{gc}(h_j{=}1|y^{(t)}, \mathbf{x}^{(t)})\, x_i^{(t)} - P_{gc}(h_j{=}1|\widetilde{y}, \widetilde{\mathbf{x}})\, \widetilde{x}_i\Big](\sigma_i^2)^{-1}, \tag{3.16}$$

$$u_{jk} \leftarrow u_{jk} + \lambda_{gen}\Big[P_{gc}(h_j{=}1|y^{(t)}, \mathbf{x}^{(t)})\, \delta_{k,y^{(t)}} - P_{gc}(h_j{=}1|\widetilde{y}, \widetilde{\mathbf{x}})\, \delta_{k,\widetilde{y}}\Big], \tag{3.17}$$

sendo que λ_{gen} é a taxa de aprendizagem generativa, e $x_i^{(t)}$ e $\widetilde{x}_i$ representam as i-ésimas componentes do vetores $\mathbf{x}^{(t)}$ e $\widetilde{\mathbf{x}}$, respectivamente. A expressão para $P_{gc}(h_j{=}1|y, \mathbf{x})$ é dada na Equação (3.9). As regras de atualização dos parâmetros σ_i^2 não são mostradas aqui pois o aprendizado dos mesmos pode ser evitado fazendo-se uma normalização prévia nas variâncias dos dados de entrada. Além disso, a fim de evitar a divergência do algoritmo de treinamento, o ajuste desses parâmetros comumente exige taxas de aprendizagem algumas ordens de grandeza menores do que aquelas usadas quando os parâmetros não são ajustados. Consequentemente, o aprendizado das variâncias acaba também por causar uma maior lentidão no treinamento. Pode-se dizer que esse problema é herdado das RBMs Gauss-Bernoulli, visto que o mesmo fenômeno é observado no treinamento desses modelos (HINTON, 2010).

3.3.4 Treinamento discriminativo

A função perda discriminativa é definida como

$$\mathcal{L}_{disc}(\boldsymbol{\theta}; y^{(t)}, \mathbf{x}^{(t)}) = -\ln P_{gc}(y^{(t)}|\mathbf{x}^{(t)}), \tag{3.18}$$

em que, novamente, $\boldsymbol{\theta}$ é um vetor cujas componentes são os vários parâmetros do modelo e o par $(y^{(t)}, \mathbf{x}^{(t)})$ é uma amostra de treinamento. $P_{gc}(y|\mathbf{x})$ é a distribuição condicional das classes dado o estado das variáveis de entrada. No treinamento discriminativo, busca-se minimizar $\mathcal{L}_{disc}(\boldsymbol{\theta}; y^{(t)}, \mathbf{x}^{(t)})$ com relação aos parâmetros do modelo, o que corresponde a maximizar a verossimilhança de se observar $y^{(t)}$ (a classe esperada) quando fixada a amostra de entrada. Percebe-se, portanto, que o objetivo desse tipo de treinamento é direcionado para a tarefa de classificação. Como citado anteriormente, uma RBM treinada com o emprego da função perda discriminativa é denominada RBM discriminativa (DRBM).

Visando o uso do SGD no treinamento discriminativo, o gradiente da função perda dada em (3.18) deve ser conhecido. Pode-se demonstrar que para os parâmetros d_k, $k{=}1, \ldots, n_c$, tem-se

$$\frac{\partial \mathcal{L}_{disc}(\boldsymbol{\theta}; y^{(t)}, \mathbf{x}^{(t)})}{\partial d_k} = -\delta_{k,y^{(t)}} + P_{gc}(y{=}k|\mathbf{x}^{(t)}), \tag{3.19}$$

enquanto que para todos os demais parâmetros vale

$$\frac{\partial \mathcal{L}_{disc}(\theta; y^{(t)}, \mathbf{x}^{(t)})}{\partial \theta} = -\sum_{j=1}^{n_h} P_{gc}(h_j{=}\,1|y^{(t)}, \mathbf{x}^{(t)}) \frac{\partial o_j(y^{(t)}, \mathbf{x}^{(t)})}{\partial \theta} \tag{3.20}$$
$$+\mathbb{E}_{P_{gc}(y|\mathbf{x})}\left[\sum_{j=1}^{n_h} P_{gc}(h_j{=}\,1|Y, \mathbf{x}^{(t)}) \frac{\partial o_j(Y, \mathbf{x}^{(t)})}{\partial \theta}\bigg|\mathbf{x}^{(t)}\right],$$

sendo que o escalar θ indica qualquer um desses parâmetros e a função auxiliar $o_j(y, \mathbf{x})$ é escrita como

$$o_j(y, \mathbf{x}) = b_j + u_{jy} + \sum_{i=1}^{n_d} w_{ji} \frac{x_i}{\sigma_i^2}. \tag{3.21}$$

Deve-se destacar que as expressões nas Equações (3.19) e (3.20) são tratáveis e podem ser avaliadas de forma exata e eficiente, não necessitando, portanto, de qualquer tipo de aproximação (LAROCHELLE; BENGIO, 2008). Assim, de posse do gradiente da função perda, chega-se às seguintes regras de atualização de parâmetros segundo o SGD:

$$b_j \leftarrow b_j + \lambda_{disc}\, \Delta b_j, \tag{3.22}$$
$$d_k \leftarrow d_k + \lambda_{disc}\, \Delta d_k, \tag{3.23}$$
$$w_{ji} \leftarrow w_{ji} + \lambda_{disc}\, \Delta b_j\, x_i^{(t)} (\sigma_i^2)^{-1}, \tag{3.24}$$
$$u_{jk} \leftarrow u_{jk} + \lambda_{disc}\, \Delta d_k\, P_{gc}(h_j{=}1|y{=}k, \mathbf{x}^{(t)}), \tag{3.25}$$

sendo que λ_{disc} é a taxa de aprendizagem discriminativa e $P_{gc}(h_j{=}1|y, \mathbf{x})$ é dada na Equação (3.9). Não são fornecidas as regras de atualização para os parâmetros c_i nem os σ_i^2, visto que os primeiros não são relevantes para o treinamento discriminativo enquanto que os últimos podem ter sua aprendizagem evitada fazendo-se uma normalização na variância dos dados de entrada. Para concisão de notação, as seguintes definições foram empregadas nas regras de atualização:

$$\Delta b_j = \left[P_{gc}(h_j{=}1|y^{(t)}, \mathbf{x}^{(t)}) - \sum_{y^*=1}^{n_c} P_{gc}(y^*|\mathbf{x}^{(t)}) P_{gc}(h_j{=}1|y^*, \mathbf{x}^{(t)})\right], \tag{3.26}$$

$$\Delta d_k = \delta_{k,y^{(t)}} - P_{gc}(y{=}k|\mathbf{x}^{(t)}). \tag{3.27}$$

É interessante perceber que no treinamento discriminativo não é utilizada a amostragem de Gibbs, que faz parte da aproximação proposta pelo algoritmo CD (e derivados) para lidar com a intratabilidade no cálculo do gradiente da função perda generativa.

3.3.5 Treinamento híbrido

No treinamento híbrido, utiliza-se uma função perda que combina as perdas discriminativa e generativa, ou seja,

$$\mathcal{L}_{hyb}(\boldsymbol{\theta}; y^{(t)}, \mathbf{x}^{(t)}) = \mathcal{L}_{disc}(\boldsymbol{\theta}; y^{(t)}, \mathbf{x}^{(t)}) + \alpha \mathcal{L}_{gen}(\boldsymbol{\theta}; y^{(t)}, \mathbf{x}^{(t)}), \tag{3.28}$$

em que α é o peso dado ao critério generativo. Vale lembrar que uma RBM treinada visando a minimização da função perda híbrida é denominada RBM discriminativa híbrida (HDRBM). Como notado em Larochelle e Bengio (2008), o termo generativo na Equação (3.28) pode ser visto como um regularizador dependente dos dados e, assim, tem como objetivo evitar o chamado *overfitting* do modelo. O *overfitting* é relativamente comum em sistemas de aprendizagem de máquina e, particularmente, em modelos com um grande número de parâmetros. Ele se refere ao fato de o classificador apresentar um bom desempenho para classificação de amostras de treinamento e um desempenho ruim para classificação de amostras não antes vistas. Em outras palavras, com o *overfitting* o classificador perde parte da sua capacidade de generalização.

Conforme (2.27), as regras de atualização de parâmetros para o treinamento híbrido com taxa de aprendizagem λ_{hyb} teriam a forma

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \lambda_{hyb} \frac{\partial \mathcal{L}_{hyb}(\boldsymbol{\theta}; y^{(t)}, \mathbf{x}^{(t)})}{\partial \boldsymbol{\theta}}. \tag{3.29}$$

Considerando a linearidade do operador gradiente ($\partial(\cdot)/\partial\boldsymbol{\theta}$), o uso da função perda híbrida dada em (3.28) nessa última expressão permite escrever as regras de atualização como

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \left[\lambda_{hyb} \frac{\partial \mathcal{L}_{disc}(\boldsymbol{\theta}; y^{(t)}, \mathbf{x}^{(t)})}{\partial \boldsymbol{\theta}} + \alpha \lambda_{hyb} \frac{\partial \mathcal{L}_{gen}(\boldsymbol{\theta}; y^{(t)}, \mathbf{x}^{(t)})}{\partial \boldsymbol{\theta}} \right], \tag{3.30}$$

ou seja, a atualização híbrida de parâmetros pode ser escrita como a soma das atualizações discriminativa, com taxa de aprendizagem $\lambda_{disc} = \lambda_{hyb}$, e generativa, com taxa de aprendizagem $\lambda_{gen} = \alpha\lambda_{hyb}$. Assim, as regras de atualização híbridas podem ser facilmente obtidas e, por isso, são aqui omitidas. Nota-se, entretanto, que na aplicação de (3.30) aos parâmetros c_i, i=1, ..., n_d, deve-se considerar que a atualização discriminativa desses parâmetros é nula, fazendo-se $\partial\mathcal{L}_{disc}(\boldsymbol{\theta}; y^{(t)}, \mathbf{x}^{(t)})/\partial c_i = 0$, visto que os mesmos não são atualizados no treinamento discriminativo.

3.3.6 Classificação

A tarefa de classificação é um tipo de inferência que emprega a distribuição condicional das classes dado o estado das variáveis de entrada,

$$P_{gc}(y|\mathbf{x}) = \frac{\exp\left[d_y + \sum_{j=1}^{n_h} \zeta\Big(b_j + u_{jy} + \sum_{i=1}^{n_d} w_{ji} \frac{x_i}{\sigma_i^2}\Big)\right]}{\sum_{y^*=1}^{n_c} \exp\left[d_{y^*} + \sum_{j=1}^{n_h} \zeta\Big(b_j + u_{jy^*} + \sum_{i=1}^{n_d} w_{ji} \frac{x_i}{\sigma_i^2}\Big)\right]}, \tag{3.31}$$

sendo que $\zeta(z) = \ln(1 + e^z)$ representa a função *softplus* já introduzida. Deve-se observar que essa última equação tem a mesma forma de (2.76) da página 64 e, portanto, pode também ser avaliada num tempo $\mathcal{O}(n_h n_d + n_h n_c)$. Vale ressaltar que $P_{gc}(y|\mathbf{x})$ é usada na classificação independentemente do tipo de treinamento (generativo, discriminativo ou

híbrido) utilizado para se treinar a RBM. Pode-se notar que a Equação (3.31) fornece, efetivamente, a probabilidade de que a amostra de entrada pertença a uma certa classe. Assim, de posse de uma amostra a ser classificada, $\mathbf{x}^{(c)}$, a estimativa da sua classe, $\widehat{y}^{(c)}$, é normalmente dada por

$$\widehat{y}^{(c)} = \arg\max_{y} P_{gc}(y|\mathbf{x}^{(c)}). \tag{3.32}$$

Em palavras, segundo essa última equação a classe estimada seria aquela mais provável segundo o modelo. Certas aplicações podem se beneficiar diretamente da Equação (3.31) para não apenas estimar a classe de uma amostra, mas também para estabelecer uma medida da confiança que o modelo teria naquela escolha. Essa medida de confiança abre também a oportunidade para que múltiplas RBMs (ou outros modelos que forneçam a mesma medida) possam ser combinadas com a técnica conhecida como *boosting* para obter classificadores mais poderosos (FREUND; SCHAPIRE, 1995).

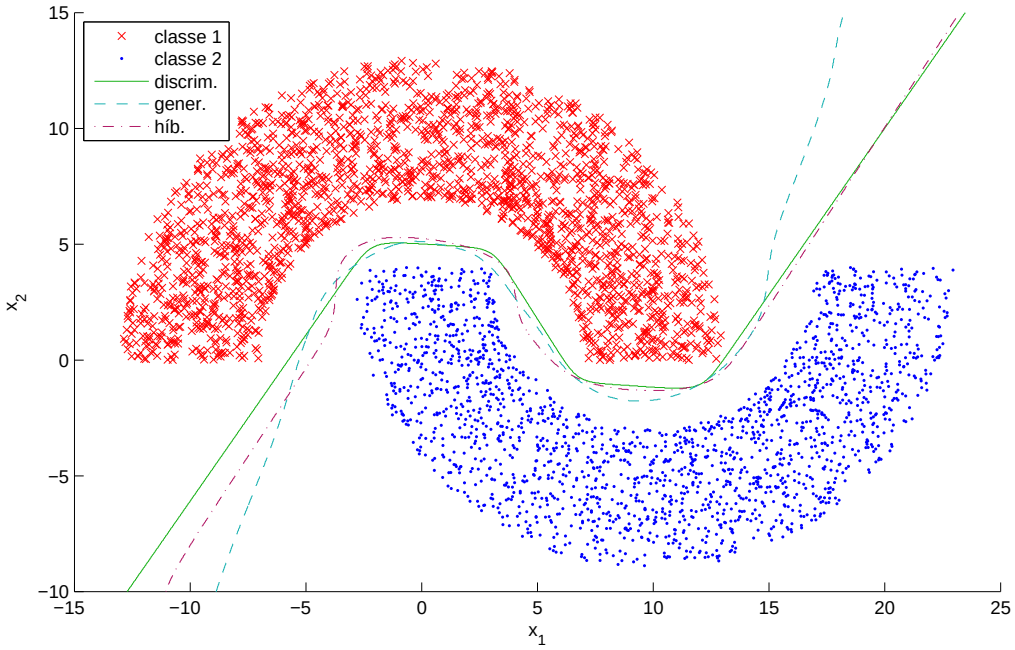
3.4 EXPERIMENTOS

Nesta seção, a variante proposta é explorada na tarefa de classificação. Especificamente, o modelo é aplicado a um problema de duas dimensões a fim de permitir uma fácil visualização dos resultados.

Na Figura 37, são mostrados os dados usados no treinamento do modelo. As coordenadas dos pontos representam os estados das variáveis de entrada, enquanto que as cores/símbolos de cada ponto ($\times$ e $\cdot$) indicam os estados da camada de classificação (ou seja, a classe atribuída ao ponto). Pode-se notar, na figura, que os dados de cada classe formam duas regiões distintas tendo a forma de semi-circunferências. Em cada uma dessas regiões, existem 2000 pontos que foram produzidos aleatoriamente. Na literatura, o caso de teste para algoritmos de aprendizagem de máquina apresentado na figura é conhecido como *double-moon* (HAYKIN, 2009). Empregando esses dados, foram então realizados os três tipos de treinamento discutidos na Seção 3.3: discriminativo (com n_h=20 e λ_{disc}=0,025), generativo (com n_h=40, λ_{gen}=0,005) e híbrido (com n_h=40, λ_{gen}=0,025 e α=0,0001)[1]. Cada um deles foi executado até que se atingisse 100% de acurácia (taxa de acertos) nos dados de treinamento. Na Figura 37, são mostradas também as linhas que separam as classes, segundo cada um dos modelos obtidos. Observa-se que os três classificadores têm a capacidade de separar adequadamente os dados de treinamento e, assim, atingir uma acurácia perfeita. Entretanto, eles apresentam linhas de decisão diferentes. Para o treinamento generativo, a linha de decisão apresenta curvas suaves, especialmente nas regiões próximas aos dados. Já as dos treinamentos discriminativo e híbrido, parecem ser formadas por segmentos que tendem a ser retos. As mesmas informações da Figura 37

[1] Em todos os casos, a quantidade de unidades ocultas empregadas foi o menor valor no conjunto $\{10, 20, 30, \ldots\}$ que permitiu ao classificador atingir 100% de acurácia em um treinamento limitado a 10.000 épocas. Os outros parâmetros foram definidos arbitrariamente.

Figura 37 – Dados de treinamento e linhas de decisão dos classificadores obtidos.



Fonte: produção do próprio autor.

são mostradas, considerando agora uma região mais ampla do espaço de amostras, na Figura 38. Essa figura evidencia que as várias linhas de decisão tendem a retas em regiões afastadas da área de concentração dos dados de treinamento.

É interessante buscar uma explicação para as características dos resultados que acabam de ser salientadas. Com esse objetivo, nota-se primeiramente que, no presente caso de teste, existem apenas duas classes e, portanto, para uma amostra $\mathbf{x}^{(d)}$ pertencente à linha de decisão do classificador, vale

$$P_{gc}(y{=}1|\mathbf{x}^{(d)}) = P_{gc}(y{=}2|\mathbf{x}^{(d)}). \tag{3.33}$$

Além disso, pode-se escrever

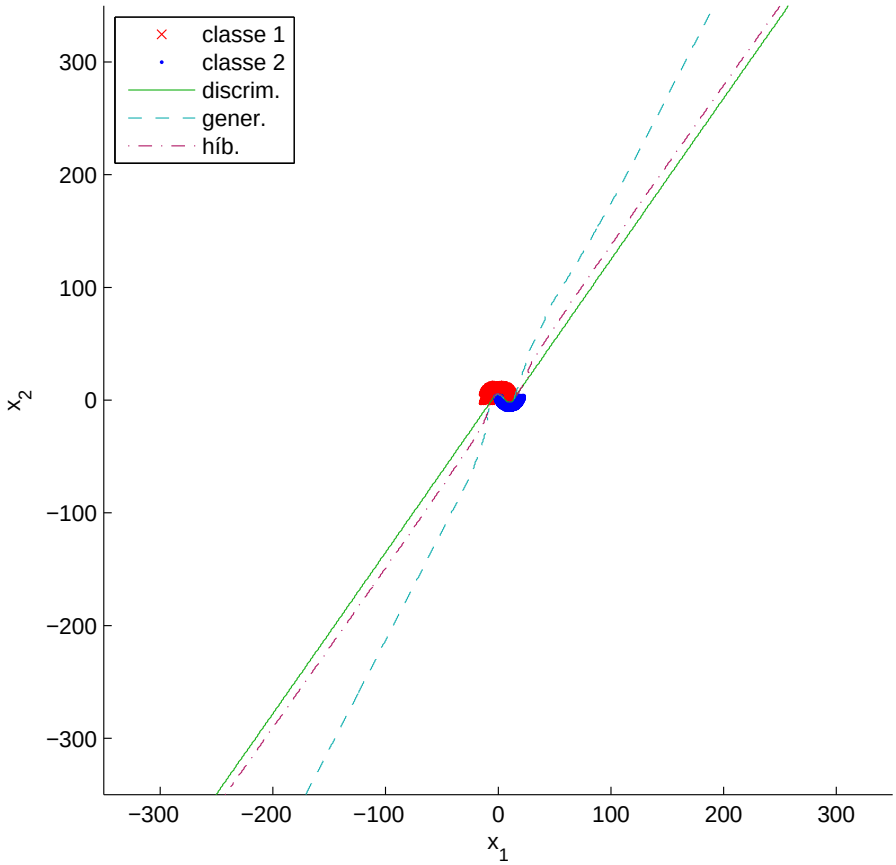
$$P_{gc}(y|\mathbf{x}) = \frac{P_{gc}(y,\mathbf{x})}{\sum\limits_{y'=1}^{2} P_{gc}(y',\mathbf{x})} = \frac{P_{gc}(\mathbf{x}|y)P_{gc}(y)}{\sum\limits_{y'=1}^{2} P_{gc}(\mathbf{x}|y')P_{gc}(y')} = \frac{P_{gc}(\mathbf{x}|y)}{\sum\limits_{y'=1}^{2} P_{gc}(\mathbf{x}|y')}, \tag{3.34}$$

sendo que na última igualdade considerou-se que as classes *são equiprováveis* (fato conhecido para os dados de treinamento) e que *essa característica tenha sido capturada pelo modelo* (no seu treinamento). Com a aplicação de (3.34) à (3.33) chega-se que

$$P_{gc}(\mathbf{x}^{(d)}|y{=}1) = P_{gc}(\mathbf{x}^{(d)}|y{=}2). \tag{3.35}$$

Considerando agora que, para $\mathbf{x}$ suficientemente distante da área de concentração de massa de probabilidade da distribuição $P_{gc}(\mathbf{x}|y)$, essa mesma distribuição possa ser aproximada

Figura 38 – Visão ampla das linhas de decisão (relativa à Figura 37).



Fonte: produção do próprio autor.

por

$$\widetilde{P}_{gc}(\mathbf{x}|y) = f(||\mathbf{x} - C_y||), \tag{3.36}$$

sendo que $||\cdot||$ representa a norma euclidiana, C_y indica o centro de massa da distribuição $P_{gc}(\mathbf{x}|y)$ para um y dado e $f(\cdot)$ é uma função qualquer que torne $\widetilde{P}_{gc}(\mathbf{x}|y)$ uma legítima função de distribuição de probabilidade. Nota-se que $\widetilde{P}_{gc}(\mathbf{x}|y)$ é simplesmente uma função com simetria radial em torno de C_y. O uso de (3.36) em (3.35) produz

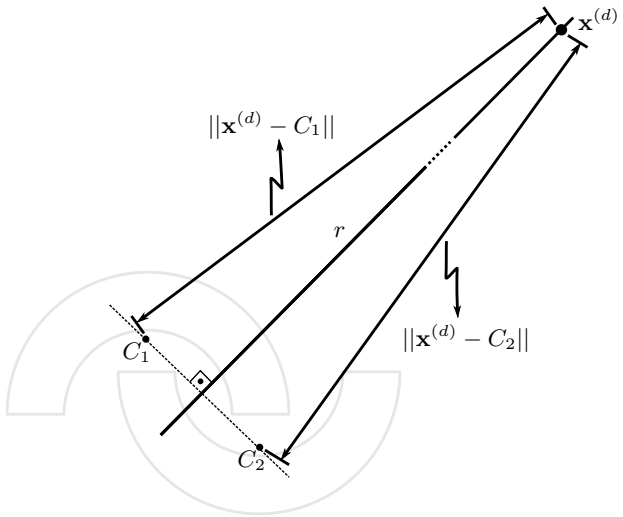
$$\widetilde{P}_{gc}(\mathbf{x}^{(d)}|y{=}1) = \widetilde{P}_{gc}(\mathbf{x}^{(d)}|y{=}2), \tag{3.37}$$

ou seja,

$$f(||\mathbf{x}^{(d)} - C_1||) = f(||\mathbf{x}^{(d)} - C_2||). \tag{3.38}$$

É fácil perceber que essa última igualdade fica satisfeita ao longo de uma reta perpendicular à linha ligando C_1 a C_2, como ilustrado na Figura 39. Apesar das diversas simplificações assumidas, a orientação da reta r na figura é consistente com as orientações observadas para as linhas de decisão, como vistas na Figura 38.

Figura 39 – $P_{gc}(\mathbf{x}|y)$ aproximada por distribuições radiais centradas em C_1 e C_2.

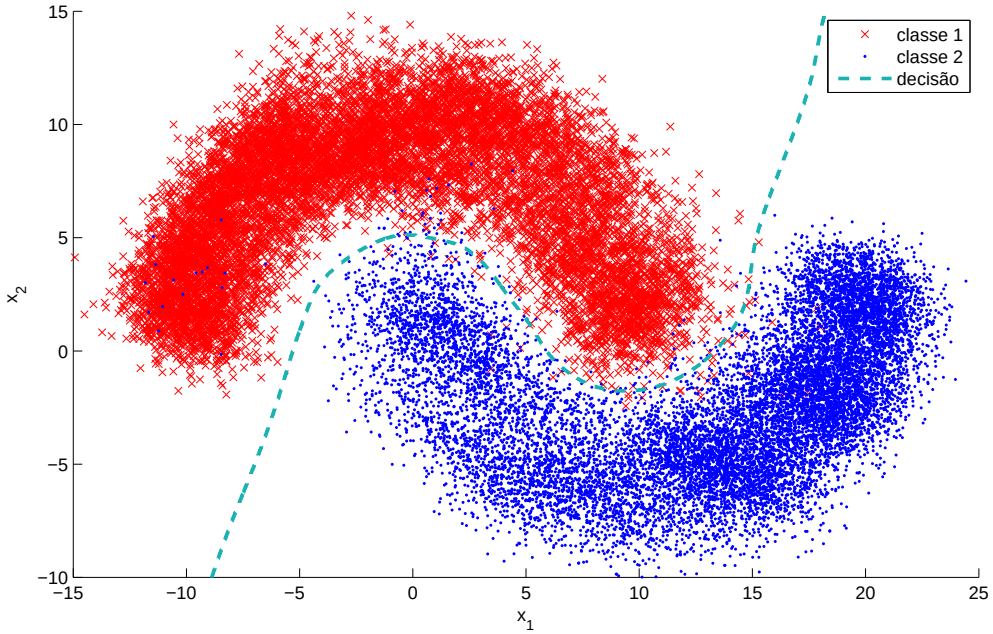


Fonte: produção do próprio autor.

Um fato notável sobre o modelo atual é que, apesar da presença da camada de classificação, ele ainda representa um distribuição de probabilidade, da qual se podem extrair amostras. A Figura 40 mostra o resultado da amostragem do modelo obtido no treinamento generativo. Fica claro que esse modelo reproduz a distribuição dos dados de treinamento, haja vista a similaridade dos mesmos com os dados amostrados. Deve-se notar que as amostras exibidas na figura representam realizações de $P_{gc}(y, \mathbf{x})$, a distribuição das variáveis visíveis da presente variante. É útil perceber também que, se as regiões contendo amostras das classes 1 e 2 forem consideradas individualmente, então essas regiões são representativas das distribuições $P_{gc}(\mathbf{x}|y{=}1)$ e $P_{gc}(\mathbf{x}|y{=}2)$, respectivamente. Tendo em vista esse último fato e considerando (3.35), na Figura 40, pode-se esperar que as linhas de decisão se localizem nas áreas em que as densidades de pontos das duas classes sejam iguais, como pode ser observado na figura e, com maior facilidade, na região entre as meias luas. As últimas observações ajudam a explicar também porque a linha de decisão produzida pelo treinamento generativo tende a ser mais suave do que a do discriminativo. No treinamento discriminativo, as atualizações de parâmetro dependem somente das amostras de treinamento. Em contrapartida, no treinamento generativo, as atualizações também dependem das amostras produzidas pelo modelo, mostradas na Figura 40. Como essas amostras são extraídas de distribuições condicionalmente gaussianas, as curvas de nível são suaves, característica que se reflete nas linhas de decisão.

Ainda com relação ao mesmo caso de teste, é interessante observar o resultado da amostragem dos modelos obtidos pelos treinamentos discriminativo e híbrido, conforme Figuras 41 e 42, respectivamente. Verifica-se que as distribuições dos dados extraídos desses modelos guardam pouca semelhança com a distribuição dos dados de treinamento.

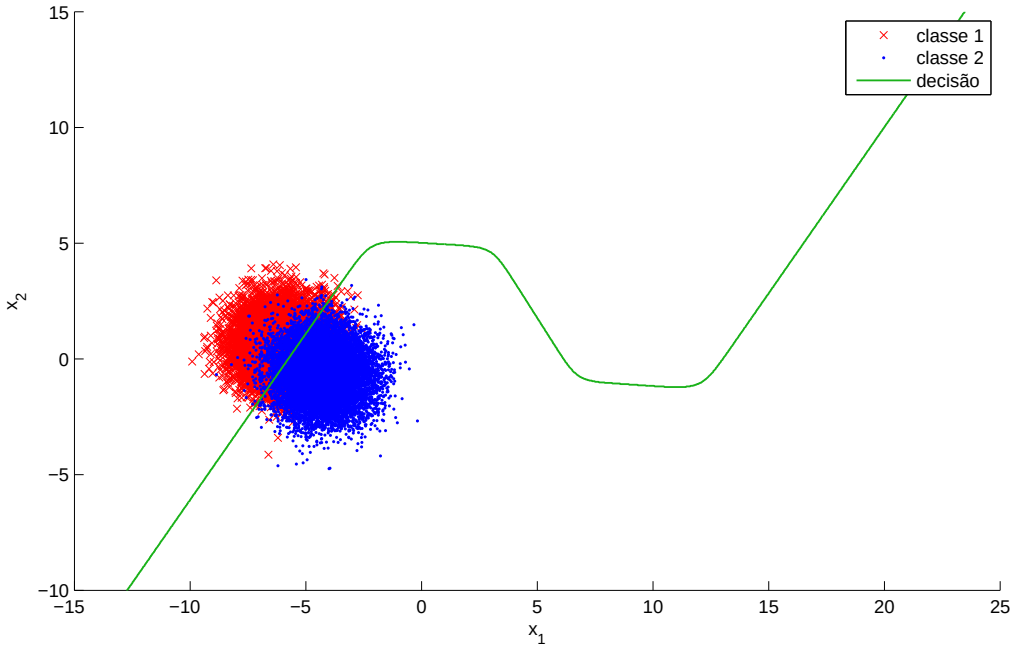
Figura 40 – Resultado da amostragem do modelo – treinamento generativo.



Fonte: produção do próprio autor.

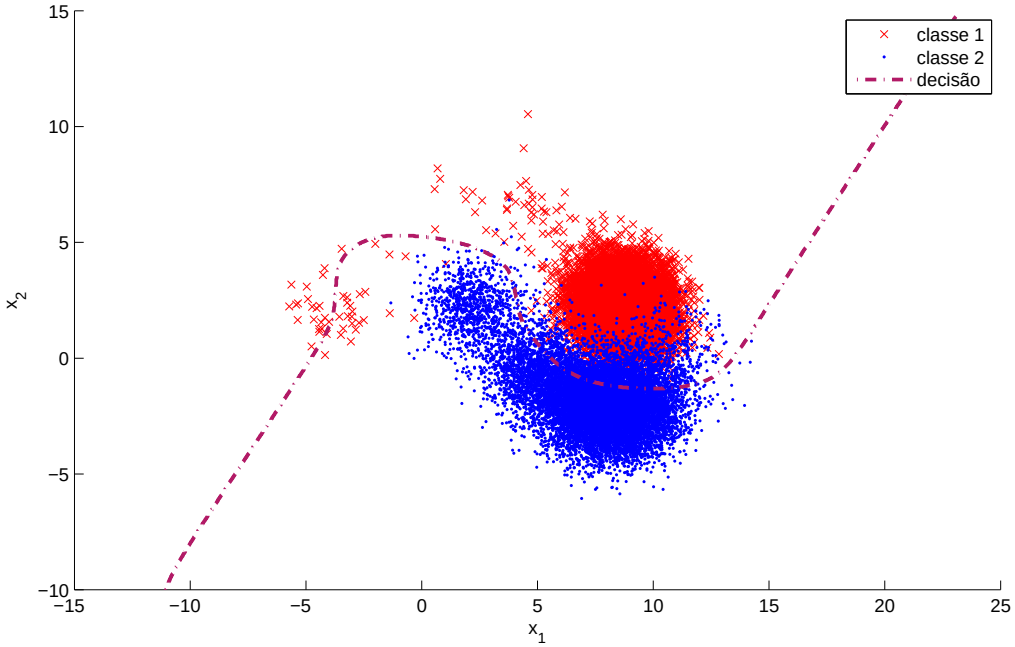
Quanto a essa constatação, vale dizer que, nas arquiteturas de aprendizagem profunda constituídas pelo empilhamento de RBMs (como DNNs construídas a partir de DBNs), é normalmente aplicado um ajuste fino discriminativo após o seu treinamento generativo, visando especializar a estrutura à tarefa de classificação sendo considerada. É um fato conhecido na literatura de que tal ajuste fino destrói a capacidade generativa do modelo original (WANG; SILVER, 2015). Analogamente, os resultados nas Figuras 41 e 42 mostram que as RBMs treinadas diretamente para classificação nem sequer apresentam a capacidade generativa. Isso se contrasta com o resultado da Figura 40, em que se verifica que o modelo treinado generativamente pode ser usado de forma bem sucedida para classificação.

Figura 41 – Resultado da amostragem do modelo – treinamento discriminativo.



Fonte: produção do próprio autor.

Figura 42 – Resultado da amostragem do modelo – treinamento híbrido.



Fonte: produção do próprio autor.

O segundo caso de teste mostrado nesta seção é similar ao anterior. Agora, entretanto, os dados de treinamento apresentam ruídos. Especificamente, 10% dos pontos em cada semi-circunferência têm suas classes trocadas (para a classe oposta) simulando ruídos ou erros de classificação. As configurações dos algoritmos de treinamento são também iguais às anteriores, exceto que, para o caso híbrido, usou-se α=0,01 a fim de ilustrar certos pontos a serem discutidos à frente. O presente teste busca verificar, em termos qualitativos, os efeitos dos erros de classificação produzidos pelos diferentes tipos de treinamento.

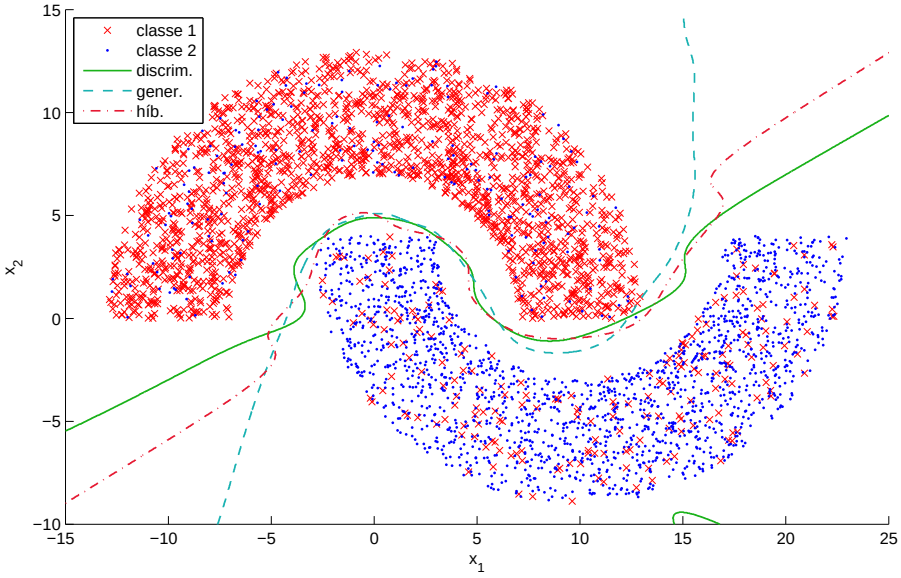
Na Figura 43, são exibidos os dados de treinamento e as linhas de decisão obtidas na condição descrita. Verifica-se novamente que todos os classificadores apresentam linhas de decisão que produziriam altas acurácias de classificação, na medida em que as semi-circunferências são bem separadas pelas referidas linhas. Portanto, considerando pontos próximos aos dados de treinamento, pode-se dizer que os três tipos de treinamento apresentam boa robustez à presença de ruídos de classificação. Contudo, comparando-se as Figuras 43 e 37, percebe-se que há sensíveis diferenças entre as linhas de decisão obtidas nas condições com e sem ruído. Essa percepção é confirmada pela visão mais ampla do espaço de amostras dada na Figura 44, a qual, quando comparada com o caso sem ruído da Figura 38, revela grandes diferenças entre os resultados das duas condições. Em especial, nota-se que:

1. Os treinamentos discriminativo e híbrido produziram linhas de decisão que parecem ser compostas aproximadamente por braços de hipérboles, em adição às retas observadas no caso sem ruído.
2. Os supostos braços de hipérboles do treinamento discriminativo se aproximam mais da região de concentração de dados do que aqueles do treinamento híbrido.

Quanto ao Item 1, uma suposição é que os efeitos constatados poderiam indicar uma tendência dos algoritmos (discriminativo e híbrido) de se ajustarem aos ruídos nos dados, algo indesejado e que não foi percebido no caso generativo. Com relação ao Item 2, pode-se entender que a parcela generativa do treinamento híbrido acaba por aumentar a robustez desse tipo de treinamento ao *overfitting*. Outros testes realizados (não apresentados aqui) mostraram que, para menores valores de α, os resultados do treinamento híbrido ficam mais parecidos com o caso discriminativo. Esses resultados corroboram a ideia de que o termo generativo funciona como um regularizador, no treinamento híbrido (LAROCHELLE; BENGIO, 2008).

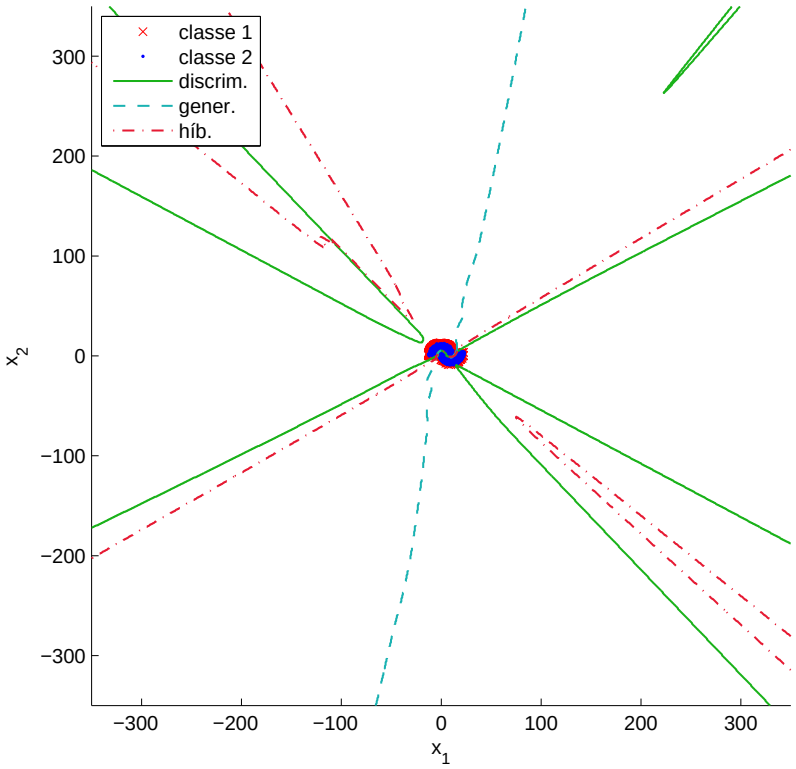
Por fim, vale dizer que as observações fornecidas aqui são válidas para o caso de teste sob avaliação e que, como frequentemente ocorre em problemas de aprendizagem de máquina, em razão das dificuldades em se analisar analiticamente os modelos empregados (dadas as não linearidades envolvidas) e até em se modelar os dados de treinamento, a determinação do algoritmo a ser empregado em um problema específico, bem como as configurações do mesmo, costumam envolver um considerável trabalho experimental.

Figura 43 – Dados de treinamento (ruidosos) e linhas de decisão dos classificadores obtidos.



Fonte: produção do próprio autor.

Figura 44 – Visão ampla das linhas de decisão (relativa à Figura 43).



Fonte: produção do próprio autor.

3.5 CONCLUSÕES

Neste capítulo, a variante de RBM proposta nesta dissertação foi descrita em detalhes. Foram apresentadas as equações necessárias à realização de inferência, amostragem e classificação com o modelo, bem como as regras de atualização de parâmetros para a realização dos seus treinamentos discriminativo, generativo e híbrido. A variante foi avaliada no problema *double-moon*, com base no qual pode-se dizer que:

1. Com o treinamento generativo, o modelo tem não apenas a capacidade de classificação, mas também a de geração de amostras similares às de treinamento. O experimento comprova que a variante é capaz de modelar a distribuição conjunta dos dados de entrada e suas correspondentes classificações. As linhas de decisão do classificador são mais suaves do que as obtidas com os demais tipos de treinamento. Os resultados parecem indicar ainda que o treinamento generativo tem uma maior robustez a ruídos nos dados de treinamento e ao problema de *overfitting*.
2. Com o treinamento discriminativo, o modelo resultante não apresenta nenhuma capacidade generativa. As linhas de decisão são mais abruptas e o classificador atinge uma acurácia de 100% com um número menor de unidades ocultas do que no caso generativo.
3. O treinamento híbrido produz um classificador cujas linhas de decisão têm características intermediárias entre os outros tipos de treinamento. A parcela generativa da função perda do treinamento híbrido funciona como um regularizador que pode auxiliar na prevenção do *overfitting* do modelo. Como o treinamento é majoritariamente discriminativo, o modelo resultante também não tem uma capacidade generativa apreciável.

Um ponto comum entre os diferentes tipos de treinamento é que em todos eles a variante conseguiu atingir 100% de acurácia de classificação. Esse fato mostra que a variante proposta é capaz de representar curvas de decisão não triviais e, portanto, o modelo pode ser considerado como um classificador de uso geral em domínios contínuos.

4 DETECÇÃO DE ATIVIDADE VOCAL EMPREGANDO RBMS

Neste capítulo, a variante de RBM proposta no presente trabalho é avaliada na tarefa de detecção de atividade vocal, o que é feito por meio de uma variedade de experimentos computacionais. O capítulo está organizado como segue. Primeiramente, nas Seções 4.1 e 4.2, tem-se os detalhes da configuração experimental e das medidas de desempenho comuns a todos os experimentos. Em seguida, a Seção 4.3 introduz os diferentes detectores usados como referência ou base de comparação. Passa-se então aos experimentos propriamente ditos, os quais são apresentados em três seções. Na Seção 4.4, estão os experimentos realizados na chamada condição única de SNR. Nessa seção, os VADs baseados em RBM são comparados com outros baseados em SVM e também com os detectores de referência. Um segundo grupo de experimentos foram executados na chamada condição múltipla de SNR e são discutidos na Seção 4.5. Por fim, a Seção 4.6 avalia diferentes combinações de medidas, as quais tiveram seu uso em VAD sugerido recentemente na literatura. Para tanto, o modelo proposto neste trabalho é utilizado no papel de classificador.

4.1 CONFIGURAÇÃO EXPERIMENTAL

A Figura 45 mostra a infraestrutura desenvolvida para permitir a realização dos diversos experimentos. Na figura, os retângulos duplos sobrepostos representam artefatos, sendo esses, em sua maioria, arquivos. As caixas com bordas arredondadas indicam processos que utilizam e produzem novos artefatos e, fisicamente, correspondem a scripts implementados e executados no MATLAB (versão 7.11). Esses processos são brevemente descritos a seguir:

1. **Preparação dos conjuntos de dados (datasets)**: Nesse processo, os arquivos de áudio de uma versão modificada do *corpus* NOIZEUS (HU; LOIZOU, 2008) sofrem eventuais pré-processamentos e, então, têm suas características extraídas. Características associadas aos segmentos de áudio são ainda casadas com as classificações esperadas para cada segmento, as quais são obtidas dos arquivos contendo as rotulações manuais. Esses resultados (características e rotulações) são divididos em conjuntos de treinamento, validação e teste para uso posterior com algoritmos baseados em aprendizagem de máquina. Detalhes sobre as modificações ao *corpus*, a rotulação manual e a extração de características são fornecidos mais à frente.
2. **Algoritmos para comparação**: Esse bloco representa a execução de quaisquer algoritmos de detecção de atividade vocal sobre os dados brutos do *corpus*. Como

resultado, são produzidos os dados de desempenho desses algoritmos que permitiriam comparações posteriores com outros detectores.

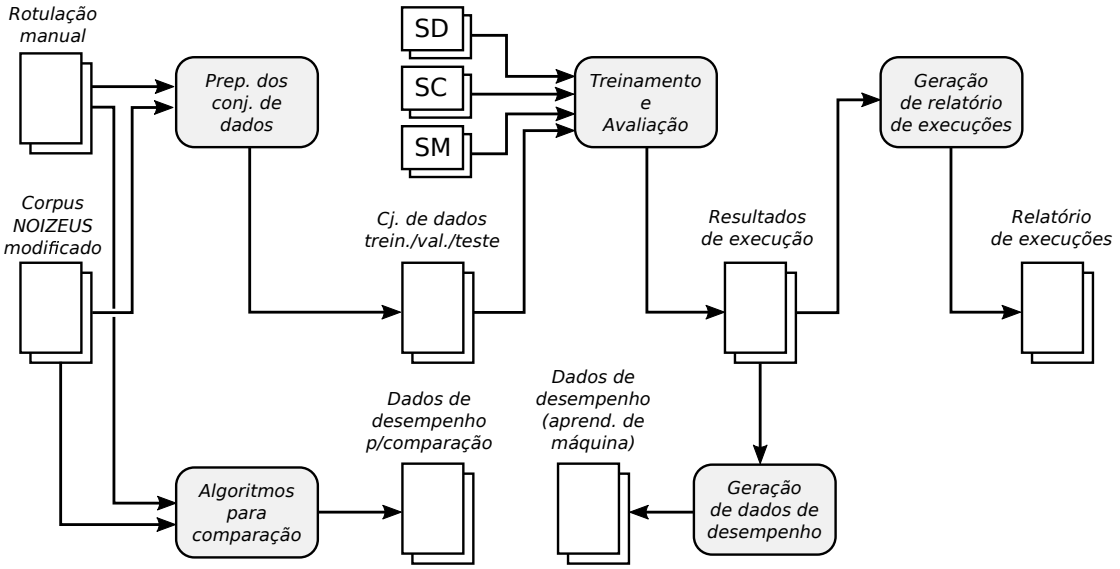
3. **Treinamento e avaliação**: Primeiramente esse bloco cria os vetores de características por meio da concatenação e possível pós-processamento das características obtidas dos conjuntos de dados (datasets). Esse passo é controlado pelos arquivos de seleção de dados (SD) e de seleção de características (SC). Em seguida, os vetores são usados no treinamento de um modelo, cuja configuração é controlada pelo arquivo de seleção de modelo (SM). Durante o treinamento, o modelo é avaliado com o uso dos dados de validação e, ao final do treinamento, o desempenho é determinado sobre os dados de teste. Para cada combinação de conjunto de dados (SD), de características (SC) e de configuração de modelo (SM) é produzido um arquivo de *resultados de execução* contendo os parâmetros do modelo obtido, os resultados das avaliações mencionadas e todas as informações necessárias para se repetir o treinamento, caso necessário.
4. **Geração de relatório de execuções**: Esse processo coleta os vários arquivos de *resultados de execução* e produz um *relatório de execuções*, no qual são fornecidas informações gerais de desempenho das configurações testadas. É interessante destacar que os resultados de execução podem ter sido produzidos em máquinas diferentes. De fato, essa possibilidade orientou a definição da arquitetura descrita aqui.
5. **Geração de dados de desempenho**: Esse bloco produz dados mais detalhados de desempenho de um subconjunto especificado de resultados de execução. A ideia é que o relatório de execuções seja usado para se determinar quais as combinações de vetores de características e modelos produzem melhores resultados para, então, se fazer uma análise mais detalhada dessas combinações. Essa segunda análise é realizada pelo bloco de *geração de dados de desempenho*.

Preparação do corpus

Como já mencionado, nos experimentos, utilizou-se uma versão modificada do *corpus* (conjunto de arquivos de áudio) denominado NOIZEUS (HU; LOIZOU, 2008), o qual é composto por 30 arquivos de áudio amostrados à taxa de 8 kHz (gravados originalmente à taxa de 25 kHz e subamostrados posteriormente), cada qual correspondendo a uma frase pronunciada por um dentre 6 locutores (3 masculinos e 3 femininos). O tempo total de áudio é de aproximadamente 90 segundos. Esse *corpus* foi escolhido por três motivos: primeiramente, o número relativamente pequeno de frases viabiliza uma rotulação manual do áudio; em segundo lugar, as frases foram concebidas de modo a conter todos os fonemas da língua inglesa; e, por fim, ele é disponibilizado gratuitamente.

Vale aqui fornecer as motivações para as mudanças feitas ao *corpus*. O áudio de cada um dos arquivos que compõem o *corpus* foi manualmente rotulado de modo a indicar os

Figura 45 – Configuração experimental.



Fonte: produção do próprio autor.

trechos contendo voz. Segundo essa rotulação, obteve-se um percentual de atividade vocal variando entre 64,9% e 91,2% entre os arquivos, e uma média de 83,4%. Tendo em vista esse percentual, pode-se dizer que essa base de informações é apreciavelmente desbalanceada, pois possui uma quantidade consideravelmente maior de exemplos positivos (presença de voz) do que negativos (ausência de voz). Assim, esse desbalanceamento representaria uma quantidade relativamente pequena de dados para a avaliação dos desempenhos dos detectores na condição de ausência de voz. Um problema mais grave é que, no caso dos detectores baseados em aprendizagem de máquina, o treinamento nessa situação tenderia ainda a produzir detectores enviesados no sentido de detectar voz. Por esses motivos, escolheu-se fazer o balanceamento do *corpus*. Além do balanceamento, considerou-se interessante que os testes de desempenho dos algoritmos pudessem ser feitos numa faixa mais larga de relações sinal-ruído do que a disponibilizada no *corpus* (originalmente, de 0 dB a 15 dB, em passos de 5 dB).

A versão modificada do *corpus* foi gerada da seguinte forma: aos arquivos sem ruído adicionou-se 0,8 s de silêncio antes e após o áudio original e, em seguida, somou-se o ruído obtido da gravação *car noise* da base AURORA-2 (HIRSCH; PEARCE, 2000), de modo a se obter as relações sinal-ruído desejadas (de −5 dB a 20 dB, em passos de 5 dB). Para determinar o nível (potência) dos áudios representando sinal e ruído, usou-se o método B de (ITU, 1993), da mesma forma como feito na geração dos arquivos ruidosos do *corpus* original (HU; LOIZOU, 2008). Deve-se destacar que alguns dos detectores usados como base de comparação (descritos na Seção 4.3) utilizam o início do áudio para estimar as características do ruído. Portanto, o procedimento descrito permite também que esses

detectores operem apropriadamente.

Rotulação manual

Como no corpus NOIZEUS são fornecidos os arquivos originais sem ruído, a rotulação foi baseada nos mesmos. A ausência de ruído é um fator facilitador que permite a identificação imediata de regiões contendo palavras. Frequentemente, palavras são pronunciadas sem que haja silêncio entre elas, formando regiões contendo voz. A determinação dos pontos iniciais e finais dessas regiões é um tanto subjetiva porque próximo a eles a potência do sinal é normalmente ordens de grandeza inferior às partes centrais das palavras e o sinal se confunde com o pequeno ruído existente mesmo no áudio denominado *sem ruído*. Em cada região do áudio em que surgiram dúvidas sobre a classificação, um trecho da região era atenuado em 20 dB e regiões próximas eram ouvidas. Se a atenuação não causasse mudanças perceptíveis no áudio, considerava-se que o trecho não continha voz. Essa técnica foi especialmente útil na determinação aproximada das regiões de voz terminadas por consoantes fricativas (sons de s e z, por exemplo). Além disso, consoantes plosivas (como sons de p e b) geram silêncios relativamente longos, mesmo em meio às palavras. Silêncios dessa natureza mais longos do que 20 ms foram considerados como ausência de voz.

Separação dos conjuntos de dados

No processo de preparação dos conjuntos de dados, 70% dos arquivos do *corpus* (escolhidos aleatoriamente) foram usados para produzir o conjunto de treinamento e os 30% restantes produziram os conjuntos de validação e teste, como feito por Zou et al. (2014).

4.2 MEDIDAS DE DESEMPENHO

A fim de permitir comparações de desempenho entre diferentes detectores, escolheu-se empregar duas medidas (escalares): a acurácia balanceada e a área sob a curva de característica de operação do receptor (ROC – *Receiver Operating Characteristic*). Para comparações mais detalhadas, utilizam-se eventualmente as próprias curvas ROC. Essas medidas e curvas serão explicadas a seguir. Vale dizer que as mesmas são apropriadas para a avaliação de classificadores binários (tendo duas classes), em que uma dada amostra pode ser considerada como positiva ou negativa. Além disso, as classes positiva e negativa representam, neste texto, a presença e a ausência de voz, respectivamente.

A acurácia balanceada (BA – *Balanced Accuracy*), de uso comum na área de aprendizagem de máquina, é dada por

$$\mathrm{BA} = \tfrac{1}{2}\,\mathrm{TPR} + \tfrac{1}{2}\,\mathrm{TNR}, \tag{4.1}$$

sendo que TPR (*True Positive Rate*) e TNR (*True Negative Rate*) são as taxas de acerto (número de amostras classificadas corretamente dividido pelo número total de amostras em um conjunto) quando fornecidas ao classificador somente amostras positivas (para TPR) ou negativas (para TNR). Deve-se perceber que, para um classificador atingir valores elevados (próximos de 100%) de acurácia balanceada (ou simplesmente acurácia, daqui em diante), ele deve ter bons desempenhos tanto na classificação de amostras positivas quanto negativas. Por exemplo, um detector totalmente enviesado, que classifique todas amostras como positivas, tem BA=50% e, assim, é visto como um detector muito pobre ou de desempenho ruim.

As curvas ROC são frequentemente utilizadas em telecomunicações e fornecem uma visão mais detalhada do desempenho de um detector do que se consegue com simples medidas escalares. A curva ROC corresponde ao traçado da probabilidade de detecção (P_D) *versus* a probabilidade de falso alarme (P_{FA}) do detector. Os pontos da curva são obtidos variando-se um parâmetro do detector que normalmente corresponde a um limiar de decisão entre as duas classes. P_D representa a probabilidade de o detector indicar como positiva uma amostra que seja verdadeiramente positiva, enquanto que P_{FA} seria a probabilidade de ele indicar como positiva uma amostra conhecidamente negativa. Tendo em vista esses significados, é fácil perceber que P_D e P_{FA} podem ser aproximadas a partir das medidas empíricas, TPR e TNR, como

$$P_D \approx \text{TPR} \tag{4.2}$$

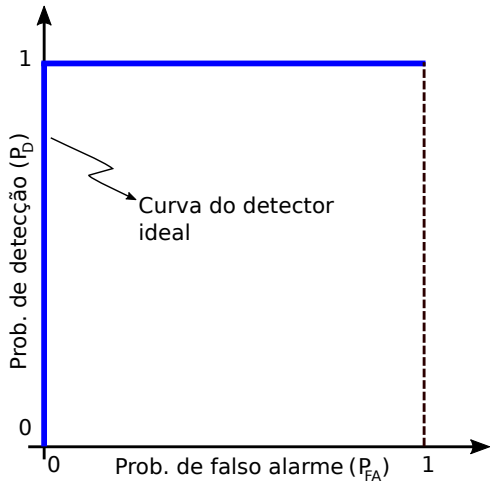
$$P_{FA} \approx 1 - \text{TNR}. \tag{4.3}$$

A área sob a curva ROC (ou, simplesmente, área ROC) é, portanto, aproximada como a área sob a curva TPR $\times$ (1 $-$ TNR). Vale lembrar que a acurácia do detector depende também das medidas TPR e TNR, conforme Equação (4.1). Sendo assim, diferentes pontos da sua curva ROC produzem diferentes acurácias. Em vista desse fato, deve-se esclarecer que, neste texto, a acurácia de um detector é considerada como o melhor valor de acurácia que se obtém ao longo de sua curva ROC. Cumpre dizer que certos detectores (como aqueles baseados em RBM) são capazes de produzir como saída uma medida de probabilidade de que uma dada amostra seja positiva. A execução desse tipo de detector em um lote de amostras gera, portanto, um conjunto de medidas de probabilidade sobre as quais pode ser aplicado o limiar variável mencionado anteriormente. Com isso, são determinados os correspondentes valores de TPR e TNR necessários à construção da curva ROC.

Na Figura 46, é mostrada a curva ROC do detector ideal. O ponto (P_{FA}, P_D)=$(0, 1)$ corresponde ao ponto de acurácia perfeita, isto é, em que todas as amostras são classificadas corretamente. É útil saber que, normalmente, considera-se que os pontos $(0, 0)$ e $(1, 1)$ fazem parte da curva ROC de qualquer detector, visto que o primeiro é obtido ignorando-se o detector e classificando todas amostras como negativas, enquanto que, para o segundo,

basta classificá-las todas como positivas. A presença dos pontos mencionados permite que a área sob a curva ROC seja calculada mesmo quando o detector não possuir um limiar de detecção configurável, ou seja, quando o detector apresentar um único ponto de operação, como ocorre com alguns dos VADs de referência que foram avaliados neste trabalho.

Figura 46 – Curva ROC do detector ideal.



Fonte: produção do próprio autor.

Além das medidas citadas, referentes ao desempenho de classificação, foram também realizadas medições de tempo de execução dos algoritmos visando comparar aproximadamente o custo computacional dos mesmos. A medida usada é chamada aqui de taxa de trabalho de tempo real e representa simplesmente o tempo de áudio processado divido pelo tempo necessário para fazê-lo (em um determinado processador). Assim, maiores valores dessa medida indicam um menor custo computacional do algoritmo.

Cumpre informar que as medidas de custo computacional que serão apresentadas nas seções posteriores foram todas realizadas no MATLAB 7.11 executado sobre o Windows 7 em uma máquina com processador Intel i7-4790 com 4 núcleos físicos operando a 3,6 GHz e com um total de 16 GB de memória (RAM).

4.3 DETECTORES DE REFERÊNCIA

Os seguintes detectores de atividade vocal foram escolhidos para servirem como base de comparação contra aqueles baseados em RBM:

1. **ITU G.729-B** (BENYASSINE et al., 1997): Esse VAD é parte opcional do *codec* de áudio ITU G.729, o qual foi adotado pela indústria de telefonia. Por isso, ele é comumente usado como base de comparação para outros VADs. Suas decisões (sobre a presença ou ausência de voz) são tomadas com base em medidas tradicionais, tais como:

a) Energia em todo o espectro;
b) Energia na parte inferior do espectro (abaixo de 1 kHz);
c) Espectro da predição linear;
d) Taxa de cruzamento por zero.

Em particular, as três primeiras medidas mencionadas são determinadas a partir dos coeficientes de autocorrelação do sinal, calculados pelo *codec* de voz. Portanto, parte do processamento do *codec* é compartilhado com o VAD. Uma implementação de referência (em linguagem C) do *codec* é disponibilizada pela ITU e foi adaptada para permitir avaliações do seu VAD. Esse VAD não possui um parâmetro equivalente a um limiar de detecção e, por isso, somente um ponto da curva ROC pode ser para ele determinado. Embora seja possível atribuir uma área ROC a esse detector (como explicado na Seção 4.2), existe a possibilidade de se subestimar seu desempenho devido ao número reduzido de pontos conhecidos. De fato, essa possibilidade foi um motivador adicional para o uso da acurácia nas comparações de desempenho.

2. **ITU G.729-II** (ITU, 2005): Esse é o VAD ITU G.729-B com as modificações especificadas no Apêndice II da recomendação ITU G.729, em que foram propostas melhorias ao VAD original.
3. **LTSD** (RAMIREZ et al., 2004): O LTSD (*Long-Term Spectral Divergence*) opera comparando a envoltória de longo prazo do espectro do sinal com uma estimativa do espectro do ruído. Esse VAD é às vezes usado como base de comparação para outros detectores dado o seu bom desempenho em uma larga faixa de relações sinal-ruído. Além disso, o mesmo mostrou ter um desempenho superior a VADs adotados em *codecs* de voz de padrões europeus (e.g., AMR – *Adaptive Multirate* e AFE – *Advanced Front-End*). O algoritmo foi implementado em MATLAB com base no artigo original. A saída desse VAD é binária e, por isso, para determinar os pontos da sua curva ROC, o algoritmo foi executado repetidas vezes variando-se um dos seus parâmetros (denominado γ no artigo original) que cumpre o papel de um limiar de detecção.
4. **Sohn** (SOHN; KIM; SUNG, 1999): O VAD Sohn é baseado em um modelo estatístico. Resumidamente, os coeficientes da DFT dos segmentos do áudio são modelados como variáveis aleatórias gaussianas assintoticamente independentes tendo parâmetros diferentes sob as hipóteses de ausência e de presença de voz. As variâncias do ruído (nos coeficientes da DFT) são estimados com base nos instantes iniciais do áudio. Os demais parâmetros do modelo gaussiano são determinados (em cada segmento de áudio) a partir das amplitudes das componentes (em frequência) do sinal, as quais, por sua vez, são estimadas pelo método do mínimo erro quadrático médio. De posse dos parâmetros do modelo, suas distribuições nas duas hipóteses (ausência e presença de voz) ficam determinadas e a decisão do detector é tomada com o emprego do teste de razão de verossimilhança (LRT – *Likelihood-Ratio Test*) (KAY, 1993). A

implementação desse VAD em MATLAB foi obtida do pacote `voicebox`[1].

5. **Ying** (YING et al., 2011): No detector Ying, os logaritmos das energias (daqui em diante referenciadas simplesmente como log-energias) em diferentes sub-bandas do espectro do sinal são analisadas independentemente ao longo do tempo. Os autores apontam que, em segmentos contendo somente ruído, a log-energia em cada sub-banda tem distribuição gaussiana simples (unimodal), enquanto que em segmentos contendo voz a distribuição é bimodal, ou, mais especificamente, uma mistura de duas gaussianas. Esse fato é usado da seguinte forma: para cada sub-banda as medidas de log-energia de uma sequência de quadros (segmentos do áudio) são ajustadas a uma mistura de duas gaussianas empregando, para tanto, uma versão modificada do algoritmo EM (*Expectation Maximization*) proposta pelos autores. Desse ajuste, são obtidas as médias, variâncias e pesos das gaussianas da mistura. Com essas informações, são então calculados limiares ótimos para detecção de voz. Assim, para um quadro a ser classificado, as suas log-energias são comparadas com os referidos limiares em cada sub-banda, produzindo um conjunto de indicações binárias de presença ou ausência de voz. A saída do detector é a média dessas indicações, a qual é vista como uma medida de probabilidade de presença de atividade vocal no quadro. A implementação desse VAD em MATLAB foi gentilmente fornecida por um dos seus autores, Dongwen Ying.
6. **Ghosh** (GHOSH; TSIARTAS; NARAYANAN, 2011): O detector Ghosh toma suas decisões a partir de uma medida denominada Variabilidade de Longo Prazo do Sinal (LTSV – *Long-Term Signal Variability*), definida pelos autores. A fim de determinar essa medida, o espectro de curto prazo do sinal é calculado. Para cada frequência do espectro (numa faixa considerada plausível para a voz), as potências do sinal num conjunto de quadros consecutivos são consideradas como uma distribuição de probabilidade, cuja entropia é calculada. Obtém-se, portanto, um conjunto de medidas de entropia, cada qual referente a uma frequência do espectro. A variância de tais medidas corresponde ao LTSV. A saída binária do detector é obtida aplicando-se um simples limiar ao LTSV. Para obter a curva ROC, varia-se esse mesmo limiar.

Vale dizer que alguns dos VADs descritos oferecem um esquema chamado na literatura de *hang-over*. Nos VADs G.729-B, G.729-II, Sohn e Ying, tal esquema é intimamente ligado à operação do detector e, por isso, foi utilizado nos experimentos. Já os detectores LTSD e Ghosh foram usados sem *hang-over*. O esquema de *hang-over* tem como objetivo evitar cortes (por parte do VAD) nas partes finais dos trechos de áudio contendo voz. Em esquemas convencionais, por exemplo, isso é feito com a inserção de um atraso na saída do detector quando a mesma transita do estado de presença para o de ausência de voz. A consequência para o detector é um aumento na sua medida de probabilidade

[1] <http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html>

de detecção à custa de um aumento na sua probabilidade de falso alarme (SOHN; KIM; SUNG, 1999). Deve-se perceber, entretanto, que, para as medidas escalares de desempenho (acurácia balanceada e área ROC) usadas no presente trabalho, os efeitos citados tendem a se cancelar. Por isso, o esquema de *hang-over* não é considerado aqui como um empecilho à comparação entre VADs.

Por fim, é interessante destacar que os detectores Sohn, Ying e Ghosh são apontados na literatura como VADs de estado da arte (DRUGMAN et al., 2015).

4.4 AVALIAÇÕES EM CONDIÇÃO ÚNICA DE RELAÇÃO SINAL-RUÍDO

Nesta seção, são apresentados os detalhes dos experimentos realizados naquela que é denominada na literatura de condição única (*single-condition*) de relação sinal-ruído. Nessa condição, um modelo é treinado e avaliado com dados de uma única SNR. Consequentemente, o modelo resultante seria normalmente utilizável apenas na SNR em que fora treinado ou valores próximos dela. Portanto, fica implícito que, numa aplicação real, a SNR deveria ser conhecida (e fixa) ou seria determinada (por algum método de estimação de SNR) e os parâmetros do modelo a ela apropriados, selecionados dinamicamente.

A seguir, a Seção 4.4.1 fornece os detalhes dos procedimentos empregados no treinamento das RBMs bem como dos vetores de características com elas utilizados. As RBMs foram comparadas aqui com os detectores de referência (descritos na Seção 4.3) e também com outro mecanismo de aprendizagem de máquina: as SVMs (FRIEDMAN; HASTIE; TIBSHIRANI, 2001). As configurações dos detectores de referência e o procedimento adotado para treinamento das SVMs são descritos ainda na mesma seção. Os resultados dos diversos experimentos são apresentados em duas partes. Na Seção 4.4.2, as RBMs têm seus resultados comparados com os VADs de referência. Então, na Seção 4.4.3, as DRBMs apenas são contrastadas com as SVMs. Por fim, as conclusões relativas ao conjunto de avaliações em condição única de SNR são expostas na Seção 4.4.4.

4.4.1 Procedimentos e configurações

Na Tabela 5, são exibidas as duas configurações de vetores de características que foram consideradas nos experimentos, sendo que as medidas que formam o conteúdo dos vetores estão descritas na Seção 2.4.2. A configuração C1 é baseada nos MFCCs (13 valores) e no log da energia de quadro, que totalizam 14 medidas escalares por segmento de áudio. O vetor de características propriamente dito é formado pela concatenação dessas medidas com as derivadas temporais de primeira e segunda ordem delas mesmas (indicadas como Δ's e $\Delta\Delta$'s na tabela) calculadas conforme explicado na página 70. Na literatura, muitos experimentos mostraram que a combinação dos MFCCs com a informação de energia de quadro seria capaz de melhorar o desempenho de sistemas de reconhecimento

de fala (ZHENG; ZHANG, 2000). Daí a escolha dessa configuração. A configuração C2 é construída de forma análoga à C1, porém são usadas as FBEs (com 23 valores) em vez dos MFCCs. Cumpre lembrar que os MFCCs são calculados a partir das FBEs. Entretanto, diferentemente dos MFCCs, as FBEs são consideradas medidas fortemente correlacionadas entre si. Assim, o objetivo de se avaliar os modelos com a configuração C2 é o de verificar a capacidade dos mesmos de lidar com dados correlacionados.

Tabela 5 – Configurações dos vetores de características.

Configuração	**Conteúdo do vetor**	**Dimensão**
C1	MFCCs, ln(FE) e respectivos Δ's e $\Delta\Delta$'s.	42
C2	FBEs, ln(FE) e respectivos Δ's e $\Delta\Delta$'s.	72

É interessante informar que, nos experimentos, o áudio de cada arquivo passou primeiramente por um filtro de pré-ênfase (com coeficiente 0,97) e o sinal resultante foi segmentado em quadros sobrepostos com duração de 25 ms e deslocamento de 10 ms entre quadros. Desses quadros, foram então extraídos os MFCCs e FBEs segundo o processo descrito na Seção 2.4.2.2. Além disso, antes de serem fornecidos aos algoritmos de treinamento, os vetores de características foram normalizados para que as médias de suas componentes fossem nulas e tivessem variâncias unitárias (sobre o conjunto de treinamento).

Na Tabela 6, tem-se os parâmetros utilizados no treinamento das RBMs. A primeira coluna da tabela dá um nome a cada configuração, identificando a combinação de tipo de treinamento e de vetor de características empregados. Deve-se lembrar que as siglas RBM, DRBM e HDRBM são usadas para indicar as RBMs produzidas pelos treinamentos generativo, discriminativo e híbrido, nessa ordem. Assim, por exemplo, a configuração DRBM-C1 denota um RBM empregando o vetor de características C1 sendo treinada discriminativamente.

Tabela 6 – Parâmetros de treinamento das RBMs.

Configuração	**Parâmetros de treinamento**
DRBM-C1	n_h=35, λ_{disc}=$0,005$
DRBM-C2	n_h=25, λ_{disc}=$0,001$
RBM-C1	n_h=350, λ_{gen}=$0,005$
RBM-C2	n_h=450, λ_{gen}=$0,001$
HDRBM-C1	n_h=350, λ_{hyb}=$0,001$, α=10^{-4}
HDRBM-C2	n_h=350, λ_{hyb}=$0,005$, α=10^{-4}

Os parâmetros de treinamento (às vezes chamados de hiperparâmetros do modelo) fornecidos na Tabela 6 são aqueles que produziram os melhores resultados de acurácia (dentre várias configurações avaliadas) e para os quais as medidas de desempenho são

apresentadas. Chegou-se a tais parâmetros por meio de buscas em grade (*grid-search*). Em particular, para o treinamento discriminativo foram feitos testes com as combinações

$$(n_h, \lambda_{disc}) \in \{10, 15, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90\} \times \{1{\times}10^{-3}, 5{\times}10^{-3}\}$$

e, para o treinamento generativo,

$$(n_h, \lambda_{gen}) \in \{30, 50, 75, 100, 150, 200, 250, 300, \ldots, 800\} \times \{1{\times}10^{-3}, 5{\times}10^{-3}\}.$$

Entretanto, deve-se esclarecer que, em razão do alto custo computacional envolvido, essas avaliações se basearam apenas nos dados do áudio com SNR de 0 dB. Para o treinamento híbrido, foram testadas (com SNR de 0 dB) as combinações

$$(n_h, \lambda_{hyb}, \alpha) \in \{350, 400, 450, 500\} \times \{1{\times}10^{-3}, 5{\times}10^{-3}\} \times \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}.$$

Nesse último caso, escolheu-se avaliar alguns dos valores de n_h que produziram bons resultados no treinamento generativo. Nota-se, entretanto, que essa é uma solução heurística para diminuir a quantidade de combinações de hiperparâmetros a avaliar a um número aceitável. Finalmente, em todos os treinamentos de RBMs, o conjunto de treinamento foi dividido em lotes contendo 70 amostras cada. Nessa situação, a média dos gradientes da função perda (calculada nas amostras no lote) é usada na atualização de parâmetros do modelo em cada passo de treinamento.

Para avaliações com SVMs, foi usada a implementação dos algoritmos que fazem parte do pacote de Estatística e Aprendizagem de Máquina do MATLAB (versão 7.11). Visando utilizar apropriadamente a capacidade do modelo (SVM), seguiu-se o procedimento de treinamento e refinamento recomendado na documentação do pacote. Segundo o referido procedimento, realiza-se inicialmente um treinamento em que o software é instruído a determinar automaticamente um dos parâmetros de treinamento do modelo, denominado `KernelScale` (aqui indicado como KS) enquanto outro parâmetro, chamado `BoxConstraint` (abreviado como BC) é definido com o valor 1. Com o valor assim obtido para KS, denotado KS_0, faz-se uma busca em grade em que

$$(\mathrm{KS}, \mathrm{BC}) \in \{\mathrm{KS}_0\, 10^{-5}, \mathrm{KS}_0\, 10^{-4}, \ldots, \mathrm{KS}_0\, 10^{4}, \mathrm{KS}_0\, 10^{5}\} \times \{10^{-5}, 10^{-4}, \ldots, 10^{4}, 10^{5}\}.$$

Então, a partir dos valores desses parâmetros $(\mathrm{KS}_1, \mathrm{BC}_1)$ que produzirem os melhores desempenhos, faz-se uma nova busca em grade em que

$$\begin{gathered}(\mathrm{KS}, \mathrm{BC}) \in \{\mathrm{KS}_1\,(1{,}6)^{-4}, \mathrm{KS}_1\,(1{,}6)^{-3}, \ldots, \mathrm{KS}_1\,(1{,}6)^{3}, \mathrm{KS}_1\,(1{,}6)^{4}\} \\ \times \\ \{(1{,}6)^{-4}, (1{,}6)^{-3}, \ldots, (1{,}6)^{3}, (1{,}6)^{4}\}.\end{gathered}$$

Esse segundo refinamento de parâmetros que foi realizado nos experimentos é considerado opcional no procedimento recomendado. O fator de 1,6 entre valores de parâmetros foi escolhido porque $(1{,}6)^5 \approx 10$, que corresponde ao passo do primeiro refinamento. Dois tipos de SVMs foram avaliadas, as lineares, indicadas no texto como SVM_ℓ, e aquelas com

um núcleo baseado numa função de base radial (núcleo gaussiano), denotadas como SVM_r. As combinações destas com diferentes vetores de características são indicadas nas seções posteriores com a mesma notação usada para as RBMs. Assim, por exemplo, SVM_ℓ-C1 representa uma SVM linear empregando o vetor de características C1.

A seguir são detalhadas as configurações dos detectores de referência na forma em que os mesmos foram aplicados nos experimentos:

1. **ITU G.729-B / ITU G.729-II**: Esses detectores não possuem parâmetros configuráveis.
2. **LTSD**: Seguindo a nomenclatura de parâmetros do artigo original (RAMIREZ et al., 2004), usou-se N=6, K=3, NFFT=256 e α=0,98. Destes, somente o fator de esquecimento (α), difere dos valores propostos no artigo original. O novo valor produziu desempenhos superiores e por isso foi selecionado. Para o LTSD, foram usadas as mesmas configurações de tamanho e deslocamento entre quadros empregadas na extração dos MFCCs, ou seja, quadros de 25 ms de comprimento com deslocamento de 10 ms entre quadros sucessivos.
3. **Sohn** (SOHN; KIM; SUNG, 1999): Para esse VAD, determinou-se que dois parâmetros tinham influência direta no desempenho. Por meio de uma busca em grade executada em cada SNR, chegou-se aos parâmetros na Tabela 7. Os símbolos `ts` (*mean talkspurt length*) e `of` (*overlap factor*) são nomes usados no pacote `voicebox` que foi utilizado nos testes. Selecionou-se ainda um deslocamento entre quadros de 10 ms. O comprimento dos quadros é dado por `of` $\times$ 10 ms e, portanto, varia conforme a SNR.

Tabela 7 – Parâmetros do VAD Sohn para cada SNR.

Parâmetro	∞ dB	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB
`ts`	0,070	0,100	0,090	0,125	0,100	0,100	0,100
`of`	2	3	4	5	6	7	8

4. **Ying**: Esses detector foi avaliado com as exatas configurações do artigo original (YING et al., 2011).
5. **Ghosh**: Seguindo a nomenclatura de parâmetros do artigo original (GHOSH; TSIARTAS; NARAYANAN, 2011), usou-se M=5 e R conforme Tabela 8, os quais foram obtidos por uma busca em grade. Novamente, foram usadas as mesmas configurações de tamanho (25 ms) e deslocamento (10 ms) entre quadros empregadas na extração dos MFCCs. É interessante dizer que o parâmetro R determina o número de quadros passados a serem empregados no cálculo da entropia, que é a base da medida LTSV utilizada pelo detector (conforme explicado na Seção 4.3). Assim, observando-se a

Tabela 8, nota-se que o detector chega a utilizar (na SNR de 0 dB) dados de até 400 ms (40 × 10 ms entre quadros) no passado para tomar suas decisões.

Tabela 8 – Parâmetros do VAD Ghosh para cada SNR.

Parâmetro	∞ dB	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB
R	10	25	20	20	25	40	30

4.4.2 Resultados para VADs baseados em RBM e VADs de referência

A Tabela 9 mostra a acurácia obtida pelos vários detectores em diferentes SNRs. Na parte superior da tabela, estão os detectores baseados em RBM. Observa-se, pelos pares de linhas tendo o mesmo tipo de treinamento, que as acurácias obtidas com as configurações C1 e C2 (de vetores de características) são similares entre si em cada uma das SNRs. Nota-se ainda que, na maioria das SNRs, os melhores resultados dentre as RBMs (destacados em negrito na parte superior da Tabela 9) são conseguidos pelas DRBMs (DRBM-C1 e DRBM-C2) muito embora estas tenham um número de unidades ocultas (n_h) cerca de uma ordem de grandeza inferior aos das outras configurações (RBM-C1, RBM-C2, HDRBM-C1 e HDRBM-C2), como pode ser visto pela Tabela 6. Como indicado na Seção 3.3.6, o custo computacional de classificação varia linearmente com n_h, de modo que o uso das DRBMs seria vantajoso também sob o ponto de vista de complexidade computacional na presente aplicação. Pelos motivos expostos, deste ponto em diante as análises considerarão somente as DRBMs.

Na parte inferior da Tabela 9, estão as acurácias dos detectores de referência. É possível observar que, para o áudio sem ruído (∞ dB), os melhores resultados são obtidos pelo VAD G.729-B, enquanto que os detectores Ying e Ghosh exibem as menores acurácias dentre os detectores de referência. Interessantemente, estes dois apresentam melhores desempenhos na SNR de 20 dB do que na ausência de ruído, indicando que um pequeno

Tabela 9 – Acurácia (%) para RBMs e VADs de referência a diferentes SNRs.

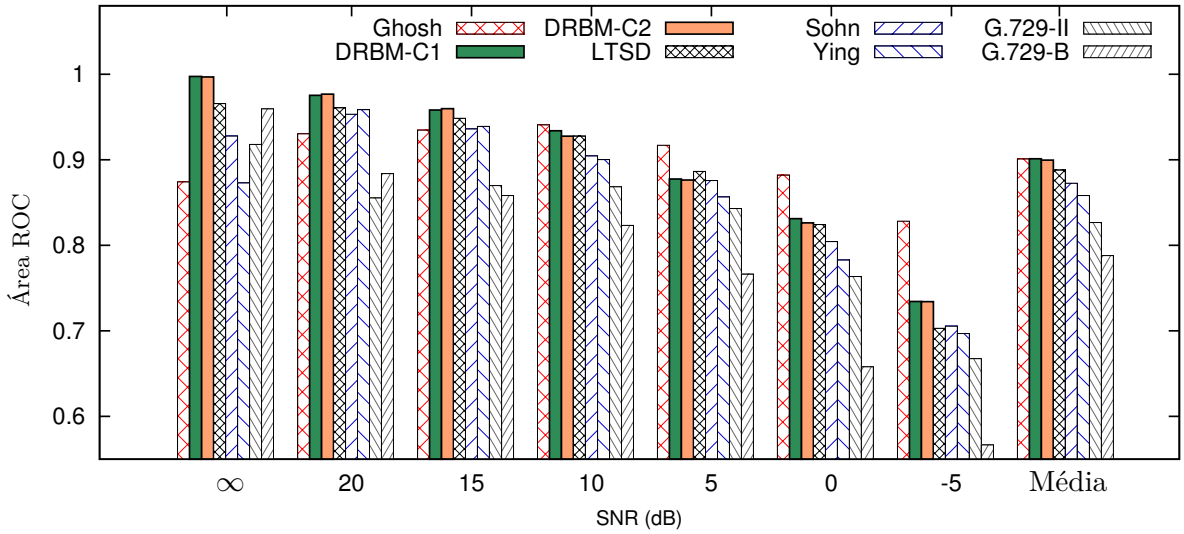
Detector	∞ dB	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	**Média**
DRBM-C1	97,65	**94,16**	91,41	**88,11**	82,54	**77,86**	68,61	**85,76**
DRBM-C2	97,69	93,98	**91,62**	87,74	**83,05**	76,80	**68,62**	85,64
RBM-C1	94,86	89,96	89,95	86,67	81,70	75,52	67,79	83,78
RBM-C2	97,40	90,11	91,10	87,30	82,39	76,17	68,13	84,66
HDRBM-C1	97,59	93,99	91,27	87,79	82,83	77,67	68,26	85,63
HDRBM-C2	**97,79**	94,04	91,51	88,01	82,95	77,20	68,57	85,73
G.729-B	**95,95**	88,38	85,82	82,32	76,63	65,80	56,68	78,80
G.729-II	91,80	85,55	86,98	86,85	84,29	76,35	66,76	82,65
LTSD	93,73	**91,07**	**89,70**	86,96	82,83	76,88	66,62	83,97
Sohn	91,00	90,93	88,96	86,43	83,63	76,89	68,70	83,79
Ying	87,31	90,95	88,44	84,25	79,63	72,78	65,62	81,28
Ghosh	85,53	88,70	89,51	**89,84**	**86,71**	**82,75**	**78,59**	**85,95**

nível de ruído é benéfico ao funcionamento deles. Os dados mostram também que o VAD Ghosh é consideravelmente superior aos demais em SNRs de 10 dB ou inferiores. De fato, na média ao longo das SNRs, esse VAD produz os melhores resultados, seguido pelos detectores LTSD e Sohn.

Na comparação das DRBMs com os detectores de referência, tem-se o seguinte. Desconsiderando o VAD Ghosh, percebe-se, pela Tabela 9 novamente, que as DRBMs oferecem acurácias superiores ou comparáveis às dos outros VADs na maioria das SNRs. Além disso, para SNRs de 15 dB ou maiores, as DRBMs fornecem desempenhos sensivelmente superiores ao do VAD Ghosh, mas essa situação se inverte nas SNRs de 10 dB ou menores. Assim, o VAD Ghosh atinge a melhor média de acurácia ao longo das SNRs, seguido pelos detectores DRBM-C1 e DRBM-C2.

As medidas de área ROC dos detectores, dadas na Figura 47, confirmam os comentários anteriores relativos às acurácias. A figura mostra que os detectores baseados em DRBM apresentam áreas ROC maiores que todos os outros, exceto o VAD Ghosh, na maioria das SNRs. Os resultados novamente evidenciam a superioridade desse VAD em SNRs mais baixas e o seu desempenho relativamente inferior em SNRs elevadas.

Figura 47 – Área sob a curva ROC para DRBMs e VADs de referência a diferentes SNRs.

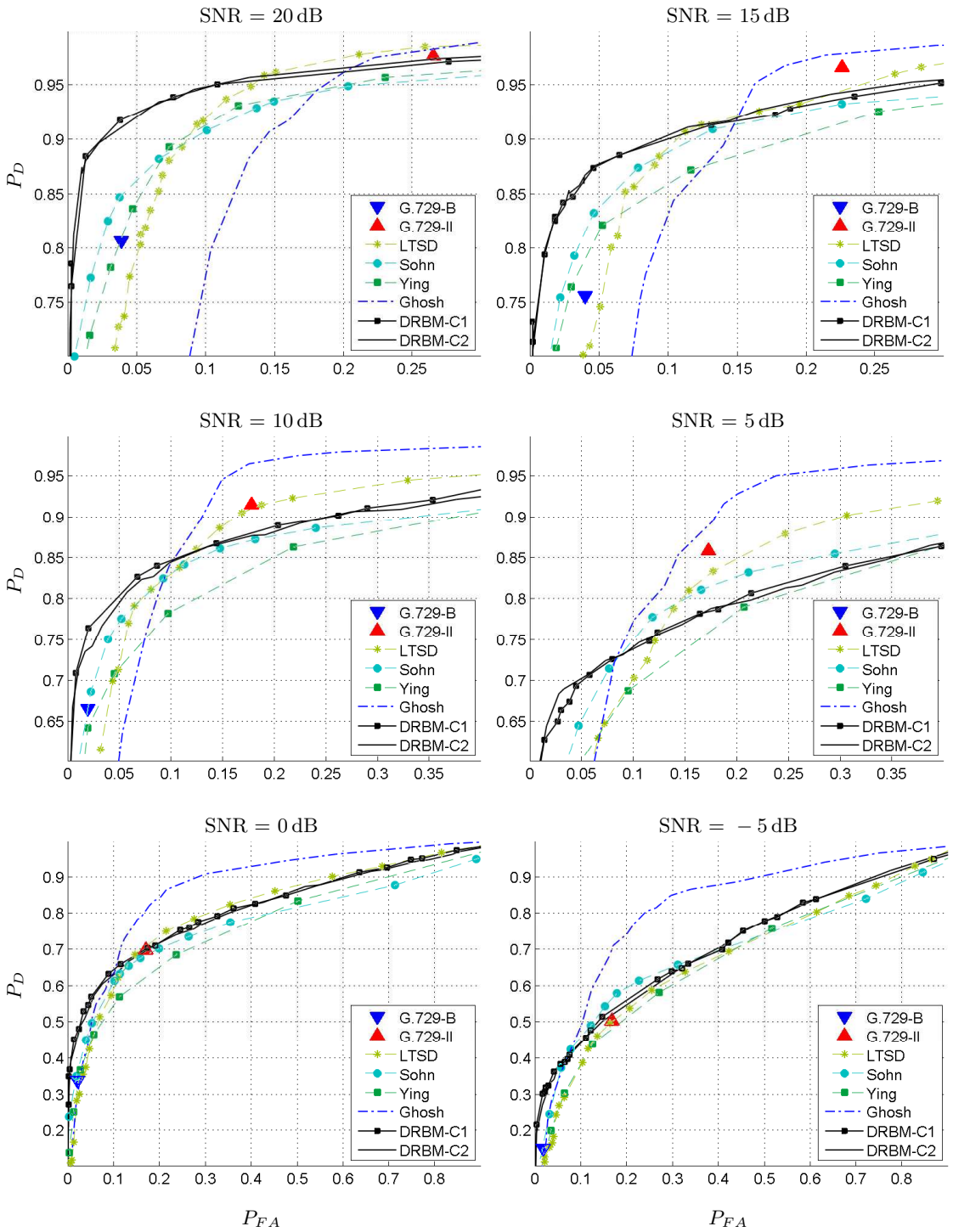


Fonte: produção do próprio autor.

Uma visão mais detalhada do desempenho dos vários detectores é fornecida pelas suas curvas ROC, dadas na Figura 48. Uma característica que é observada em todas as SNRs é que, para uma probabilidade de falso alarme (P_{FA}) suficientemente baixa, os VADs baseados em DRBM oferecem uma probabilidade de detecção (P_D) maior que as dos demais detectores e, portanto, seu uso seria vantajoso nessa condição. Isso é constatado até mesmo quando as DRBMs são comparadas com o VAD Ghosh. Por exemplo, para a SNR de 10 dB, em que a área ROC do VAD Ghosh é superior às demais, percebe-se que,

para $P_{FA} < 0{,}1$, o uso das configurações baseadas em DRBM é mais interessante uma vez que elas oferecem valores de P_D consideravelmente maiores do que os fornecidos pelo VAD Ghosh.

Figura 48 – Curvas ROC (treinamento em condição única de SNR).



Fonte: produção do próprio autor.

Agora que os desempenhos de classificação dos detectores foram apresentados, é interessante observar os seus custos computacionais, obtidos empiricamente e exibidos na

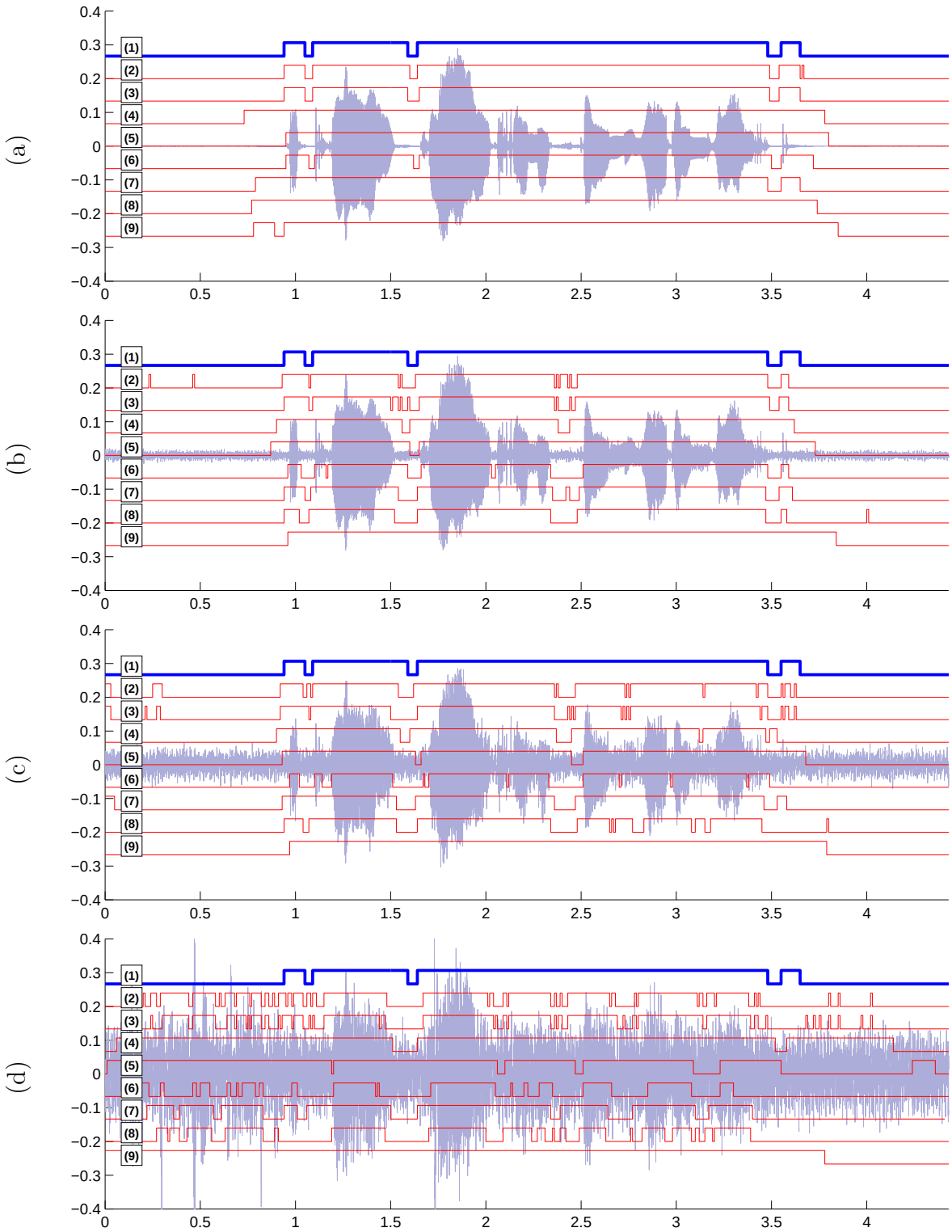
Tabela 10 – Taxa de trabalho de tempo real (s/s) para DRBMs e VADs de referência.

DRBM-C1	DRBM-C2	LTSD	SOHN	YING	GHOSH	G.729-B	G.729-II
509,3	560,4	322,5	110,9	64,5	99,7	48,6	47,5

Tabela 10. Para os detectores baseados em DRBM, as medidas mostradas englobam o trabalho de extração de características (vetores C1 ou C2) seguido da classificação pela respectiva DRBM. A partir da referida tabela, verifica-se que os detectores DRBM-C1 e DRBM-C2 apresentam os menores custos computacionais, sendo eles capazes de processar mais de 500 segundos de áudio em um segundo do processador (conforme interpretação da medida de taxa de trabalho de tempo real introduzida na Seção 4.2). Em contrapartida, os VADs G.729-B e G.729-II conseguem processar menos de 50 segundos de áudio no mesmo tempo de uso do processador. Deve-se ressalvar, entretanto, que estes VADs tomam suas decisões com base em informações produzidas por um *codec* de voz e, portanto, parte do tempo gasto por eles corresponde à codificação de voz. Comparando agora o detector DRBM-C1 (que apresentou acurácia média maior do que a do DRBM-C2) com os dois VADs com melhor acurácia média dentre os de referência – Ghosh e LTSD – pode-se concluir que o detector baseado em DRBM é capaz de processar uma quantidade de áudio aproximadamente 5 vezes maior que o primeiro e 1,6 vezes maior que o segundo VAD de referência, por unidade de tempo.

Finalmente, a título de ilustração, a Figura 49 mostra a operação dos vários VADs em algumas SNRs. Na condição sem ruído, os detectores DRBM-C1, DRBM-C2 e G.729-B apresentam as saídas que mais se assemelham à rotulação manual. Os detectores LTSD, Sohn e Ying assumem a existência de um ruído no áudio e, portanto, operam em condições não ideais na ausência dele. Nas SNRs de 20 dB e 10 dB, os detectores exibem saídas mais semelhantes entre si. Percebe-se que o VAD Ghosh tende a produzir uma saída estável durante todo o trecho contendo voz. Isso pode ser atribuído ao uso (por parte desse VAD) de uma janela relativamente grande (aproximadamente 400 ms) na suas tomadas de decisão. Pode-se entender que uma janela maior aumenta a confiança do detector em indicar a presença ou ausência de voz num dado trecho de áudio. Em contrapartida, as referidas decisões são menos localizadas no tempo. Pode-se notar, por exemplo, que o VAD Ghosh não foi capaz de detectar os curtos trechos de ausência de voz ao longo da frase pronunciada no áudio. Por fim, os gráficos obtidos na condição de 0 dB de SNR tornam evidentes as dificuldades dos VADs em determinar os trechos contendo voz.

Figura 49 – Exemplo de operação dos detectores a diferentes SNRs: (a) Sem ruído, (b) 20 dB, (c) 10 dB e (d) 0 dB. Para cada SNR, são mostradas, sobrepostas à imagem do áudio, a (1) rotulação manual e as saídas dos detectores, sendo: (2) DRBM-C1, (3) DRBM-C2, (4) LTSD, (5) G.729-II, (6) G.729-B, (7) Sohn, (8) Ying e (9) Ghosh.



Fonte: produção do próprio autor.

4.4.3 Resultados para VADs baseados em DRBM e em SVM

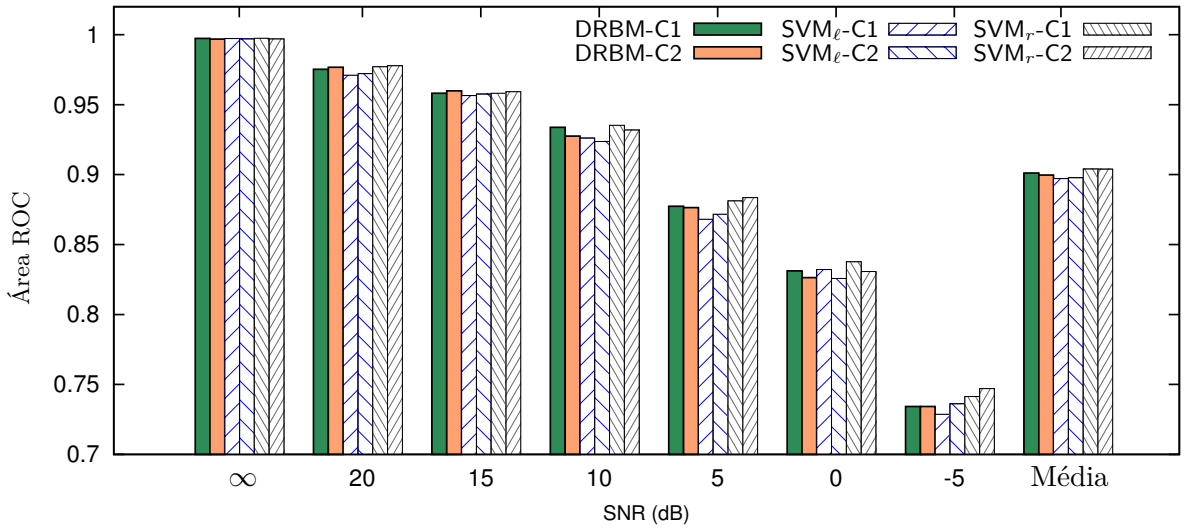
Na tarefa de VAD, as DRBMs produziram resultados superiores aos das outras RBMs e ainda conseguiram tais resultados empregando um menor número de unidades ocultas. Por isso, apenas elas são comparadas agora com outro mecanismo de aprendizagem de máquina, as SVMs, as quais foram treinadas com os mesmos dados que as DRBMs.

Na Tabela 11, pode-se observar as acurácias das DRBMs e SVMs quando empregando as configurações C1 e C2 de vetores de características (conforme Tabela 5). Percebe-se que, tanto na configuração C1 quanto na C2, as DRBMs fornecem, na maioria das SNRs, desempenhos levemente superiores aos das SVM lineares (SVM_ℓ), mas também levemente inferiores aos das SVMs com núcleo gaussiano (SVM_r). Verifica-se que, em termos da acurácia média (ao longo das SNRs), a diferença entre as DRBMs e as correspondentes SVMs é inferior a 0,6%. As similaridades nos desempenhos dos modelos são confirmadas pelas suas medidas de área ROC, as quais estão representadas na Figura 50.

Tabela 11 – Acurácia (%) para DRBMs e SVMs (treinamento em condição única de SNR).

Detector	∞ dB	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	**Média**
DRBM-C1	97,65	94,16	91,41	88,11	82,54	77,86	68,61	85,76
SVM_ℓ-C1	97,78	92,17	91,33	87,42	82,21	77,74	68,43	85,29
SVM_r-C1	97,82	94,71	91,46	88,72	83,24	78,63	69,13	86,24
DRBM-C2	97,69	93,98	91,62	87,74	83,05	76,80	68,62	85,64
SVM_ℓ-C2	97,66	92,29	91,37	88,02	82,72	77,25	68,08	85,34
SVM_r-C2	97,75	94,69	91,53	88,67	83,44	78,27	69,16	86,21

Figura 50 – Área sob a curva ROC para DRBMs e SVMs (treinamento em condição única de SNR).



Fonte: produção do próprio autor.

As taxas de trabalho de tempo real dos vários detectores baseados em aprendizagem de máquina foram medidas e são mostradas na Tabela 12. Para as SVMs, os valores na

tabela representam médias ao longo das SNRs. De fato, SVMs treinadas em condições diversas produzem classificadores com diferentes custos computacionais, o que é explicado pelo fato de as mesmas apresentarem quantidades também diferentes de vetores de suporte. Para o vetor de características C1, vê-se que o detector baseado em DRBM seria capaz de processar (em um segundo do processador) aproximadamente 2,5 e 3,4 vezes mais áudio do que as SVMs linear e com núcleo gaussiano, respectivamente. Para o vetor C2, a relação é ainda maior: 4,3 e 5,3 vezes, em cada caso.

Tabela 12 – Taxa de trabalho de tempo real (s/s) para DRBMs e SVMs.

DRBM-C1	DRBM-C2	SVM$_\ell$-C1	SVM$_\ell$-C2	SVM$_r$-C1	SVM$_r$-C2
509,3	560,4	204,5	130,0	149,6	106,7

Pode-se ainda descontar os tempos despendidos na extração de características do áudio a fim de comparar apenas os custos computacionais dos classificadores que fazem parte de cada um dos detectores. Para facilitar a comparação, calculou-se a razão entre o tempo de classificação de cada um dos classificadores e o tempo do mais rápido deles (DRBM-C2). Os resultados são apresentados na Tabela 13. Para os detectores empregando o vetor C1, verifica-se que as SVMs linear e com núcleo gaussiano despendem um tempo aproximadamente 12,3 e 21,2 vezes maior que a DRBM para classificar o mesmo conjunto de amostras, respectivamente. No caso do vetor C2, esses números aumentam para 27,3 e 35,2, na mesma ordem.

Tabela 13 – Razão de tempo de classificação (s/s) em relação à DRBM-C2.

DRBM-C1	DRBM-C2	SVM$_\ell$-C1	SVM$_\ell$-C2	SVM$_r$-C1	SVM$_r$-C2
1,36	1,00	16,7	27,3	28,8	35,2

4.4.4 Conclusões

Nos experimentos, constata-se que as DRBMs empregando os vetores de características C1 (baseado em MFCCs) e C2 (baseado em FBEs) produzem desempenhos parecidos entre si, independentemente do tipo de treinamento usado para obtenção dos parâmetros de cada modelo (discriminativo, generativo ou híbrido). Como mencionado anteriormente, as FBEs são vistas como medidas fortemente correlacionadas entre si, enquanto os MFCCs são considerados pouco correlacionados. Dessa forma, há razões para concluir que o modelo proposto seja capaz de lidar apropriadamente com dados correlacionados. Interessantemente, essa capacidade já foi verificada na literatura para as DNNs construídas a partir de DBNs, as quais, por sua vez, são formadas pelo empilhamento de RBMs (HINTON et al., 2012).

Na comparação entre os diferentes tipos de treinamento de RBMs, pode-se dizer, a partir dos resultados, que o treinamento discriminativo seria o mais vantajoso na aplicação

de VAD por produzir desempenhos melhores ou comparáveis aos dos outros, com um custo computacional consideravelmente inferior (em razão do número pelo menos 10 vezes menor de unidades ocultas). De fato, esse ponto já foi destacado na Seção 4.4.2 e motivou o foco dado às DRBMs no presente trabalho.

Comparando os detectores baseados em DRBM com os VADs de referência, é possível dizer que as DRBMs oferecem, nas condições dos experimentos, resultados (em termos de acurácia e área ROC média) bastante razoáveis, sendo eles superiores aos de todos VADs de referência, exceto o Ghosh. O detector DRBM-C1, por exemplo, fornece, na média ao longo das SNRs, uma acurácia 1,79% superior à do LTSD, o segundo melhor detector de referência. Além disso, os detectores baseados em DRBM se mostram superiores ao próprio VAD Ghosh quando em SNRs elevadas (15 dB ou acima). Em SNRs de 10 dB e abaixo, o uso das DRBMs também poderia ser mais interessante do que o uso do Ghosh, caso se desejasse operar com probabilidades de falso alarme muito baixas (à custa de uma menor probabilidade de detecção), como verificado pelas curvas ROC.

Quanto ao custo computacional, as medidas realizadas indicam que as DRBMs podem produzir VADs mais eficientes do que aqueles tomados como base de comparação. Por exemplo, segundo os resultados apresentados, o detector DRBM-C1 é capaz de processar pelo menos 5 vezes mais áudio do que os VADs de estado da arte (Sohn, Ying e Ghosh), por unidade de tempo.

Considerando agora os resultados obtidos com as SVMs, tem-se a seguinte situação. As DRBMs oferecem valores médios de acurácia e de área ROC levemente superiores às SVMs lineares e levemente inferiores às SVMs com núcleo gaussiano, sendo que a diferença de acurácia média entre todos os modelos não chega a 1%. Assim, em termos das medidas de desempenho, as SVMs com núcleo gaussiano parecem produzir detectores moderadamente mais poderosos do que as DRBMs. Entretanto, para o vetor C1, por exemplo, a SVM com núcleo gaussiano apresenta um custo computacional de classificação 21,2 vezes maior do que a DRBM. Em vista desses fatos, é seguro dizer que a escolha da DRBM no lugar da SVM poderia ser justificada em muitas aplicações.

4.5 AVALIAÇÕES EM CONDIÇÃO MÚLTIPLA DE RELAÇÃO SINAL-RUÍDO

Nesta seção, são apresentados os experimentos realizados na chamada condição múltipla (*multi-condition*) de relação sinal-ruído. Nessa condição, um modelo é treinado com uma mistura de dados de diferentes SNRs. Entretanto, o modelo resultante de tal treinamento é normalmente avaliado separadamente em cada uma das SNRs (como feito neste trabalho). Um vantagem do treinamento na condição mencionada seria a obtenção de um detector usável em uma larga faixa de SNRs, dispensando o trabalho de estimação de SNR do canal, por exemplo. Nota-se que a adição da SNR como um fator de variabilidade

aos dados, pode ser vista como um aumento na dificuldade do problema de classificação a ser enfrentado pelos modelos. Em contrapartida, a maior quantidade de dados a eles fornecida no treinamento (pela combinação dos dados de SNRs diversas) pode auxiliá-los a capturar melhor as características que distinguem a voz de outros sons. Portanto, os experimentos nesta seção têm como objetivo verificar o impacto gerado pelo treinamento em condição múltipla de SNR nos desempenhos dos modelos, e daí a viabilidade de um detector produzido dessa maneira. As seções seguintes estão organizadas assim: na Seção 4.5.1 são dados os detalhes dos procedimentos e das configurações dos modelos testados. Os resultados dos experimentos e as correspondentes conclusões são fornecidos nas Seções 4.5.2 e 4.5.3, respectivamente.

4.5.1 Procedimentos e configurações

Nas avaliações em condição múltipla de SNR, utilizou-se apenas a configuração C1 (conforme Tabela 5) para o vetor de características. O áudio foi pré-processado e os vetores de características foram extraídos e normalizados da mesma forma como explicado na Seção 4.4.1. Os resultados mostrados na próxima seção são relativos ao treinamento discriminativo (DRBM) com parâmetros

$$n_h{=}80, \lambda_{disc}{=}5 \times 10^{-3}.$$

Estes foram obtidos por uma busca em grade, em que

$$(n_h, \lambda_{disc}) \in \{10, 15, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90\} \times \{1{\times}10^{-3}, 5{\times}10^{-3}\}.$$

Para comparação, foram avaliadas também as SVMs linear e com núcleo gaussiano, tendo como entrada os mesmos dados fornecidos para as DRBMs. O procedimento para treinamento das SVMs, bem como a nomenclatura dos modelos, seguem também aqueles apresentados na Seção 4.4.1.

4.5.2 Resultados

Na Tabela 14, são mostradas as acurácias dos detectores avaliados, cujos classificadores foram treinados na condição múltipla de SNR. A partir da tabela, percebe-se que as acurácias fornecidas pela DRBM são superiores àquelas da SVM linear e inferiores às da SVM com núcleo gaussiano. Isto é verificado consistentemente em todas SNRs.

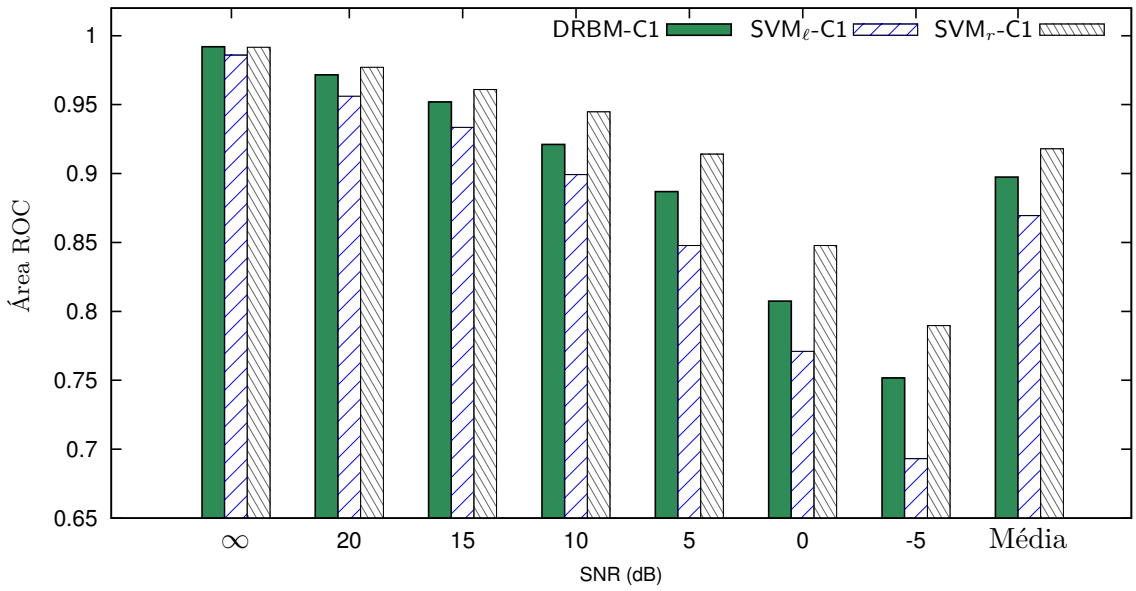
Tabela 14 – Acurácia (%) para DRBMs e SVMs (treinamento em condição múltipla de SNR).

Detector	∞ dB	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	**Média**
DRBM-C1	96,92	93,48	91,30	88,28	84,51	76,82	70,45	85,97
SVM$_\ell$-C1	95,42	91,80	89,13	85,07	79,19	71,84	64,68	82,45
SVM$_r$-C1	95,76	94,25	92,12	89,50	86,50	78,95	73,92	87,29

É interessante comparar as acurácias da Tabela 14 com aquelas da Tabela 11, relativas ao treinamento em condição única de SNR. Primeiramente, verifica-se que, com o treinamento em condição múltipla de SNR, houve um aumento na acurácia média dos detectores DRBM-C1 (de 85,76% para 85,97%) e SVM$_r$-C1 (de 86,24% para 87,29%), enquanto que o SVM$_\ell$-C1 apresentou uma diminuição na referida medida (de 85,29% para 82,45%). Em segundo lugar, nota-se que também ocorreu um distanciamento entre os desempenhos dos detectores. Em condição única de SNR, a diferença absoluta entre acurácias médias do detector DRBM-C1 para os detectores SVM$_\ell$-C1 e SVM$_r$-C1 eram, respectivamente, 0,47% e 0,48%, as quais passaram, em condição múltipla, para 3,52% e 1,32% (nessa ordem). Assim, além do distanciamento, observa-se que, em termos de acurácia, o desempenho da DRBM fica mais próximo da SVM com núcleo gaussiano do que da SVM linear.

As medidas de área ROC dos detectores, representadas na Figura 51, têm características qualitativamente semelhantes às das acurácias, já comentadas. A visualização da figura evidencia que, em SNRs elevadas, os detectores DRBM-C1 e SVM$_r$-C1 têm desempenhos mais próximos entre si, e superiores ao do SVM$_\ell$-C1. Já em SNRs mais baixas, há um considerável afastamento nas medidas de área ROC dos detectores, com o DRBM-C1 fornecendo desempenhos intermediários entre os outros dois. Esses desempenhos intermediários ficam claros também nas curvas ROC, mostradas na Figura 52 para algumas SNRs.

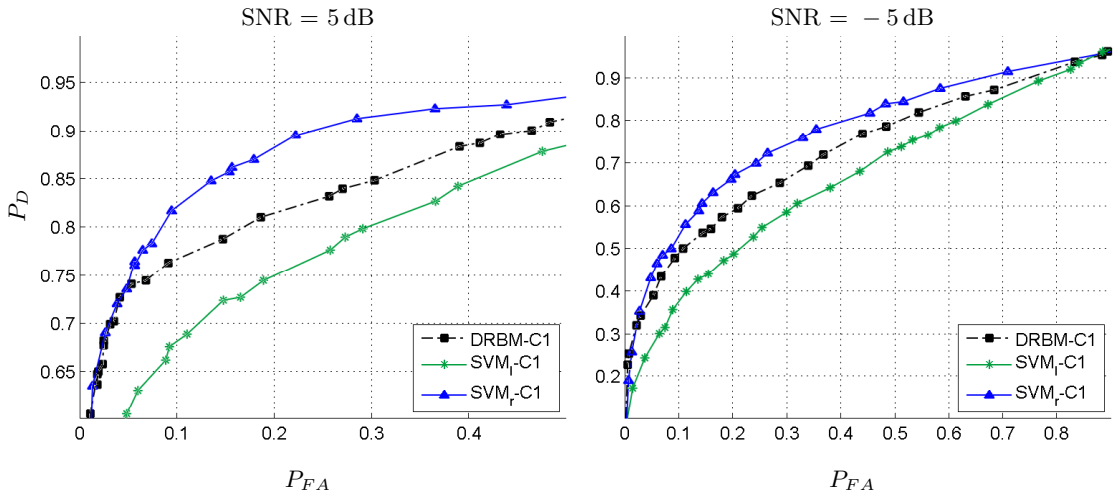
Figura 51 – Área sob a curva ROC para DRBMs e SVMs (treinamento em condição múltipla de SNR).



Fonte: produção do próprio autor.

Vale agora comparar os custos computacionais dos detectores. Considerando inicialmente as taxas de trabalho de tempo real, dadas na Tabela 15, nota-se que o detector

Figura 52 – Curvas ROC (treinamento em condição múltipla de SNR).



Fonte: produção do próprio autor.

DRBM-C1 seria capaz de processar cerca de 400 segundos de áudio por segundo em um certo processador. Esse valor é cerca de 26,7 e 20,2 vezes maior do que o conseguido com os detectores SVM$_\ell$-C1 e SVM$_r$-C1, respectivamente.

Tabela 15 – Taxa de trabalho de tempo real (s/s) (treinamento em condição múltipla de SNR).

DRBM-C1	SVM$_\ell$-C1	SVM$_r$-C1
400,7	15,0	19,9

Como feito na condição única de SNR, pode-se ainda comparar os tempos dos classificadores que compõem os detectores. Assim, os tempos de cada um deles normalizados pelo tempo do mais rápido (DRBM-C1) são mostrados na Tabela 16. Vê-se agora que as SVMs linear e com núcleo gaussiano despendem, respectivamente, um tempo que é 65,8 e 48,1 vezes maior do que a DRBM para classificar o mesmo conjunto de amostras. Como mencionado anteriormente, o custo computacional relativamente elevado de cada uma das SVMs pode ser explicado pelo grande número de vetores de suporte por elas utilizados para atingir bons desempenhos.

Tabela 16 – Razão de tempo de classificação (s/s) em relação à DRBM-C1 (treinamento em condição múltipla de SNR).

DRBM-C1	SVM$_\ell$-C1	SVM$_r$-C1
1,0	65,8	48,1

4.5.3 Conclusões

Como observado nos resultados da seção anterior, o treinamento em condição múltipla de SNR produz, em relação à condição única, um pequeno ganho de acurácia

média (0,21%) para o detector baseado em DRBM, e um ganho maior (1,02%) para aquele baseado em SVM com núcleo gaussiano. Em contrapartida, a SVM linear apresenta uma sensível perda de acurácia (3,52%) com o treinamento em condição múltipla de SNR. Pode-se interpretar que o novo fator de variabilidade nos dados, adicionado por essa condição de treinamento, representa um obstáculo para a operação da SVM linear. Já para a DRBM e a SVM com núcleo gaussiano, a hipótese é que o aumento na dificuldade do problema em razão do novo fator de variabilidade seja compensado pelos ganhos fornecidos pela maior quantidade de dados de treinamento (da qual esses modelos conseguiram tirar vantagem). Sendo assim, conclui-se que o treinamento em condição múltipla de SNR seria benéfico a esses dois modelos e, mais ainda, seria interessante na prática, por produzir, como já mencionado, detectores capazes de operar em uma ampla faixa de SNRs.

Comparando a DRBM com a SVM com núcleo gaussiano, vê-se que esse último produz desempenhos superiores, fornecendo um acurácia média 1,32% acima da DRBM. Na SNR de $-5\,\mathrm{dB}$, a diferença nas acurácias é de quase 3,5%. Esses ganhos podem ser importantes em algumas aplicações. Entretanto, com relação ao custo computacional, o VAD baseado em DRBM se mostra substancialmente vantajoso. Pelos resultados constatados, o detector DRBM-C1 é capaz de processar 20,2 vezes mais áudio do que o SVM_r-C1, por unidade de tempo. Além disso, comparando os tempos de classificação apenas, o VAD SVM_r-C1 despende um tempo 48,1 vezes maior do que o DRBM-C1 para processar uma mesma quantidade de amostras. Em muitas aplicações, tais números poderiam justificar a escolha da DRBM no lugar da SVM com núcleo gaussiano, apesar do desempenho um pouco inferior da primeira.

4.6 AVALIAÇÕES DE MEDIDAS PROPOSTAS RECENTEMENTE

Como discutido na Seção 2.4.2, Sadjadi e Hansen (2013) propuseram um VAD empregando medidas sensíveis à presença no áudio de sons vocalizados. Drugman et al. (2015) utilizaram as mesmas medidas com outras mais relacionadas com a fonte sonora e obtiveram resultados de estado da arte. Nesse último trabalho, muitos detalhes do detector apresentado pelos autores foram omitidos. Entretanto, os autores atribuem os ótimos resultados às medidas (características) cujo uso em VAD foi por eles proposto. Sendo assim, nesta seção, as diferentes combinações de medidas testadas em Drugman et al. (2015) são avaliadas na tarefa de VAD usando a variante de RBM proposta no presente trabalho. Com isso, objetiva-se verificar o potencial dessas medidas para elevar o desempenho de VADs baseados em aprendizagem de máquina. As seções à frente estão organizadas como segue. Inicialmente, o trabalho de referência já mencionado é brevemente descrito na Seção 4.6.1. Em seguida, a Seção 4.6.2 detalha o procedimento de treinamento do modelo e as configurações usadas na extração de características do áudio. Os resultados e as conclusões sobre os experimentos são então fornecidos nas Seções 4.6.3 e 4.6.4,

respectivamente.

4.6.1 Trabalho de referência

Como descrito na página 68, Drugman et al. (2015) propuseram um VAD empregando 3 conjuntos de medidas (todas descritas na Seção 2.4.2):

1. Os coeficientes mel-ceptrais (MFCCs).
2. As medidas propostas por Sadjadi e Hansen (2013): harmonicidade, clareza, ganho de predição e periodicidade.
3. Medidas propostas pelos próprios autores: CPP, SRH_1 e SRH_2.

Os melhores resultados foram conseguidos com a técnica que os autores chamam de fusão de decisão (*decision fusion*), em que cada conjunto de medidas (da lista anterior) define a entrada de uma ANN. A saída combinada das ANNs é dada pela média geométrica das suas saídas individuais. Os vetores de características são construídos concatenando as medidas mencionadas com suas derivadas temporais de primeira e segunda ordem. Antes de adentrar uma ANN, esses vetores são pré-processados, passando por um filtro de mediana (ao longo do tempo) com uma largura de 11 quadros (5 de cada lado do quadro atual). A saída combinada das ANNs é convertida num valor binário (por um simples limiar) e a sequência desses valores no tempo é pós-processada por uma operação de fechamento morfológico com constante de tempo de 600 ms. Os autores comparam os resultados do detector descrito com os detectores G.729-B, Sohn, Ying e Ghosh. Segundo a métrica usada no trabalho, a medida F_1 (F_1-score), o VAD proposto é bem superior aos demais. Ele obtém F_1=93,0% na média, enquanto o segundo melhor VAD (Ghosh), consegue F_1=69,1%. A medida F_1 não está diretamente relacionada com a acurácia balanceada usada no presente trabalho. Entretanto, pode-se demonstrar que, para um conjunto de dados balanceado, vale

$$\mathrm{BA} = 1 + \left(1 - \frac{1}{F_1}\right)\mathrm{TPR}, \tag{4.4}$$

sendo que TPR (definida na Seção 4.2) não é conhecida, mas é um número entre 0 e 1. Assim, para F_1=93,0%, como reportado do artigo, BA $\geq$ 92,4%. Esse número dá uma ideia da acurácia média que seria esperada nos experimentos.

4.6.2 Procedimentos e configurações

As configurações básicas de vetores de características são fornecidas na Tabela 17, sendo que os símbolos nela utilizados foram introduzidos na Seção 2.4.2. Nota-se que todas as configurações são compostas de um conjunto de medidas juntamente com suas derivadas de primeira e segunda ordem (Δ's e $\Delta\Delta$'s). A configuração chamada MFCC

é naturalmente baseada nos MFCCs e aquela denominada C1 (já introduzida) adiciona a eles a informação de energia de quadro. A configuração S diz respeito ao conjunto de medidas propostas por Sadjadi e Hansen (2013), enquanto que a configuração D se refere àquelas incorporadas por Drugman et al. (2015). Além das configurações na tabela, foram testadas ainda combinações das mesmas, a saber: S+D, C1+S, C1+D e C1+S+D. Nessa notação, o sinal de soma denota a concatenação dos respectivos vetores de características. A operação de log em algumas das medidas indicadas na Tabela 17 foi motivada pelos experimentos na Seção 2.4.2.9. Basicamente, essa operação produz novas medidas cujas distribuições são aproximadamente gaussianas e, portanto, mais apropriadas a hipótese de distribuição condicionalmente gaussiana do modelo usado como classificador. Os vetores de características foram ainda normalizados como explicado na Seção 4.4.1.

Tabela 17 – Configurações básicas dos vetores de características (medidas recentes).

Configuração	**Conteúdo do vetor**	**Dimensão**
MFCC	MFCCs e respectivos Δ's e $\Delta\Delta$'s.	39
C1	MFCCs, ln(FE) e respectivos Δ's e $\Delta\Delta$'s.	42
S	$\ln(h)$, c, G_p, P_{hps} e respectivos Δ's e $\Delta\Delta$'s.	12
D	CPP, $\ln(\mathrm{SRH}_1)$, SRH_2 e respectivos Δ's e $\Delta\Delta$'s.	9

Na extração de características, foram usadas as seguintes configurações (conforme nomenclatura de parâmetros dada na Seção 2.4.2):

1. MFCC, C1: Quadros de 25 ms com deslocamento de 10 ms entre quadros.
2. S: Quadros de 32 ms com deslocamento de 10 ms entre quadros, N_{hps}=8 (para HPS) e p=10 (para G_p) como feito em (SADJADI; HANSEN, 2013).
3. D: Quadros de 40 ms com deslocamento de 5 ms entre quadros. O tamanho do quadro se baseou naquele usado em (HILLENBRAND; CLEVELAND; ERICKSON, 1994) (originalmente 41 ms). Deve-se dizer que a medida SRH foi proposta inicialmente para a determinação da frequência principal da voz (tom). Em tal uso, foram utilizados quadros de 100 ms. Entretanto, testes preliminares mostraram que quadros de 40 ms seriam apropriados para o uso em VAD. No cálculo das duas variantes de SRH, usou-se N_{srh}=5, como em (DRUGMAN; ALWAN, 2011).

Como classificador, em todos os testes foram usadas DRBMs (treinamento discriminativo) tendo

$$n_h{=}30, \lambda_{disc}{=}5 \times 10^{-3}.$$

4.6.3 Resultados

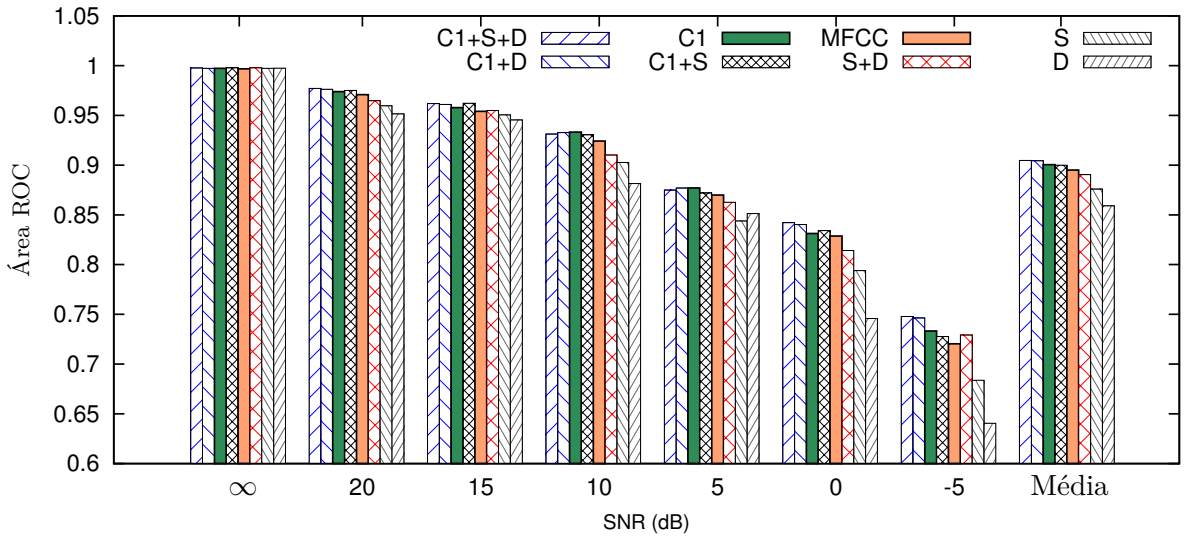
Na Tabela 18, são exibidas as acurácias obtidas pelas DRBMs empregando diferentes configurações de vetores de características. As linhas da tabela estão organizadas em ordem

decrescente de acurácias médias. Nota-se, primeiramente, que as configurações envolvendo apenas as medidas propostas recentemente (S, D, S+D) obtêm desempenhos inferiores aos da configuração chamada MFCC. Essa, por sua vez, é inferior à configuração C1, a qual foi empregada nos vários experimentos das seções anteriores. Verifica-se ainda que a adição das medidas da configuração S àquelas da C1 (produzindo a C1+S) não fornece ganhos significativos. Já a adição das medidas da configuração D, ou mesmo da D+S, produz pequenos ganhos na média. De fato, comparando as configurações C1 e C1+S+D, nota-se que há um aumento de 0,57% de acurácia na média ao longo das SNRs pelo uso das medidas recentes. Além disso, o maior aumento individual de acurácia constatado é de 1,51% e ocorre na SNR de -5 dB. As medidas de área ROC representadas na Figura 53 são qualitativamente semelhantes às medidas de acurácia.

Tabela 18 – Acurácia (%) para DRBMs empregando diferentes vetores de características.

Configuração	∞ dB	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	**Média**
C1+S+D	97,85	94,30	92,09	87,67	83,07	78,67	69,33	86,14
C1+D	97,55	93,94	92,00	87,77	83,19	78,38	69,38	86,03
C1+S	97,87	94,02	91,80	87,53	82,49	77,49	67,90	85,59
C1	97,62	93,62	91,25	88,06	82,72	77,88	67,82	85,57
MFCC	97,46	93,34	90,67	87,01	81,63	77,11	66,80	84,86
S+D	97,88	91,69	91,28	85,39	82,00	74,73	67,00	84,28
S	97,60	90,74	90,47	84,78	80,27	72,31	63,77	82,85
D	97,85	90,01	90,58	82,60	80,25	69,65	61,23	81,74

Figura 53 – Área sob a curva ROC para DRBMs empregando diferentes vetores de características.



Fonte: produção do próprio autor.

4.6.4 Conclusões

Nos resultados apresentados na seção anterior, constata-se que os vetores de características baseados apenas nas medidas recentes (S, D, S+D) produzem desempenhos

inferiores ao dos tradicionais MFCCs. Verifica-se também que a simples adição da informação de energia de quadro à configuração MFCC, produzindo aquela chamada C1, causa um aumento de 0,71% na acurácia média do detector, sendo, portanto, vantajosa, como reportado na literatura. A combinação das medidas recentes ao vetor C1, dando origem à configuração C1+S+D, oferece ganhos pouco expressivos na acurácia do VAD (0,57%). Embora ocorram ganhos um pouco maiores em SNRs específicas, os comentários fornecidos até esse ponto são suficientes para se concluir que os resultados obtidos não são consistentes com aqueles reportados no trabalho de referência (DRUGMAN; ALWAN, 2011) e brevemente descritos na Seção 4.6.1, em que os desempenhos conseguidos com as medidas recentes são muito superiores àqueles obtidos pelos VADs de estado da arte (inclusive aos do VAD Ghosh que também foi avaliado no trabalho). Sendo assim, a seguir são apresentadas algumas hipóteses que poderiam justificar a inconsistência observada:

1. Uma primeira possibilidade é que o modelo (DRBM) não tenha sido capaz de usufruir dos potenciais ganhos que as medidas recentes teriam a oferecer. Trabalhos posteriores poderiam (como sugestão) refutar essa possibilidade por meio de experimentos com outros modelos. Entretanto, é interessante apontar que a observação da Tabela 18 mostra que, em todos os casos em que vetores de características são concatenados, são produzidos aumentos de acurácia. Tal fato pode indicar que o fator limitante para o desempenho do detector não seja o modelo usado como classificador.
2. Os fortes ganhos de desempenho reportados no trabalho de referência podem estar relacionados a características específicas do *corpus* proprietário empregado pelos autores e podem não refletir os verdadeiros desempenhos do detector por eles proposto.
3. Como explicado na Seção 4.6.1, os autores do trabalho de referência realizam o pré-processamento dos vetores de características e o pós-processamento da saída (bruta) do detector. Estes poderiam eventualmente explicar os ganhos observados. Deve-se dizer que tais processamentos não foram executados nos experimentos porque o objetivo deles era a avaliação das medidas em si. Além disso, os próprios autores atribuem os ótimos resultados por eles obtidos às medidas propostas.

Em resumo, o simples uso das medidas propostas recentemente não parece produzir os excelentes ganhos reportados no trabalho de referência. As razões para essa inconsistência poderiam ser investigadas em trabalhos futuros.

5 CONCLUSÕES E TRABALHOS FUTUROS

O principal motivo para a popularidade das RBMs nos dias atuais é o seu uso na construção de estruturas com múltiplas camadas, tais como as DBNs e DNNs, as quais têm relevância no ramo da aprendizagem de máquina denominado aprendizagem profunda. Nessas estruturas, RBMs são empilhadas e treinadas camada a camada de forma *não supervisionada*. Apesar disso, Larochelle e Bengio (2008) mostraram que as RBMs poderiam ser usadas como classificadores isolados, sendo, para tanto, treinadas de forma *supervisionada*, um uso que foi pouco explorado na literatura. Na presente dissertação, a versão de RBM proposta pelos referidos autores passou por uma adaptação e foi então aplicada ao problema de VAD.

Os tópicos à frente estão organizados da seguinte forma. Na Seção 5.1, os objetivos gerais deste trabalho, dentre os quais a adaptação do modelo, são revisitados e as conclusões a eles relacionadas são apresentadas. Em seguida, na Seção 5.2, são expostas algumas ideias para trabalhos futuros.

5.1 CONCLUSÕES

Na Seção 1.3, página 17, foram enumerados os objetivos da pesquisa, a partir dos quais vários experimentos foram elaborados e realizados. Ao longo da dissertação esses experimentos foram analisados isoladamente e agora os seus resultados são interpretados no contexto dos objetivos gerais do trabalho, que são listados a seguir.

1. **Adaptação do modelo a entradas contínuas**.
   A RBM com camada de classificação apresentada por Larochelle e Bengio (2008) foi adaptada para permitir o seu uso com dados de entrada definidos num domínio contínuo, dando origem à variante proposta. Na avaliação dessa variante no problema *double-moon*, conseguiu-se 100% de acurácia, independentemente to tipo de treinamento utilizado. Esse fato indica que a variante pode ser considerada como um classificador de uso geral, visto que a mesma é capaz de representar superfícies de decisão não triviais.

2. **Aplicação do modelo proposto ao problema de VAD empregando medidas clássicas**.
   A variante proposta foi avaliada em diferentes experimentos, com base nos quais são respondidas as seguintes perguntas:

   a) *Como um classificador baseado no modelo proposto lida com medidas não (ou pouco) correlacionadas, como é o caso dos MFCCs, e com medidas fortemente*

correlacionadas, como as FBEs?

Nas simulações, foram conseguidos desempenhos similares entre os detectores empregando (como vetores de características) MFCCs e aqueles utilizando FBEs. Essa situação foi verificada nos diferentes tipos de treinamento do classificador constituído pela variante proposta de RBM. Há, portanto, indícios para se concluir que o modelo proposto é capaz de lidar de forma adequada tanto com dados de entrada não correlacionados quanto com dados fortemente correlacionados. Essa observação reforça ainda a usabilidade da variante como um classificador de uso geral.

b) *Como um VAD baseado no modelo proposto se compara com outros VADs conhecidos na literatura*?

Considerando os *valores médios* (ao longo das SNRs) de acurácia e área ROC, os VADs baseados na variante proposta obtiveram desempenhos superiores ou comparáveis aos dos VADs tomados como referência. Além disso, dentre os diferentes tipos de treinamento do modelo, o discriminativo (que dá origem à DRBM) produziu os melhores resultados com o menor custo computacional e, por isso, recebeu um maior foco nos experimentos e comparações feitos neste trabalho. Sendo assim, com relação às DRBMs e aos VADs de referência, foram feitas as seguintes constatações:

i. As DRBMs são capazes de produzir detectores com desempenhos superiores ou similares, em todas as SNRs, aos dos VADs G.729-B, G.729-II, LTSD, Sohn e Ying.

ii. As DRBMs oferecem melhores acurácias e áreas ROC que o VAD Ghosh em SNRs elevadas (15 dB ou acima), enquanto este último consegue desempenhos sensivelmente melhores que as DRBMs (e todos os demais VADs) em SNRs mais baixas (10 dB ou abaixo).

iii. Para probabilidades de falso alarme suficientemente baixas, as DRBMs fornecem probabilidades de detecção maiores do que os VADs de referência, em todas as SNRs.

iv. Em razão do número relativamente pequeno de unidades ocultas que produziram seus melhores resultados, as DRBMs apresentam um baixo custo computacional. Por exemplo, a configuração do detector baseada em MFCCs se mostrou capaz de processar pelo menos 5 vezes mais áudio por unidade de tempo do que os VADs de estado da arte (Sohn, Ying e Ghosh).

Com base nessas observações, conclui-se que os detectores baseados em DRBM podem, em muitos casos, ser vantajosos em relação aos VADs de referência (inclusive aos considerados estado da arte) tanto em termos de desempenho quanto de custo computacional.

c) *Como um VAD baseado no modelo proposto se compara com um baseado em SVM (outro modelo empregado em aprendizagem de máquina)*?

Para comparação da DRBM com as SVMs, foram consideradas duas condições de treinamento, a saber:

i. **Condição única de SNR**: O detector resultante é usável em uma faixa estreita de SNRs. Nessa condição, a DRBM e as SVMs linear e com núcleo gaussiano ofereceram desempenhos (acurácia e área ROC) similares entre si, sendo que a DRBM exibiu um custo computacional inferior aos demais. Em particular, para o vetor de características baseado em MFCCs, o detector empregando DRBM foi capaz de processar 2,5 e 3,4 vezes mais áudio por unidade de tempo do que as SVMs linear e com núcleo gaussiano, respectivamente. Sendo assim, o uso da DRBM pode ser considerado vantajoso em comparação com as SVMs.

ii. **Condição múltipla de SNR**: O detector produzido é utilizável em uma ampla faixa de SNRs. Nessa condição de treinamento, ocorreu um afastamento entre os desempenhos dos modelos, com a DRBM conseguindo valores de acurácia e área ROC intermediários entre as duas versões de SVM. A SVM linear mostrou desempenho pobre e alto custo computacional, de modo que o seu treinamento em condição múltipla de SNR não seria de interesse. Já a SVM com núcleo gaussiano exibiu acurácia média 1,32% superior à da DRBM e, na SNR de $-5\,\mathrm{dB}$, a diferença nas acurácias foi de quase 3,5%. Em contrapartida, a DRBM conseguiu processar 20,2 vezes mais áudio por unidade de tempo do que a mesma SVM. Conclui-se, portanto, que esses dois modelos oferecem diferentes compromissos entre desempenho de classificação e custo computacional. Dependendo da importância dada a cada um desses fatores em uma certa aplicação, a escolha de um modelo ou do outro poderia ser mais interessante.

3. **Aplicação do modelo proposto ao problema de VAD empregando medidas propostas recentemente**.

Como parte da pesquisa, foram realizados testes empregando a DRBM com vetores de características compostos por diferentes combinações de medidas relativamente recentes, cujo uso em VAD foi proposto por Drugman et al. (2015). Os desempenhos conseguidos com o uso exclusivo de tais medidas foram inferiores àqueles obtidos com os tradicionais MFCCs. Além disso, combinações das medidas recentes com os MFCCs produziram ganhos pouco significativos para o VAD. Sendo assim, pode-se dizer que não há evidências para se concluir que o uso das medidas propostas seja especialmente interessante na aplicação de VAD.

5.2 TRABALHOS FUTUROS

Algumas ideias para trabalhos futuros são explicadas a seguir:

1. Os experimentos de VAD realizados neste trabalho se basearam no áudio contendo ruído veicular (*car noise*) do *corpus* NOIZEUS (HU; LOIZOU, 2008). Sendo assim, as conclusões do trabalho são, a princípio, válidas para esse tipo de ruído. Um trabalho futuro poderia investigar o uso da variante proposta com outros tipos de ruído.
2. Foi observado, nos experimentos, que o VAD Ghosh consegue ótimos resultados em SNRs baixas, enquanto que, em SNRs altas, apresenta desempenhos pobres em comparação com outros detectores avaliados. Sob a hipótese de que a vantagem do VAD Ghosh em SNRs baixas esteja no uso de informações de longo prazo do sinal, um trabalho futuro poderia tentar fundir informações de curto prazo, como MFCCs ou FBEs, com informações de longo prazo, como as entropias ao longo do tempo usadas pelo VAD Ghosh. Essa fusão poderia ser realizada naturalmente por meio de mecanismos de aprendizagem de máquina e teria como objetivo produzir um VAD que fosse superior em todas as SNRs.
3. Como os resultados relativos às medidas propostas recentemente não ofereceram uma conclusão definitiva sobre as mesmas, uma sugestão para trabalhos futuros seria reavaliá-las em outros experimentos. Na reavaliação, poderiam ser testados outros mecanismos de aprendizagem de máquina ou mesmo outras bases de dados de áudio.
4. Os resultados obtidos nos problemas *double-moon* e na detecção de atividade vocal indicam que a variante proposta pode ser considerada como um classificador de uso geral. Sendo assim, trabalhos futuros poderiam avaliá-la em outros tipos de problemas práticos de classificação.

REFERÊNCIAS BIBLIOGRÁFICAS

ACKLEY, D. H.; HINTON, G. E.; SEJNOWSKI, T. J. A learning algorithm for boltzmann machines. **Cognitive science**, Elsevier, v. 9, n. 1, p. 147–169, 1985.

BENGIO, Y. Learning deep architectures for ai. **Foundations and trends® in Machine Learning**, Now Publishers Inc., v. 2, n. 1, p. 1–127, 2009.

BENGIO, Y. et al. Greedy layer-wise training of deep networks. **Advances in neural information processing systems**, MIT; 1998, v. 19, p. 153, 2007.

BENYASSINE, A. et al. Itu-t recommendation g. 729 annex b: a silence compression scheme for use with g. 729 optimized for v. 70 digital simultaneous voice and data applications. **IEEE Communications Magazine**, IEEE, v. 35, n. 9, p. 64–73, 1997.

BISHOP, C. **Bishop Pattern Recognition and Machine Learning**. [S.l.]: Springer, New York, 2001.

BISHOP, C. M. **Pattern Recognition and Machine Learning**. [S.l.]: Springer Science+Business Media, LLC, 2006.

CASELLA, G.; BERGER, R. L. **Statistical inference**. 2. ed. [S.l.]: Duxbury Pacific Grove, CA, 2002.

CHANG, J.-H.; KIM, N. S.; MITRA, S. K. Voice activity detection based on multiple statistical models. **IEEE Transactions on Signal Processing**, IEEE, v. 54, n. 6, p. 1965–1976, 2006.

CHO, K.; ILIN, A.; RAIKO, T. Improved learning of gaussian-bernoulli restricted boltzmann machines. In: **Artificial Neural Networks and Machine Learning–ICANN 2011**. [S.l.]: Springer, 2011. p. 10–17.

DENG, L.; YU, D. Deep learning: Methods and applications. **Foundations and Trends in Signal Processing**, Now Publishers Inc., v. 7, n. 3–4, p. 197–387, 2014.

DESJARDINS, G. et al. Tempered markov chain monte carlo for training of restricted boltzmann machines. In: **International Conference on Artificial Intelligence and Statistics**. [S.l.: s.n.], 2010. p. 145–152.

DRUGMAN, T.; ALWAN, A. Joint robust voicing detection and pitch estimation based on residual harmonics. In: **Interspeech**. [S.l.: s.n.], 2011. p. 1973–1976.

DRUGMAN, T. et al. Voice activity detection: Merging source and filter-based information. **Signal Processing Letters, IEEE**, IEEE, v. 23, n. 2, p. 252–256, 2015.

EARL, D. J.; DEEM, M. W. Parallel tempering: Theory, applications, and new perspectives. **Physical Chemistry Chemical Physics**, Royal Society of Chemistry, v. 7, n. 23, p. 3910–3916, 2005.

FISCHER, A.; IGEL, C. Training restricted boltzmann machines: An introduction. **Pattern Recognition**, Elsevier, v. 47, n. 1, p. 25–39, 2014.

FREUND, Y.; HAUSSLER, D. **Unsupervised learning of distributions of binary vectors using two layer networks**. [S.l.]: Computer Research Laboratory [University of California, Santa Cruz], 1994.

FREUND, Y.; SCHAPIRE, R. E. A desicion-theoretic generalization of on-line learning and an application to boosting. In: SPRINGER. **Computational learning theory**. [S.l.], 1995. p. 23–37.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. **The elements of statistical learning**. [S.l.]: Springer series in statistics Springer, Berlin, 2001. v. 1.

GAZOR, S.; ZHANG, W. A soft voice activity detector based on a laplacian-gaussian model. **IEEE Transactions on Speech and Audio Processing**, IEEE, v. 11, n. 5, p. 498–505, 2003.

GELFAND, A. E.; SMITH, A. Sampling-based approaches to calculating marginal densities. **Journal of the American statistical association**, Taylor & Francis Group, v. 85, n. 410, p. 398–409, 1990.

GEMAN, S.; GEMAN, D. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, IEEE, n. 6, p. 721–741, 1984.

GHOSH, P. K.; TSIARTAS, A.; NARAYANAN, S. Robust voice activity detection using long-term signal variability. **IEEE Transactions on Audio, Speech, and Language Processing**, IEEE, v. 19, n. 3, p. 600–613, 2011.

HAIGH, J.; MASON, J. Robust voice activity detection using cepstral features. In: IEEE. **TENCON'93. Proceedings. IEEE Region 10 Conference on Computer, Communication, Control and Power Engineering.** [S.l.], 1993. v. 3, p. 321–324.

HASAN, M. R.; JAMIL, M.; RAHMAN, M. G. R. M. S. Speaker identification using mel frequency cepstral coefficients. **variations**, v. 1, p. 4, 2004.

HAYKIN, S. **Neural networks and learning machines**. 3. ed. [S.l.]: Pearson Education Upper Saddle River, 2009.

HILLENBRAND, J.; CLEVELAND, R. A.; ERICKSON, R. L. Acoustic correlates of breathy vocal quality. **Journal of Speech, Language, and Hearing Research**, ASHA, v. 37, n. 4, p. 769–778, 1994.

HINTON, G. A practical guide to training restricted boltzmann machines. **Momentum**, v. 9, n. 1, p. 926, 2010.

HINTON, G. et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. **Signal Processing Magazine, IEEE**, IEEE, v. 29, n. 6, p. 82–97, 2012.

HINTON, G. E. Training products of experts by minimizing contrastive divergence. **Neural computation**, MIT Press, v. 14, n. 8, p. 1771–1800, 2002.

HINTON, G. E.; OSINDERO, S.; TEH, Y.-W. A fast learning algorithm for deep belief nets. **Neural computation**, MIT Press, v. 18, n. 7, p. 1527–1554, 2006.

HINTON, G. E.; SEJNOWSKI, T. J. Parallel distributed processing: Explorations in the microstructure of cognition. In: RUMELHART, D. E.; MCCLELLAND, J. L.; GROUP, C. P. R. (Ed.). **Learning and Relearning in Boltzmann Machines**. Cambridge, MA, USA: MIT Press, 1986. v. 1, cap. 7, p. 282–317. ISBN 0-262-68053-X.

HIRSCH, H.-G.; PEARCE, D. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: **ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)**. [S.l.: s.n.], 2000.

HOPFIELD, J. J. Neural networks and physical systems with emergent collective computational abilities. **Proceedings of the national academy of sciences**, National Acad Sciences, v. 79, n. 8, p. 2554–2558, 1982.

HOYT, J. D.; WECHSLER, H. Detection of human speech in structured noise. In: IEEE. **IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP-94.** [S.l.], 1994. v. 2, p. II–237.

HU, Y.; LOIZOU, P. C. Evaluation of objective quality measures for speech enhancement. **IEEE Transactions on Audio, Speech, and Language Processing**, IEEE, v. 16, n. 1, p. 229–238, 2008.

ITTICHAICHAREON, C.; SUKSRI, S.; YINGTHAWORNSUK, T. Speech recognition using mfcc. In: **International Conference on Computer Graphics, Simulation and Modeling (ICGSM'2012) July**. [S.l.: s.n.], 2012. p. 28–29.

ITU. **ITU-T Recommendation P.56: Objective measurement of active speech level**. 1993.

ITU. **Appendix II – ITU-T G.729 Annex B enhancements in voice-over-IP applications – Option 1**. 2005.

JUNQUA, J.-C.; REAVES, B.; MAK, B. A study of endpoint detection algorithms in adverse conditions: incidence on a dtw and hmm recognizer. In: **Second European Conference on Speech Communication and Technology**. [S.l.: s.n.], 1991.

KAILATH, T. The divergence and bhattacharyya distance measures in signal selection. **IEEE transactions on communication technology**, IEEE, v. 15, n. 1, p. 52–60, 1967.

KAY, S. M. Fundamentals of statistical signal processing. **PTR Prentice-Hall, Englewood Cliffs, NJ**, 1993.

KENNY, P. **Notes on Boltzmann Machines**. [S.l.], 2011.

LANDAU, L. D.; LIFSHITZ, E. M. **Statistical Physics**. 3. ed. [S.l.]: Pergamon, 1980. v. 5.

LAROCHELLE, H. **Restricted Boltzmann machine - definition**. 2013. Disponível em: <https://youtu.be/p4Vh_zMw-HQ>.

LAROCHELLE, H.; BENGIO, Y. Classification using discriminative restricted boltzmann machines. In: ACM. **Proceedings of the 25th international conf. on Machine learning**. [S.l.], 2008. p. 536–543.

LI, K.; SWAMY, M.; AHMAD, M. O. An improved voice activity detection using higher order statistics. **IEEE Transactions on Speech and Audio Processing**, IEEE, v. 13, n. 5, p. 965–974, 2005.

MARTIN, A.; DAMNATI, G.; MAUUARY, L. Robust speech/non-speech detection using lda applied to mfcc for continuous speech recognition. In: **INTERSPEECH**. [S.l.: s.n.], 2001. p. 885–888.

MEYER, C. D. **Matrix analysis and applied linear algebra**. [S.l.]: Siam, 2000. v. 2.

MUDA, L.; BEGAM, M.; ELAMVAZUTHI, I. Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques. **arXiv preprint arXiv:1003.4083**, 2010.

NAKAGAWA, S.; ASAKAWA, K.; WANG, L. Speaker recognition by combining mfcc and phase information. **spectrum**, v. 60, n. 700Hz, p. 76–4, 2007.

NEAL, R. M. Connectionist learning of belief networks. **Artificial intelligence**, Elsevier, v. 56, n. 1, p. 71–113, 1992.

NEMER, E.; GOUBRAN, R.; MAHMOUD, S. Robust voice activity detection using higher-order statistics in the lpc residual domain. **IEEE Transactions on Speech and Audio Processing**, IEEE, v. 9, n. 3, p. 217–231, 2001.

NORTON, M. P.; KARCZUB, D. G. **Fundamentals of noise and vibration analysis for engineers**. [S.l.]: Cambridge university press, 2003.

OPPENHEIM, A. V.; SCHAFER, R. W. **Discrete-time signal processing**. [S.l.]: Pearson Higher Education, 2010.

PAPA, J. P. et al. Model selection for discriminative restricted boltzmann machines through meta-heuristic techniques. **Journal of Computational Science**, Elsevier, v. 9, p. 14–18, 2015.

PAPOULIS, A.; PILLAI, S. U. **Probability, random variables, and stochastic processes**. [S.l.]: Tata McGraw-Hill Education, 2002.

PHUNG, D. Q.; VENKATESH, S. et al. Ordinal boltzmann machines for collaborative filtering. In: AUAI PRESS. **Proceedings of the Twenty-fifth Conference on Uncertainty in Artificial Intelligence**. [S.l.], 2009. p. 548–556.

RABINER, L.; SAMBUR, M. Voiced-unvoiced-silence detection using the itakura lpc distance measure. In: IEEE. **IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'77.** [S.l.], 1977. v. 2, p. 323–326.

RAMIREZ, J. et al. Efficient voice activity detection algorithms using long-term speech information. **Speech communication**, Elsevier, v. 42, n. 3, p. 271–287, 2004.

RENEVEY, P.; DRYGAJLO, A. Entropy based voice activity detection in very noisy conditions. **threshold**, v. 5, n. 5.5, p. 6, 2001.

ROSALES, R.; SCLAROFF, S. Combining generative and discriminative models in a framework for articulated pose estimation. **International Journal of Computer Vision**, Springer, v. 67, n. 3, p. 251–276, 2006.

ROUX, N. L.; BENGIO, Y. Representational power of restricted boltzmann machines and deep belief networks. **Neural computation**, MIT Press, v. 20, n. 6, p. 1631–1649, 2008.

SADJADI, S. O.; HANSEN, J. H. Unsupervised speech activity detection using voicing measures and perceptual spectral flux. **Signal Processing Letters, IEEE**, IEEE, v. 20, n. 3, p. 197–200, 2013.

SALAKHUTDINOV, R.; MNIH, A.; HINTON, G. Restricted boltzmann machines for collaborative filtering. In: ACM. **Proceedings of the 24th international conference on Machine learning**. [S.l.], 2007. p. 791–798.

SALAKHUTDINOV, R.; MURRAY, I. On the quantitative analysis of deep belief networks. In: ACM. **Proceedings of the 25th international conference on Machine learning**. [S.l.], 2008. p. 872–879.

SMOLENSKY, P. Parallel distributed processing: Explorations in the microstructure of cognition. In: RUMELHART, D. E.; MCCLELLAND, J. L.; GROUP, C. P. R. (Ed.). **Information Processing in Dynamical Systems: Foundations of Harmony Theory**. Cambridge, MA, USA: MIT Press, 1986. v. 1, cap. 6, p. 194–281. ISBN 0-262-68053-X.

SOHN, J.; KIM, N. S.; SUNG, W. A statistical model-based voice activity detection. **IEEE Signal Processing Letters**, IEEE, v. 6, n. 1, p. 1–3, 1999.

SWENDSEN, R. H.; WANG, J.-S. Nonuniversal critical dynamics in monte carlo simulations. **Physical review letters**, APS, v. 58, n. 2, p. 86, 1987.

TIELEMAN, T. Training restricted boltzmann machines using approximations to the likelihood gradient. In: ACM. **Proceedings of the 25th international conference on Machine learning**. [S.l.], 2008. p. 1064–1071.

TUCKER, R. Voice activity detection using a periodicity measure. In: IET. **IEE Proceedings on Communications, Speech and Vision**. [S.l.], 1992. v. 139, n. 4, p. 377–380.

WANG, T.; SILVER, D. L. Learning paired-associate images with an unsupervised deep learning architecture. In: SPRINGER. **Canadian Conference on Artificial Intelligence**. [S.l.], 2015. p. 250–263.

WELLING, M.; ROSEN-ZVI, M.; HINTON, G. E. Exponential family harmoniums with an application to information retrieval. In: **Advances in neural information processing systems**. [S.l.: s.n.], 2004. p. 1481–1488.

WU, Z.; CAO, Z. Improved mfcc-based feature for robust speaker identification. **Tsinghua Science & Technology**, Elsevier, v. 10, n. 2, p. 158–161, 2005.

YING, D. et al. Voice activity detection based on an unsupervised learning framework. **IEEE Transactions on Audio, Speech, and Language Processing**, IEEE, v. 19, n. 8, p. 2624–2633, 2011.

YOUNG, S. et al. The htk book. **Cambridge university engineering department**, v. 3, p. 175, 2002.

ZHENG, F.; ZHANG, G. Integrating the energy information into mfcc. In: **INTERSPEECH**. [S.l.: s.n.], 2000. p. 389–392.

ZOU, Y. et al. Improved voice activity detection based on support vector machine with high separable speech feature vectors. In: IEEE. **19th International Conference on Digital Signal Processing (DSP)**. [S.l.], 2014. p. 763–767.

ZULFIQAR, A.; MUHAMMAD, A.; ENRIQUEZ, A. M. A speaker identification system using mfcc features with vq technique. In: IEEE. **Intelligent Information Technology Application, 2009. IITA 2009. Third International Symposium on**. [S.l.], 2009. v. 3, p. 115–118.

ZWEIG, G.; RUSSELL, S. Speech recognition with dynamic bayesian networks. In: CITESEER. **Fifteenth National Conference on Artificial Intelligence (AAAI'98)**. [S.l.], 1998. p. 173–180.

APÊNDICE A – EQUAÇÕES RELATIVAS ÀS RBMS BERNOULLI-BERNOULLI

A.1 EXPRESSÃO EFICIENTE PARA A ENERGIA LIVRE

Expandindo-se $E(\mathbf{v}, \mathbf{h})$, como definida em (2.8), em (2.11), obtém-se

$$
\begin{aligned}
P(\mathbf{v}) &= \frac{1}{Z} \sum_{\mathbf{h}\in\{0,1\}^{n_h}} \exp\left(\sum_{j=1}^{n_h} b_j h_j + \sum_{i=1}^{n_v} c_i v_i + \sum_{i=1}^{n_v}\sum_{j=1}^{n_h} h_j w_{ji} v_i\right) \\
&\overset{(a)}{=} \frac{1}{Z} \exp\left(\sum_{i=1}^{n_v} c_i v_i\right) \sum_{h_1=0}^{1} \cdots \sum_{h_{n_h}=0}^{1} \exp\left(\sum_{j=1}^{n_h} b_j h_j + \sum_{j=1}^{n_h}\sum_{i=1}^{n_v} h_j w_{ji} v_i\right) \\
&\overset{(b)}{=} \frac{1}{Z} \exp\left(\sum_{i=1}^{n_v} c_i v_i\right) \sum_{h_1=0}^{1} \cdots \sum_{h_{n_h}=0}^{1} \prod_{j=1}^{n_h} \exp\left(b_j h_j + \sum_{i=1}^{n_v} h_j w_{ji} v_i\right) \\
&\overset{(c)}{=} \frac{1}{Z} \exp\left(\sum_{i=1}^{n_v} c_i v_i\right) \prod_{j=1}^{n_h} \left[\sum_{h_j=0}^{1} \exp\left(b_j h_j + \sum_{i=1}^{n_v} h_j w_{ji} v_i\right)\right] \\
&\overset{(d)}{=} \frac{1}{Z} \exp\left(\sum_{i=1}^{n_v} c_i v_i\right) \prod_{j=1}^{n_h} \left[1 + \exp\left(b_j + \sum_{i=1}^{n_v} w_{ji} v_i\right)\right] \\
&\overset{(e)}{=} \frac{1}{Z} \exp\left(\sum_{i=1}^{n_v} c_i v_i\right) \prod_{j=1}^{n_h} \left[\exp\left(\ln\left(1 + \exp\left(b_j + \sum_{i=1}^{n_v} w_{ji} v_i\right)\right)\right)\right] \\
&\overset{(f)}{=} \frac{1}{Z} \exp\left[\sum_{i=1}^{n_v} c_i v_i + \sum_{j=1}^{n_h} \ln\left(1 + \exp\left(b_j + \sum_{i=1}^{n_v} w_{ji} v_i\right)\right)\right].
\end{aligned}
\tag{A.1}
$$

Comparando-se a última expressão obtida para $P(\mathbf{v})$ com aquela apresentada na Equação (2.12), chega-se que

$$
\mathcal{F}(\mathbf{v}) = -\sum_{i=1}^{n_v} c_i v_i - \sum_{j=1}^{n_h} \ln\left(1 + \exp\left(b_j + \sum_{i=1}^{n_v} w_{ji} v_i\right)\right). \tag{A.2}
$$

Aqui tem-se uma soma contendo um número de termos que é linear na quantidade de variáveis ocultas, sendo, portanto, considerada tratável, em oposição àquela apresentada na Equação (2.13). As passagens em (A.1) são comentadas aqui:

(a) O somatório envolvendo apenas as variáveis visíveis no interior da função $\exp(\cdot)$ é constante com relação ao somatório mais externo (aquele envolvendo as variáveis ocultas) sendo movido para fora do mesmo;

(b) Converte-se o exponencial de um somatório em um produtório de exponenciais;

(c) Notando-se que cada elemento do produtório somente depende de uma variável oculta, utiliza-se do seguinte fato

$$\begin{aligned}\sum_{h_1}\cdots\sum_{h_{n_h}}\prod_{j=1}^{n_h}\phi(h_j) &= \sum_{h_1}\cdots\sum_{h_{n_h}}\phi(h_1)\cdots\phi(h_{n_h})\\ &= \sum_{h_1}\phi(h_1)\cdots\sum_{h_{n_h}}\phi(h_{n_h}) = \prod_{j=1}^{n_h}\left(\sum_{h_j}\phi(h_j)\right);\end{aligned} \tag{A.3}$$

(d) Avalia-se o somatório envolvendo os dois valores possíveis para h_j;

(e) Usa-se o fato de que $x = \exp(\ln(x))$ para reescrever os elementos do produtório;

(f) Unem-se o primeiro exponencial com aqueles do produtório, de modo a obter um único exponencial (no nível mais externo da expressão).

A Equação (A.2) pode ser escrita de forma mais sucinta usando-se a função denominada *softplus* definida como $\zeta(x) = \ln(1 + \exp(x))$:

$$\mathcal{F}(\mathbf{v}) = -\sum_{i=1}^{n_v} c_i v_i - \sum_{j=1}^{n_h}\zeta\left(b_j + \sum_{i=1}^{n_v} w_{ji} v_i\right). \tag{A.4}$$

Deve-se mencionar que $\zeta(x)$ é a primitiva de $\varphi(x)$, ou seja, $d\zeta(x)/dx = \varphi(x)$, algo que será útil em derivações posteriores. Por fim, considerando-se que a aplicação de uma função escalar a um vetor corresponde a aplicar essa função a cada componente do vetor, pode-se reescrever (A.4) como

$$\mathcal{F}(\mathbf{v}) = -\mathbf{c}^T\mathbf{v} - \sum_{j=1}^{n_h}\zeta(b_j + \mathbf{W}_{j*}\mathbf{v}), \tag{A.5}$$

sendo que a notação $\mathbf{W}_{j*}$ representa a j-ésima linha da matriz $\mathbf{W}$.

A.2 DISTRIBUIÇÕES CONDICIONAIS

A.2.1 Derivação de $P(\mathbf{v}|\mathbf{h})$

$$
\begin{aligned}
P(\mathbf{v}|\mathbf{h}) = \frac{P(\mathbf{v},\mathbf{h})}{P(\mathbf{h})} &= \frac{P(\mathbf{v},\mathbf{h})}{\sum\limits_{\mathbf{v}'\in\{0,1\}^{n_v}} P(\mathbf{v}',\mathbf{h})} = \frac{\exp(-E(\mathbf{v},\mathbf{h})/Z}{\sum\limits_{\mathbf{v}'\in\{0,1\}^{n_v}} \exp(-E(\mathbf{v}',\mathbf{h})/Z} \\
&\overset{(a)}{=} \frac{\exp\left(\sum\limits_{j=1}^{n_h} b_j h_j + \sum\limits_{i=1}^{n_v} c_i v_i + \sum\limits_{i=1}^{n_v}\sum\limits_{j=1}^{n_h} h_j w_{ji} v_i\right)/Z}{\sum\limits_{v'_1=0}^{1}\cdots\sum\limits_{v'_{n_v}=0}^{1} \exp\left(\sum\limits_{j=1}^{n_h} b_j h_j + \sum\limits_{i=1}^{n_v} c_i v'_i + \sum\limits_{i=1}^{n_v}\sum\limits_{j=1}^{n_h} h_j w_{ji} v'_i\right)/Z} \\
&\overset{(b)}{=} \frac{\exp\left(\sum\limits_{i=1}^{n_v} c_i v_i + \sum\limits_{i=1}^{n_v}\sum\limits_{j=1}^{n_h} h_j w_{ji} v_i\right)}{\sum\limits_{v'_1=0}^{1}\cdots\sum\limits_{v'_{n_v}=0}^{1} \exp\left(\sum\limits_{i=1}^{n_v} c_i v'_i + \sum\limits_{i=1}^{n_v}\sum\limits_{j=1}^{n_h} h_j w_{ji} v'_i\right)} \\
&\overset{(c)}{=} \frac{\prod\limits_{i=1}^{n_v} \exp\left(c_i v_i + \sum\limits_{j=1}^{n_h} h_j w_{ji} v_i\right)}{\sum\limits_{v'_1=0}^{1}\cdots\sum\limits_{v'_{n_v}=0}^{1} \prod\limits_{i=1}^{n_v} \exp\left(c_i v'_i + \sum\limits_{j=1}^{n_h} h_j w_{ji} v'_i\right)} \\
&\overset{(d)}{=} \frac{\prod\limits_{i=1}^{n_v} \exp\left(c_i v_i + \sum\limits_{j=1}^{n_h} h_j w_{ji} v_i\right)}{\prod\limits_{i=1}^{n_v}\left(\sum\limits_{v'_i=0}^{1} \exp\left(c_i v'_i + \sum\limits_{j=1}^{n_h} h_j w_{ji} v'_i\right)\right)} \\
&\overset{(e)}{=} \prod_{i=1}^{n_v} \frac{\exp\left(c_i v_i + \sum\limits_{j=1}^{n_h} h_j w_{ji} v_i\right)}{\left(1+\exp\left(c_i + \sum\limits_{j=1}^{n_h} h_j w_{ji}\right)\right)}
\end{aligned}
\tag{A.6}
$$

As passagens em (A.6) são comentadas aqui:

(a) Substitui-se a Equação (2.8) na definição de probabilidade condicional. No denominador, usa-se um sobrescrito para diferenciar a variável do somatório ($\mathbf{v}'$) da variável livre ($\mathbf{v}$) do numerador;

(b) Eliminam-se fatores comuns do numerador e do denominador;

(c) Converte-se o exponencial de um somatório em um produtório de exponenciais tanto no numerador quanto no denominador;

(d) Usa-se resultado análogo ao obtido na Equação (A.3), agora para as variáveis visíveis;

(e) Coloca-se ambos numerador e denominador dentro de um mesmo produtório. Deve-se notar que cada um dos fatores desse produtório depende somente de uma variável visível. Mais ainda, cada fator é, por si só, uma função massa de probabilidade (i.e., é não negativo e o somatório ao longo do seu domínio é 1). Conclui-se, portanto, que

$$P(v_i|\mathbf{h}) = \frac{\exp\left(c_i v_i + \sum_{j=1}^{n_h} h_j w_{ji} v_i\right)}{1 + \exp\left(c_i + \sum_{j=1}^{n_h} h_j w_{ji}\right)}, \tag{A.7}$$

e

$$P(\mathbf{v}|\mathbf{h}) = \prod_{i=1}^{n_v} P(v_i|\mathbf{h}). \tag{A.8}$$

Esse resultado implica a independência das variáveis visíveis quando fixado o valor das variáveis ocultas. Fazendo-se v_i=1 na Equação (A.7), chega-se à probabilidade de ativação de uma unidade visível dado $\mathbf{h}$, isto é,

$$\begin{aligned} P(v_i{=}1|\mathbf{h}) &= \Pr(V_i{=}1|\mathbf{H}{=}\mathbf{h}) \\ &= \frac{\exp\left(c_i + \sum_{j=1}^{n_h} h_j w_{ji}\right)}{1 + \exp\left(c_i + \sum_{j=1}^{n_h} h_j w_{ji}\right)} = \varphi\left(c_i + \sum_{j=1}^{n_h} w_{ji} h_j\right). \end{aligned} \tag{A.9}$$

A.2.2 Derivação de $P(\mathbf{h}|\mathbf{v})$

Dada a simetria do modelo (e das equações subjacentes), resultados análogos aos da seção anterior podem ser obtidos por simples substituições de nomes. Assim, as derivações não serão aqui repetidas.

A.3 GRADIENTE DA FUNÇÃO PERDA

De (2.22) e (2.12) tem-se a seguinte expressão para a função perda:

$$\mathcal{L}(\boldsymbol{\theta}; \mathbf{v}^{(t)}) = \mathcal{F}(\mathbf{v}^{(t)}) + \ln Z, \tag{A.10}$$

sendo deixada implícita a dependência de $\mathcal{F}(\cdot)$ e Z com relação aos parâmetros do modelo. Diferenciado-se (A.10) com relação a $\boldsymbol{\theta}$ e tendo em vista que $Z = \sum_{\mathbf{v}\in\{0,1\}^{n_v}} \exp(-\mathcal{F}(\mathbf{v}))$, obtém-se

$$\begin{aligned}
\frac{\partial \mathcal{L}(\boldsymbol{\theta}; \mathbf{v}^{(t)})}{\partial \boldsymbol{\theta}} &= \frac{\partial \mathcal{F}(\mathbf{v}^{(t)})}{\partial \boldsymbol{\theta}} + \frac{1}{Z}\frac{\partial Z}{\partial \boldsymbol{\theta}} \\
&= \frac{\partial \mathcal{F}(\mathbf{v}^{(t)})}{\partial \boldsymbol{\theta}} - \frac{1}{Z}\sum_{\mathbf{v}\in\{0,1\}^{n_v}} \exp(-\mathcal{F}(\mathbf{v}))\frac{\partial \mathcal{F}(\mathbf{v})}{\partial \boldsymbol{\theta}} \\
&= \frac{\partial \mathcal{F}(\mathbf{v}^{(t)})}{\partial \boldsymbol{\theta}} - \sum_{\mathbf{v}\in\{0,1\}^{n_v}} P(\mathbf{v})\frac{\partial \mathcal{F}(\mathbf{v})}{\partial \boldsymbol{\theta}} \\
&= \frac{\partial \mathcal{F}(\mathbf{v}^{(t)})}{\partial \boldsymbol{\theta}} - \mathbb{E}_{P(\mathbf{v})}\left[\frac{\partial \mathcal{F}(\mathbf{V})}{\partial \boldsymbol{\theta}}\right],
\end{aligned} \tag{A.11}$$

sendo que, na última passagem, utilizou-se a definição de esperança de uma transformação de variável aleatória. Fica, assim, demonstrado o resultado apresentado na Equação (2.23).

É útil escrever o gradiente da função de energia livre (com relação ao vetor de parâmetros) em termos da função de energia global. Diferenciando-se (2.13), escreve-se

$$\begin{aligned}
\frac{\partial \mathcal{F}(\mathbf{v})}{\partial \boldsymbol{\theta}} &= -\frac{1}{\sum\limits_{\mathbf{h}'\in\{0,1\}^{n_h}} \exp(-E(\mathbf{v},\mathbf{h}'))}\sum_{\mathbf{h}\in\{0,1\}^{n_h}} \exp(-E(\mathbf{v},\mathbf{h}))\left(-\frac{\partial E(\mathbf{v},\mathbf{h})}{\partial \boldsymbol{\theta}}\right) \\
&= \sum_{\mathbf{h}\in\{0,1\}^{n_h}} \frac{\exp(-E(\mathbf{v},\mathbf{h}))}{\sum\limits_{\mathbf{h}'\in\{0,1\}^{n_h}} \exp(-E(\mathbf{v},\mathbf{h}'))}\left(\frac{\partial E(\mathbf{v},\mathbf{h})}{\partial \boldsymbol{\theta}}\right) \\
&= \sum_{\mathbf{h}\in\{0,1\}^{n_h}} P(\mathbf{h}|\mathbf{v})\frac{\partial E(\mathbf{v},\mathbf{h})}{\partial \boldsymbol{\theta}} \\
&= \mathbb{E}_{P(\mathbf{h}|\mathbf{v})}\left[\frac{\partial E(\mathbf{v},\mathbf{H})}{\partial \boldsymbol{\theta}}\middle|\mathbf{v}\right].
\end{aligned} \tag{A.12}$$

De (A.11) e (A.12), demonstra-se a expressão na Equação (2.24):

$$\begin{aligned}
\frac{\partial \mathcal{L}(\boldsymbol{\theta}; \mathbf{v}^{(t)})}{\partial \boldsymbol{\theta}} &= \mathbb{E}_{P(\mathbf{h}|\mathbf{v})}\left[\frac{\partial E(\mathbf{v}^{(t)},\mathbf{H})}{\partial \boldsymbol{\theta}}\middle|\mathbf{v}^{(t)}\right] - \mathbb{E}_{P(\mathbf{v})}\left[\mathbb{E}_{P(\mathbf{h}|\mathbf{v})}\left[\frac{\partial E(\mathbf{V},\mathbf{H})}{\partial \boldsymbol{\theta}}\middle|\mathbf{V}\right]\right] \\
&= \mathbb{E}_{P(\mathbf{h}|\mathbf{v})}\left[\frac{\partial E(\mathbf{v}^{(t)},\mathbf{H})}{\partial \boldsymbol{\theta}}\middle|\mathbf{v}^{(t)}\right] - \sum_{\mathbf{v}\in\{0,1\}^{n_v}} P(\mathbf{v})\sum_{\mathbf{h}\in\{0,1\}^{n_h}} P(\mathbf{h}|\mathbf{v})\frac{\partial E(\mathbf{v},\mathbf{h})}{\partial \boldsymbol{\theta}} \\
&= \mathbb{E}_{P(\mathbf{h}|\mathbf{v})}\left[\frac{\partial E(\mathbf{v}^{(t)},\mathbf{H})}{\partial \boldsymbol{\theta}}\middle|\mathbf{v}^{(t)}\right] - \sum_{\mathbf{v}\in\{0,1\}^{n_v}}\sum_{\mathbf{h}\in\{0,1\}^{n_h}} P(\mathbf{v},\mathbf{h})\frac{\partial E(\mathbf{v},\mathbf{h})}{\partial \boldsymbol{\theta}} \\
&= \mathbb{E}_{P(\mathbf{h}|\mathbf{v})}\left[\frac{\partial E(\mathbf{v}^{(t)},\mathbf{H})}{\partial \boldsymbol{\theta}}\middle|\mathbf{v}^{(t)}\right] - \mathbb{E}_{P(\mathbf{v},\mathbf{h})}\left[\frac{\partial E(\mathbf{V},\mathbf{H})}{\partial \boldsymbol{\theta}}\right].
\end{aligned} \tag{A.13}$$

A.4 REGRAS DE ATUALIZAÇÃO DE PARÂMETROS

As regras de atualização de parâmetros podem ser obtidas diferenciado-se a expressão para $\mathcal{F}(\mathbf{v})$ apresentada na Equação (A.4) com relação aos vários parâmetros do modelo, ou seja, determinando-se as componentes de $\partial\mathcal{F}(\mathbf{v})/\partial\boldsymbol{\theta}$. Primeiramente, para os vieses das unidades ocultas, deriva-se

$$\begin{aligned}\frac{\partial\mathcal{F}(\mathbf{v})}{\partial b_j} &= \frac{\partial}{b_j}\left[-\sum_{i=1}^{n_v} c_i v_i - \sum_{j=1}^{n_h}\zeta\left(b_j + \sum_{i=1}^{n_v} w_{ji}v_i\right)\right] \\ &= -\varphi\left(b_j + \sum_{i=1}^{n_v} w_{ji}v_i\right) \\ &= -P(h_j{=}1|\mathbf{v}),\end{aligned} \tag{A.14}$$

e, para os vieses das unidades visíveis,

$$\begin{aligned}\frac{\partial\mathcal{F}(\mathbf{v})}{\partial c_i} &= \frac{\partial}{c_i}\left[-\sum_{i=1}^{n_v} c_i v_i - \sum_{j=1}^{n_h}\zeta\left(b_j + \sum_{i=1}^{n_v} w_{ji}v_i\right)\right] \\ &= -v_i.\end{aligned} \tag{A.15}$$

Finalmente, para os pesos de conexão,

$$\begin{aligned}\frac{\partial\mathcal{F}(\mathbf{v})}{\partial w_{ji}} &= \frac{\partial}{w_{ji}}\left[-\sum_{i=1}^{n_v} c_i v_i - \sum_{j=1}^{n_h}\zeta\left(b_j + \sum_{i=1}^{n_v} w_{ji}v_i\right)\right] \\ &= -\varphi\left(b_j + \sum_{i=1}^{n_v} w_{ji}v_i\right) v_i \\ &= -P(h_j{=}1|\mathbf{v})\, v_i.\end{aligned} \tag{A.16}$$

Usam-se, então, esses resultados na versão escalar de (2.28), dada por

$$\theta \leftarrow \theta - \lambda\left(\frac{\partial\mathcal{F}(\mathbf{v}^{(t)})}{\partial\theta} - \frac{\partial\mathcal{F}(\widetilde{\mathbf{v}})}{\partial\theta}\right), \tag{A.17}$$

em que θ é alguma componente do vetor $\boldsymbol{\theta}$. Fazendo-se as substituições mencionadas, chega-se às Equações (2.29), (2.30) e (2.31).

APÊNDICE B – EQUAÇÕES RELATIVAS ÀS RBMS GAUSS-BERNOULLI

B.1 DISTRIBUIÇÃO MARGINAL DAS VARIÁVEIS VISÍVEIS

Como nas RBMs, a distribuição marginal das variáveis visíveis é dada por

$$P_g(\mathbf{v}) = \sum_{\mathbf{h}\in\{0,1\}^{n_h}} \frac{\exp(-E_g(\mathbf{v},\mathbf{h}))}{Z_g}. \tag{B.1}$$

B.2 EXPRESSÃO EFICIENTE PARA A ENERGIA LIVRE

A energia livre, $\mathcal{F}_g(\mathbf{v})$, é definida de modo a satisfazer a igualdade

$$P_g(\mathbf{v}) = \frac{\exp(-\mathcal{F}_g(\mathbf{v}))}{Z_g}. \tag{B.2}$$

Expandindo-se $E_g(\mathbf{v},\mathbf{h})$ conforme (2.36) em (B.1), obtém-se

$$\begin{aligned}
P_g(\mathbf{v}) &= \frac{1}{Z_g} \sum_{\mathbf{h}\in\{0,1\}^{n_h}} \exp\left(\sum_{j=1}^{n_h} b_j h_j - \sum_{i=1}^{n_v} \frac{(v_i-c_i)^2}{2\sigma_i^2} + \sum_{i=1}^{n_v}\sum_{j=1}^{n_h} h_j w_{ji} \frac{v_i}{\sigma_i^2}\right) \\
&= \frac{1}{Z_g} \exp\left(-\sum_{i=1}^{n_v} \frac{(v_i-c_i)^2}{2\sigma_i^2}\right) \sum_{h_1=0}^{1} \cdots \sum_{h_{n_h}=0}^{1} \exp\left(\sum_{j=1}^{n_h} b_j h_j + \sum_{j=1}^{n_h}\sum_{i=1}^{n_v} h_j w_{ji} \frac{v_i}{\sigma_i^2}\right) \\
&= \frac{1}{Z_g} \exp\left(-\sum_{i=1}^{n_v} \frac{(v_i-c_i)^2}{2\sigma_i^2}\right) \sum_{h_1=0}^{1} \cdots \sum_{h_{n_h}=0}^{1} \prod_{j=1}^{n_h} \exp\left(b_j h_j + \sum_{i=1}^{n_v} h_j w_{ji} \frac{v_i}{\sigma_i^2}\right) \\
&= \frac{1}{Z_g} \exp\left(-\sum_{i=1}^{n_v} \frac{(v_i-c_i)^2}{2\sigma_i^2}\right) \prod_{j=1}^{n_h} \left[\sum_{h_j=0}^{1} \exp\left(b_j h_j + \sum_{i=1}^{n_v} h_j w_{ji} \frac{v_i}{\sigma_i^2}\right)\right] \\
&= \frac{1}{Z_g} \exp\left(-\sum_{i=1}^{n_v} \frac{(v_i-c_i)^2}{2\sigma_i^2}\right) \prod_{j=1}^{n_h} \left[1 + \exp\left(b_j + \sum_{i=1}^{n_v} w_{ji} \frac{v_i}{\sigma_i^2}\right)\right] \\
&= \frac{1}{Z_g} \exp\left(-\sum_{i=1}^{n_v} \frac{(v_i-c_i)^2}{2\sigma_i^2}\right) \prod_{j=1}^{n_h} \left[\exp\left(\ln\left(1 + \exp\left(b_j + \sum_{i=1}^{n_v} w_{ji} \frac{v_i}{\sigma_i^2}\right)\right)\right)\right] \\
&= \frac{1}{Z_g} \exp\left[-\sum_{i=1}^{n_v} \frac{(v_i-c_i)^2}{2\sigma_i^2} + \sum_{j=1}^{n_h} \ln\left(1 + \exp\left(b_j + \sum_{i=1}^{n_v} w_{ji} \frac{v_i}{\sigma_i^2}\right)\right)\right].
\end{aligned} \tag{B.3}$$

As passagens são análogas àquelas apresentadas e explicadas para as RBMs na Seção A.1. Comparando-se a última expressão obtida para $P_g(\mathbf{v})$ com aquela apresentada na Equação (B.2), chega-se que

$$\mathcal{F}_g(\mathbf{v}) = \sum_{i=1}^{n_v} \frac{(v_i-c_i)^2}{2\sigma_i^2} - \sum_{j=1}^{n_h} \ln\left(1 + \exp\left(b_j + \sum_{i=1}^{n_v} w_{ji} \frac{v_i}{\sigma_i^2}\right)\right). \tag{B.4}$$

Assim como fora feito para RBMs, pode-se escrever a Equação (B.4) de forma mais sucinta usando-se a função *softplus* (indicada por $\zeta(\cdot)$), resultando em

$$\mathcal{F}_g(\mathbf{v}) = \sum_{i=1}^{n_v} \frac{(v_i - c_i)^2}{2\sigma_i^2} - \sum_{j=1}^{n_h} \zeta\left(b_j + \sum_{i=1}^{n_v} w_{ji} \frac{v_i}{\sigma_i^2}\right). \tag{B.5}$$

B.3 DISTRIBUIÇÕES CONDICIONAIS

B.3.1 Derivação de $P_g(\mathbf{v}|\mathbf{h})$

$$\begin{aligned}
P_g(\mathbf{v}|\mathbf{h}) &= \frac{P_g(\mathbf{v},\mathbf{h})}{P_g(\mathbf{h})} = \frac{P_g(\mathbf{v},\mathbf{h})}{\int P_g(\mathbf{v}',\mathbf{h})\,\mathbf{dv}'} = \frac{\exp(-E_g(\mathbf{v},\mathbf{h}))/Z_g}{\int \exp(-E_g(\mathbf{v}',\mathbf{h}))/Z_g\,\mathbf{dv}'} \\
&= \frac{\exp\left(\sum_{j=1}^{n_h} b_j h_j - \sum_{i=1}^{n_v} \frac{(v_i - c_i)^2}{2\sigma_i^2} + \sum_{i=1}^{n_v}\sum_{j=1}^{n_h} h_j w_{ji} \frac{v_i}{\sigma_i^2}\right)}{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \exp\left(\sum_{j=1}^{n_h} b_j h_j - \sum_{i=1}^{n_v} \frac{(v_i' - c_i)^2}{2\sigma_i^2} + \sum_{i=1}^{n_v}\sum_{j=1}^{n_h} h_j w_{ji} \frac{v_i'}{\sigma_i^2}\right) dv_1'\cdots dv_{n_v}'} \\
&= \frac{\exp\left(-\sum_{i=1}^{n_v} \frac{(v_i - c_i)^2}{2\sigma_i^2} + \sum_{i=1}^{n_v}\sum_{j=1}^{n_h} h_j w_{ji} \frac{v_i}{\sigma_i^2}\right)}{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \exp\left(-\sum_{i=1}^{n_v} \frac{(v_i' - c_i)^2}{2\sigma_i^2} + \sum_{i=1}^{n_v}\sum_{j=1}^{n_h} h_j w_{ji} \frac{v_i'}{\sigma_i^2}\right) dv_1'\cdots dv_{n_v}'} \\
&= \frac{\prod_{i=1}^{n_v} \exp\left(-\frac{(v_i - c_i)^2}{2\sigma_i^2} + \sum_{j=1}^{n_h} h_j w_{ji} \frac{v_i}{\sigma_i^2}\right)}{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \prod_{i=1}^{n_v} \exp\left(-\frac{(v_i' - c_i)^2}{2\sigma_i^2} + \sum_{j=1}^{n_h} h_j w_{ji} \frac{v_i'}{\sigma_i^2}\right) dv_1'\cdots dv_{n_v}'} \\
&= \frac{\prod_{i=1}^{n_v} \exp\left(-\frac{(v_i - c_i)^2}{2\sigma_i^2} + \sum_{j=1}^{n_h} h_j w_{ji} \frac{v_i}{\sigma_i^2}\right)}{\prod_{i=1}^{n_v}\left[\int_{-\infty}^{\infty} \exp\left(-\frac{(v_i' - c_i)^2}{2\sigma_i^2} + \sum_{j=1}^{n_h} h_j w_{ji} \frac{v_i'}{\sigma_i^2}\right) dv_i'\right]}.
\end{aligned} \tag{B.6}$$

O exponencial no numerador pode ser desenvolvido numa forma mais conveniente completando-se quadrados. Para concisão de notação, define-se

$$\gamma_i(\mathbf{h}) = \sum_{j=1}^{n_h} h_j w_{ji},$$

de modo que o exponencial mencionado pode ser escrito como

$$\begin{aligned}
\exp\left(-\frac{(v_i-c_i)^2}{2\sigma_i^2}+\gamma_i(\mathbf{h})\frac{v_i}{\sigma_i^2}\right) &= \exp\left(-\frac{v_i^2-2v_ic_i+c_i^2-2\gamma_i(\mathbf{h})v_i}{2\sigma_i^2}\right)\\
&= \exp\left(-\frac{v_i^2-2v_i(c_i+\gamma_i(\mathbf{h}))+c_i^2+(c_i+\gamma_i(\mathbf{h}))^2-(c_i+\gamma_i(\mathbf{h}))^2}{2\sigma_i^2}\right)\\
&= \exp\left(-\frac{(v_i-(c_i+\gamma_i(\mathbf{h})))^2+c_i^2-(c_i+\gamma_i(\mathbf{h}))^2}{2\sigma_i^2}\right)\\
&= \exp\left[-\frac{(v_i-(c_i+\gamma_i(\mathbf{h})))^2}{2\sigma_i^2}\right]\exp\left[\frac{(c_i+\gamma_i(\mathbf{h}))^2-c_i^2}{2\sigma_i^2}\right].
\end{aligned}$$

O exponencial no denominador, por sua vez, seria escrito de forma idêntica, exceto que no lugar de v_i ter-se-ia v_i'. Usando esses resultados na última expressão para $P_g(\mathbf{v}|\mathbf{h})$ pode-se continuar o desenvolvimento

$$\begin{aligned}
P_g(\mathbf{v}|\mathbf{h}) &= \frac{\prod_{i=1}^{n_v}\exp\left[-\frac{(v_i-(c_i+\gamma_i(\mathbf{h})))^2}{2\sigma_i^2}\right]\exp\left[\frac{(c_i+\gamma_i(\mathbf{h}))^2-c_i^2}{2\sigma_i^2}\right]}{\prod_{i=1}^{n_v}\left[\int_{-\infty}^{\infty}\exp\left[-\frac{(v_i'-(c_i+\gamma_i(\mathbf{h})))^2}{2\sigma_i^2}\right]\exp\left[\frac{(c_i+\gamma_i(\mathbf{h}))^2-c_i^2}{2\sigma_i^2}\right]dv_i'\right]}\\
&= \frac{\prod_{i=1}^{n_v}\exp\left[-\frac{(v_i-(c_i+\gamma_i(\mathbf{h})))^2}{2\sigma_i^2}\right]}{\prod_{i=1}^{n_v}\left[\int_{-\infty}^{\infty}\exp\left[-\frac{(v_i'-(c_i+\gamma_i(\mathbf{h})))^2}{2\sigma_i^2}\right]dv_i'\right]}\\
&= \prod_{i=1}^{n_v}\left[\frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left[-\frac{(v_i-(c_i+\gamma_i(\mathbf{h})))^2}{2\sigma_i^2}\right]\right].
\end{aligned} \tag{B.7}$$

Nota-se, portanto, que $P_g(\mathbf{v}|\mathbf{h})$ é um produtório de funções de densidade de probabilidade Gaussiana, cada qual tendo média $c_i+\gamma_i(\mathbf{h})$ e variância σ_i^2, ou seja,

$$P_g(\mathbf{v}|\mathbf{h}) = \prod_{i=1}^{n_v} P_g(v_i|\mathbf{h}), \tag{B.8}$$

sendo que

$$P_g(v_i|\mathbf{h}) = \mathcal{N}\left(v_i\middle|c_i+\sum_{j=1}^{n_h}h_jw_{ji},\sigma_i^2\right), \tag{B.9}$$

em que $\mathcal{N}(\cdot|\mu,\sigma^2)$ denota a função de densidade de probabilidade Gaussiana com média μ e variância σ^2.

B.3.2 Derivação de $P_g(\mathbf{h}|\mathbf{v})$

$$
\begin{aligned}
P_g(\mathbf{h}|\mathbf{v}) &= \frac{P_g(\mathbf{v},\mathbf{h})}{P_g(\mathbf{v})} = \frac{P_g(\mathbf{v},\mathbf{h})}{\sum\limits_{\mathbf{h}'\in\{0,1\}^{n_h}} P_g(\mathbf{v},\mathbf{h}')} = \frac{\exp(-E_g(\mathbf{v},\mathbf{h}))/Z_g}{\sum\limits_{\mathbf{h}'\in\{0,1\}^{n_h}} \exp(-E_g(\mathbf{v},\mathbf{h}'))/Z_g} \\
&= \frac{\exp\left(\sum\limits_{j=1}^{n_h} b_j h_j - \sum\limits_{i=1}^{n_v} \frac{(v_i-c_i)^2}{2\sigma_i^2} + \sum\limits_{i=1}^{n_v}\sum\limits_{j=1}^{n_h} h_j w_{ji}\frac{v_i}{\sigma_i^2}\right)}{\sum\limits_{h'_1=0}^{1}\cdots\sum\limits_{h'_{n_h}=0}^{1} \exp\left(\sum\limits_{j=1}^{n_h} b_j h'_j - \sum\limits_{i=1}^{n_v} \frac{(v_i-c_i)^2}{2\sigma_i^2} + \sum\limits_{i=1}^{n_v}\sum\limits_{j=1}^{n_h} h'_j w_{ji}\frac{v_i}{\sigma_i^2}\right)} \\
&= \frac{\prod\limits_{j=1}^{n_h} \exp\left(b_j h_j + \sum\limits_{i=1}^{n_v} h_j w_{ji}\frac{v_i}{\sigma_i^2}\right)}{\sum\limits_{h'_1=0}^{1}\cdots\sum\limits_{h'_{n_h}=0}^{1}\prod\limits_{j=1}^{n_h} \exp\left(b_j h'_j + \sum\limits_{i=1}^{n_v} h'_j w_{ji}\frac{v_i}{\sigma_i^2}\right)} \\
&= \frac{\prod\limits_{j=1}^{n_h} \exp\left(b_j h_j + \sum\limits_{i=1}^{n_v} h_j w_{ji}\frac{v_i}{\sigma_i^2}\right)}{\prod\limits_{j=1}^{n_h}\left(\sum\limits_{h'_j=0}^{1} \exp\left(b_j h'_j + \sum\limits_{i=1}^{n_v} h'_j w_{ji}\frac{v_i}{\sigma_i^2}\right)\right)} \\
&= \prod_{j=1}^{n_h} \frac{\exp\left(b_j h_j + \sum\limits_{i=1}^{n_v} h_j w_{ji}\frac{v_i}{\sigma_i^2}\right)}{\left(1+\exp\left(b_j + \sum\limits_{i=1}^{n_v} w_{ji}\frac{v_i}{\sigma_i^2}\right)\right)}.
\end{aligned}
\tag{B.10}
$$

Como nas RBMs, nota-se que cada um dos fatores do produtório na última passagem depende somente de uma variável oculta. Além disso, cada fator é, por si só, uma função massa de probabilidade (i.e., é não negativo e o somatório ao longo do seu domínio é 1). Conclui-se que

$$
P_g(\mathbf{h}|\mathbf{v}) = \prod_{j=1}^{n_h} P_g(h_j|\mathbf{v}), \tag{B.11}
$$

com

$$
P_g(h_j|\mathbf{v}) = \frac{\exp\left(b_j h_j + \sum\limits_{i=1}^{n_v} h_j w_{ji}\frac{v_i}{\sigma_i^2}\right)}{1+\exp\left(b_j + \sum\limits_{i=1}^{n_v} w_{ji}\frac{v_i}{\sigma_i^2}\right)}. \tag{B.12}
$$

Fazendo-se h_j=1 na Equação (B.12), chega-se à probabilidade de ativação de uma unidade oculta dado $\mathbf{v}$, isto é,

$$
\begin{aligned}
P_g(h_j=1|\mathbf{v}) &= \Pr(H_j{=}1|\mathbf{V}{=}\mathbf{v}) \\
&= \frac{\exp\left(b_j + \sum\limits_{i=1}^{n_v} w_{ji}\frac{v_i}{\sigma_i^2}\right)}{1+\exp\left(b_j + \sum\limits_{i=1}^{n_v} w_{ji}\frac{v_i}{\sigma_i^2}\right)} = \varphi\left(b_j + \sum_{i=1}^{n_v} w_{ji}\frac{v_i}{\sigma_i^2}\right).
\end{aligned}
\tag{B.13}
$$

B.4 GRADIENTE DA FUNÇÃO PERDA

Como nas RBMs, tem-se a função perda associada à amostra de treinamento $\mathbf{v}^{(t)}$, dada por

$$\mathcal{L}_g(\boldsymbol{\theta}; \mathbf{v}^{(t)}) = -\ln P_g(\mathbf{v}^{(t)}), \tag{B.14}$$

sendo $\boldsymbol{\theta}$ um vetor em que cada componente representa um dos parâmetros do modelo. De (B.14) e (B.2) obtém-se a seguinte expressão para a função perda:

$$\mathcal{L}_g(\boldsymbol{\theta}; \mathbf{v}^{(t)}) = \mathcal{F}_g(\mathbf{v}^{(t)}) + \ln Z_g. \tag{B.15}$$

sendo deixada implícita a dependência de $\mathcal{F}_g(\cdot)$ e Z_g com relação aos parâmetros do modelo. Diferenciado-se (B.15) com relação a $\boldsymbol{\theta}$ e tendo em vista que $Z_g = \int_{\mathbf{v}\in\mathbb{R}^{n_v}} \exp(-\mathcal{F}_g(\mathbf{v}))\, \mathbf{dv}$, obtém-se

$$\begin{aligned}
\frac{\partial \mathcal{L}_g(\boldsymbol{\theta}; \mathbf{v}^{(t)})}{\partial \boldsymbol{\theta}} &= \frac{\partial \mathcal{F}_g(\mathbf{v}^{(t)})}{\partial \boldsymbol{\theta}} + \frac{1}{Z_g}\frac{\partial Z_g}{\partial \boldsymbol{\theta}} \\
&= \frac{\partial \mathcal{F}_g(\mathbf{v}^{(t)})}{\partial \boldsymbol{\theta}} - \frac{1}{Z_g}\int_{\mathbf{v}\in\mathbb{R}^{n_v}} \exp(-\mathcal{F}_g(\mathbf{v}))\frac{\partial \mathcal{F}_g(\mathbf{v})}{\partial \boldsymbol{\theta}}\, \mathbf{dv} \\
&= \frac{\partial \mathcal{F}_g(\mathbf{v}^{(t)})}{\partial \boldsymbol{\theta}} - \int_{\mathbf{v}\in\mathbb{R}^{n_v}} P_g(\mathbf{v})\frac{\partial \mathcal{F}_g(\mathbf{v})}{\partial \boldsymbol{\theta}}\, \mathbf{dv} \\
&= \frac{\partial \mathcal{F}_g(\mathbf{v}^{(t)})}{\partial \boldsymbol{\theta}} - \mathbb{E}_{P_g(\mathbf{v})}\left[\frac{\partial \mathcal{F}(\mathbf{V})}{\partial \boldsymbol{\theta}}\right].
\end{aligned} \tag{B.16}$$

Tem-se, portanto, uma base análoga à das RBMs para o desenvolvimento dos diversos algoritmos de treinamento.

B.5 REGRAS DE ATUALIZAÇÃO DE PARÂMETROS

As regras de atualização de parâmetros podem ser obtidas diferenciado-se a expressão para $\mathcal{F}_g(\mathbf{v})$ apresentada na Equação (B.5) com relação aos vários parâmetros do modelo, ou seja, determinando-se as componentes de $\partial\mathcal{F}_g(\mathbf{v})/\partial\boldsymbol{\theta}$. Em seguidas essas derivadas seriam usadas na versão escalar de (2.28) de modo a produzir as regras de atualização propriamente ditas.

Primeiramente, para os vieses das unidades ocultas, deriva-se

$$\begin{aligned}
\frac{\partial \mathcal{F}_g(\mathbf{v})}{\partial b_j} &= \frac{\partial}{b_j}\left[\sum_{i=1}^{n_v}\frac{(v_i - c_i)^2}{2\sigma_i^2} - \sum_{j=1}^{n_h}\zeta\left(b_j + \sum_{i=1}^{n_v} w_{ji}\frac{v_i}{\sigma_i^2}\right)\right] \\
&= -\varphi\left(b_j + \sum_{i=1}^{n_v} w_{ji}\frac{v_i}{\sigma_i^2}\right) \\
&= -P_g(h_j{=}1|\mathbf{v})
\end{aligned} \tag{B.17}$$

e, para os vieses das unidades visíveis,

$$\begin{aligned}\frac{\partial \mathcal{F}_g(\mathbf{v})}{\partial c_i} &= \frac{\partial}{c_i}\left[\sum_{i=1}^{n_v}\frac{(v_i-c_i)^2}{2\sigma_i^2} - \sum_{j=1}^{n_h}\zeta\left(b_j + \sum_{i=1}^{n_v} w_{ji}\frac{v_i}{\sigma_i^2}\right)\right] \\ &= -\frac{1}{\sigma_i^2}(v_i - c_i).\end{aligned} \tag{B.18}$$

Para os pesos de conexão, tem-se

$$\begin{aligned}\frac{\partial \mathcal{F}_g(\mathbf{v})}{\partial w_{ji}} &= \frac{\partial}{w_{ji}}\left[\sum_{i=1}^{n_v}\frac{(v_i-c_i)^2}{2\sigma_i^2} - \sum_{j=1}^{n_h}\zeta\left(b_j + \sum_{i=1}^{n_v} w_{ji}\frac{v_i}{\sigma_i^2}\right)\right] \\ &= -\varphi\left(b_j + \sum_{i=1}^{n_v} w_{ji}\frac{v_i}{\sigma_i^2}\right)\frac{v_i}{\sigma_i^2} \\ &= -\frac{1}{\sigma_i^2}P_g(h_j{=}1|\mathbf{v})\, v_i.\end{aligned} \tag{B.19}$$

Por fim, para os parâmetros correspondentes às variâncias, esses devem receber um tratamento especial de modo a garantir que seus valores sejam sempre não negativos. O cumprimento de tal restrição pode ser obtido naturalmente definindo-se, como proposto em Cho, Ilin e Raiko (2011), um conjunto de parâmetros auxiliares (z_i), tais que

$$\exp(z_i) = \sigma_i^2, \tag{B.20}$$

para $i = 1, \ldots, n_v$. Diferenciado-se $\mathcal{F}_g(\mathbf{v})$ com relação a um dos parâmetros auxiliares obtém-se

$$\begin{aligned}\frac{\partial \mathcal{F}_g(\mathbf{v})}{\partial z_i} &= \frac{\partial \mathcal{F}_g(\mathbf{v})}{\partial(\sigma_i^2)}\frac{d(\sigma_i^2)}{dz_i} = \frac{\partial}{\partial(\sigma_i^2)}\left[\sum_{i=1}^{n_v}\frac{(v_i-c_i)^2}{2\sigma_i^2} - \sum_{j=1}^{n_h}\zeta\left(b_j + \sum_{i=1}^{n_v} w_{ji}\frac{v_i}{\sigma_i^2}\right)\right]\frac{d(\sigma_i^2)}{dz_i} \\ &= \left[-\frac{(v_i-c_i)^2}{2(\sigma_i^2)^2} + \varphi\left(b_j + \sum_{i=1}^{n_v} w_{ji}\frac{v_i}{\sigma_i^2}\right)w_{ji}\frac{v_i}{(\sigma_i^2)^2}\right](\sigma_i^2) \\ &= \frac{1}{\sigma_i^2}\left[-\frac{(v_i-c_i)^2}{2} + P_g(h_j{=}1|\mathbf{v})\, w_{ji}v_i\right].\end{aligned} \tag{B.21}$$

Portanto, a atualização dos parâmetros σ_i^2 seria dividida em três passos:

1. Faz-se $z_i \leftarrow \ln(\sigma_i^2)$.
2. Atualiza-se z_i de acordo com o método SGD.
3. Faz-se $\sigma_i^2 \leftarrow z_i$.

APÊNDICE C – EQUAÇÕES RELATIVAS À VARIANTE PROPOSTA

C.1 DISTRIBUIÇÕES CONDICIONAIS

Nesta seção, todas as distribuições condicionais do modelo proposto que foram mostradas no texto principal são demonstradas. As demonstrações partem da distribuição conjunta dada na Equação (3.3) com a definição de energia global na Equação (3.1).

C.1.1 $\mathrm{P}_{gc}(y|\mathbf{h})$

$$
\begin{aligned}
P_{gc}(y|\mathbf{h}) &= \frac{P_{gc}(y,\mathbf{h})}{P_{gc}(\mathbf{h})} = \frac{\int\limits_{\mathbb{R}^{n_d}} P_{gc}(y,\mathbf{x}',\mathbf{h})\,\mathbf{dx}'}{\sum\limits_{y'=1}^{n_c}\int\limits_{\mathbb{R}^{n_d}} P_{gc}(y',\mathbf{x}',\mathbf{h})\,\mathbf{dx}'} \qquad \text{(C.1)}\\
&= \frac{\int\limits_{\mathbb{R}^{n_d}} \exp\left[\sum\limits_{j=1}^{n_h} b_j h_j - \sum\limits_{i=1}^{n_d}\frac{(x_i'-c_i)^2}{2\sigma_i^2} + \sum\limits_{i=1}^{n_d}\sum\limits_{j=1}^{n_h} h_j w_{ji}\frac{x_i'}{\sigma_i^2} + \sum\limits_{k=1}^{n_c} d_k\delta_{k,y} + \sum\limits_{k=1}^{n_c}\sum\limits_{j=1}^{n_h} h_j u_{jk}\delta_{k,y}\right]\mathbf{dx}'}{\sum\limits_{y'=1}^{n_c}\int\limits_{\mathbb{R}^{n_d}} \exp\left[\sum\limits_{j=1}^{n_h} b_j h_j - \sum\limits_{i=1}^{n_d}\frac{(x_i'-c_i)^2}{2\sigma_i^2} + \sum\limits_{i=1}^{n_d}\sum\limits_{j=1}^{n_h} h_j w_{ji}\frac{x_i'}{\sigma_i^2} + \sum\limits_{k=1}^{n_c} d_k\delta_{k,y'} + \sum\limits_{k=1}^{n_c}\sum\limits_{j=1}^{n_h} h_j u_{jk}\delta_{k,y'}\right]\mathbf{dx}'}\\
&= \frac{\exp\left(d_y + \sum\limits_{j=1}^{n_h} h_j u_{jy}\right)\left\{\int\limits_{\mathbb{R}^{n_d}} \exp\left[-\sum\limits_{i=1}^{n_d}\frac{(x_i'-c_i)^2}{2\sigma_i^2} + \sum\limits_{i=1}^{n_d}\sum\limits_{j=1}^{n_h} h_j w_{ji}\frac{x_i'}{\sigma_i^2}\right]\mathbf{dx}'\right\}}{\left[\sum\limits_{y'=1}^{n_c}\exp\left(d_{y'} + \sum\limits_{j=1}^{n_h} h_j u_{jy'}\right)\right]\left\{\int\limits_{\mathbb{R}^{n_d}} \exp\left[-\sum\limits_{i=1}^{n_d}\frac{(x_i'-c_i)^2}{2\sigma_i^2} + \sum\limits_{i=1}^{n_d}\sum\limits_{j=1}^{n_h} h_j w_{ji}\frac{x_i'}{\sigma_i^2}\right]\mathbf{dx}'\right\}}\\
&= \frac{\exp\left(d_y + \sum\limits_{j=1}^{n_h} h_j u_{jy}\right)}{\sum\limits_{y'=1}^{n_c}\exp\left(d_{y'} + \sum\limits_{j=1}^{n_h} h_j u_{jy'}\right)}.
\end{aligned}
$$

C.1.2 $\mathrm{P}_{gc}(\mathbf{x}|\mathbf{h})$

$$
\begin{aligned}
P_{gc}(\mathbf{x}|\mathbf{h}) &= \frac{P_{gc}(\mathbf{x},\mathbf{h})}{P_{gc}(\mathbf{h})} = \frac{\sum\limits_{y'=1}^{n_c} P_{gc}(y',\mathbf{x},\mathbf{h})}{\sum\limits_{y'=1}^{n_c}\int\limits_{\mathbb{R}^{n_d}} P_{gc}(y',\mathbf{x}',\mathbf{h})\,\mathbf{dx}'} \qquad \text{(C.2)}\\
&= \frac{\sum\limits_{y'=1}^{n_c}\exp\left[\sum\limits_{j=1}^{n_h} b_j h_j - \sum\limits_{i=1}^{n_d}\frac{(x_i-c_i)^2}{2\sigma_i^2} + \sum\limits_{i=1}^{n_d}\sum\limits_{j=1}^{n_h} h_j w_{ji}\frac{x_i}{\sigma_i^2} + \sum\limits_{k=1}^{n_c} d_k\delta_{k,y'} + \sum\limits_{k=1}^{n_c}\sum\limits_{j=1}^{n_h} h_j u_{jk}\delta_{k,y'}\right]}{\sum\limits_{y'=1}^{n_c}\int\limits_{\mathbb{R}^{n_d}} \exp\left[\sum\limits_{j=1}^{n_h} b_j h_j - \sum\limits_{i=1}^{n_d}\frac{(x_i'-c_i)^2}{2\sigma_i^2} + \sum\limits_{i=1}^{n_d}\sum\limits_{j=1}^{n_h} h_j w_{ji}\frac{x_i'}{\sigma_i^2} + \sum\limits_{k=1}^{n_c} d_k\delta_{k,y'} + \sum\limits_{k=1}^{n_c}\sum\limits_{j=1}^{n_h} h_j u_{jk}\delta_{k,y'}\right]\mathbf{dx}'}\\
&= \frac{\left[\sum\limits_{y'=1}^{n_c}\exp\left(d_{y'} + \sum\limits_{j=1}^{n_h} h_j u_{jy'}\right)\right]\left\{\exp\left[-\sum\limits_{i=1}^{n_d}\frac{(x_i-c_i)^2}{2\sigma_i^2} + \sum\limits_{i=1}^{n_d}\sum\limits_{j=1}^{n_h} h_j w_{ji}\frac{x_i}{\sigma_i^2}\right]\right\}}{\left[\sum\limits_{y'=1}^{n_c}\exp\left(d_{y'} + \sum\limits_{j=1}^{n_h} h_j u_{jy'}\right)\right]\left\{\int\limits_{\mathbb{R}^{n_d}} \exp\left[-\sum\limits_{i=1}^{n_d}\frac{(x_i'-c_i)^2}{2\sigma_i^2} + \sum\limits_{i=1}^{n_d}\sum\limits_{j=1}^{n_h} h_j w_{ji}\frac{x_i'}{\sigma_i^2}\right]\mathbf{dx}'\right\}}\\
&= \frac{\exp\left[-\sum\limits_{i=1}^{n_d}\frac{(x_i-c_i)^2}{2\sigma_i^2} + \sum\limits_{i=1}^{n_d}\sum\limits_{j=1}^{n_h} h_j w_{ji}\frac{x_i}{\sigma_i^2}\right]}{\int\limits_{\mathbb{R}^{n_d}} \exp\left[-\sum\limits_{i=1}^{n_d}\frac{(x_i'-c_i)^2}{2\sigma_i^2} + \sum\limits_{i=1}^{n_d}\sum\limits_{j=1}^{n_h} h_j w_{ji}\frac{x_i'}{\sigma_i^2}\right]\mathbf{dx}'}.
\end{aligned}
$$

Pode-se continuar o desenvolvimento da última expressão, convertendo-se o exponencial de um somatório (indexado por i) em um produtório de exponenciais, tanto no numerador quanto no denominador, produzindo

$$P_{gc}(\mathbf{x}|\mathbf{h}) = \prod_{i=1}^{n_d} \frac{\exp\left[-\frac{(x_i-c_i)^2}{2\sigma_i^2} + \sum\limits_{j=1}^{n_h} h_j w_{ji} \frac{x_i}{\sigma_i^2}\right]}{\int\limits_{\mathbb{R}^{n_d}} \exp\left[-\frac{(x_i'-c_i)^2}{2\sigma_i^2} + \sum\limits_{j=1}^{n_h} h_j w_{ji} \frac{x_i'}{\sigma_i^2}\right] \mathbf{dx}'}. \tag{C.3}$$

A última expressão permite dizer que $P_{gc}(\mathbf{x}|\mathbf{h})$ pode ser escrita como um produto de funções de densidade de probabilidade gaussianas. Isso é percebido notando-se que no numerador de cada fator do produtório tem-se uma função gaussiana (exponencial de um polinômio de segundo grau) enquanto que no denominador aparece a integral da expressão no numerador, garantindo, portanto, que cada fator tenha área unitária (em x_i). Essa observação sugere a definição

$$P_{gc}(x_i|\mathbf{h}) = \frac{\exp\left[-\frac{(x_i-c_i)^2}{2\sigma_i^2} + \sum\limits_{j=1}^{n_h} h_j w_{ji} \frac{x_i}{\sigma_i^2}\right]}{\int\limits_{\mathbb{R}^{n_d}} \exp\left[-\frac{(x_i'-c_i)^2}{2\sigma_i^2} + \sum\limits_{j=1}^{n_h} h_j w_{ji} \frac{x_i'}{\sigma_i^2}\right] \mathbf{dx}'}, \tag{C.4}$$

de modo que

$$P_{gc}(\mathbf{x}|\mathbf{h}) = \prod_{i=1}^{n_d} P_{gc}(x_i|\mathbf{h}). \tag{C.5}$$

É possível determinar a média e variância da distribuição gaussiana representada pela Equação (C.4). Para tanto, basta completar quadrados no polinômio que aparece dentro das exponenciais no numerador e denominador, como feito na página 141 para as RBMs Gauss-Bernoulli. Fazendo-se isso, chega-se à expressão para $P_{gc}(x_i|\mathbf{h})$ dada na Equação (3.6).

C.1.3 $\mathrm{P}_{gc}(\mathbf{h}|y,\mathbf{x})$

$$\begin{aligned}
P_{gc}(\mathbf{h}|y,\mathbf{x}) &= \frac{P_{gc}(y,\mathbf{x},\mathbf{h})}{P_{gc}(y,\mathbf{x})} = \frac{P_{gc}(y,\mathbf{x},\mathbf{h})}{\sum\limits_{\mathbf{h}'\in\{0,1\}^{n_h}} P_{gc}(y,\mathbf{x},\mathbf{h}')} \\
&= \frac{\exp\left[\sum\limits_{j=1}^{n_h} b_j h_j - \sum\limits_{i=1}^{n_d} \frac{(x_i-c_i)^2}{2\sigma_i^2} + \sum\limits_{i=1}^{n_d}\sum\limits_{j=1}^{n_h} h_j w_{ji} \frac{x_i}{\sigma_i^2} + \sum\limits_{k=1}^{n_c} d_k \delta_{k,y} + \sum\limits_{k=1}^{n_c}\sum\limits_{j=1}^{n_h} h_j u_{jk} \delta_{k,y}\right]}{\sum\limits_{\mathbf{h}'\in\{0,1\}^{n_h}} \exp\left[\sum\limits_{j=1}^{n_h} b_j h_j' - \sum\limits_{i=1}^{n_d} \frac{(x_i-c_i)^2}{2\sigma_i^2} + \sum\limits_{i=1}^{n_d}\sum\limits_{j=1}^{n_h} h_j' w_{ji} \frac{x_i}{\sigma_i^2} + \sum\limits_{k=1}^{n_c} d_k \delta_{k,y} + \sum\limits_{k=1}^{n_c}\sum\limits_{j=1}^{n_h} h_j' u_{jk} \delta_{k,y}\right]} \\
&= \frac{\prod\limits_{j=1}^{n_h} \exp\left(b_j h_j + \sum\limits_{i=1}^{n_d} h_j w_{ji} \frac{x_i}{\sigma_i^2} + h_j u_{jy}\right)}{\sum\limits_{\mathbf{h}'\in\{0,1\}^{n_h}} \prod\limits_{j=1}^{n_h} \exp\left(b_j h_j' + \sum\limits_{i=1}^{n_d} h_j' w_{ji} \frac{x_i}{\sigma_i^2} + h_j' u_{jy}\right)}.
\end{aligned} \tag{C.6}$$

Pode-se utilizar o resultado dado na Equação (A.3) para converter os somatórios de produtórios em produtórios de somatórios no denominador da última expressão, chegando-se a

$$P_{gc}(\mathbf{h}|y,\mathbf{x}) = \frac{\prod_{j=1}^{n_h} \exp\left(b_j h_j + \sum_{i=1}^{n_d} h_j w_{ji} \frac{x_i}{\sigma_i^2} + h_j u_{jy}\right)}{\prod_{j=1}^{n_h} \sum_{h'_j=0}^{1} \exp\left(b_j h'_j + \sum_{i=1}^{n_d} h'_j w_{ji} \frac{x_i}{\sigma_i^2} + h'_j u_{jy}\right)} \tag{C.7}$$

$$= \prod_{j=1}^{n_h} \frac{\exp\left(b_j h_j + \sum_{i=1}^{n_d} h_j w_{ji} \frac{x_i}{\sigma_i^2} + h_j u_{jy}\right)}{\sum_{h'_j=0}^{1} \exp\left(b_j h'_j + \sum_{i=1}^{n_d} h'_j w_{ji} \frac{x_i}{\sigma_i^2} + h'_j u_{jy}\right)}.$$

Lembrando que h_j assume apenas os valores 0 ou 1, nota-se que cada fator do produtório na última expressão para $P_{gc}(\mathbf{h}|y,\mathbf{x})$ representa um distribuição de Bernoulli, o que sugere a definição

$$P_{gc}(h_j|y,\mathbf{x}) = \frac{\exp\left(b_j h_j + \sum_{i=1}^{n_d} h_j w_{ji} \frac{x_i}{\sigma_i^2} + h_j u_{jy}\right)}{\sum_{h'_j=0}^{1} \exp\left(b_j h'_j + \sum_{i=1}^{n_d} h'_j w_{ji} \frac{x_i}{\sigma_i^2} + h'_j u_{jy}\right)} \tag{C.8}$$

$$= \frac{\exp\left(b_j h_j + \sum_{i=1}^{n_d} h_j w_{ji} \frac{x_i}{\sigma_i^2} + h_j u_{jy}\right)}{1 + \exp\left(b_j + \sum_{i=1}^{n_d} w_{ji} \frac{x_i}{\sigma_i^2} + u_{jy}\right)}, \tag{C.9}$$

de modo que

$$P_{gc}(\mathbf{h}|y,\mathbf{x}) = \prod_{j=1}^{n_h} P_{gc}(h_j|y,\mathbf{x}). \tag{C.10}$$

A probabilidade de ativação da variável H_j é obtida avaliando-se $P_{gc}(h_j|y,\mathbf{x})$ em h_j=1, o que resulta em

$$P_{gc}(h_j{=}1|y,\mathbf{x}) = \frac{\exp\left(b_j + \sum_{i=1}^{n_d} w_{ji} \frac{x_i}{\sigma_i^2} + u_{jy}\right)}{1 + \exp\left(b_j + \sum_{i=1}^{n_d} w_{ji} \frac{x_i}{\sigma_i^2} + u_{jy}\right)} \tag{C.11}$$

$$= \varphi\left(b_j + \sum_{i=1}^{n_d} w_{ji} \frac{x_i}{\sigma_i^2} + u_{jy}\right) \tag{C.12}$$

$$= \varphi\left(o_j(y,\mathbf{x})\right). \tag{C.13}$$

Na Equação (C.12), empregou-se a definição da função sigmóide e, na Equação (C.13), a definição da função auxiliar, $o_j(y,\mathbf{x})$, dada na Equação (3.21).

C.1.4 $\mathrm{P}_{gc}(y|\mathbf{x})$

$$
\begin{aligned}
P_{gc}(y|\mathbf{x}) = \frac{P_{gc}(y,\mathbf{x})}{P_{gc}(\mathbf{x})} &= \frac{\sum\limits_{\mathbf{h}'\in\{0,1\}^{n_h}} P_{gc}(y,\mathbf{x},\mathbf{h}')}{\sum\limits_{y'=1}^{n_c}\sum\limits_{\mathbf{h}'\in\{0,1\}^{n_h}} P_{gc}(y',\mathbf{x},\mathbf{h}')} \qquad \text{(C.14)}\\
&= \frac{\sum\limits_{\mathbf{h}'\in\{0,1\}^{n_h}} \exp\left[\sum\limits_{j=1}^{n_h} b_j h'_j - \sum\limits_{i=1}^{n_d}\frac{(x_i-c_i)^2}{2\sigma_i^2} + \sum\limits_{i=1}^{n_d}\sum\limits_{j=1}^{n_h} h'_j w_{ji}\frac{x_i}{\sigma_i^2} + \sum\limits_{k=1}^{n_c} d_k\delta_{k,y} + \sum\limits_{k=1}^{n_c}\sum\limits_{j=1}^{n_h} h'_j u_{jk}\delta_{k,y}\right]}{\sum\limits_{y'=1}^{n_c}\sum\limits_{\mathbf{h}'\in\{0,1\}^{n_h}} \exp\left[\sum\limits_{j=1}^{n_h} b_j h'_j - \sum\limits_{i=1}^{n_d}\frac{(x_i-c_i)^2}{2\sigma_i^2} + \sum\limits_{i=1}^{n_d}\sum\limits_{j=1}^{n_h} h'_j w_{ji}\frac{x_i}{\sigma_i^2} + \sum\limits_{k=1}^{n_c} d_k\delta_{k,y'} + \sum\limits_{k=1}^{n_c}\sum\limits_{j=1}^{n_h} h'_j u_{jk}\delta_{k,y'}\right]}\\
&= \frac{\exp(d_y)\left[\sum\limits_{\mathbf{h}'\in\{0,1\}^{n_h}}\prod\limits_{j=1}^{n_h}\exp\left(b_j h'_j + \sum\limits_{i=1}^{n_d} h'_j w_{ji}\frac{x_i}{\sigma_i^2} + h'_j u_{jy}\right)\right]}{\sum\limits_{y'=1}^{n_c}\exp(d_{y'})\left[\sum\limits_{\mathbf{h}'\in\{0,1\}^{n_h}}\prod\limits_{j=1}^{n_h}\exp\left(b_j h'_j + \sum\limits_{i=1}^{n_d} h'_j w_{ji}\frac{x_i}{\sigma_i^2} + h'_j u_{jy'}\right)\right]}.
\end{aligned}
$$

Pode-se usar novamente o resultado da Equação (A.3) para converter os somatórios de produtórios em produtórios de somatórios na última expressão, o que leva a

$$
\begin{aligned}
P_{gc}(y|\mathbf{x}) &= \frac{\exp(d_y)\left\{\prod\limits_{j=1}^{n_h}\left[\sum\limits_{h'_j=0}^{1}\exp\left(b_j h'_j + \sum\limits_{i=1}^{n_d} h'_j w_{ji}\frac{x_i}{\sigma_i^2} + h'_j u_{jy}\right)\right]\right\}}{\sum\limits_{y'=1}^{n_c}\exp(d_{y'})\left\{\prod\limits_{j=1}^{n_h}\left[\sum\limits_{h'_j=0}^{1}\exp\left(b_j h'_j + \sum\limits_{i=1}^{n_d} h'_j w_{ji}\frac{x_i}{\sigma_i^2} + h'_j u_{jy'}\right)\right]\right\}} \qquad \text{(C.15)}\\
&= \frac{\exp(d_y)\left\{\prod\limits_{j=1}^{n_h}\left[1+\exp\left(b_j + \sum\limits_{i=1}^{n_d} w_{ji}\frac{x_i}{\sigma_i^2} + u_{jy}\right)\right]\right\}}{\sum\limits_{y'=1}^{n_c}\exp(d_{y'})\left\{\prod\limits_{j=1}^{n_h}\left[1+\exp\left(b_j + \sum\limits_{i=1}^{n_d} w_{ji}\frac{x_i}{\sigma_i^2} + u_{jy'}\right)\right]\right\}}\\
&= \frac{\exp(d_y)\left\{\prod\limits_{j=1}^{n_h}\left[1+\exp\left(o_j(y,\mathbf{x})\right)\right]\right\}}{\sum\limits_{y'=1}^{n_c}\exp(d_{y'})\left\{\prod\limits_{j=1}^{n_h}\left[1+\exp\left(o_j(y',\mathbf{x})\right)\right]\right\}},
\end{aligned}
$$

sendo que na última passagem usou-se novamente a definição da função auxiliar $o_j(y,\mathbf{x})$, fornecida na Equação (3.21). Nota-se agora que os produtórios que ocorrem na última expressão têm a forma $1+\exp(z)$ e podem ser reescritos em termos da função *softplus* como

$$1+\exp(z) = \exp(\ln(1+\exp(z))) = \exp(\zeta(z)).$$

O emprego desse resultado na última expressão para $P_{gc}(y|\mathbf{x})$ produz

$$
\begin{aligned}
P_{gc}(y|\mathbf{x}) &= \frac{\exp(d_y)\prod\limits_{j=1}^{n_h}\exp\left[\zeta(o_j(y,\mathbf{x}))\right]}{\sum\limits_{y'=1}^{n_c}\exp(d_{y'})\prod\limits_{j=1}^{n_h}\exp\left[\zeta(o_j(y',\mathbf{x}))\right]} \qquad \text{(C.16)}\\
&= \frac{\exp\left[d_y + \sum\limits_{j=1}^{n_h}\zeta(o_j(y,\mathbf{x}))\right]}{\sum\limits_{y'=1}^{n_c}\exp\left[d_{y'} + \sum\limits_{j=1}^{n_h}\zeta(o_j(y',\mathbf{x}))\right]}, \qquad \text{(C.17)}
\end{aligned}
$$

a qual será mais conveniente para as derivações das regras de atualização de parâmetros no treinamento discriminativo.

C.2 GRADIENTE DA FUNÇÃO PERDA GENERATIVA

De (3.10) e (3.3), chega-se que

$$\begin{aligned}
\mathcal{L}_{gen}(\boldsymbol{\theta}; y^{(t)}, \mathbf{x}^{(t)}) &= -\ln P_{gc}(y^{(t)}, \mathbf{x}^{(t)}) \\
&= -\ln \left[\sum_{\mathbf{h}' \in \{0,1\}^{n_h}} P_{gc}(y^{(t)}, \mathbf{x}^{(t)}, \mathbf{h}') \right] \\
&= -\ln \left[\sum_{\mathbf{h}' \in \{0,1\}^{n_h}} \frac{\exp(-E_{gc}(y^{(t)}, \mathbf{x}^{(t)}, \mathbf{h}'))}{Z_{gc}} \right] \\
&= -\ln \left[\sum_{\mathbf{h}' \in \{0,1\}^{n_h}} \exp(-E_{gc}(y^{(t)}, \mathbf{x}^{(t)}, \mathbf{h}')) \right] + \ln Z_{gc}.
\end{aligned} \tag{C.18}$$

Assim, a derivada de $\mathcal{L}_{gen}(\boldsymbol{\theta}; y^{(t)}, \mathbf{x}^{(t)})$ com relação a um parâmetro qualquer, θ, componente do vetor $\boldsymbol{\theta}$, pode ser escrita como

$$\begin{aligned}
\frac{\partial \mathcal{L}_{gen}(\boldsymbol{\theta}; y^{(t)}, \mathbf{x}^{(t)})}{\partial \theta} &= -\sum_{\mathbf{h}'' \in \{0,1\}^{n_h}} \left[\frac{\exp(-E_{gc}(y^{(t)}, \mathbf{x}^{(t)}, \mathbf{h}''))}{\sum\limits_{\mathbf{h}' \in \{0,1\}^{n_h}} \exp(-E_{gc}(y^{(t)}, \mathbf{x}^{(t)}, \mathbf{h}'))} \left(-\frac{\partial E_{gc}(y^{(t)}, \mathbf{x}^{(t)}, \mathbf{h}'')}{\partial \theta} \right) \right] \\
&\quad + \frac{1}{Z_{gc}} \frac{\partial}{\partial \theta} \left[\sum_{y''=1}^{n_c} \int_{\mathbb{R}^{n_d}} \sum_{\mathbf{h}'' \in \{0,1\}^{n_h}} \exp(-E_{gc}(y'', \mathbf{x}'', \mathbf{h}'')) \, \mathbf{dx}'' \right] \\
&= \left[\sum_{\mathbf{h}'' \in \{0,1\}^{n_h}} \frac{\exp(-E_{gc}(y^{(t)}, \mathbf{x}^{(t)}, \mathbf{h}''))}{\sum\limits_{\mathbf{h}' \in \{0,1\}^{n_h}} \exp(-E_{gc}(y^{(t)}, \mathbf{x}^{(t)}, \mathbf{h}'))} \left(\frac{\partial E_{gc}(y^{(t)}, \mathbf{x}^{(t)}, \mathbf{h}'')}{\partial \theta} \right) \right] \\
&\quad - \left[\sum_{y''=1}^{n_c} \int_{\mathbb{R}^{n_d}} \sum_{\mathbf{h}'' \in \{0,1\}^{n_h}} \frac{\exp(-E_{gc}(y'', \mathbf{x}'', \mathbf{h}''))}{Z_{gc}} \left(\frac{\partial E_{gc}(y'', \mathbf{x}'', \mathbf{h}'')}{\partial \theta} \right) \mathbf{dx}'' \right].
\end{aligned} \tag{C.19}$$

Notando-se que

$$P_{gc}(h''|y^{(t)}, \mathbf{x}^{(t)}) = \frac{\exp(-E_{gc}(y^{(t)}, \mathbf{x}^{(t)}, \mathbf{h}''))}{\sum\limits_{\mathbf{h}' \in \{0,1\}^{n_h}} \exp(-E_{gc}(y^{(t)}, \mathbf{x}^{(t)}, \mathbf{h}'))}$$

e

$$P_{gc}(y'', \mathbf{x}'', \mathbf{h}'') = \frac{\exp(-E_{gc}(y'', \mathbf{x}'', \mathbf{h}''))}{Z_{gc}},$$

a derivada da função perda generativa com relação à θ pode ser escrita como

$$\begin{aligned}
\frac{\partial \mathcal{L}_{gen}(\boldsymbol{\theta}; y^{(t)}, \mathbf{x}^{(t)})}{\partial \theta} &= \left[\sum_{\mathbf{h}'' \in \{0,1\}^{n_h}} P_{gc}(h''|y^{(t)}, \mathbf{x}^{(t)}) \frac{\partial E_{gc}(y^{(t)}, \mathbf{x}^{(t)}, \mathbf{h}'')}{\partial \theta} \right] \\
&\quad - \left[\sum_{y''=1}^{n_c} \int_{\mathbb{R}^{n_d}} \sum_{\mathbf{h}'' \in \{0,1\}^{n_h}} P_{gc}(y'', \mathbf{x}'', \mathbf{h}'') \frac{\partial E_{gc}(y'', \mathbf{x}'', \mathbf{h}'')}{\partial \theta} \mathbf{dx}'' \right].
\end{aligned} \tag{C.20}$$

Identifica-se que ambas as expressões no interior de colchetes representam esperanças com relação a diferentes distribuições definidas pelo modelo. Finalmente, reescrevendo a última equação com o uso do operador esperança e lembrando que os resultados valem para qualquer θ, chega-se ao gradiente da função perda generativa fornecida na Equação (3.11).

C.3 GRADIENTE DA FUNÇÃO PERDA DISCRIMINATIVA

De (3.18) e (C.17), pode-se escrever

$$
\begin{aligned}
-\mathcal{L}_{disc}(\boldsymbol{\theta}; y^{(t)}, \mathbf{x}^{(t)}) &= \ln P_{gc}(y^{(t)}|\mathbf{x}^{(t)}) \\
&= d_{y^{(t)}} + \sum_{j=1}^{n_h} \zeta\left(o_j(y^{(t)}, \mathbf{x}^{(t)})\right) - \ln\left[\sum_{y'=1}^{n_c} \exp\left(d_{y'} + \sum_{j=1}^{n_h} \zeta\left(o_j(y', \mathbf{x}^{(t)})\right)\right)\right].
\end{aligned} \tag{C.21}
$$

Com a função perda escrita nessa última forma, percebe-se que os únicos parâmetros que aparecem explicitamente em (C.21) são os vieses da camada de classificação (especificamente $d_{y^{(t)}}$). Entretanto, $\mathcal{L}_{disc}(\boldsymbol{\theta}; y^{(t)}, \mathbf{x}^{(t)})$ depende da função $o_j(y, \mathbf{x})$, a qual engloba os parâmetros que compõem $\mathbf{W}$, $\mathbf{b}$, $\boldsymbol{\sigma}^2$ e $\mathbf{U}$. Por esse motivo, é interessante desenvolver a derivada da função perda em termos de $\partial o_j(y, \mathbf{x})/\partial\theta$, em que θ seria um parâmetro escalar qualquer englobado por $o_j(y, \mathbf{x})$. Cumpre lembrar que $\zeta(\cdot)$ (função *softplus*) é a primitiva de $\varphi(\cdot)$ (função sigmóide), o que facilita os cálculos que seguem:

$$
\begin{aligned}
-\frac{\partial \mathcal{L}_{disc}(\boldsymbol{\theta}; y^{(t)}, \mathbf{x}^{(t)})}{\partial\theta} &= \frac{\partial}{\partial\theta}\Bigg\{ d_{y^{(t)}} + \sum_{j=1}^{n_h} \zeta\left(o_j(y^{(t)}, \mathbf{x}^{(t)})\right) \\
&\qquad - \ln\left[\sum_{y'=1}^{n_c} \exp\left(d_{y'} + \sum_{j=1}^{n_h} \zeta\left(o_j(y', \mathbf{x}^{(t)})\right)\right)\right]\Bigg\}
\end{aligned} \tag{C.22}
$$

$$
\begin{aligned}
&= \sum_{j=1}^{n_h} \varphi\left(o_j(y^{(t)}, \mathbf{x}^{(t)})\right) \frac{\partial o_j(y^{(t)}, \mathbf{x}^{(t)})}{\partial\theta} \\
&- \sum_{y''=1}^{n_c} \frac{\exp\left(d_{y''} + \sum_{j=1}^{n_h} \zeta\left(o_j(y'', \mathbf{x}^{(t)})\right)\right)}{\sum_{y'=1}^{n_c} \exp\left(d_{y'} + \sum_{j=1}^{n_h} \zeta\left(o_j(y', \mathbf{x}^{(t)})\right)\right)} \sum_{j=1}^{n_h} \varphi\left(o_j(y'', \mathbf{x}^{(t)})\right) \frac{\partial o_j(y'', \mathbf{x}^{(t)})}{\partial\theta}
\end{aligned} \tag{C.23}
$$

$$
\begin{aligned}
&= \sum_{j=1}^{n_h} P_{gc}(h_j{=}1|y^{(t)}, \mathbf{x}^{(t)}) \frac{\partial o_j(y^{(t)}, \mathbf{x}^{(t)})}{\partial\theta} \\
&- \sum_{y''=1}^{n_c} P_{gc}(y''|\mathbf{x}^{(t)}) \sum_{j=1}^{n_h} P_{gc}(h_j{=}1|y'', \mathbf{x}^{(t)}) \frac{\partial o_j(y'', \mathbf{x}^{(t)})}{\partial\theta}.
\end{aligned} \tag{C.24}
$$

Na passagem de (C.23) para (C.24), utilizou-se (C.13) e (C.17). Notando-se ainda que somatório mais externo na segunda linha de (C.24) representa um cálculo de esperança segundo a distribuição $P_{gc}(y|\mathbf{x})$, chega-se à expressão dada na Equação (3.20).

C.4 REGRAS DE ATUALIZAÇÃO PARA TREINAMENTO GENERATIVO

O uso da aproximação (3.12) em (3.11) produz

$$\frac{\partial \mathcal{L}_{gen}(\boldsymbol{\theta}; y^{(t)}, \mathbf{x}^{(t)})}{\partial \boldsymbol{\theta}} \approx \mathbb{E}_{P_{gc}(\mathbf{h}|y,\mathbf{x})}\left[\frac{\partial E_{gc}(y^{(t)}, \mathbf{x}^{(t)}, \mathbf{H})}{\partial \boldsymbol{\theta}}\bigg| y^{(t)}, \mathbf{x}^{(t)}\right] - \mathbb{E}_{P_{gc}(\mathbf{h}|y,\mathbf{x})}\left[\frac{\partial E_{gc}(\widetilde{y}, \widetilde{\mathbf{x}}, \mathbf{H})}{\partial \boldsymbol{\theta}}\bigg| \widetilde{y}, \widetilde{\mathbf{x}}\right], \tag{C.25}$$

a partir da qual são determinadas as regras de atualização de parâmetros para a variante. Para tanto, nota-se que é necessário o conhecimento das derivadas da função de energia global, $E_{gc}(y, \mathbf{x}, \mathbf{h})$ (Equação (3.1)), com relação a cada um dos parâmetros do modelo. É simples mostrar que

$$\frac{\partial E_{gc}(y, \mathbf{x}, \mathbf{h})}{\partial w_{ji}} = -h_j \frac{x_i}{\sigma_i^2}, \tag{C.26}$$

$$\frac{\partial E_{gc}(y, \mathbf{x}, \mathbf{h})}{\partial u_{jk}} = -h_j \, \delta_{k,y}, \tag{C.27}$$

$$\frac{\partial E_{gc}(y, \mathbf{x}, \mathbf{h})}{\partial b_j} = -h_j, \tag{C.28}$$

$$\frac{\partial E_{gc}(y, \mathbf{x}, \mathbf{h})}{\partial c_i} = -\frac{(x_i - c_i)}{\sigma_i^2}, \tag{C.29}$$

$$\frac{\partial E_{gc}(y, \mathbf{x}, \mathbf{h})}{\partial d_k} = -\delta_{k,y}. \tag{C.30}$$

Para as variâncias, são empregados parâmetros auxiliares, $z_i = \ln(\sigma_i^2)$, como feito com as RBMs Gauss-Bernoulli (conforme explicado na página 50) e, assim

$$\frac{\partial E_{gc}(y, \mathbf{x}, \mathbf{h})}{\partial z_i} = -\left[\frac{(x_i - c_i)^2}{2\sigma_i^2} - \left(\frac{x_i}{\sigma_i^2}\right)\left(\sum_{j=1}^{n_h} h_j w_{ji}\right)\right]. \tag{C.31}$$

Observa-se que o gradiente da função perda generativa, como escrito em (C.25), é composto por duas parcelas, ambas envolvendo uma esperança condicional com relação à distribuição $P_{gc}(\mathbf{h}|y, \mathbf{x})$, para y e $\mathbf{x}$ fixos. Lembrando que as variáveis ocultas são independentes entre si dado o estado das visíveis (conforme Equação (3.5)), então vale

$$\mathbb{E}_{P_{gc}(\mathbf{h}|y,\mathbf{x})}\left[H_j|y, \mathbf{x}\right] = \mathbb{E}_{P_{gc}(h_j|y,\mathbf{x})}\left[H_j|y, \mathbf{x}\right] \tag{C.32}$$

e, levando em conta que as variáveis ocultas são binárias (Bernoulli), tem-se ainda

$$\mathbb{E}_{P_{gc}(\mathbf{h}|y,\mathbf{x})}\left[H_j|y, \mathbf{x}\right] = P_{gc}(h_j{=}1|y, \mathbf{x}). \tag{C.33}$$

Esse resultado, em conjunto com a linearidade do operador esperança, pode ser usado para se avaliar, para cada parâmetro θ do modelo, a esperança de $\partial E_{gc}(y, \mathbf{x}, \mathbf{H})/\partial\theta$ que

aparece em ambas as parcelas de (C.25) (para y e $\mathbf{x}$ quaisquer). Chega-se que

$$\mathbb{E}_{P_{gc}(\mathbf{h}|y,\mathbf{x})}\left[\frac{\partial E_{gc}(y,\mathbf{x},\mathbf{H})}{\partial w_{ji}}\middle| y,\mathbf{x}\right] = -P_{gc}(h_j{=}1|y,\mathbf{x})\,\frac{x_i}{\sigma_i^2}, \tag{C.34}$$

$$\mathbb{E}_{P_{gc}(\mathbf{h}|y,\mathbf{x})}\left[\frac{\partial E_{gc}(y,\mathbf{x},\mathbf{H})}{\partial u_{jk}}\middle| y,\mathbf{x}\right] = -P_{gc}(h_j{=}1|y,\mathbf{x})\,\delta_{k,y}, \tag{C.35}$$

$$\mathbb{E}_{P_{gc}(\mathbf{h}|y,\mathbf{x})}\left[\frac{\partial E_{gc}(y,\mathbf{x},\mathbf{H})}{\partial b_j}\middle| y,\mathbf{x}\right] = -P_{gc}(h_j{=}1|y,\mathbf{x}), \tag{C.36}$$

$$\mathbb{E}_{P_{gc}(\mathbf{h}|y,\mathbf{x})}\left[\frac{\partial E_{gc}(y,\mathbf{x},\mathbf{H})}{\partial c_i}\middle| y,\mathbf{x}\right] = -\frac{(x_i-c_i)}{\sigma_i^2}, \tag{C.37}$$

$$\mathbb{E}_{P_{gc}(\mathbf{h}|y,\mathbf{x})}\left[\frac{\partial E_{gc}(y,\mathbf{x},\mathbf{H})}{\partial d_k}\middle| y,\mathbf{x}\right] = -\delta_{k,y}, \tag{C.38}$$

$$\mathbb{E}_{P_{gc}(\mathbf{h}|y,\mathbf{x})}\left[\frac{\partial E_{gc}(y,\mathbf{x},\mathbf{H})}{\partial z_i}\middle| y,\mathbf{x}\right] = -\left[\frac{(x_i-c_i)^2}{2\sigma_i^2} - \left(\frac{x_i}{\sigma_i^2}\right)\left(\sum_{j=1}^{n_h} P_{gc}(h_j{=}1|y,\mathbf{x})\,w_{ji}\right)\right]. \tag{C.39}$$

Segundo o SGD, as regras de atualização de parâmetros para o treinamento generativo teriam a forma

$$\theta \leftarrow \theta - \lambda_{gen}\frac{\partial \mathcal{L}_{gen}(\boldsymbol{\theta}; y^{(t)}, \mathbf{x}^{(t)})}{\partial \theta}. \tag{C.40}$$

Portanto, para obtê-las, bastaria aplicar à (C.25) cada um dos resultados nas Equações (C.34) a (C.38). Com isso, chega-se às regras (3.13) a (3.17) apresentadas no texto principal, em que omitiu-se a regra relativa aos parâmetros σ_i^2. Esses são atualizados por intermédio de z_i, executando-se a sequência de regras:

$$z_i \leftarrow \ln(\sigma_i^2), \tag{C.41}$$

$$z_i \leftarrow z_i + \lambda_{gen}\left\{\left[\frac{(x_i^{(t)}-c_i)^2}{2\sigma_i^2} - \left(\frac{x_i^{(t)}}{\sigma_i^2}\right)\left(\sum_{j=1}^{n_h} P_{gc}(h_j{=}1|y^{(t)},\mathbf{x}^{(t)})\,w_{ji}\right)\right]\right. \tag{C.42}$$

$$\left. - \left[\frac{(\widetilde{x}_i-c_i)^2}{2\sigma_i^2} - \left(\frac{\widetilde{x}_i}{\sigma_i^2}\right)\left(\sum_{j=1}^{n_h} P_{gc}(h_j{=}1|\widetilde{y},\widetilde{\mathbf{x}})\,w_{ji}\right)\right]\right\}, \tag{C.43}$$

$$\sigma_i^2 \leftarrow \exp(z_i). \tag{C.44}$$

C.5 REGRAS DE ATUALIZAÇÃO PARA TREINAMENTO DISCRIMINATIVO

As regras de atualização de parâmetros para o treinamento discriminativo podem ser determinadas derivando-se (C.21) com relação a cada um dos parâmetros do modelo e aplicando o resultado ao SGD, cujas regras de atualização teriam o formato

$$\theta \leftarrow \theta - \lambda_{disc} \frac{\partial \mathcal{L}_{disc}(\boldsymbol{\theta}; y^{(t)}, \mathbf{x}^{(t)})}{\partial \theta}. \tag{C.45}$$

Para o viés da camada de classificação de índice k, d_k, tem-se

$$\begin{aligned} -\frac{\partial \mathcal{L}_{disc}(\boldsymbol{\theta}; y^{(t)}, \mathbf{x}^{(t)})}{\partial d_k} &= \frac{\partial}{\partial d_k} \Bigg\{ d_{y^{(t)}} + \sum_{j=1}^{n_h} \zeta\left(o_j(y^{(t)}, \mathbf{x}^{(t)})\right) \qquad (C.46) \\ &\qquad - \ln\left[\sum_{y'=1}^{n_c} \exp\left(d_{y'} + \sum_{j=1}^{n_h} \zeta\left(o_j(y', \mathbf{x}^{(t)})\right)\right)\right]\Bigg\} \\ &= \delta_{k,y^{(t)}} - \sum_{y''=1}^{n_c} \frac{\exp\left(d_{y''} + \sum_{j=1}^{n_h} \zeta\left(o_j(y'', \mathbf{x}^{(t)})\right)\right)}{\sum_{y'=1}^{n_c} \exp\left(d_{y'} + \sum_{j=1}^{n_h} \zeta\left(o_j(y', \mathbf{x}^{(t)})\right)\right)} \delta_{y'',k} \qquad (C.47) \\ &= \delta_{k,y^{(t)}} - P_{gc}(y = k|\mathbf{x}^{(t)}) \triangleq \Delta d_k. \qquad (C.48) \end{aligned}$$

A aplicação do último resultado ao SGD, conforme (C.45), produz

$$d_k \leftarrow d_k + \lambda_{disc}\left[\delta_{k,y^{(t)}} - P_{gc}(y{=}k|\mathbf{x}^{(t)})\right], \tag{C.49}$$

que é apresentada em (3.23) com a ajuda de (3.27).

Como explicado na página 150, para os parâmetros que formam $\mathbf{W}$, $\mathbf{b}$, $\mathbf{U}$ e $\boldsymbol{\sigma}^2$, as derivadas da função perda são obtidas mais facilmente a partir de (C.24). Para as variâncias, são empregados parâmetros auxiliares, $z_i = \ln(\sigma_i^2)$, como feito com as RBMs Gauss-Bernoulli (conforme explicado na página 50). Além disso, para evitar conflito com os símbolos i, j e k, normalmente usados para indexar os parâmetros, aqui eles são temporariamente substituídos por i', j' e k', respectivamente. Com isso, as derivadas relevantes de $o_j(y, \mathbf{x})$ são:

$$\frac{\partial o_j(y, \mathbf{x})}{\partial w_{j'i'}} = \delta_{j,j'}\left(\frac{x_{i'}}{\sigma_{i'}^2}\right), \tag{C.50}$$

$$\frac{\partial o_j(y, \mathbf{x})}{\partial b_{j'}} = \delta_{j,j'}, \tag{C.51}$$

$$\frac{\partial o_j(y, \mathbf{x})}{\partial u_{j'k'}} = \delta_{j,j'}\,\delta_{y,k'}, \tag{C.52}$$

$$\frac{\partial o_j(y, \mathbf{x})}{\partial z_{i'}} = -w_{ji'}\left(\frac{x_{i'}}{e^{z_{i'}}}\right) = -w_{ji'}\left(\frac{x_{i'}}{\sigma_{i'}^2}\right). \tag{C.53}$$

A seguir, cada um dos resultados indo de (C.50) a (C.53) será aplicado à (C.24) na ordem mais interessante.

Para $b_{j'}$:

$$-\frac{\partial \mathcal{L}_{disc}(\boldsymbol{\theta}; y^{(t)}, \mathbf{x}^{(t)})}{\partial b_{j'}} = \sum_{j=1}^{n_h} P_{gc}(h_j{=}1|y^{(t)}, \mathbf{x}^{(t)}) \frac{\partial o_j(y^{(t)}, \mathbf{x}^{(t)})}{\partial b_{j'}} \tag{C.54}$$

$$- \sum_{y''=1}^{n_c} P_{gc}(y''|\mathbf{x}^{(t)}) \sum_{j=1}^{n_h} P_{gc}(h_j{=}1|y'', \mathbf{x}^{(t)}) \frac{\partial o_j(y'', \mathbf{x}^{(t)})}{\partial b_{j'}}$$

$$= \sum_{j=1}^{n_h} P_{gc}(h_j{=}1|y^{(t)}, \mathbf{x}^{(t)})\ \delta_{j,j'}$$

$$- \sum_{y''=1}^{n_c} P_{gc}(y''|\mathbf{x}^{(t)}) \sum_{j=1}^{n_h} P_{gc}(h_j{=}1|y'', \mathbf{x}^{(t)})\ \delta_{j,j'}$$

$$= \left[P_{gc}(h_{j'}{=}1|y^{(t)}, \mathbf{x}^{(t)}) - \sum_{y''=1}^{n_c} P_{gc}(y''|\mathbf{x}^{(t)}) P_{gc}(h_{j'}{=}1|y'', \mathbf{x}^{(t)}) \right] \triangleq \Delta b_{j'}. \tag{C.55}$$

Para $w_{j'i'}$:

$$-\frac{\partial \mathcal{L}_{disc}(\boldsymbol{\theta}; y^{(t)}, \mathbf{x}^{(t)})}{\partial w_{j'i'}} = \sum_{j=1}^{n_h} P_{gc}(h_j{=}1|y^{(t)}, \mathbf{x}^{(t)}) \frac{\partial o_j(y^{(t)}, \mathbf{x}^{(t)})}{\partial w_{j'i'}} \tag{C.56}$$

$$- \sum_{y''=1}^{n_c} P_{gc}(y''|\mathbf{x}^{(t)}) \sum_{j=1}^{n_h} P_{gc}(h_j{=}1|y'', \mathbf{x}^{(t)}) \frac{\partial o_j(y'', \mathbf{x}^{(t)})}{\partial w_{j'i'}}$$

$$= \sum_{j=1}^{n_h} P_{gc}(h_j{=}1|y^{(t)}, \mathbf{x}^{(t)})\ \delta_{j,j'} \left(\frac{x_{i'}^{(t)}}{\sigma_{i'}^2} \right)$$

$$- \sum_{y''=1}^{n_c} P_{gc}(y''|\mathbf{x}^{(t)}) \sum_{j=1}^{n_h} P_{gc}(h_j{=}1|y'', \mathbf{x}^{(t)})\ \delta_{j,j'} \left(\frac{x_{i'}^{(t)}}{\sigma_{i'}^2} \right)$$

$$= \left[P_{gc}(h_{j'}{=}1|y^{(t)}, \mathbf{x}^{(t)}) - \sum_{y''=1}^{n_c} P_{gc}(y''|\mathbf{x}^{(t)}) P_{gc}(h_{j'}{=}1|y'', \mathbf{x}^{(t)}) \right] \left(\frac{x_{i'}^{(t)}}{\sigma_{i'}^2} \right) \tag{C.57}$$

$$= \Delta b_{j'} \left(\frac{x_{i'}^{(t)}}{\sigma_{i'}^2} \right). \tag{C.58}$$

Para $u_{j'k'}$:

$$-\frac{\partial \mathcal{L}_{disc}(\boldsymbol{\theta}; y^{(t)}, \mathbf{x}^{(t)})}{\partial u_{j'k'}} = \sum_{j=1}^{n_h} P_{gc}(h_j{=}1|y^{(t)}, \mathbf{x}^{(t)}) \frac{\partial o_j(y^{(t)}, \mathbf{x}^{(t)})}{\partial u_{j'k'}} \tag{C.59}$$

$$- \sum_{y''=1}^{n_c} P_{gc}(y''|\mathbf{x}^{(t)}) \sum_{j=1}^{n_h} P_{gc}(h_j{=}1|y'', \mathbf{x}^{(t)}) \frac{\partial o_j(y'', \mathbf{x}^{(t)})}{\partial u_{j'k'}}$$

$$= \sum_{j=1}^{n_h} P_{gc}(h_j{=}1|y^{(t)}, \mathbf{x}^{(t)})\ \delta_{j,j'}\ \delta_{y^{(t)},k'}$$

$$- \sum_{y''=1}^{n_c} P_{gc}(y''|\mathbf{x}^{(t)}) \sum_{j=1}^{n_h} P_{gc}(h_j{=}1|y'', \mathbf{x}^{(t)})\ \delta_{j,j'}\ \delta_{y'',k'}$$

$$= P_{gc}(h_{j'}{=}1|y{=}k', \mathbf{x}^{(t)})\ \delta_{y^{(t)},k'} - P_{gc}(y{=}k'|\mathbf{x}^{(t)}) P_{gc}(h_{j'}{=}1|y{=}k', \mathbf{x}^{(t)})$$

$$= \left[\delta_{y^{(t)},k'} - P_{gc}(y{=}k'|\mathbf{x}^{(t)}) \right] P_{gc}(h_{j'}{=}1|y{=}k', \mathbf{x}^{(t)}) \tag{C.60}$$

$$= \Delta d_{k'}\ P_{gc}(h_{j'}{=}1|y{=}k', \mathbf{x}^{(t)}). \tag{C.61}$$

A aplicação de (C.55), (C.58) e (C.61) ao SGD, conforme (C.45), leva imediatamente a (3.22), (3.24) e (3.25), respectivamente.

Finalmente, apesar das ressalvas quanto ao ajuste das variâncias, as regras relativas à atualização das mesmas são aqui determinadas. Assim, para $z_{i'}$, pode-se desenvolver:

$$-\frac{\partial \mathcal{L}_{disc}(\boldsymbol{\theta}; y^{(t)}, \mathbf{x}^{(t)})}{\partial z_{i'}} = \sum_{j=1}^{n_h} P_{gc}(h_j{=}1|y^{(t)}, \mathbf{x}^{(t)}) \frac{\partial o_j(y^{(t)}, \mathbf{x}^{(t)})}{\partial z_{i'}} \tag{C.62}$$

$$- \sum_{y''=1}^{n_c} P_{gc}(y''|\mathbf{x}^{(t)}) \sum_{j=1}^{n_h} P_{gc}(h_j{=}1|y'', \mathbf{x}^{(t)}) \frac{\partial o_j(y'', \mathbf{x}^{(t)})}{\partial z_{i'}}$$

$$= \sum_{j=1}^{n_h} P_{gc}(h_j{=}1|y^{(t)}, \mathbf{x}^{(t)})\ w_{ji'} \left(-\frac{x_{i'}^{(t)}}{\sigma_{i'}^2}\right) \tag{C.63}$$

$$- \sum_{y''=1}^{n_c} P_{gc}(y''|\mathbf{x}^{(t)}) \sum_{j=1}^{n_h} P_{gc}(h_j{=}1|y'', \mathbf{x}^{(t)})\ w_{ji'} \left(-\frac{x_{i'}^{(t)}}{\sigma_{i'}^2}\right)$$

$$= \left(-\frac{x_{i'}^{(t)}}{\sigma_{i'}^2}\right) \left\{ \sum_{j=1}^{n_h} w_{ji'} \left[P_{gc}(h_j{=}1|y^{(t)}, \mathbf{x}^{(t)}) - \sum_{y''=1}^{n_c} P_{gc}(y''|\mathbf{x}^{(t)}) P_{gc}(h_j{=}1|y'', \mathbf{x}^{(t)}) \right] \right\} \tag{C.64}$$

$$\triangleq \Delta z_{i'}. \tag{C.65}$$

Similarmente ao explicado na página 50, as variâncias das variáveis de entrada seriam atualizadas pela execução em sequência das regras:

$$z_{i'} \leftarrow \ln(\sigma_{i'}^2), \tag{C.66}$$

$$z_{i'} \leftarrow z_{i'} + \lambda_{disc}\, \Delta z_{i'}, \tag{C.67}$$

$$\sigma_{i'}^2 \leftarrow \exp(z_{i'}). \tag{C.68}$$