Redes probabilísticas: aprendendo estruturas e atualizando probabilidades

Rodrigo Candido Faria

DISSERTAÇÃO APRESENTADA
AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA
OBTENÇÃO DO TÍTULO
DE
MESTRE EM CIÊNCIAS

Programa: Estatística
Orientador: Prof. Dr. Sergio Wechsler

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da CAPES

São Paulo, junho de 2014

Redes probabilísticas: aprendendo estruturas e atualizando probabilidades

Esta versão da dissertação contém as correções e alterações sugeridas pela Comissão Julgadora durante a defesa da versão original do trabalho, realizada em 28/05/2014. Uma cópia da versão original está disponível no Instituto de Matemática e Estatística da Universidade de São Paulo.

Comissão Julgadora:

- Prof. Dr. Sergio Wechsler (orientador) - IME-USP
- Prof. Dr. Nikolai Valtchev Kolev - IME-USP
- Prof. Dr. Marcio Alves Diniz - UFSCar

Agradecimentos

Agradeço à toda minha família, em especial aos meus pais, Marcos e Rosangela, e ao meu irmão, Allan, por todo o apoio que sempre me deram. Dedico um agradecimento especial ao meu orientador e amigo Sergio Wechsler, que com toda sua sabedoria e conhecimento me ajudou a realizar este trabalho e a entender melhor o que realmente é a Estatística. Agradeço aos professores Marcio Alves Diniz e Nikolai Valtchev Kolev por terem dado valiosas sugestões para a melhoria deste trabalho. Gostaria de agradecer também aos professores Carlos Alberto de Bragança Pereira, Luis Gustavo Esteves, Vladimir Belitsky e Florencia Graciela Leonardi, que direta ou indiretamente me ajudaram na conclusão do mestrado. Por último mas não menos importante, agradeço à todos meus amigos do IME-USP, desde os mais antigos, ainda da época da graduação, até os mais recentes, por estarem sempre prontos para ajudar.

Resumo

FARIA, R. C. **Redes probabilísticas: aprendendo estruturas e atualizando probabilidades**. 2014. Dissertação (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2014.

Redes probabilísticas são modelos muito versáteis, com aplicabilidade crescente em diversas áreas. Esses modelos são capazes de estruturar e mensurar a interação entre variáveis, permitindo que sejam realizados vários tipos de análises, desde diagnósticos de causas para algum fenômeno até previsões sobre algum evento, além de permitirem a construção de modelos de tomadas de decisões automatizadas. Neste trabalho são apresentadas as etapas para a construção dessas redes e alguns métodos usados para tal, dando maior ênfase para as chamadas redes bayesianas, uma subclasse de modelos de redes probabilísticas. A modelagem de uma rede bayesiana pode ser dividida em três etapas: seleção de variáveis, construção da estrutura da rede e estimação de probabilidades. A etapa de seleção de variáveis é usualmente feita com base nos conhecimentos subjetivos sobre o assunto estudado. A construção da estrutura pode ser realizada manualmente, levando em conta relações de causalidade entre as variáveis selecionadas, ou semi-automaticamente, através do uso de algoritmos. A última etapa, de estimação de probabilidades, pode ser feita seguindo duas abordagens principais: uma frequentista, em que os parâmetros são considerados fixos, e outra bayesiana, na qual os parâmetros são tratados como variáveis aleatórias. Além da teoria contida no trabalho, mostrando as relações entre a teoria de grafos e a construção probabilística das redes, também são apresentadas algumas aplicações desses modelos, dando destaque a problemas nas áreas de marketing e finanças.

Palavras-chave: Redes probabilísticas, redes bayesianas, diagnósticos, previsões, tomadas de decisões automatizadas.

Abstract

FARIA, R. C. **Probabilistic networks: learning structures and updating probabilities**. 2014. Dissertação (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2014.

Probabilistic networks are very versatile models, with growing applicability in many areas. These models are capable of structuring and measuring the interaction among variables, making possible various types of analyses, such as diagnoses of causes for a phenomenon and predictions about some event, besides allowing the construction of automated decision-making models. This work presents the necessary steps to construct those networks and methods used to doing so, emphasizing the so called Bayesian networks, a subclass of probabilistic networks. The Bayesian network modeling is divided in three steps: variables selection, structure learning and estimation of probabilities. The variables selection step is usually based on subjective knowledge about the studied topic. The structure learning can be performed manually, taking into account the causal relations among variables, or semi-automatically, through the use of algorithms. The last step, of probabilities estimation, can be treated following two main approaches: by the frequentist approach, where parameters are considered fixed, and by the Bayesian approach, in which parameters are treated as random variables. Besides the theory contained in this work, showing the relations between graph theory and the construction of probabilistic networks, applications of these models are presented, highlighting problems in marketing and finance.

Keywords: Probabilistic networks, Bayesian networks, diagnoses, predictions, automated decision-making.

Sumário

Capítulo 1

Introdução

1.1 Objetivos

A intenção deste trabalho é de apresentar como é elaborada a modelagem de redes probabilísticas, dando maior ênfase para os modelos de redes bayesianas. São apresentadas as etapas de seleção de variáveis, de estruturação da rede e da estimação de seus parâmetros. Após a exposição desses pontos, são mostradas algumas aplicações práticas onde as redes bayesianas podem ser usadas, como em problemas de diagnóstico, em casos de previsão de eventos e em cenários de tomada de decisão.

1.2 Algumas Considerações

Redes probabilísticas são modelos que mostram a interação entre variáveis aleatórias. Através desses modelos podemos estudar o quanto uma variável afeta, probabilisticamente, o comportamento das outras. Uma das classes mais usadas de redes probabilísticas são as redes bayesianas, que, apesar de não necessariamente serem tratadas com um enfoque realmente bayesiano, recebem esse nome devido a sua forte conexão com o uso do teorema de Bayes. Em alguns casos é possível, através da ajuda de um especialista, obter a estrutura de uma rede bayesiana "manualmente", porém nem sempre dispomos da presença de um *expert* no assunto estudado. Além disso, algumas vezes nem mesmo um especialista da área possui conhecimento suficiente para construir a estrutura da rede, já que em muitas vezes elas possuem centenas de variáveis. É justamente nesse tipo de cenário que são necessários métodos de aprendizagem automática, ou *machine learning*, para que a estrutura da rede seja encontrada. Após a construção da estrutura, devemos encontrar os chamados "parâmetros" da rede. Esses parâmetros podem ser considerados fixos e desconhecidos, como é usual na estatística clás-

sica/frequentista, ou podem ser tratados como variáveis aleatórias, o que torna possível refinarmos nosso conhecimento sobre eles à medida que mais dados sejam coletados.

1.3 Organização do Trabalho

Este trabalho está organizado da seguinte forma: no capítulo 2 são apresentadas algumas das possíveis interpretações de probabilidade, cujo entendimento é um dos pilares da teoria de redes probabilísticas; o capítulo 3 é um resumo básico sobre teoria de probabilidades, contendo algumas definições importantes e uma pequena explicação sobre o procedimento de inferência bayesiana; o capítulo 4 apresenta definições sobre a teoria de grafos e sua relação com as redes bayesianas, dando uma pequena introdução ao uso desses modelos na resolução de problemas; no capítulo 5 são mostrados métodos de construção de redes bayesianas, abordando a seleção das variáveis a serem estudadas, métodos para montar a estrutura da rede e formas de se estimar seus parâmetros; no capítulo 6 são apresentadas algumas aplicações práticas de redes bayesianas, mostrando sua capacidade de resolver problemas de diagnóstico, previsão e de tomada de decisão; o capítulo 7 contém as considerações finais do trabalho, onde há um resumo do conteúdo apresentado e onde são mostradas sugestões para futuras pesquisas; no capítulo 8 são listadas as referências bibliográficas utilizadas no trabalho.

Capítulo 2

Interpretações de Probabilidade

O conceito de probabilidade é usado frequentemente em nosso dia-a-dia, mas geralmente de maneira informal. Comumente ouvimos qual é a probabilidade de chover amanhã ou qual é a chance do time A vencer o time B, entretanto, não há consenso algum sobre qual interpretação é a mais adequada para cada aplicação, ou se pode haver uma interpretação absoluta, que se sobreponha às outras.

Há diversas formas de segmentar indivíduos baseando-se em suas interpretações sobre probabilidade, sendo todas, claro, cheias de controvérsias. Freedman (2010), por exemplo, divide os estatísticos em duas escolas: objetivista (ou frequentista) e subjetivista (ou bayesiana), porém, assume que ainda está longe de chegar numa conclusão: *"...statisticians agree amongst themselves about as well as philosophers; many shades of opinion will be represented in each school"*.

A seguir serão apresentadas, simplificadamente, as principais interpretações de probabilidade, com base nos trabalhos de DeGroot e Schervish (2012) e Hájek (2011). Não entraremos em muitos detalhes, pois a discussão sobre qual seria a "melhor" interpretação de probabilidade já renderia um trabalho inteiro dedicado a ela, e mesmo assim, certamente, seria insuficiente para gerar conclusões definitivas.

2.1 Interpretação Clássica de Probabilidade

A interpretação clássica de probabilidade é baseada no conceito de resultados equiprováveis, introduzido por Laplace (1814). Como exemplo podemos citar o lançamento de uma moeda, que pode ter dois resultados: cara ou coroa. Se assumirmos que ambos resultados são equiprováveis, então a probabilidade de cada um será a mesma: 1/2. Generalizando, temos que, para um processo com n resultados equiprováveis, cada um dos resultados tem probabilidade de $1/n$ de ocorrer.

O principal problema nesta interpretação é a ausência de um método para

eventos não equiprováveis, o que torna sua aplicabilidade restrita a um pequeno número de problemas.

2.2 Interpretação Frequentista de Probabilidade

Em alguns casos podemos interpretar probabilidade como sendo a frequência relativa com que um evento ocorreria se repetido um grande número de vezes sob condições similares. Um exemplo é o arremesso de uma moeda honesta, em que esperaríamos que a probabilidade de cara fosse 1/2, já que a frequência relativa de cara, ao arremessarmos a moeda várias vezes, deve supostamente se aproximar de 1/2. Essa interpretação foi introduzida por Venn (1876) e posteriormente desenvolvida e popularizada por Von Mises (1928) e Reichenbach (1949).

Obviamente essa interpretação possui problemas, já que é difícil, se não impossível, definir quantas vezes o experimento deve ser repetido para ser considerado "um grande número de vezes". Outro problema é definir o que são as "condições similares" sob as quais o experimento deve ser realizado, pois experimentos feitos sob condições idênticas gerariam sempre o mesmo resultado, o que tiraria a incerteza sobre o processo. Sendo assim, a interpretação frequentista só pode ser aplicada à problemas repetíveis sob condições similares, que são circunstâncias pouco comuns em aplicações reais.

2.3 Interpretação Subjetiva de Probabilidade

A interpretação subjetiva de probabilidade é baseada na ideia de que a probabilidade que uma pessoa atribui à ocorrência de um evento representa seu próprio, e pessoal, julgamento sobre a chance do evento ocorrer. Sendo tal julgamento baseado nas crenças e informações que cada pessoa tem sobre o processo, fazendo com que não haja uma probabilidade "verdadeira e desconhecida" a ser estimada, mas sim uma probabilidade individual sobre o evento. Essas ideias foram desenvolvidas, de forma independente, por Ramsey (1926) e De Finetti (1931).

Podemos ilustrar o caráter subjetivo de probabilidade através de um simples exemplo: Suponha que o indivíduo A arremesse uma moeda e observe cara, sendo assim, para ele $P(cara) = 1$. Em seguida, o indivíduo B, que não viu o resultado da moeda, é perguntado sobre qual a probabilidade do resultado ter sido cara, e, baseando-se em suas experiências anteriores, diz que $P(cara) = 0{,}5$. Sendo assim, cada indivíduo atribuiu a probabilidade à ocorrência do evento conforme sua informação sobre o processo.

Existem críticos sobre essa interpretação de probabilidade, que dizem ser impossível uma pessoa ser totalmente livre de preconceitos e contradições ao atri-

buir probabilidade a um determinado evento, além disso, dizem ser um empecilho para pesquisas conduzidas por mais de um pesquisador, já que uma conclusão seria impossibilitada por eles terem diferentes opiniões. Por outro lado, essa interpretação permite que a incerteza e subjetividade intrínsecas da ciência sejam adicionadas nos processos de decisão, e além disso, nada impede que opiniões baseadas em frequências relativas ou em eventos equiprováveis sejam incorporadas aos métodos e técnicas desenvolvidos com base na interpretação subjetiva, também chamada de bayesiana.

2.4 Interpretação Lógica de Probabilidade

A interpretação lógica de probabilidade foi desenvolvida por Keynes (1921) e Jeffreys (1939), tendo sido mais tarde adotada pelo filósofo Carnap (1950). A ideia básica da interpretação lógica é de que uma probabilidade é uma medida de "vínculo parcial" entre uma evidência e uma hipótese, com os valores 0 e 1 sendo os casos limites. Com essa interpretação podemos, através da chamada lógica indutiva, deduzir o que sucede uma certa premissa.

Para exemplificar o uso da lógica indutiva podemos partir de duas premissas: "toda maçã é vermelha" e "fotografei uma maçã". Concluímos logicamente que "a maçã que fotografei é vermelha". A interpretação lógica de probabilidade é proposta para casos em que haja incerteza nas afirmações. Podemos ter, por exemplo, as premissas "todas as maçãs vistas até agora são vermelhas" e "fotografei uma maçã", o que nos permite dizer que "a maçã que fotografei provavelmente é vermelha".

Nesta interpretação as probabilidades são vistas como sendo relações objetivas entre proposições, não dependendo da crença que cada indivíduo possui sobre um evento. Essa característica é um dos fatores que diferenciam as interpretações lógica e subjetiva de probabilidade.

Interpretação de Probabilidade Adotada neste Trabalho

A teoria e o cálculo de probabilidades não dependem da interpretação adotada, porém, devido a aplicabilidade das técnicas bayesianas em todo e qualquer tipo de problema e pela coerência de seus aspectos filosóficos, aqui será adotada a interpretação subjetiva de probabilidade.

Capítulo 3

Alguns Conceitos e Resultados em Probabilidade

A teoria de probabilidades desempenha um papel fundamental nas ciências, sendo possível aplicá-la nos mais diversos campos de pesquisa. Utilizando conceitos probabilísticos somos capazes de resolver problemas em áreas como genética, economia, inteligência artificial e finanças. Sendo assim, a compreensão de tais conceitos torna-se uma necessidade para a realização de praticamente qualquer tipo de estudo ou análise.

Este capítulo apresenta um breve resumo sobre a teoria de probabilidades, deixando de lado maiores formalismos da chamada teoria da medida. São mostradas definições e teoremas importantes para as construções realizadas no próximo capítulo, com o intuito de facilitar seu entendimento para um maior público.

3.1 Teoria de Conjuntos: Conceitos Básicos

Um conjunto S é uma coleção de objetos distintos s. Quando s é um membro do conjunto S escreve-se $s \in S$, por outro lado, quando s não está em S escrevemos $s \notin S$. Quando $S^{'}$ é um subconjunto de S escreve-se $S^{'} \subseteq S$, ou seja, para todo $s \in S^{'}$ temos $s \in S$. $S^{'}$ é chamado subconjunto próprio de S, escrito como $S^{'} \subset S$, se $S^{'} \subseteq S$ e existir $s \in S$ tal que $s \notin S^{'}$. Geralmente, conjuntos são grafados em letras maiúsculas e elementos de conjuntos em letras minúsculas.

Quando um conjunto não possui nenhum elemento ele é chamado conjunto vazio, sendo denotado por $\varnothing$. Dois conjuntos A_1 e A_2 são disjuntos se $A_1 \cap A_2 = \varnothing$, e são ditos iguais se $A_1 \subseteq A_2$ e $A_2 \subseteq A_1$. Os conjuntos A_j , $j = 1,2,...$ são mutuamente exclusivos, ou mutuamente disjuntos, se $A_i \cap A_j = \varnothing \ \ \forall i \neq j$.

Operações de Conjuntos

Conjuntos podem ser combinados de diferentes formas para produzir outro conjunto. Nesta subseção são apresentadas algumas operações básicas nas quais S será considerado um conjunto universal, ou espaço, sendo todos os outros conjuntos em questão somente seus subconjuntos.

Complementar: $A^c = \{s \in S : s \notin A\}$

União finita: $\bigcup_{j=1}^{n} A_j = \{s \in S : s \in A_j \text{ para pelo menos um } j = 1,2,...,n\}$

União infinita: $\bigcup_{j=1}^{\infty} A_j = \{s \in S : s \in A_j \text{ para pelo menos um } j = 1,2,...\}$

Intersecção finita: $\bigcap_{j=1}^{n} A_j = \{s \in S : s \in A_j \text{ para todo } j = 1,2,...,n\}$

Intersecção infinita: $\bigcap_{j=1}^{\infty} A_j = \{s \in S : s \in A_j \text{ para todo } j = 1,2,...\}$

Diferença: $A_1 - A_2 = \{s \in S : s \in A_1, s \notin A_2\}$

Diferença simétrica: $A_1 \Delta A_2 = (A_1 - A_2) \cup (A_2 - A_1)$

Propriedades das Operações

Nesta subseção são apresentadas algumas propriedades básicas das operações de conjuntos, sendo todas facilmente estendidas das operações apresentadas acima. Estas e outras propriedades são discutidas com mais detalhes por Roussas (1997).

1 $S^c = \varnothing$, $\varnothing^c = S$, $(A^c)^c = A$

2 $S \cup A = S$, $\varnothing \cup A = A$, $A \cup A^c = S$, $A \cup A = A$

3 $S \cap A = A$, $\varnothing \cap A = \varnothing$, $A \cap A^c = \varnothing$, $A \cap A = A$

4 (Associatividade) $A_1 \cup (A_2 \cup A3) = (A_1 \cup A_2) \cup A3$, $A_1 \cap (A_2 \cap A3) = (A_1 \cap A_2) \cap A3$

5 (Comutatividade) $A_1 \cup A_2 = A_2 \cup A_1$, $A_1 \cap A_2 = A_2 \cap A_1$

6 (Distributividade) $A \cap (\cup_j A_j) = \cup_j (A \cap A_j)$, $A \cup (\cap_j A_j) = \cap_j (A \cup A_j)$

3.2 Álgebras e σ-álgebras

Definição 3.2.1. Uma classe de subconjuntos de S é uma *álgebra*, escrita como $\mathcal{A}$, se:

(A1) $\mathcal{A}$ é uma classe não vazia.

(A2) $A \in \mathcal{A}$ implica que $A^c \in \mathcal{A}$.

(A3) $A_1, A_2 \in \mathcal{A}$ implica que $A_1 \cup A_2 \in \mathcal{A}$.

Como consequências da definição anterior temos que:

(i) $S, \varnothing \in \mathcal{A}$.

(ii) Se $A_j \in \mathcal{A}$, $j = 1,2,...,n$ então $\bigcup_{j=1}^{n} A_j \in \mathcal{A}$ e $\bigcap_{j=1}^{n} A_j \in \mathcal{A}$ para qualquer n finito.

Prova (Roussas (1997), p. 8)**.**

Definição 3.2.2. Uma classe de subconjuntos de S é uma *σ-álgebra*, escrita como $\mathcal{F}$, se:

(F1) $\mathcal{F}$ é uma classe não vazia.

(F2) $A \in \mathcal{F}$ implica que $A^c \in \mathcal{F}$.

(F3) $A_j \in \mathcal{F}$, $j = 1,2,...$ implica que $\bigcup_{j=1}^{\infty} A_j \in \mathcal{F}$.

As principais consequências da definição de σ-álgebra são:

(i) $A_j \in \mathcal{F}$, $j = 1,2,...$ implica que $\bigcap_{j=1}^{\infty} A_j \in \mathcal{F}$.

(ii) Toda σ-álgebra é, por definição, uma *álgebra*, porém o contrário não é verdadeiro.

Teorema 3.2.1. Seja C uma classe arbitrária de subconjuntos de S. Então há uma única *σ-álgebra* minimal $\mathcal{F}$ contendo C, e dizemos que $\mathcal{F}$ é a *σ-álgebra gerada* por C, ou $\mathcal{F} = \sigma(C)$. O par $(S, \mathcal{F})$ é chamado *espaço mensurável*.

Prova (Roussas (1997), p. 11)**.**

Experimentos e Eventos

Um *experimento* é qualquer processo no qual todos os possíveis resultados podem ser identificados antes de sua realização. Um *evento* é um subconjunto bem definido de possíveis resultados do *experimento*.

O conjunto de todos os resultados possíveis de um experimento é chamado *espaço amostral*, que aqui denotamos por S. Os elementos s do *espaço amostral* são chamados de *pontos amostrais*, e, como dito anteriormente, os subconjuntos de S são os chamados *eventos*. S e $\varnothing$ sempre são eventos. Se A é um evento, então A^c também é evento. Se A_j, $j = 1,2,...$ são eventos então $\cup_j^\infty A_j$ também é. Formalmente, os eventos associados ao *espaço amostral* formam uma σ-álgebra de subconjuntos desse espaço.

3.3 Probabilidade

Definição 3.3.1. Uma *medida de probabilidade* P é uma função que atribui para cada evento A um número denotado por $P(A)$, chamada de probabilidade de A. Tal função satisfaz as seguintes condições:

(P1) P é não negativa; $P(A) \geq 0$ para todo evento A.

(P2) P é normalizada; $P(S) = 1$.

(P3) P é σ-aditiva; $P\left(\cup_{j=1}^\infty A_j\right) = \sum_{j=1}^{\infty} P\left(A_j\right)$ para toda coleção de eventos mutuamente disjuntos A_j, $j = 1,2,....$

É simples mostrar que se S é finito então (P3) se reduz à:

(P3´) P é finitamente aditiva; $P\left(\cup_{j=1}^n A_j\right) = \sum_{j=1}^{n} P\left(A_j\right)$ para toda coleção de eventos mutuamente disjuntos A_j, $j = 1,2,...,n$.

A definição dada acima é conhecida como a definição axiomática de probabilidades, proposta por Kolmogorov, sendo a tripla $(S,\mathcal{F},P)$ chamada de *espaço de probabilidade*.

3.4 Probabilidade Condicional

Suponha que sabemos que um evento B ocorreu e que queremos saber a probabilidade de outro evento A acontecer, mas levando em conta nosso conhecimento sobre a ocorrência de B. Sendo assim, o que realmente queremos é atualizar nosso conhecimento sobre A baseando-se na ocorrência de outro evento. Essa atualização pode ser calculada através da *probabilidade condicional de A dado que B ocorreu*.

Definição 3.4.1. Seja B um evento tal que $P(B) > 0$. Então a *probabilidade condicional*, dado B, é a função escrita como $P(\cdot|B)$ e definida para todo evento A da seguinte forma:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

A função $P(\cdot|B)$ satisfaz os axiomas de Kolmogorov, portanto, é uma medida de probabilidade (Roussas (1997), p. 22).

Em alguns experimentos é relativamente fácil atribuir probabilidades condicionais, o que torna possível calcular a ocorrência simultânea dos eventos através da regra da multiplicação, apresentada abaixo.

Teorema 3.4.1 (Regra da Multiplicação)**.** Suponha que $A_1, A_2, ..., A_n$ sejam eventos tais que $P(A_1 \cap A_2 \cap ... \cap A_{n-1}) > 0$. Então

$$P(A_1 \cap A_2 \cap ... \cap A_n) = P(A_1)\, P(A_2|A_1)\, P(A_3|A_1 \cap A_2) ... P(A_n|A_1 \cap A_2 \cap ... \cap A_{n-1})$$

Prova. Pela definição de probabilidade condicional, podemos reescrever o lado direito da equação acima como

$$P(A_1).\frac{P(A_1 \cap A_2)}{P(A_1)}.\frac{P(A_1 \cap A_2 \cap A_3)}{P(A_1 \cap A_2)} ... \frac{P(A_1 \cap A_2 \cap ... \cap A_n)}{P(A_1 \cap A_2 \cap ... \cap A_{n-1})}$$

Como $P(A_1 \cap A_2 \cap ... \cap A_{n-1}) > 0$, cada um dos denominadores neste produto são positivos. Todos os termos do produto se cancelam, exceto o último numerador $P(A_1 \cap A_2 \cap ... \cap A_n)$, o que prova o teorema.

Definição 3.4.2. Seja S o espaço amostral de um experimento, e considere n eventos $B_1, ..., B_n \in S$ tais que $B_1, ..., B_n$ são disjuntos e $\bigcup_{j=1}^{n} B_j = S$. Diz-se que esses eventos formam uma *partição* de S.

Teorema 3.4.2 (Lei da Probabilidade Total)**.** Suponha que os eventos $B_1, ..., B_n$ formem uma partição do espaço S e $P(B_j) > 0 \ \ \forall j = 1, 2, ..., n$. Então, para todo evento $A \in S$,

$$P(A) = \sum_{j=1}^{n} P(B_j) P(A|B_j)$$

Prova (DeGroot & Schervish (2012), p. 60)**.**

3.5 Teorema de Bayes

Suponhamos que nosso interesse seja de calcular a probabilidade de cada um dos eventos disjuntos $B_1,...,B_n$ acontecer, e que temos a possibilidade de observar a ocorrência de outro evento A. Se tivermos $P(A|B_j)$ para cada j podemos calcular, usando o Teorema de Bayes, as probabilidades $P(B_j|A)$, também para cada j.

Teorema 3.5.1 (Teorema de Bayes)**.** Se $\{B_j, j = 1,2,...,n\}$ é uma partição de S, se $P(B_j) > 0 \ \forall j = 1,2,...,n$ e se existe um evento A com $P(A) > 0$, então

$$P(B_j|A) = \frac{P(A|B_j)\,P(B_j)}{P(A)}$$

Prova (DeGroot & Schervish (2012), p. 77)**.**

O teorema de Bayes nos dá uma forma de computar a probabilidade condicional de cada evento numa partição, dada a observação de um evento A.

3.6 Variáveis Aleatórias

O conceito de variável aleatória é um dos mais importantes para análises estatísticas. Trata-se de uma ferramenta para modelar quantidades desconhecidas, sendo que para cada variável aleatória X e cada conjunto C de números reais, podemos calcular a probabilidade de X assumir valores de C. A coleção de todas essas probabilidades é a distribuição de X.

Definição 3.6.1. Considere um experimento com espaço amostral S. Uma *variável aleatória* X é uma função que atribui um valor real para cada resultado de S. Para qualquer conjunto de números reais C, a probabilidade de X assumir um valor contido em C é a mesma do resultado do experimento estar contida em $X^{-1}(C)$. Ou seja,

$$P(X \in C) = P(X^{-1}(C))$$

onde $X^{-1}(C)$ é o evento que consiste de todos os pontos $s \in S$ tais que $X(s) \in C$. Matematicamente escrevemos que $X^{-1}(C) = \{s \in S : X(s) \in C\}$.

Função de Distribuição Acumulada

A *função de distribuição (acumulada)* F de uma variável aleatória é definida para todo valor real x como sendo

$$F(x) = P(X \leq x) = P(X \in (-\infty, x]).$$

A *função de distribuição (acumulada) conjunta* das variáveis aleatórias X e Y é definida como

$$F(x,y) = P(X \leq x, Y \leq y).$$

Sendo que $F_x(x) = \lim_{y\to\infty} F(x,y)$ e, analogamente, $F_y(y) = \lim_{x\to\infty} F(x,y)$.

Variáveis Aleatórias Discretas

Dizemos que uma variável aleatória (v.a.) X é *discreta* se o seu conjunto de possíveis valores for contável. Se uma v.a. X é discreta, a *função de probabilidade* de X é definida como a função f tal que para cada valor x,

$$f(x) = P(X = x).$$

Teorema 3.6.1. Se X é um v.a. discreta e S o espaço amostral, a probabilidade de cada subconjunto $C \in S$ pode ser calculada por

$$P(X \in C) = \sum_{x_i \in C} f(x_i).$$

Além disso,

$$P(X \in S) = \sum_{x_i \in S} f(x_i) = 1.$$

Variáveis Aleatórias Contínuas

Uma variável aleatória X é *contínua* se o conjunto de seus possíveis valores for não-enumerável. Em outras palavras, X é uma v.a. *contínua* se existir uma função f, chamada *função densidade de probabilidade*, tal que

$$P(X \in C) = \int_C f(x)\,dx$$

para todo conjunto C. Além disso, como $F(x) = \int_{-\infty}^{x} f(x)\,dx$, temos que

$$f(x) = \frac{d}{dx}F(x).$$

Distribuições Conjuntas

Sejam X_1, X_2, ..., X_n variáveis aleatórias. A *distribuição conjunta* de X_1, X_2, ..., X_n é a coleção de todas as probabilidades da forma $P\{(X_1, X_2, ..., X_n) \in C\}$ para todo conjunto C, de vetores de números reais, tal que $\{(X_1, X_2, ..., X_n) \in C\}$ seja um evento.

A *função de probabilidade conjunta* de duas variáveis aleatórias discretas X e Y é definida como a função f tal que, para todo ponto (x, y) no plano xy,

$$f(x, y) = P(X = x, Y = y).$$

No caso de X e Y serem variáveis aleatórias contínuas, dizemos que elas possuem uma *distribuição conjunta contínua* se houver uma função não-negativa f tal que a integral

$$P\{(X, Y) \in C\} = \int_C \int f(x, y)\, dx\, dy$$

exista. Neste caso, f é chamada de *função de densidade conjunta* de X e Y. O conjunto definido como $\{(x, y) : f(x, y) > 0\}$ é chamado de *suporte da distribuição* f.

Distribuições Condicionais

A distribuição condicional de uma variável aleatória X dada a observação de um valor y, de outra variável aleatória Y, é a distribuição que usaríamos para X se soubéssemos que $Y = y$. Nesse caso é comum considerar que Y comporta-se como uma constante de valor y.

Sejam X e Y variáveis aleatórias discretas com função de probabilidade conjunta $f_{X,Y}$. Seja f_Y a função de probabilidade (marginal) de Y. Para cada y tal que $f_Y(y) > 0$ definimos

$$f_{X|Y} = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

como sendo a *função de distribuição condicional de X dado $Y = y$*.

Analogamente, sejam X e Y variáveis aleatórias contínuas com função densidade conjunta $f_{X,Y}$ e marginais f_X e f_Y, respectivamente. Seja y um valor tal que $f_Y(y) > 0$. Definimos a função

$$f_{X|Y} = \frac{f_{X,Y}(x, y)}{f_Y(y)} \qquad \text{para } -\infty < x < \infty$$

como sendo a *função densidade de probabilidade condicional de X dado $Y = y$*.

3.7 Independência

O conceito de independência é de fundamental importância nas áreas de probabilidade e estatística. Utilizando esse conceito é possível construir modelos que tenham flexibilidade para agregar informações de grandes bases de dados e/ou partindo do conhecimento de um especialista.

Definição 3.7.1 (Independência entre eventos)**.** Dois eventos A e B são independentes se

$$P(A \cap B) = P(A)\,P(B)$$

ou se, equivalentemente,

$$P(A|B) = P(A)$$

Definição 3.7.2 (Independência entre variáveis aleatórias)**.** Duas variáveis aleatórias X e Y são independentes se a densidade conjunta do par fatorar como sendo o produto de suas marginais:

$$f_{X,Y}(x,y) = f_X(x)\,f_Y(y)$$

ou se, equivalentemente, a densidade condicional de, por exemplo, Y dado $X = x$ não for uma função de X, nos permitindo escrever que

$$f_{Y|X}(y|x) = f_Y(y).$$

3.8 Independência Condicional

Antes de introduzirmos as ideias centrais da teoria de grafos, apresentadas no próximo capítulo, precisamos conhecer o conceito de *independência condicional*, que é um dos pilares de todo o ferramental que será apresentado neste trabalho.

Definição 3.8.1 (Independência condicional entre variáveis aleatórias)**.** Sejam $\mathbf{X}$, $\mathbf{Y}$ e $\mathbf{Z}$ conjuntos de variáveis aleatórias. Se, para cada valor $\mathbf{z}$, as distribuições condicionais de $\mathbf{X}$ e $\mathbf{Y}$ dado $\mathbf{Z} = \mathbf{z}$ forem independentes, então dizemos que $\mathbf{X}$ e $\mathbf{Y}$ são condicionalmente independentes dado $\mathbf{Z}$, ou seja, $\mathbf{X}$ e $\mathbf{Y}$ são condicionalmente independentes dado $\mathbf{Z}$ se, e somente se,

$$f_{X,Y|Z}(x,y|z) = f_{X|Z}(x|z)\,f_{Y|Z}(y|z).$$

Existem diversos símbolos para indicar independência entre variáveis, tendo cada autor suas próprias preferências. Aqui será adotada a notação proposta por Dawid (1979):

X e Y (marginalmente) independentes: $X \perp\!\!\!\perp Y$

X e Y independentes dado Z: $X \perp\!\!\!\perp Y \mid Z$

Uma forma de enxergar a independência condicional é encarando-a como uma indicadora de irrelevância, no sentido de que podemos interpretar a expressão $X \perp\!\!\!\perp Y \mid Z$ como se ela estivesse dizendo: se temos conhecimento sobre Z, uma informação sobre Y não altera nosso conhecimento sobre X.

3.9 Procedimento Bayesiano

Suponha que n variáveis aleatórias $X_1, ..., X_n$ formem uma amostra de uma distribuição $f(x|\theta)$ e que $\pi(\theta)$ seja nossa crença, a priori, sobre θ. Então, aplicando o teorema de Bayes, nossa distribuição a posteriori de θ é

$$\pi(\theta \mid x_1, ..., x_n) = \frac{f(x_1, ..., x_n \mid \theta)\, \pi(\theta)}{m(x_1, ..., x_n)}$$

onde

$$m(x_1, ..., x_n) = \int f(x_1, ..., x_n \mid \theta)\, \pi(\theta) d\theta\,.$$

A distribuição a priori, $\pi(\theta)$, expressa nossa crença inicial sobre θ, e a distribuição a posteriori, $\pi(\theta \mid x_1, ..., x_n)$, expressa nossa crença revisada sobre θ, depois da observação dos dados. Como $m(x_1, ..., x_n)$ é uma constante, podemos simplificar o procedimento como sendo

$$\pi(\theta \mid x_1, ..., x_n) \propto L(\theta)\, \pi(\theta)$$

no qual $L(\theta) = f(x_1, ..., x_n \mid \theta)$ é a chamada *função de verossimilhança* de θ dado $x_1, ..., x_n$.

O procedimento bayesiano pode ser usado, por exemplo, na avaliação de modelos na área de aprendizagem automática, mais conhecida pelo nome em inglês, *machine learning*, em que temos uma distribuição a priori para a v.a. M, que representa um conjunto de possíveis modelos, e uma função de verossimilhança $P(D = d|M)$, onde D é uma v.a. expressando os dados. Sendo assim, é possível avaliar a aderência de diferentes modelos à um conjunto de dados através da regra de Bayes

$$P(M|d) \propto P(M)P(d|M).$$

3.10 Coerência

A tomada de decisão sob incerteza, observada em praticamente todas as atividades humanas, baseia-se na escolha de uma ação cujas consequências não são totalmente previsíveis, pois os eventos que ocorrerão no futuro podem afetar as consequências das ações tomadas agora. De Finetti (1931) propõe uma justificativa para usar o cálculo de probabilidades como forma de quantificar a incerteza em situações de tomada de decisão. Ele propõe que os axiomas de probabilidade sejam justificados, diferentemente das ideias frequentistas, pelo uso de um requisito de racionalidade conhecido como coerência. O conceito de coerência está relacionado com decisões que evitem a chamada "perda certa". De Finetti defende sua definição de coerência através do chamado *Dutch Book argument*.

Como comentado por Robert (2011), há diferentes definições formais de coerência, mas segundo Robins e Wasserman (2000), que compararam definições surgidas entre as décadas de 50 e 90, a conclusão de vários autores foi sempre a mesma: um procedimento é coerente se, e somente se, adotar a interpretação subjetiva de probabilidade. A seguir será apresentada, resumidamente, a versão de coerência defendida pelo *Dutch Book argument* de De Finetti (1931), analisada com maior detalhamento por Loschi e Wechsler (2002).

Dutch Book Argument

Suponha que Θ seja o conjunto de todos os possíveis valores θ_j para um estado da natureza[1] θ, e que, por simplicidade, Θ seja finito. De Finetti diz que a medida de incerteza de uma pessoa sobre o valor θ_j para θ é o número $P_{(\theta_j)}$, com a qual ela sente-se indiferente entre possuir uma quantia monetária de $P_{(\theta_j)}.S_{(\theta_j)}$ e receber um bilhete que valerá $S_{(\theta_j)}$ se for revelado que $\theta = \theta_j$, ou que, caso $\theta \neq \theta_j$, valerá zero.

Para ilustrar melhor a situação, perceba que $\Theta = \{\theta_1, \theta_2, ..., \theta_n\}$ e que há $r = 2^n$ subconjuntos[2] de Θ, ou seja, há r eventos possíveis. Suponha que exista um bilhete E_i para cada um desses r eventos, com $i = 1, 2, ..., r$, e que um bilhete do tipo E_i dá direito a um prêmio no valor de $S_{(E_i)}$ caso ocorra o evento E_i. Sendo assim, se você possuir o bilhete E_i você receberá o prêmio $S_{(E_i)}$ ou não receberá nada, caso o evento E_i não ocorra. Sendo B o valor do bilhete, temos que seu ganho será:

$$B = \begin{cases} S_{(E_i)} & \text{se } E_i \text{ ocorrer} \\ 0 & \text{caso contrário.} \end{cases}$$

[1] A expressão "estado da natureza" é utilizada aqui como sinônimo de variável aleatória.

[2] $r = \sum_{i=0}^{n} \binom{n}{i} = 2^n$.

O jogo usado no *Dutch Book Argument* segue da seguinte forma: um agenciador de apostas força-o a participar de um jogo, no qual você deve escolher c bilhetes entre r disponíveis, com $0 < c \leq r$. Você deve atribuir taxas $P_{C_1}, P_{C_2}, ..., P_{C_c}$ para cada um dos bilhetes escolhidos, de forma que você ache justo pagar o preço $P_{C_i}.S_{C_i}$ pelo bilhete C_i. Lembrando que tanto as taxas P_{C_i} escolhidas por você quanto os prêmios S_{C_i} podem ser negativos, então você pode, ao invés de pagar, receber um valor para possuir o bilhete, pois ele também pode representar um ônus, e não apenas um bônus. Porém, os valores S_{C_i} só serão revelados pelo agenciador após você anunciar as taxas P_{C_i}, só então você irá pagar, ou receber, as quantias $P_{C_i}.S_{C_i}$. Com isso, De Finetti apresenta uma metáfora para situações de tomada de decisões sob incerteza, e daí vem sua definição de coerência.

Definição 3.10.1 (Coerência)**.** Você é coerente no tempo t sempre que as sua taxas $P_{C_1}, P_{C_2}, ..., P_{C_c}$ sejam tais que, para cada conjunto de valores $\{S_{C_i}, i = 1, 2, ..., c\}$ que possam ser escolhidos pelo agenciador, haja ao menos um θ_j para o qual seu saldo final seja não negativo.

Definição 3.10.2 (Incoerência)**.** Você é incoerente no tempo t sempre que não for coerente nesse tempo.

Exemplo 3.10.1. Suponha que o jogo do agenciador de apostas seja baseado no arremesso de um dado de seis faces, numeradas de 1 a 6. Agora suponha que o jogador escolha o evento E_2 referente ao dado aterrissar com a face de número 2 voltada para cima, e atribua a taxa $P_{(E_2)} = 0,5$. Depois disso o agenciador anuncia o valor do seu bilhete, $S_{(E_2)} = \$10$, então o jogador paga \$5 para obter o bilhete. Neste caso, se o resultado do arremesso for a face de número 2 do dado, o jogador ganha \$10, tendo lucro de \$5, porém, se o resultado do dado for qualquer outro diferente de 2, o jogador ganha \$0, tendo prejuízo de \$5.

No exemplo acima a escolha de $P_{(E_2)} = 0,5$ foi coerente, já que existe uma chance do saldo final ser não negativo. Mais do que isso, mesmo que o valor declarado pelo agenciador tivesse sido $S_{(E_2)} = -\$10$, ainda assim o jogador teria sido coerente, pois ele teria recebido \$5 para ficar com o bilhete e, no caso do dado ter resultado diferente de 2, o jogador teria saldo final positivo. Sendo assim, podemos dizer que ser coerente não depende do agenciador de apostas, mas sim somente do jogador.

Teorema 3.10.1. Você é coerente no tempo t se, e somente se, suas medidas de incerteza satisfizerem os seguintes axiomas:

(1) $P_{(\theta_j)} \geq 0 \quad \forall j = 1, ..., n.$

(2) $\sum_{j=1}^{n} P_{(\theta_j)} = 1.$

(3) $P_{(E)} = \sum_{j:\theta_j \in E} P_{(\theta_j)}$, para todo $E \subseteq \Theta$.

Prova (Loschi & Wechsler (2002), p. 174 - 176)**.**

É fácil perceber que, para Θ finito, os axiomas acima são equivalentes aos da seção 3.3. Sendo assim, esse teorema traz um resultado muito forte, afirmando que você é coerente no tempo t se, e somente se, suas medidas de incerteza declaradas em t respeitarem os axiomas de Kolmogorov.

Há uma construção, muito similar à mostrada acima, para a coerência condicional, que conclui que você é condicionalmente coerente no tempo t se, e somente se, satisfizer os axiomas de Kolmogorov e a regra da multiplicação, o que justifica o uso do teorema de Bayes como forma coerente de conectar probabilidades condicionais e incondicionais declaradas no mesmo tempo t (Loschi e Wechsler (2002), p. 178).

3.11 Atualização Temporal de Probabilidades

Como visto na seção anterior, podemos justificar o uso do teorema de Bayes para conectar probabilidades condicionais e incondicionais através da definição de coerência. Entretanto, não existem definições de coerência temporal que sugiram o uso dessa mesma regra. Sendo assim, nada garante que usar o teorema de Bayes para conectar probabilidades a priori e a posteriori seja uma forma coerente de atualização de probabilidades.

Devido à inexistência de uma definição de coerência temporal, você pode escolher, arbitrariamente, qualquer procedimento de atualização de probabilidades para obter uma distribuição a posteriori de sua priori. A possibilidade mais utilizada é a própria aplicação da regra de Bayes, porém, nada o impede de usar outros métodos ou, até mesmo, "começar do zero", construindo sua posteriori a partir de uma completa reavaliação de suas probabilidades (Diaconis e Zabell (1982)).

Ao utilizar a regra de Bayes para a atualização de probabilidades você estará usando o chamado *Procedimento de Condicionamento Bayesiano*. Denotemos por P e P^* as taxas anunciadas, respectivamente, nos tempos t e t^*, com $t < t^*$, e assumamos coerência estática em ambos os momentos t e t^*. Seja $\mathcal{F}$ a σ-álgebra das partes de Θ.

Definição 3.11.1 (Procedimento de Condicionamento Bayesiano)**.** Realizar o procedimento de condicionamento bayesiano é anunciar, no tempo t^*, que

$$P^*(E) = P(E|A), \ \forall E \in \mathcal{F},$$

onde A é $\mathcal{F}$-mensurável com $P(A) > 0$.

O condicionamento bayesiano é a base de praticamente todas as aplicações da Inferência Bayesiana, porém, seu uso só faz sentido quando, no tempo t^*, $P^*(A) = 1$ e $P^*(.|A) = P(.|A)$, como mostra o teorema seguinte, proposto por Jeffrey (1992).

Teorema 3.11.1. Seja $A \in \mathcal{F}$ com $P(A) > 0$. Realizar o procedimento de condicionamento bayesiano é anunciar que

$$P^*(A) = 1 \quad \text{e} \quad P^*(.|A) = P(.|A)$$

no tempo t^*.

Prova (Loschi & Wechsler (2002), p. 179)**.**

Outra possibilidade para construir probabilidades a posteriori é utilizando a chamada *Regra de Jeffrey*, que permite a atualização temporal mesmo em casos onde $0 \leq P^*(A) \leq 1$.

Definição 3.11.2 (Regra de Jeffrey)**.** Seja $\mathcal{A} = \{A_i\}_{i \geq 1}$ uma partição de Θ. Aplicar a regra de Jeffrey é anunciar, no tempo t^*, que

$$P^*(E) = \sum_{i \geq 1} P(E|A_i)\, P^*(A_i)\,, \quad E \in \mathcal{F},$$

onde $P^*(A_i) \geq 0$ para todo $i \geq 1$ e $\sum_{i \geq 1} P^*(A_i) = 1$.

Teorema 3.11.2. Seja $\mathcal{A} = \{A_i\}_{i \geq 1}$ uma partição de Θ, e assumamos que $P^*(A_i) \geq 0$ para todo $i \geq 1$ e $\sum_{i \geq 1} P^*(A_i) = 1$. Aplicar a regra de Jeffrey é equivalente a anunciar o chamado J-Condicionamento:

$$P^*(E|A_i) = P(E|A_i), \;\; E \in \mathcal{F}, \;\; \forall i$$

Prova (Loschi & Wechsler (2002), p. 181)**.**

Para facilitar a compreensão dos métodos expostos acima, consideremos dois exemplos dados por Jeffrey (1965) e explorados com mais detalhes por Loschi e Wechsler (2002).

Exemplo 3.11.1. Você está num quarto escuro e não se lembra da cor de seu pijama, construindo um espaço amostral $\Theta = \{\text{verde, azul, branco}\}$ com as seguintes probabilidades a priori: $P(\text{verde}) = P(\text{branco}) = 0,3$ e $P(\text{azul}) = 0,4$. Considere o evento $A = \{\text{azul, branco}\}$. Então, o teorema de Bayes nos dá as seguintes probabilidades condicionais $P(.|A)$:

$$\begin{gathered} P(\text{verde}|A) = 0 \\ P(\text{branco}|A) = \tfrac{3}{7} \\ P(\text{azul}|A) = \tfrac{4}{7} \end{gathered}$$

Suponha que após um relâmpago iluminar o quarto por uma fração de segundo, você considere que $P^*(A) = 1$, então, neste caso, você pode realizar o condicionamento bayesiano, que dá nas seguintes probabilidades a posteriori:

$$P^*(\text{verde}) = 0$$
$$P^*(\text{branco}) = \frac{3}{7}$$
$$P^*(\text{azul}) = \frac{4}{7}$$

que são exatamente iguais às probabilidades condicionais antes do relâmpago acontecer.

Exemplo 3.11.2. Agora suponha que o relâmpago do exemplo anterior não tenha sido de tanta ajuda, e que você tenha considerado $P^*(A) = 0,8$. Neste caso, você não pode realizar o condicionamento bayesiano, porém, aplicando a regra de Jeffrey temos as seguintes probabilidades a posteriori:

$$P^*(\text{verde}) = \frac{7}{35}$$
$$P^*(\text{branco}) = \frac{12}{35}$$
$$P^*(\text{azul}) = \frac{16}{35}$$

Pelos exemplos acima podemos ver que diferentes procedimentos de atualização temporal podem gerar diferentes distribuições a posteriori, ainda que exatamente as mesmas informações estejam disponíveis. Além disso, já que não existe uma formulação normativa para atualizações temporais de probabilidades, uma possibilidade no exemplo acima seria uma total reavaliação de suas crenças, fazendo, após o relâmpago, a distribuição a posteriori como sendo, por exemplo, $P^*(\text{verde}) = 0,1$, $P^*(\text{branco}) = 0,2$ e $P^*(\text{azul}) = 0,7$, sem conectá-la matematicamente com a priori.

Capítulo 4

Redes Probabilísticas

Redes probabilísticas são modelos gráficos que representam interações entre variáveis aleatórias, podendo tais relações serem vistas como simples conjuntos de dependências ou como associações de causa-efeito, dependendo da construção e interpretação de cada modelo.

Em geral, a construção de redes probabilísticas faz uso da chamada teoria de grafos, na qual as variáveis aleatórias são representadas como vértices e as interações entre elas como arcos. Distribuições de probabilidades conjuntas podem ser representadas naturalmente através desses modelos, onde as presenças ou ausências de arcos representam as relações de dependência ou independência entre as variáveis.

4.1 Grafos

A teoria de grafos pode ser tratada como um assunto puramente matemático e abstrato, porém, seu grande trunfo é sua capacidade de representar opiniões e conhecimentos de forma visual, mostrando-se uma teoria muito útil para vários tipos de aplicações, o que a coloca como tema central na construção de redes probabilísticas.

4.1.1 Grafos Não Direcionados

Um grafo $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ é uma estrutura que consiste de um conjunto finito de vértices $\mathcal{V}$, representando as variáveis do modelo, e de um conjunto finito de arcos $\mathcal{E}$, posicionados entre os vértices.

Os vértices de um grafo são escritos como letras (X, Y, A, B, etc.) e os arcos são representados como pares da forma $[XY]$. No caso de *grafos não direcionados*, como os mostrados nas figuras 4.1 e 4.2, a expressão $[XY]$ é equivalente à $[YX]$,

pois o arco em questão não possui um direcionamento 'de X para Y' nem 'de Y para X', somente conecta dois vértices de forma igualitária.

É útil manter em mente que a nomenclatura usada na literatura nem sempre se repete. Muitas vezes o termo 'vértice' é substituído por 'nó', e a expressão 'arco' pode ser trocada por 'aresta' ou 'conexão'.

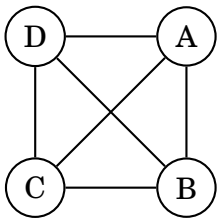


Figura 4.1: Um grafo completo

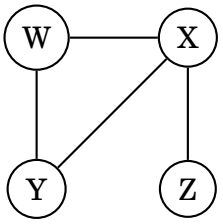


Figura 4.2: Um grafo incompleto

A construção adotada neste trabalho, proposta por Edwards (2000), não permite a existência de arcos da forma $[XX]$, ou seja, um vértice não pode estar conectado a si mesmo (figura 4.3). Outra característica desta construção é a existência de, no máximo, um arco entre dois vértices, não sendo permitida uma conexão dupla (figura 4.4).

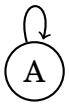


Figura 4.3: Um grafo mal construído, com vértice "auto-conectado"

Figura 4.4: Um grafo mal construído, com dupla conexão

Dizemos que dois vértices $X, Y \in \mathcal{V}$ são *adjacentes*, escrevendo $X \sim Y$, se existe um arco entre eles. Um grafo é chamado *completo* se há arcos conectando cada par de vértices, como na figura 4.1. Um grafo *incompleto* (figura 4.2) é aquele que tem pelo menos um par de vértices não adjacentes.

Qualquer subconjunto $u \subseteq \mathcal{V}$ induz um *subgrafo* de $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, isto é, o grafo $\mathcal{G}_u = (u, \mathcal{F})$ no qual o conjunto $\mathcal{F}$ consiste de todos os arcos em $\mathcal{E}$ em que ambas as extremidades estão em u. Um subconjunto $u \in \mathcal{V}$ é dito *completo* se ele induz um subgrafo completo. Em outras palavras, se todos os vértices em u são mutuamente adjacentes, então u é completo. Na figura 4.2 o subconjunto $\{X, Y, W\}$ é completo.

Um subconjunto $u \subseteq \mathcal{V}$ é chamado de *clique* se for completo maximal, ou seja, u é um clique se for completo e se $u \subset w$ implicar em w ser incompleto. No grafo da figura 4.2 há dois cliques: $\{X, Y, W\}$ e $\{X, Z\}$.

Uma sequência de vértices $X_0, X_1, ..., X_n$ tal que $X_{i-1} \sim X_i$ para $i = 1, ..., n$ é chamada de *caminho* entre X_0 e X_n de comprimento n. Tomando o exemplo da

figura 4.2, temos que Z, X, Y, W é um caminho de comprimento 3 entre Z e W e que Z, X, W é uma caminho de comprimento 2 entre os mesmos vértices Z e W. Um grafo é dito *conectado* se houver um caminho entre todo par de vértices.

Para três subconjuntos A, B e C de $\mathcal{V}$, dizemos que C *separa* A de B se todos os caminhos de A para B passarem por C. Por exemplo, na figura 4.2, $\{X\}$ separa $\{Y, W\}$ de $\{Z\}$.

A *fronteira* de um subconjunto $u \subseteq \mathcal{V}$, escrita[1] como $bd(u)$, é definida como sendo o conjunto de vértices pertencentes à $\mathcal{V} \setminus u$ que são adjacentes aos vértices em u, ou seja, $bd(u) = \{W \in \mathcal{V} \setminus u : W \sim X, X \in u\}$. Como exemplo, na figura 4.2, temos que $bd(\{X, Y, W\}) = \{Z\}$.

Segundo Edwards (2000), a utilização de grafos não direcionados tem como um de seus principais objetivos modelar relações de independência condicional da forma $X \perp\!\!\!\perp Y \mid$ (todo o resto), na qual com a expressão "todo o resto", nos referimos à todas as outras variáveis no modelo. Tais relações são indicadas no grafo da seguinte maneira: para todos os pares $\{X, Y\}$ tal que $X \perp\!\!\!\perp Y \mid$ (todo o resto), os vértices X e Y são não adjacentes. Sendo assim, do grafo não direcionado resultante, podemos ver que se duas variáveis são não adjacentes então elas são condicionalmente independentes dado todo o resto. Dizemos que um grafo com essa característica satisfaz a *propriedade markoviana pareada para grafos não direcionados*.

Outra característica dessa modelagem é a chamada *propriedade markoviana local para grafos não direcionados*, que diz que cada variável X é condicionalmente independente de seus não-vizinhos dados seus vizinhos (onde podemos substituir 'vizinhos' por 'vértices da fronteira de X'). Formalmente, essa propriedade é satisfeita se, para todo $X \in \mathcal{V}$, $X \perp\!\!\!\perp \mathcal{V} \setminus \{X \cup bd(X)\} \mid bd(X)$.

Uma outra propriedade, mais abrangente do que as markoviana pareada e local, é a chamada *propriedade markoviana global para grafos não direcionados*, que é satisfeita quando valer a seguinte relação: se dois conjuntos de variáveis $\mathbf{U}$ e $\mathbf{V}$ são separados por um terceiro conjunto $\mathbf{W}$, então $\mathbf{U} \perp\!\!\!\perp \mathbf{V} \mid \mathbf{W}$.

Sintetizando, como feito por Lauritzen (1996), temos que uma distribuição de probabilidade P sobre um grafo não direcionado $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ satisfaz:

[1] Para evitar confusões, será mantida a nomenclatura proveniente da palavra inglesa *boundary*, motivo da notação $bd(u)$.

(P) a *propriedade markoviana pareada para grafos não direcionados* se

$$X \nsim Y \Rightarrow X \perp\!\!\!\perp Y \mid \mathcal{V} \setminus \{X, Y\}$$

(L) a *propriedade markoviana local para grafos não direcionados* se

$$\forall X \in \mathcal{V} : X \perp\!\!\!\perp \mathcal{V} \setminus \{X \cup bd(X)\} \mid bd(X)$$

(G) a *propriedade markoviana global para grafos não direcionados* se

$$(X, Y, Z) \in \mathcal{V} \text{ disjuntos com } X, Y \text{ separados por } Z \Rightarrow X \perp\!\!\!\perp Y \mid Z$$

Lauritzen (1996) mostra que, para toda distribuição de probabilidade, vale a seguinte relação:

Proposição 4.1.1. Para qualquer grafo não direcionado $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ e qualquer distribuição de probabilidade sobre $\mathcal{V}$ vale que

$$\text{(G)} \Rightarrow \text{(L)} \Rightarrow \text{(P)}$$

Prova (Lauritzen (1996), p. 33)**.**

Pearl e Paz (1987) vão ainda mais longe mostrando que, sob certa condições, as propriedades markovianas pareada, local e global são equivalentes.

Proposição 4.1.2 (Pearl e Paz)**.** Sejam $A, B, C, D \in \mathcal{V}$ conjuntos disjuntos. Se uma distribuição de probabilidade sobre $\mathcal{V}$ é tal que vale a relação

$$A \perp\!\!\!\perp B \mid (C \cup D) \ \wedge \ A \perp\!\!\!\perp C \mid (B \cup D) \ \Rightarrow \ A \perp\!\!\!\perp (B \cup C) \mid D$$

Então

$$\text{(G)} \Leftrightarrow \text{(L)} \Leftrightarrow \text{(P)}$$

Prova (Lauritzen (1996), p. 34)**.**

4.1.2 Grafos Direcionados

Modelos de grafos direcionados são usados para representar relações de influência e, sob certas suposições, relações de causa-efeito. Para ilustrar e motivar o uso desses modelos consideremos o seguinte exemplo, adaptado de Edwards (2000):

Exemplo 4.1.1 (Pesquisa de mercado)**.** Suponhamos que temos uma consultoria de pesquisa de mercado, e que um fabricante de iogurte de frutas nos contrate para analisar a chance de sucesso de um novo produto, nos pedindo que perguntemos para diversos indivíduos a seguinte questão: 'você gosta de iogurte?' - com

apenas duas possíveis respostas, sim ou não. Outro pedido do cliente é que usemos somente as variáveis binárias *Gênero* (masculino ou feminino) e *Faixa Etária* (criança ou adulto) para estudar a aceitação do produto.

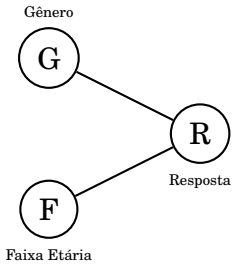


Figura 4.5: Um grafo não direcionado mostrando que $G \perp\!\!\!\perp F \mid R$

Sendo assim, poderíamos usar o grafo não direcionado da figura 4.5 para representar esse modelo, que nos diria, sob a suposição de satisfazer as condições markovianas da seção anterior, que {Gênero $\perp\!\!\!\perp$ Faixa Etária | Resposta}, o que, obviamente, não faz sentido, já que saber se uma pessoa gosta ou não de iogurte de frutas dificilmente traria alguma informação sobre seu gênero ou faixa etária. As variáveis desse modelo podem ser consideradas marginalmente independentes, mas não condicionalmente independentes, e é aí que o uso de grafos direcionados será útil, por serem modelos capazes de exprimir independências incondicionais. A relação {Gênero $\perp\!\!\!\perp$ Faixa Etária} pode ser representada através de um grafo direcionado, como na figura 4.6.

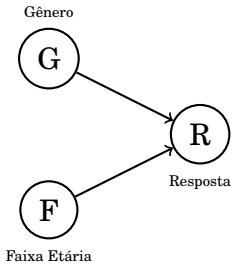


Figura 4.6: Um grafo direcionado mostrando que $G \perp\!\!\!\perp F$

Grafos direcionados contém apenas arcos com sentidos definidos, sendo representados por setas. Sendo assim, podemos definir um grafo direcionado como sendo um par $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, onde $\mathcal{V}$ é um conjunto de vértices e $\mathcal{E}$ é um conjunto de arestas direcionadas, ou setas, que serão representadas por pares ordenados de vértices. Se houver uma seta apontando de X para Y escreveremos $X \rightarrow Y$ ou,

equivalentemente, $[XY] \in \mathcal{E}$. Diferentemente dos grafos não direcionados, agora $[XY]$ não é mais o mesmo que $[YX]$.

Se $X \to Y$ ou $Y \to X$ dizemos que X e Y são *adjacentes*, podendo tal relação ser representada, como no caso não direcionado, por $X \sim Y$. Chamamos uma sequência de vértices $\{X_1,...,X_k\}$ de *caminho* quando $X_i \sim X_{i+1}$ para cada $i = 1,...,k-1$. Um *caminho direcionado* de X_1 até X_k, por sua vez, é um caminho no qual, para cada $i = 1,...,k-1$, tivermos $X_i \to X_{i+1}$. Quando o primeiro e o último vértices coincidirem num caminho direcionado, ou seja, quando $X_1 = X_k$, diremos se tratar de um *ciclo direcionado*.

Concentraremos nossa atenção nos grafos direcionados sem a presença de ciclos, chamados de *grafos acíclicos direcionados*, que a partir daqui chamaremos pela já famosa sigla 'DAGs', do inglês, *directed acyclic graphs*. Na figura 4.7 é apresentado um DAG, já na figura 4.8 é mostrado um grafo direcionado cíclico.

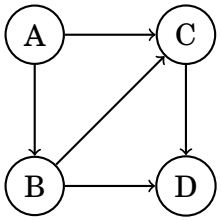


Figura 4.7: Um grafo direcionado acíclico

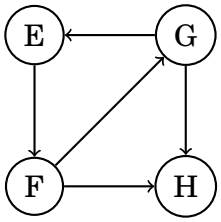


Figura 4.8: Um grafo direcionado cíclico

Se $A \to B$ então A é chamado de *pai* de B, e B, por sua vez, é chamado de filho de A. O conjunto de pais de B é escrito como $pa(B)$ e o conjunto de filhos[2] como $ch(B)$. Seguindo a mesma semântica, se houver um caminho direcionado de A para B, dizemos que A é um *antepassado* de B, e B é um *descendente* de A. O conjunto de antepassados de B é escrito como $an(B)$ e o conjunto de descendentes como $de(B)$. O caso do conjunto de não-descendentes de B é escrito como $nd(B)$.

As definições de pais, filhos, antepassados e descendentes podem ser facilmente estendidas para conjuntos de vértices. Para um conjunto $S \subseteq \mathcal{V}$ definimos, por exemplo, $pa(S) = \{\bigcup_{v \in S} pa(v)\} \setminus S$, ou seja, $pa(S)$ é o conjunto de vértices que não estão em S que possuem ao menos um filho em S. Outras definições podem ser estendidas similarmente, como o conjunto chamado de *ancestral* de S, que é definido por $an^+(S) = S \cup an(S)$.

Seja $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ um DAG com cada vértice representando uma das variáveis aleatórias $\{X_1,...,X_n\}$. Dizemos que uma distribuição de probabilidade conjunta P admite:

[2] O uso da expressão $ch(B)$ vem do inglês, *children*.

(FR) uma *fatoração recursiva* sobre o DAG $\mathcal{G}$ se puder ser escrita como

$$P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i \mid pa(X_i))$$

A importância da fatoração recursiva será evidenciada mais a frente, porém, para isso, precisamos conhecer os conceitos de moralização e d-separação, que serão explicados a seguir.

Definição 4.1.1 (Moralização)**.** Dado um DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, construímos um grafo não direcionado, $\mathcal{G}^m$, 'casando' os pais e apagando os sentidos dos arcos de $\mathcal{G}$, ou seja,

(1) Para todo $v \in \mathcal{V}$, conectamos os nós em $pa(v)$ com linhas não direcionadas

(2) Transformamos as setas de $\mathcal{G}$ em arcos não direcionados

chamamos $\mathcal{G}^m$ de *grafo moral correspondente a* $\mathcal{G}$

Para entendermos melhor a definição acima, consideremos o seguinte exemplo, adaptado de Koski e Noble (2009):

Exemplo 4.1.2 (Moralizando um DAG)**.** Considere o DAG da figura 4.9, o resultado de sua moralização é apresentado na figura 4.10, onde os vértices pais foram 'casados' e as setas foram substituídas por linhas não direcionadas.

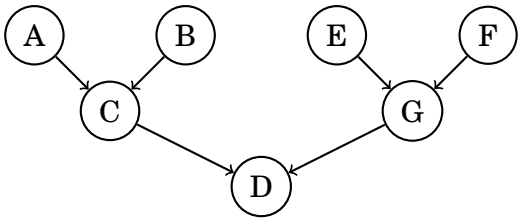


Figura 4.9: DAG $\mathcal{G}$

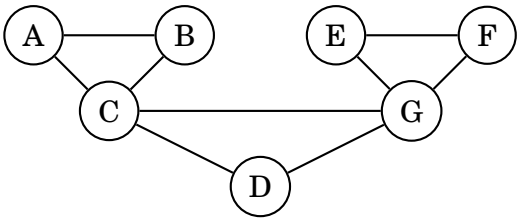


Figura 4.10: O grafo moral $\mathcal{G}^m$

Com o conceito de moralização em mente, somos capazes de escrever uma das definições mais importantes para a teoria de DAGs, a chamada d-separação.

Definição 4.1.2 (d-Separação)**.** Seja $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ um DAG e sejam $X, Y, S \subseteq \mathcal{V}$ conjuntos disjuntos de vértices. Consideremos o conjunto ancestral de $X \cup Y \cup S$, ou seja,

$$A = an^+(X \cup Y \cup S) = (X \cup Y \cup S) \cup an(X \cup Y \cup S)$$

Seja $\mathcal{G}_A$ o subgrafo gerado pelo conjunto A e $\mathcal{G}_A^m$ o grafo moral correspondente a $\mathcal{G}_A$. Se X e Y forem separados por S em $\mathcal{G}_A^m$, dizemos que X e Y são *d-separados* por S. Quando isso ocorre, escrevemos $X \perp^d Y \,|\, S$.

A primeira vista é difícil entender onde a definição de d-separação pode ser útil, porém, há um resultado de extrema importância ligando o conceito de independência condicional ao de d-separação.

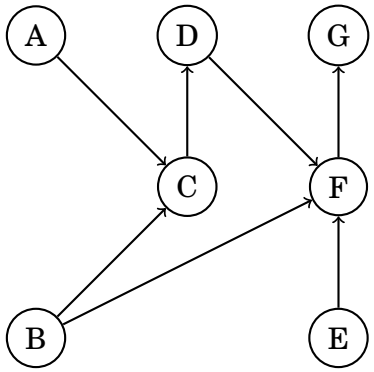


Figura 4.11: DAG original $\mathcal{G}$

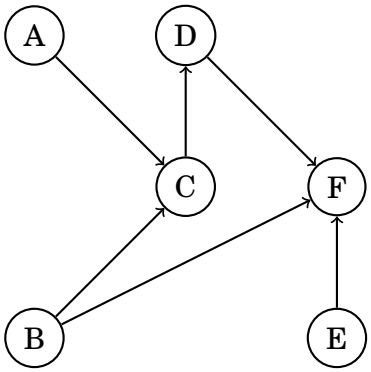


Figura 4.12: Ancestral de $C \cup F \cup D$

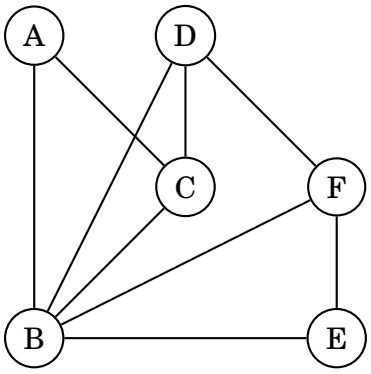


Figura 4.13: Ancestral moralizado

Teorema 4.1.1 (d-Separação implica em independência condicional)**.** Sejam $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ um DAG e P uma distribuição de probabilidade conjunta que fatore recursivamente sobre $\mathcal{G}$. Então, para quaisquer subconjuntos disjuntos $A, B, S \subseteq \mathcal{V}$, vale que

$$A \perp^d B \,|\, S \;\Rightarrow\; A \perp\!\!\!\perp B \,|\, S$$

Ou seja, se A e B são d-separados por S, com P fatorando recursivamente sobre $\mathcal{G}$, então A e B são condicionalmente independentes dado S.

Prova (Koski e Noble (2009), p. 66)**.**

Exemplo 4.1.3 (Buscando independências condicionais)**.** Vamos supor que temos o DAG da figura 4.11 com uma distribuição P que fatora recursivamente sobre suas variáveis, e que queremos saber se a relação de independência condicional $C \perp\!\!\!\perp F \mid D$ vale ou não. Para termos a resposta, basta sabermos se C e F são d-separados por D. Sendo assim, primeiro devemos encontrar o conjunto ancestral de $C \cup F \cup D$, como é mostrado na figura 4.12, e então moralizar esse subgrafo, que nos dá o grafo da figura 4.13, que mostra que C e F não são separados por D, sendo assim, $C \not\perp\!\!\!\perp F \mid D$.

O seguinte lema justifica o exemplo do início da seção (4.1.1), mostrando que se dois subconjuntos de vértices são d-separados por um conjunto vazio, então eles são marginalmente independentes.

Lema 4.1.2. Sejam $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ um DAG e P uma distribuição de probabilidade conjunta que fatore recursivamente sobre $\mathcal{G}$. Então, para quaisquer subconjuntos disjuntos $A, B \subseteq \mathcal{V}$, vale que

$$A \perp^d B \mid \emptyset \;\Rightarrow\; A \perp\!\!\!\perp B$$

Esse lema apresenta uma das principais diferenças entre grafos não direcionados e DAGs. Em ambos os casos a ausência de conexão entre dois vértices implica independência condicional. Porém, dois vértices não conectados num grafo não direcionado são condicionalmente independentes dadas **todas** as outras variáveis restantes do modelo, enquanto num DAG eles são condicionalmente independentes dados seus **antecessores**. Sendo assim, o exemplo 4.1.1 no início desta seção ilustra muito bem essa diferença, já que as variáveis G e F do DAG não possuem antecessores, ou seja, $an(G) = an(F) = \emptyset$, e como $G \neq F$ então $G \perp\!\!\!\perp F$.

Com os conceitos de d-separação e fatoração recursiva conhecidos, agora somos capazes de apresentar as condições markovianas para DAGs. Para isso usaremos a construção feita por Lauritzen (1996), que diz que, para um grafo acíclico direcionado $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, uma distribuição de probabilidade conjunta P satisfaz:

(LD) a *propriedade markoviana direcionada local* se

$$\forall\, v \in \mathcal{V}\ ,\ v \perp\!\!\!\perp (nd(v) \setminus pa(v)) \mid pa(v)$$

(GD) a *propriedade markoviana direcionada global* se

$$\forall\, (U, W, Z) \subseteq \mathcal{V} \text{ disjuntos com } U \perp^d W \mid Z \;\Rightarrow\; U \perp\!\!\!\perp W \mid Z$$

(OD) a *propriedade markoviana direcionada ordenada* se

$$\forall\, v \in \mathcal{V}\ ,\ v \perp\!\!\!\perp (an(v) \setminus pa(v)) \mid pa(v)$$

Um resultado muito importante apresentado por Cowell et al. (1999) relaciona as propriedades (FR), (LD), (GD) e (OD) para grafos acíclicos direcionados.

Teorema 4.1.3. Sejam $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ um grafo acíclico direcionado e P uma distribuição de probabilidade conjunta, temos que as seguintes condições são equivalentes:

- P admite uma fatoração recursiva sobre $\mathcal{G}$
- P obedece à propriedade markoviana direcionada global em relação à $\mathcal{G}$
- P obedece à propriedade markoviana direcionada local em relação à $\mathcal{G}$
- P obedece à propriedade markoviana direcionada ordenada em relação à $\mathcal{G}$

Prova (Cowell et al. (1999), p. 74)**.**

O teorema acima trás o forte resultado de que (FR) $\Leftrightarrow$ (LD) $\Leftrightarrow$ (GD) $\Leftrightarrow$ (OD), que é um ingrediente importante para apresentarmos duas definições das mais recorrentes na literatura de DAGs, apresentadas a seguir com base nos trabalhos de Dawid (2010) e Neapolitan (2003).

Definição 4.1.3 (Condição Markoviana)**.** Sejam $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ um DAG e P uma distribuição de probabilidade conjunta sobre $\mathcal{V}$. Se cada d-separação em $\mathcal{G}$ representar uma relação de independência de P, ou seja, se para cada $(A, B, S) \subseteq \mathcal{V}$ disjuntos valer que

$$A \perp^d B \,|\, S \;\Rightarrow\; A \perp\!\!\!\perp B \,|\, S$$

dizemos que a dupla $(\mathcal{G}, P)$ satisfaz a *Condição Markoviana*.

Se uma distribuição P fatora recursivamente sobre o DAG $\mathcal{G}$ temos, pelo teorema 4.1.1, que a dupla $(\mathcal{G}, P)$ satisfaz a condição markoviana. Mais a frente, o teorema 4.2.1 apresenta a volta, mostrando que se a dupla $(\mathcal{G}, P)$ satisfaz a condição markoviana então P fatora recursivamente sobre $\mathcal{G}$.

Definição 4.1.4 (Fidelidade)**.** Seja P uma distribuição de probabilidade conjunta sobre o DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Se cada relação de independência de P representar uma d-separação em $\mathcal{G}$, ou seja, se para cada $(A, B, S) \subseteq \mathcal{V}$ disjuntos valer que

$$A \perp\!\!\!\perp B \,|\, S \;\Rightarrow\; A \perp^d B \,|\, S$$

dizemos que a dupla $(\mathcal{G}, P)$ possui *fidelidade*.

É muito comum na literatura os autores assumirem que a condição markoviana e a fidelidade são satisfeitas, obtendo a relação

$$A \perp^d B \,|\, S \;\Leftrightarrow\; A \perp\!\!\!\perp B \,|\, S.$$

O que em muitos casos é válido, como é mostrado no exemplo 4.1.4. Entretanto, nem sempre tais hipóteses podem ser assumidas, como ilustrado no exemplo 4.1.5. Ambos exemplos são apresentados por Dawid (2010).

Exemplo 4.1.4 (Dawid (2010))**.** Suponha que temos as variáveis aleatórias (Z,U,X,Y) com as seguintes relações de independência condicional: $U \perp\!\!\!\perp Z$ e $Y \perp\!\!\!\perp Z \,|\, (X,U)$. Podemos então representar tais relações através do DAG (único) da figura 4.14, satisfazendo tanto a condição markoviana quanto a fidelidade com relação a P.

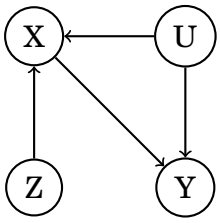


Figura 4.14: DAG satisfazendo as condições markoviana e de fidelidade

Exemplo 4.1.5 (Dawid (2010))**.** Suponha que temos as variáveis aleatórias (X,Y,Z) com as seguintes relações de independência: $X \perp\!\!\!\perp Y$ e $X \perp\!\!\!\perp Y \mid Z$. Neste caso, não é possível representar tais relações através de um DAG, o que nos mostra um exemplo onde é inviável assumir as hipóteses de condição markoviana e fidelidade. Para esse tipo de problema o ideal é usar alguma outra forma de modelagem.

Para encerrarmos a discussão sobre DAGs é importante conhecermos o conceito de *equivalência*, que nos permite extrair um importante resultado baseado na estrutura visual dos grafos.

Definição 4.1.5 (Equivalência)**.** Sejam $\mathcal{G}_1 = (\mathcal{V}, \mathcal{E}_1)$ e $\mathcal{G}_2 = (\mathcal{V}, \mathcal{E}_2)$ dois DAGs contendo o mesmo conjunto de variáveis $\mathcal{V}$. Dizemos que $\mathcal{G}_1$ e $\mathcal{G}_2$ são *equivalentes* se, para cada $(A,B,C) \subseteq \mathcal{V}$ disjuntos, A e B forem d-separados por C em $\mathcal{G}_1$ se, e somente se, A e B forem d-separados por C em $\mathcal{G}_2$. Ou seja

$$A \perp^d_{\mathcal{G}_1} B \,|\, C \;\Leftrightarrow\; A \perp^d_{\mathcal{G}_2} B \,|\, C.$$

Sendo assim, se tivermos uma distribuição P que fatore recursivamente sobre dois DAGs $\mathcal{G}_1$ e $\mathcal{G}_2$ equivalentes, podemos dizer pelo teorema 4.1.1 que ambos representam as mesmas relações de independência condicional, como explicado em mais detalhes por Neapolitan (2003). Uma forma de verificar a equivalência entre dois DAGs de uma forma visual é através do teorema 4.1.4, que usa as definições de esqueleto e v-estrutura.

Definição 4.1.6 (Esqueleto)**.** O *esqueleto* de um DAG $\mathcal{G}$ é o grafo não direcionado obtido pela substituição das setas de $\mathcal{G}$ por arcos não direcionado.

Exemplo 4.1.6 (Esqueleto de um DAG)**.** Suponha que temos o DAG $\mathbb{G}$ da figura 4.15. O esqueleto de $\mathbb{G}$ é apresentado na figura 4.16.

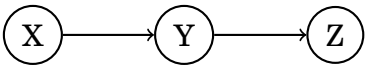


Figura 4.15: DAG $\mathbb{G}$

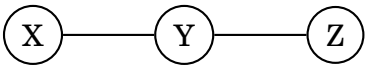


Figura 4.16: Esqueleto de $\mathbb{G}$

Definição 4.1.7 (V-estrutura)**.** Uma *v-estrutura* em um DAG $\mathcal{G}$ é uma tripla de vértices (X,Y,Z) tal que $\mathcal{G}$ contém as setas $X \rightarrow Z$ e $Y \rightarrow Z$, com X e Y não adjacentes.

Exemplo 4.1.7 (Uma v-estrutura)**.** No grafo da figura 4.17, a tripla (X,Y,Z) é uma v-estrutura.

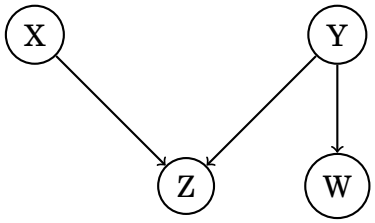


Figura 4.17: A tripla (X,Y,Z) é uma v-estrutura

Com as definições dadas acima podemos escrever o seguinte teorema, proposto por Verma e Pearl (1991):

Teorema 4.1.4 (Equivalência entre DAGs)**.** Dois DAGs são equivalentes se, e somente se, tiverem o mesmo esqueleto e as mesmas v-estruturas.

Prova (Verma e Pearl (1991), p. 224)**.**

Como vimos anteriormente, uma forma de caracterizar a equivalência entre dois DAGs $\mathcal{G}_1$ e $\mathcal{G}_2$ é através de uma distribuição de probabilidade que fatore recursivamente sobre ambos. Se, usando o teorema de Bayes, conseguirmos obter de $\mathcal{G}_1$ as mesmas probabilidades condicionais expressas por $\mathcal{G}_2$ então as condições do teorema 4.1.4 são satisfeitas e os DAGs $\mathcal{G}_1$ e $\mathcal{G}_2$ são equivalentes.

Exemplo 4.1.8 (DAGs equivalentes)**.** Consideremos os DAGs $\mathcal{G}_1$ e $\mathcal{G}_2$ representados nas figuras 4.18 e 4.19, respectivamente. Suponhamos que uma distribuição P fatore sobre os DAGs $\mathcal{G}_1$ e $\mathcal{G}_2$ da seguinte forma:

$$\text{Para o DAG } \mathcal{G}_1\text{: } P(A,B,C) = P(A)\,P(B|A)\,P(C|B)$$

$$\text{Para o DAG } \mathcal{G}_2\text{: } P(A,B,C) = P(B)\,P(A|B)\,P(C|B)$$

Aplicando a regra de Bayes na fatoração do DAG $\mathcal{G}_1$ podemos escrever que $P(B|A) = P(A|B)P(B)/P(A)$, ou seja, podemos reescrever a fatoração do DAG $\mathcal{G}_1$ como

$$\begin{aligned} P(A,B,C) &= P(A)\ P(B|A)\ P(C|B) \\ &= \cancel{P(A)}\ \frac{P(A|B)P(B)}{\cancel{P(A)}}\ P(C|B) \\ &= P(B)\ P(A|B)\ P(C|B) \end{aligned}$$

Com isso, através da aplicação da regra de Bayes conseguimos, a partir das probabilidades condicionais da fatoração do DAG $\mathcal{G}_1$, chegar na fatoração inicial do DAG $\mathcal{G}_2$. Sendo assim, dizemos que os DAGs $\mathcal{G}_1$ e $\mathcal{G}_2$ são equivalentes.

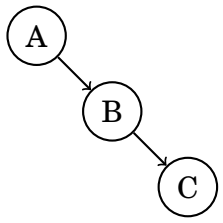


Figura 4.18: DAG $\mathcal{G}_1$: A → B → C

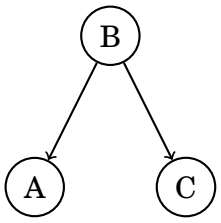


Figura 4.19: DAG $\mathcal{G}_2$: A ← B → C

Pelo que vimos acima, vários DAGs podem representar o mesmo conjunto de independências condicionais, sendo então chamados de DAGs equivalentes. Dizemos que DAGs equivalentes pertencem a uma *classe de equivalência*, que pode ser representada por um grafo acíclico parcialmente direcionado, ou PDAG, do inglês *partially directed acyclic graph*. Em um PDAG, $\mathcal{D} = (\mathcal{F}, \mathcal{V})$, representando uma classe de equivalência, as setas em $\mathcal{F}$ são comuns a todos os DAGs pertencentes à classe, arcos não direcionados de $\mathcal{F}$ correspondem à setas orientadas para um sentido em alguns DAGs e em sentido contrário em outros. A ausência de arcos no PDAG indica que em todos os DAGs pertencentes à classe de equivalência tais arcos também não existem.

Exemplo 4.1.9 (Classe de equivalência)**.** Consideremos os DAGs $\mathcal{D}_1$ e $\mathcal{D}_2$ mostrados, respectivamente, nas figuras 4.20 e 4.21. $\mathcal{D}_1$ e $\mathcal{D}_2$ possuem o mesmo esqueleto e mesmas v-estruturas, logo, são equivalentes. A classe de equivalência à qual $\mathcal{D}_1$ e $\mathcal{D}_2$ pertencem é representada pelo PDAG da figura 4.22.

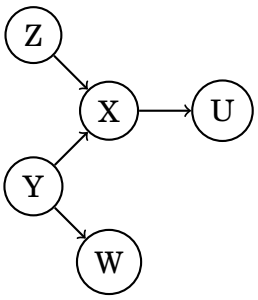


Figura 4.20: DAG $\mathcal{D}_1$

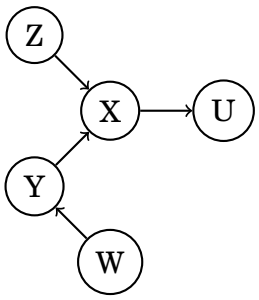


Figura 4.21: DAG $\mathcal{D}_2$

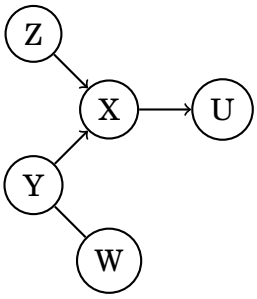


Figura 4.22: PDAG representando a classe de equivalência de $\mathcal{D}_1$ e $\mathcal{D}_2$

4.2 Redes Bayesianas

Uma rede bayesiana é um modelo de representação de conhecimentos sobre algum tipo de situação ou problema, consistindo de duas partes: (i) um componente qualitativo, que mostra a estrutura da rede na forma de um grafo acíclico direcionado e (ii) um componente quantitativo, na forma de probabilidades condicionais. O DAG de uma rede bayesiana representa, através de seus vértices, as variáveis aleatórias de interesse, e, por seus arcos, as influências diretas entre tais variáveis. As probabilidades condicionais da rede quantificam a dependência entre cada variável e seus pais no DAG. Formalmente, temos a seguinte definição:

Definição 4.2.1 (Rede Bayesiana)**.** Sejam $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ um DAG e P uma distribuição de probabilidade conjunta sobre $\mathcal{V}$. Se a dupla $(\mathcal{G}, P)$ satisfizer a condição markoviana (definição 4.1.3), dizemos que $(\mathcal{G}, P)$ é uma *rede bayesiana*.

O verdadeiro poder de uma rede bayesiana está em sua capacidade de síntese, pois ela admite que a representação de uma distribuição de probabilidade conjunta seja feita de forma compacta, permitindo que façamos inferências de forma simplificada, sendo desnecessário o uso explícito da distribuição conjunta. Esse grande trunfo das redes bayesianas é decorrente do teorema apresentado abaixo.

Teorema 4.2.1 (Redes bayesianas fatoram recursivamente)**.** Se a dupla $(\mathcal{G}, P)$ satisfaz a condição markoviana, então P é igual ao produto das probabilidades condicionais de cada vértice em $\mathcal{G}$ dados os valores de seus respectivos pais.

Prova (Neapolitan (2003), p. 34 - 35)**.**

Pelo teorema 4.2.1 apresentado acima, temos que a distribuição conjunta de uma rede bayesiana pode ser escrita como

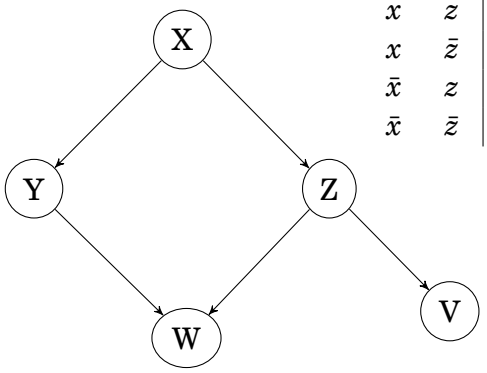
$$P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i \mid pa(X_i)) \ .$$

Sendo assim, recordando sobre o teorema 4.1.3, temos que um DAG de uma rede bayesiana obedece à todas as propriedades markovianas direcionadas, o que torna sua utilização uma boa saída para modelar problemas com muitas variáveis, já que, dados os pais de uma variável, não existe dependência entre ela e seus não descendentes.

A nomenclatura mais comum na literatura chama o DAG e as probabilidades condicionais de cada vértice de, respectivamente, *estrutura* e *parâmetros* da rede. Sendo assim, a construção do modelo depende, principalmente, desses dois ingredientes. A estrutura de uma rede bayesiana pode ser obtida de diversas formas, podendo ser feita uma construção manual ou utilizando-se métodos que levem em conta bancos de dados. O mesmo serve para os parâmetros, que podem ser obtidos por vários tipos de métodos. Esses assuntos serão abordados nos próximos capítulos. Agora consideramos um exemplo em que tanto a estrutura quanto os parâmetros da rede são conhecidos.

Exemplo 4.2.1 (Estrutura e parâmetros conhecidos)**.** Suponhamos que as variáveis aleatórias binárias X, Y, Z, W e V formem uma rede bayesiana com estrutura e parâmetros conhecidos. A estrutura dessa rede, na forma de um DAG, e os parâmetros, apresentados em tabelas, são mostrados na figura 4.23.

No exemplo anterior são apresentadas tabelas, cada uma referente a uma variável aleatória, contendo probabilidades condicionais. Uma tabela desse tipo, referente a uma variável A, é chamada de *tabela de probabilidades condicionais (TPC) para* A.



X	$P(X)$
x	0.2
$\bar{x}$	0.8

X	Z	$P(Z\|X)$
x	z	0.8
x	$\bar{z}$	0.2
$\bar{x}$	z	0.1
$\bar{x}$	$\bar{z}$	0.9

X	Y	$P(Y\|X)$
x	y	0.3
x	$\bar{y}$	0.7
$\bar{x}$	y	0.5
$\bar{x}$	$\bar{y}$	0.5

Z	V	$P(V\|Z)$
z	v	0.4
z	$\bar{v}$	0.6
$\bar{z}$	v	0.7
$\bar{z}$	$\bar{v}$	0.3

Y	Z	W	$P(W\|Y,Z)$
y	z	w	0.95
y	z	$\bar{w}$	0.05
y	$\bar{z}$	w	0.9
y	$\bar{z}$	$\bar{w}$	0.1
$\bar{y}$	z	w	0.8
$\bar{y}$	z	$\bar{w}$	0.2
$\bar{y}$	$\bar{z}$	w	0
$\bar{y}$	$\bar{z}$	$\bar{w}$	1

Figura 4.23: Rede bayesiana apresentada com estrutura e parâmetros

Neste trabalho trataremos apenas redes bayesianas aplicadas à variáveis aleatórias discretas, já que é o tipo de utilização com literatura mais rica. Para casos em que usaríamos, normalmente, variáveis contínuas, como salário ou altura, usaremos a famosa *discretização*, em que são escolhidos intervalos das variáveis contínuas originais para serem tratados como casos de variáveis discretas e finitas. Para modelos que englobem variáveis aleatórias contínuas explicitamente, direcionamos o leitor ao trabalho de Cobb et al. (2007). A seguir apresentamos exemplos de possíveis discretizações.

(a) Uma variável aleatória contínua para representar a altura de pessoas, em centímetros, pode ser definida como $A \in [0, \infty)$. Uma possível discretização para esse caso seria $\bar{A} = \{(0, 110], (111, 130], (131, 150], (151, 170], (171, 190], (191, \infty)\}$, de forma que cada intervalo representa um caso da variável discreta $\bar{A}$.

(b) Uma variável contínua representando a renda de indivíduos pode ser definida como $R \in [0,\infty)$. Uma possível discretização seria considerar que rendas entre 0 e 2000 sejam baixas, entre 2001 e 4000 sejam médias e que de 4001 para cima sejam altas, sendo possível definir uma variável discreta $\bar{R}$ = {baixa, média, alta}. Obviamente, é possível tornar esses intervalos menores, nos permitindo escrever $\bar{R}$ = {baixa, média baixa, média alta, alta}, ou algo semelhante. A escolha do tamanho dos intervalos depende do tipo de problema e da necessidade de precisão em cada aplicação.

A definição de rede bayesiana está intrinsecamente ligada às relações de independências, marginais e condicionais, entre as variáveis aleatórias de interesse. Com isso, poderíamos dizer que uma rede bayesiana não passa de uma representação gráfica de tais relações. Entretanto, em diversos casos esses modelos são tratados como redes causais, em que as variáveis pais são interpretadas como causas e as filhas como efeitos. Na seção seguinte faremos uma breve discussão sobre a interpretação causal de redes bayesianas, com base nos trabalhos de Pearl (2000), Spirtes et al. (2000), Koller e Friedman (2009) e Gopnik e Schulz (2007).

4.2.1 Causalidade

O tema da causalidade é um tanto controverso, sendo um assunto comum à diversas áreas da ciência. Existem trabalhos sobre causalidade feitos por filósofos, psicólogos, estatísticos, matemáticos, cientistas da computação entre outros pesquisadores de diferentes campos de atuação, o que mostra a importância do assunto e revela a dificuldade em tratá-lo, por ser claramente um tema multi-disciplinar.

Mesmo que assumamos apenas um enfoque estatístico, é difícil definir formalmente o que é uma causa e o que é um efeito, ou como um dado evento influencia o acontecimento de outro, sendo assim, vamos apenas apresentar, brevemente, os resultados mais aceitos sobre o assunto, buscando mostrar as principais diferenças entre um modelo probabilístico e um causal.

Um *modelo causal*, chamado por alguns de *rede bayesiana causal*, consiste de uma grafo acíclico direcionado com os vértices representando as variáveis de interesse. Neste modelo, cada variável X é regida por um mecanismo causal que determina, estocasticamente, seus valores baseando-se em suas variáveis pais, neste contexto chamadas de *causas diretas*. Com isso, podemos dizer que se uma rede bayesiana tiver sua estrutura representando relações de causa-efeito, ela é um modelo causal.

A causalidade é um ingrediente importante no processo de construção de redes bayesianas, porém, não é estritamente necessário que os arcos direcionados de um DAG recebam uma interpretação causal. Como vimos, o modelo $X \to Y$ é probabilisticamente equivalente ao modelo $Y \to X$, já que podem ser representados

pelo mesmo PDAG $X \sim Y$, porém, Kjærulff e Madsen (2013) mostram que arcos direcionados no sentido *(efeito)* $\rightarrow$ *(causa)* podem acabar prejudicando a representação das dependências condicionais da rede, o que torna o direcionamento causal dos arcos uma questão importante mesmo quando a interpretação adotada para a rede for não-causal.

Os modelos causais mais simples são os de *causa comum*, *corrente causal* e *efeito comum*, representados, respectivamente, pelas figuras 4.24, 4.25 e 4.26. Tendo tais modelos as seguintes fatorações:

$$\text{Causa comum:} \quad P(X,Y,Z) = P(Y|X)\,P(Z|X)\,P(X)$$

$$\text{Corrente causal:} \quad P(X,Y,Z) = P(Z|Y)\,P(Y|X)\,P(X)$$

$$\text{Efeito comum:} \quad P(X,Y,Z) = P(Z|Y,X)\,P(Y)\,P(X)$$

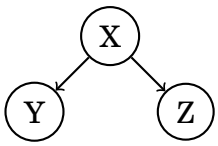


Figura 4.24: Causa comum

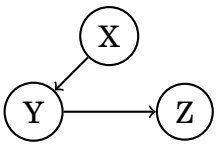


Figura 4.25: Corrente causal

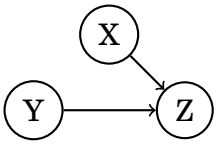


Figura 4.26: Efeito comum

Para muitas aplicações, não há diferença entre interpretarmos uma rede bayesiana como causal ou não, já que com ambas as construções podemos aplicar o teorema de Bayes para fazermos inferências. A diferença entre modelos causais e probabilísticos aparece quando estamos interessados nos efeitos de *intervenções* no modelo, que são situações onde não apenas observamos os valores que as variáveis assumem, mas sim onde possamos tomar ações que manipulem essas variáveis.

A modelagem de intervenções pode ser feita através da chamada *intervenção ideal*, que com o operador $do(X = x)$ força a variável (ou conjunto de variáveis) X a assumir o valor x, sem mais ser afetada por outras causas. Sendo assim, se tomarmos como exemplo o modelo de corrente causal (figura 4.25), podemos ver a diferença entre uma observação e uma intervenção sobre uma variável da seguinte forma:

Observação de $Y = 1$: $P(X, Y = 1, Z) = P(Z|Y = 1)\,P(Y = 1|X)\,P(X)$

Intervenção $do(Y = 1)$: $P(X, do(Y = 1), Z) = P(Z|Y = 1)\,P(X)$

O efeito de uma intervenção pode ser representado pela eliminação de arcos na rede bayesiana em questão. Para o exemplo acima, temos que a intervenção $do(Y = 1)$ nos dá o DAG da figura 4.27.

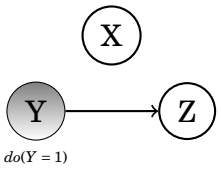


Figura 4.27: Modelo de corrente causal com a intervenção $do(Y = 1)$

Generalizando, temos que para um conjunto de variáveis aleatórias $\mathcal{V} = (X_1, X_2, ..., X_n)$, a intervenção sobre o subconjunto $T \subseteq \mathcal{V}$ da forma $T = t$ nos dá a seguinte distribuição pós-intervenção:

$$P(\mathcal{V}) = \prod_{X_i \in \mathcal{V} \setminus T} P(X_i \mid pa(X_i))$$

Através dessa modelagem podemos responder diversos tipos de perguntas probabilísticas que não são tão bem respondidas sem a análise de intervenções, como nos exemplos propostos por Koller e Friedman (2009):

> *Diagnóstico e tratamento*: "Se fizermos esse paciente tomar esta medicação, quais suas chances de melhorar?"
>
> *Marketing*: "Se baixarmos o preço dos hambúrgueres, será que os clientes comprarão mais mostarda"
>
> *Política econômica*: "Se baixarmos a taxa de juros, será que a inflação vai subir?"
>
> *Descobertas científicas*: "Fumar causa câncer?"

Outro tipo de análise, diferente tanto da observacional quanto da intervencional, é a chamada análise de *contrafatuais*, que surge em situações nas quais já temos a informação do verdadeiro estado da natureza, mas onde gostaríamos de saber o que teria acontecido caso tivéssemos feito algum tipo de intervenção em uma ou mais variáveis. Este tipo de análise responde à perguntas do tipo: "Será que o estado de intoxicação do motorista causou o acidente?", ou seja, já sabemos

que o acidente ocorreu, mas o que teria acontecido caso tivéssemos manipulado o estado de intoxicação do motorista?

Não nos alongaremos na discussão sobre contrafatuais por ser um tema muito discutível e com abordagem fora do escopo deste trabalho. Para mais informações sobre esse assunto indicamos Pearl (2000).

Para fecharmos o assunto de causalidade é importante termos em mente que a construção de modelos causais baseados em dados de intervenções é muito difícil de ser feita, já que algumas intervenções são complexas de serem executadas e, em muitos casos, até mesmo ilegais. Sendo assim, para que o aprendizado de modelos causais a partir de dados observacionais seja consistente, devemos assumir algumas hipóteses sobre o modelo:

> Seja $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ um DAG cujos arcos representem relações causais e seja P uma distribuição de probabilidade sobre $\mathcal{V}$.
>
> *Condição Markoviana Causal*: A dupla $(\mathcal{G}, P)$ satisfaz a condição markoviana causal se, e somente se, $\forall W \in \mathcal{V}$, $W \perp\!\!\!\perp \mathcal{V} \setminus (de(W) \cup pa(W)) \mid pa(W)$.
>
> *Condição de Minimalidade Causal*: A dupla $(\mathcal{G}, P)$ satisfaz a condição de minimalidade causal se, e somente se, para todo subgrafo próprio $\mathcal{H} \subset \mathcal{G}$ a dupla $(\mathcal{H}, P)$ não satisfizer a condição markoviana causal.
>
> *Condição de Fidelidade*: A dupla $(\mathcal{G}, P)$ satisfaz a condição de fidelidade se, e somente se, cada relação de independência condicional verdadeira em P for representada por uma d-separação em $\mathcal{G}$.

Para mais detalhes sobre a motivação dessas suposições, direcionamos o leitor aos trabalhos de Spirtes et al. (2000) e Koller e Friedman (2009).

Capítulo 5

Construção de Redes Bayesianas

A construção de uma Rede Bayesiana pode ser divida em três etapas. Primeiro, é necessário que selecionemos as variáveis de interesse para a aplicação, definindo seus possíveis valores e como proceder com eventuais discretizações. Em segundo lugar, precisamos construir a estrutura da rede, conectando as variáveis selecionadas de forma que seja montado um grafo acíclico direcionado. Por último, precisamos obter os parâmetros da rede, definindo uma tabela de probabilidades condicionais para cada variável de interesse.

A etapa de seleção das variáveis é geralmente feita com base em análises de um especialista da área em questão. Sendo assim, se o problema estiver relacionado com o diagnóstico de uma doença, a escolha das variáveis pode ser auxiliada por um profissional da saúde, se o problema for relacionado com o marketing de uma empresa, um executivo da área pode ajudar na identificação das variáveis de interesse, e assim sucessivamente. Já a construção da estrutura da rede pode ser feita tanto manualmente, com base também nos conhecimentos causais de um especialista da área, quanto através de algoritmos, que aplicados sobre os dados disponíveis, geram uma rede probabilística de forma computacional. A etapa de cálculo dos parâmetros também pode ser feita de diversas formas, tanto com base em conhecimentos subjetivos, através de uma ótica mais próxima da puramente bayesiana, quanto por métodos de máxima verossimilhança.

5.1 Seleção de Variáveis

O conjunto das variáveis escolhidas para uma Rede Bayesiana forma a base de todo o modelo. Podemos dividir essas variáveis em dois grandes grupos: variáveis aleatórias e variáveis de decisão. Neste capítulo iremos tratar casos em que hajam somente variáveis aleatórias, deixando os casos em que apareçam variáveis de decisão mais à frente.

Vamos supor que temos um problema a ser resolvido, e que julguemos que o uso de uma Rede Bayesiana seja uma boa opção para tratar esse problema. Para sabermos se uma variável aleatória deve ser incluída ou não em nosso modelo, podemos usar o chamado *teste de claridade*, proposto por Kjærulff e Madsen (2013):

> Para testar se uma variável X, candidata a entrar no modelo, é realmente apropriada e útil para o problema, ela deve ser condizente com os seguintes princípios:
>
> 1. O espaço de estados de X deve consistir de um conjunto exaustivo e mutuamente exclusivo de valores possíveis.
>
> 2. Preferencialmente, X deve representar um conjunto único de eventos, de forma que nenhuma outra variável do modelo seja sua "concorrente". Se o modelo já tiver, por exemplo, uma variável *Temperatura*, com espaço de estados {*frio, quente*}, não devemos ter uma variável *Temperatura Quente* com espaço de estados {*sim, não*}, pois teríamos informações redundantes.
>
> 3. X deve ser claramente definida, sem deixar dúvidas sobre sua semântica.

A identificação das variáveis de um problema pode tornar-se uma tarefa confusa, por isso é importante seguirmos alguns princípios, assim como os apresentados acima, que sirvam como auxílio para nossas modelagens. Kjærulff e Madsen (2013) apresentam um exemplo (reproduzido abaixo), conhecido como Problema de *Monty Hall*[1], demonstrando como o teste de claridade pode ajudar na construção de uma rede bayesiana.

Exemplo 5.1.1 (Teste de claridade)**.** Suponhamos que temos três portas, A, B e C, e que atrás de somente uma delas haja um grande prêmio. Se escolhermos a porta certa, o prêmio é nosso. Decidimos qual porta queremos, mas antes da porta escolhida ser aberta, uma outra pessoa (que já sabe onde está o prêmio) abre uma das outras duas portas, não contendo o prêmio. Depois disso, somos permitidos a mudar nossa decisão inicial. A pergunta é: devemos escolher a outra porta ou devemos manter a primeira escolha?

Poderíamos tentar modelar esse problema com uma rede bayesiana, na qual cada porta seria representada por uma variável com espaço de estados {*com prêmio, sem prêmio*}, porém, dessa forma violaríamos o segundo princípio do teste de

[1] O Problema de *Monty Hall* é baseado num programa de TV americano, da década de 1970, no qual um dos apresentadores chamava-se Monty Hall. A situação é descrita no exemplo 5.1.1.

claridade, não sendo uma boa escolha de modelagem. Uma boa escolha deve levar em conta as informações disponíveis no problema, podendo ser obtida através das seguintes perguntas:

> Problema: Onde está o prêmio? Essa pergunta nos leva à variável *Local do Prêmio*, com espaço de estados $\{A, B, C\}$.
>
> Informação 1: Qual foi a primeira porta escolhida por nós? O que permite termos uma variável *Primeira Escolha*, também com espaço de estados $\{A, B, C\}$.
>
> Informação 2: Qual porta foi aberta pela outra pessoa? O que nos dá uma variável *Escolha da Outra Pessoa*, com o mesmo espaço de estados $\{A, B, C\}$.

Com as variáveis *Local do Prêmio*, *Primeira Escolha* e *Escolha da Outra Pessoa* passando no teste de claridade.

Obviamente há diversas formas de se fazer a seleção das variáveis de um modelo além do teste de claridade apresentado acima. Uma visão mais geral sobre esse assunto é apresentada por George (2000), que mostra os princípios por trás de alguns métodos de seleção de variáveis.

5.2 Construção de Estruturas

Com as variáveis do modelo devidamente selecionadas, passamos a ter o problema de construir a estrutura da rede bayesiana, que nada mais é do que identificarmos entre quais variáveis haverão arcos e qual direcionamento deve ser dado a cada um deles. A construção dessa estrutura pode ser feita de duas formas: manualmente ou semi-automaticamente.

5.2.1 Construção Manual

A construção manual de uma rede bayesiana pode ser feita através das relações de causa-efeito identificadas pelo especialista da área a ser estudada. Entretanto, caso esse especialista não esteja disponível para auxiliar na construção da rede, ou caso ele precise de algum guia para ajudá-lo em suas análises, é útil manter em mente os tipos de variáveis que podem surgir no problema:

> *Variáveis problema:* São as variáveis de interesse, das quais queremos calcular distribuições a posteriori dados os valores das variáveis de informação. Este tipo de variável pode ser não observável, como, por exemplo, um parâmetro de uma distribuição.

Variáveis de informação: São variáveis observáveis que contém informações relevantes para a resolução do problema, podendo ser divididas em duas subcategorias:

Variáveis de background: São informações disponíveis antes da ocorrência do problema, com influência causal sobre as variáveis problema e variáveis de sintomas. São geralmente as variáveis 'raízes' das redes bayesianas.

Variáveis de sintomas: São as consequências do problema, sendo disponíveis somente após sua ocorrência, ou seja, variáveis de sintomas são causadas pelas variáveis problema. Geralmente, as variáveis problema e de background são pais das variáveis de sintomas.

Variáveis mediadoras: São variáveis não observáveis cujas probabilidades a posteriori não são de interesse imediato, mas que ajudam a manter as relações de independência corretas na rede. Estas variáveis geralmente são filhas de variáveis problema e de background e pais de variáveis de sintomas.

Os tipos de variáveis dados acima, propostos por Kjærulff e Madsen (2013), são uma forma de auxiliar a construção de redes bayesianas feitas manualmente, pois, após a identificação do tipo de uma variável, é muito mais fácil conectá-la às outras, já que geralmente o fluxo causal apresentado na figura 5.1 é satisfeito.

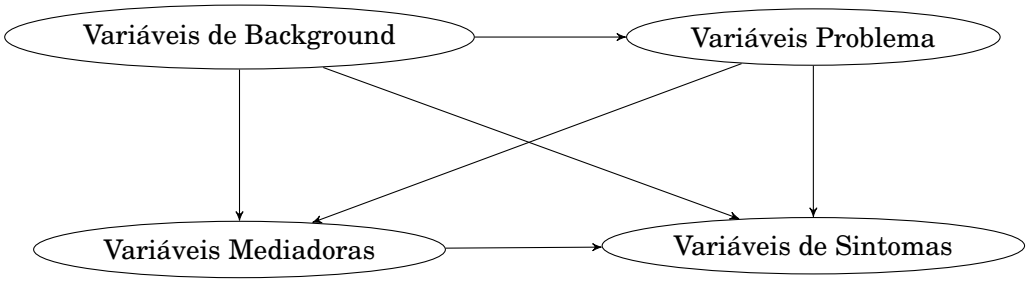


Figura 5.1: Fluxo causal mais comum em redes bayesianas

5.2.2 Construção Semi-Automática

A construção semi-automática de redes bayesianas, também chamada de "baseada em dados", busca identificar a estrutura do modelo através da aplicação de algoritmos que utilizam bancos de dados para construir DAGs. Esse tipo de técnica surgiu da necessidade de incluir um número elevado de variáveis nos modelos, o

que dificulta em muito sua construção manual, já que mesmo especialistas de uma determinada área têm dificuldades em conectar as variáveis de forma satisfatória.

Para que a aplicação de algoritmos seja uma boa saída na construção de redes bayesianas, assumimos que o processo ou fenômeno a ser estudado segue uma distribuição de probabilidade P_0, chamada de *distribuição de probabilidade subjacente do processo*. Com isso, na verdade estamos assumindo que a fonte dos dados pode ser representada por uma amostra proveniente de P_0. O objetivo da construção baseada em dados é identificar um modelo que fique adequado à essa distribuição.

Spirtes et al. (2000) propõe algumas suposições para que o algoritmo escolhido para a construção do modelo descubra um DAG com estrutura equivalente à do DAG de P_0, sendo elas apresentadas à seguir:

- O DAG original de P_0 satisfaz a condição de fidelidade.
- O conjunto de dados consiste de uma amostra de casos independentes e identicamente distribuídos.
- A base de dados é infinitamente grande.
- Não há variáveis latentes, ou omitidas, no modelo.
- Os dados devem ser completos, ou seja, não podem haver entradas em branco na base de dados.
- Os testes estatísticos não possuem erros.

Obviamente, na prática, é impossível que todas as suposições apresentadas acima sejam satisfeitas, servindo apenas como um suporte teórico para alguns dos algoritmos desenvolvidos para a construção de redes bayesianas. Entretanto, é importante lembrarmos que mesmo não satisfazendo essas suposições para um determinado problema, redes bayesianas ainda podem servir como uma ótima aproximação para sua modelagem.

Os algoritmos de aprendizado de estrutura podem ser agrupados em duas categorias principais: algoritmos baseados em restrições e algoritmos baseados em escores. Para facilitar o uso de tais algoritmos, introduzimos aqui uma notação alternativa para representar relações de independência entre variáveis: se duas variáveis aleatórias X e Y são independentes, escrevemos $I(X,Y|\emptyset)$, mas se X e Y são condicionalmente independentes dado Z, escrevemos $I(X,Y|Z)$.

5.2.2.1 Algoritmos Baseados em Restrições

Algoritmos baseados em restrições aprendem a estrutura de uma rede bayesiana baseando-se em testes de independência, através dos quais é possível encontrar o esqueleto da rede em questão. Após o esqueleto ter sido encontrado, são propostos métodos para orientar os arcos não direcionados do esqueleto de forma que não sejam gerados ciclos, fazendo com que o resultado final seja a estrutura da rede, na forma de um DAG.

A principal inspiração para os métodos baseados em restrições é o algoritmo IC (*Inductive Causation*), proposto por Verma e Pearl (1991), que pode ser resumido em três passos (Scutari (2010)):

(1) Descobrir o esqueleto da rede.

(2) Encontrar as v-estruturas e fixar o direcionamento de seus arcos.

(3) Direcionar todos os outros arcos de forma que o grafo seja acíclico.

Um grande problema nos passos descritos acima, e que justifica a nomenclatura de algoritmos "baseados em restrições", é a quantidade de testes de independência a serem feitos para o esqueleto da rede ser encontrado. Cada dupla de variáveis $X, Y \in \mathcal{V}$ pode ter suas relações de independência testadas para cada $\mathcal{C} \subseteq \mathcal{V} \setminus \{X, Y\}$, ou seja, a relação $I(X, Y|\mathcal{C})$ seria testada para cada subconjunto $\mathcal{C}$ possível. Se tivéssemos, por exemplo, selecionado 20 variáveis a serem modeladas, poderíamos construir $2,35 \cdot 10^{72}$ DAGs diferentes (Jensen e Nielsen (2007)), o que tornaria o teste de todas as possíveis relações de independência uma tarefa computacionalmente inviável. Sendo assim, uma alternativa é restringir o número de combinações a serem testadas, podendo tais restrições serem feitas com o auxílio de um especialista da área ou usando um critério puramente técnico, como restringir os testes somente às variáveis da vizinhança da dupla testada.

Para ilustrar o funcionamento de um algoritmo baseado em restrições, vamos considerar um dos representantes mais famosos dessa classe de métodos, o algoritmo PC, proposto por Spirtes et al. (2000). Porém, antes de apresentá-lo, consideremos o teorema dado abaixo, que facilita a construção do algoritmo.

Teorema 5.2.1. As variáveis aleatórias $X, Y \in \mathcal{V}$ não são conectadas se, e somente se, $I(X, Y \mid pa(X))$ ou $I(X, Y \mid pa(Y))$.

Prova (Jensen e Nielsen (2007), p. 234)**.**

O algoritmo PC recebe como entrada um grafo não-direcionado completo, com todas as variáveis a serem incluídas no modelo, e segue uma lógica realmente muito próxima à do algoritmo IC, resumido acima. Seu funcionamento divide-se

nas três etapas mostradas abaixo, na forma resumida e irrestrita do algoritmo, como construído por Jensen e Nielsen (2007) e Mahmood (2011).

Algoritmo PC

Entrada: Grafo não-direcionado completo entre todas as variáveis de $\mathcal{V}$.

Saída: A estrutura da rede na forma de um DAG.

(1) Deletar os arcos entre cada dupla $\{X,Y\}$ que tiver um $Z \subseteq \mathcal{V} \setminus \{X,Y\}$ tal que $I(X,Y|Z)$ e armazenar todas essas relações de independência. Por trás deste passo está o fato de que, se houver uma conexão entre X e Y, eles não podem ser condicionalmente independentes para nenhum Z dado. Este passo constrói o esqueleto do DAG procurado.

(2) Para toda tripla $X \sim Y \sim W$ com $X \not\sim W$ em que soubermos que $I(X,W|Z)$ e que $Y \notin Z$, devemos construir a v-estrutura $X \rightarrow Y \leftarrow W$. Esta orientação deve ser seguida pois qualquer outra implicaria na relação $I(X,W|Y)$, o que seria inconsistente com as relações de independência armazenadas no passo **(1)**. Esta etapa do algoritmo constrói um PDAG, representando a classe de equivalência do DAG procurado.

(3) Aplicar as seguintes regras para obter o DAG procurado:

1. *(Evitar novas v-estruturas):* Se houver triplas $X \rightarrow Y \sim W$ com $X \not\sim W$, orientar $Y \rightarrow W$.
2. *(Evitar ciclos):* Se houver uma dupla $X \sim Y$ cujo direcionamento $X \rightarrow Y$ introduzir um ciclo no grafo, fazer $X \leftarrow Y$.
3. *(Escolher aleatoriamente):* Se nenhuma das duas regras anteriores puder ser aplicada, direcionar as conexões restantes aleatoriamente ou com base em conhecimento causais.

Um fator de extrema importância para os algoritmos baseados em restrições é a forma de testar as relações de independência, como as armazenadas no passo **(1)** do algoritmo PC, que são, obviamente, desconhecidas em aplicações reais. Para lidar com essa tarefa são comumente usados testes de hipóteses clássicos (Kjærulff e Madsen (2013)) sobre as hipóteses

$$H_0 : P(X,Y) = P(X)P(Y)$$

$$H_1 : P(X,Y) \neq P(X)P(Y)$$

onde podemos usar a estatística G^2, que, sob a hipótese nula, tem distribuição assintótica χ^2 com os graus de liberdade apropriados. A estatística G^2 é calculada como

$$G^2 = 2\sum_{x,y} N_{xy} \log\left(\frac{N_{xy}}{E_{xy}}\right)$$

onde N_{xy} representa o número de vezes em que aparece no banco de dados a entrada $(X = x, Y = y)$, N é o número total de entradas na base e $E_{xy} = \frac{N_x N_y}{N}$.

Para testar a independência condicional entre duas variáveis dada uma terceira, pode ser usada uma forma semelhante à mostrada acima, onde as hipóteses testadas são

$$H_0 : P(X,Y|Z) = P(X|Z)P(Y|Z)$$

$$H_1 : P(X,Y|Z) \neq P(X|Z)P(Y|Z)$$

com a estatística G^2 calculada como

$$G^2 = 2\sum_{x,y,z} N_{xyz} \log\left(\frac{N_{xyz}}{E_{xyz}}\right)$$

onde $E_{xyz} = \frac{N_{xz} N_{yz}}{N_z}$.

Se a estatística G^2 for pequena o suficiente, com $G^2 < \alpha$ para algum α escolhido, então a hipótese H_0 não pode ser rejeitada, estando aí a principal deficiência dos algoritmos baseados em restrições: a sensibilidade à falhas nos testes de hipóteses. Basta que apenas um dos testes dê uma resposta errada para prejudicarmos todo o processo de construção da rede (Koller e Friedman (2009)).

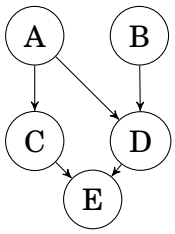


Figura 5.2: Rede geradora dos dados

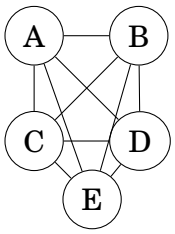


Figura 5.3: Grafo completo de entrada

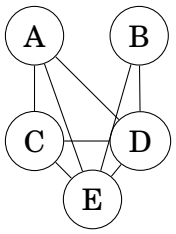


Figura 5.4: Grafo após testes incondicionais

Exemplo 5.2.1 (Jensen e Nielsen (2007))**.** Suponha que tenhamos gerado uma amostra através da rede bayesiana da figura 5.2. A entrada do algoritmo PC é da forma mostrada na figura 5.3, mas após os testes de independência incondicional,

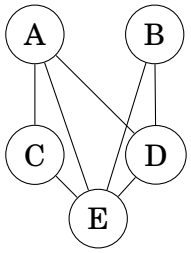


Figura 5.5: Após testes condicionados a uma v.a.

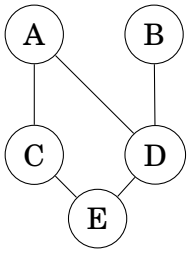


Figura 5.6: Após testes condicionados a duas v.a.'s

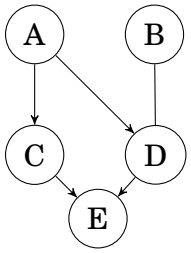


Figura 5.7: PDAG obtido após o passo (2)

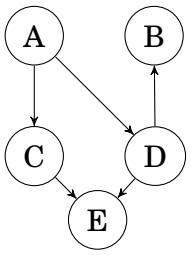


Figura 5.8: Possível rede obtida após o passo (3)

obtemos o grafo da figura 5.4. Em seguida, após os testes de independência dada uma variável, obtemos a figura 5.5 e após os testes de independência dadas duas variáveis, a figura 5.6. Após o passo **(2)** do algoritmo PC, obtemos o PDAG da figura 5.7, e após o passo **(3)** podemos ter tanto o DAG da figura 5.2 quanto o da figura 5.8.

5.2.2.2 Algoritmos Baseados em Escores

Os algoritmos baseados em escores atribuem um valor, ou escore, para cada estrutura candidata a formar uma rede bayesiana. Esse escore reflete a proximidade de cada possível estrutura ao DAG da rede "verdadeira", formada a partir da hipotética distribuição subjacente P_0. Ao escolhermos uma função de escore, que recebe a estrutura de uma rede bayesiana como argumento e retorna um valor, temos que a tarefa do algoritmo resume-se a um problema de busca, em que é procurada a estrutura com maior escore. Sendo assim, podemos descrever completamente um algoritmo desse tipo apenas especificando dois componentes: (i) uma função de escore e (ii) um procedimento de busca (Jensen e Nielsen (2007)).

Uma das funções de escore mais usadas é o chamado Critério de Informação Bayesiano, ou BIC (de *Bayesian Information Criterion*), que é construído da seguinte forma:

$$\text{BIC}\,(S \mid \mathcal{D}) \;=\; \log P(\mathcal{D} \mid \hat{\theta}_S, S) - \frac{\dim(S)}{2}\,\log(N)$$

onde $\hat{\theta}_S$ é o estimador de máxima verossimilhança (que será mostrado em maiores detalhes na seção seguinte) dos parâmetros da estrutura S, N é o número de casos no banco de dados, $dim(S)$ é o número de parâmetros livres, ou graus de liberdade, na rede S e $\mathcal{D}$ são os dados contidos na base.

A característica que torna o BIC uma opção muito comum como função de escore é seu perfil parcimonioso, no sentido de dar maiores escores para estruturas mais simples. No primeiro termo do critério BIC temos $\log P(\mathcal{D} \mid \hat{\theta}_S, S)$, que é uma forma de medir o quão bem o modelo candidato se adequa aos dados, enquanto no segundo termo temos $\frac{\dim(S)}{2}\log(N)$, que penaliza modelos muito complexos. Porém, o peso dessa penalidade cai a medida que o número de casos no banco de dados sobe. Toda a derivação e justificativa do BIC pode ser encontrada no trabalho de Koller e Friedman (2009).

Há diversas funções de escore, entre elas os famosos critério de Akaike e o escore BD (de *Bayesian Dirichlet*), entre muitos outros, não sendo possível explorá-los neste trabalho por constituírem uma área muito abrangente da estatística, focada na seleção de modelos. Para mais informações sobre esse assunto, indicamos o trabalho de Kadane e Lazar (2004).

Após a função de escore ter sido escolhida, o problema da seleção de estrutura resume-se à um processo de busca pela opção com melhor escore. Sendo que, para realizar tal tarefa, existem diversos algoritmos que consideram um *espaço de busca*, em que cada ponto contido nesse espaço corresponde à um DAG.

Uma abordagem muito comum na literatura é a chamada *busca heurística*, em que o espaço de busca é percorrido de forma que não seja necessário testar cada uma das possíveis estruturas, já que, em muitos casos, percorrer todas as possibilidades é computacionalmente inviável.

Para que o espaço de busca seja explorado são usados vários tipos de métodos, sendo uma opção muito conhecida a chamada *busca local*, que se baseia na aplicação de *operadores de busca*, descritos abaixo (Jensen e Nielsen (2007)):

- Adição de arco: um arco é inserido entre dois vértices não adjacentes.
- Remoção de arco: um arco é removido entre dois vértices.
- Reversão de arco: o sentido de um arco é revertido.

Com a aplicação desses operadores é possível partir de uma estrutura inicial e armazenar o escore obtido após uma operação, o que possibilita a busca de uma estrutura com escore maior do que a anterior. Um algoritmo de busca local pode

ser descrito da seguinte forma:

Algoritmo de Busca Local

Entrada: Uma estrutura $G \in \mathcal{G}$ inicial, priores para os parâmetros (caso necessário) e uma função de escore.

Saída: A estrutura da rede na forma de um DAG.

(1) Gerar um subconjunto $\bar{\mathcal{G}} \in \mathcal{G}$ de estruturas candidatas, usando os operadores de busca sobre a estrutura G.

(2) Calcular o escore de cada estrutura candidata obtida no passo **(1)** e armazenar aquela $G^\star \in \bar{\mathcal{G}}$ com maior escore.

(3) Repetir os passos **(1)** e **(2)** usando o grafo $G^\star$ como o de entrada, ou seja, fazendo $G := G^\star$.

(4) Repetir os passos **(1)**, **(2)** e **(3)** até que o escore não seja mais melhorado.

(5) Retornar o último $G^\star$ encontrado, que é o DAG com melhor escore entre os testados.

Como em todo tipo de método, há diversas outras formas de se buscar o DAG que maximize a função de escore escolhida, sendo possível aplicar vários tipos de algoritmos para realizar essa tarefa. Outras opções muito eficientes são os algoritmos de arrefecimento simulado (*simulated annealing*) e de programação genética (*genetic programming*).

5.3 Estimação de Parâmetros

A estimação de parâmetros é a parte final da modelagem por redes bayesianas. Após a seleção das variáveis ter sido feita e a estrutura da rede ter sido fixada, precisamos agora conhecer as probabilidades condicionais de cada vértice dados os valores de seus antecessores. Para essa tarefa existem várias abordagens, sendo aqui mostradas duas das mais utilizadas: estimação por máxima verossimilhança e estimação por aprendizado sequencial, também conhecida como método bayesiano.

Nesta seção, assim como nas anteriores, vamos considerar que nossa base de dados é *completa*, ou seja, vamos considerar que a base de dados $\mathcal{D}$ a ser utilizada terá todos os casos configurados para todas as variáveis do modelo, não sendo

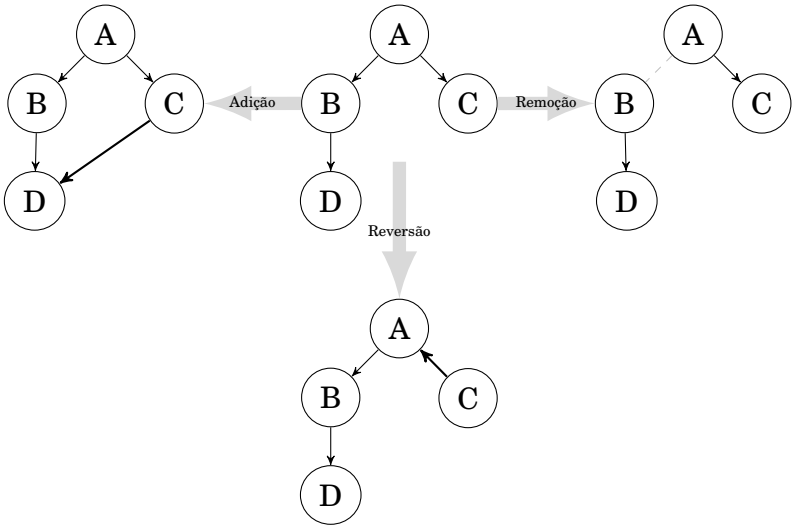


Figura 5.9: Exemplos de operações para busca local

permitida, por exemplo, uma amostra como $\{x = x^1, y = ?, z = z^0\}$, em que o valor de y não foi observado. Mais formalmente, dizemos que um conjunto de dados $\mathcal{D}$ para um conjunto de variáveis $\mathcal{V}$ é um vetor $(d_1, ..., d_N)$ onde cada d_i é um *caso*, e representa uma instância de $\mathcal{V}$.

Outras suposições importantes a serem adotadas aqui são as de *meta-independência global* e *meta-independência local*. Um modelo, digamos $X \to Y$, com $X \sim N(\alpha_1, \alpha_2)$ e $Y \sim N(\beta_1, \beta_2)$, possui meta-independência global se $(\alpha_1, \alpha_2) \perp\!\!\!\perp (\beta_1, \beta_2)$, mas se o modelo tiver meta-independência local teremos também que $(\alpha_1 \perp\!\!\!\perp \alpha_2)$ e $(\beta_1 \perp\!\!\!\perp \beta_2)$, nos permitindo escrever

$$f_1(\alpha_1, \alpha_2, \beta_1, \beta_2) = f_2(\alpha_1) f_3(\alpha_2) f_4(\beta_1) f_5(\beta_2).$$

5.3.1 Estimação por Máxima Verossimilhança

Seja $\mathcal{B} = (\mathcal{G}, P)$ uma rede bayesiana com estrutura $\mathcal{G}$ conhecida e distribuição de probabilidade conjunta P desconhecida. Queremos estimar os parâmetros de $\mathcal{B}$, definidos como

$$\theta_{i,k|j} = P(X_i = k \mid pa(X_i) = j).$$

Consideremos então a chamada *distribuição empírica* de nosso conjunto de dados $\mathcal{D}$, definida por

$$P_{\mathcal{D}}(\alpha) = \frac{\mathcal{D}\#(\alpha)}{N}$$

onde $\mathcal{D}\#(\alpha)$ é o número de vezes em que o evento α aparece no banco de dados e onde N é o número total de casos nessa base. Podemos definir também o cálculo de probabilidades condicionais empíricas como

$$P_{\mathcal{D}}(x \mid \mathbf{u}) = \frac{\mathcal{D}\#(x, \mathbf{u})}{\mathcal{D}\#(\mathbf{u})}$$

onde $\mathcal{D}\#(x, \mathbf{u})$ representa o número de vezes em que aparecem casos iguais a $(x, \mathbf{u})$ na base de dados. Além disso, se θ é o conjunto de todos os parâmetros $\theta_{i,k|j}$, temos que a função de verossimilhança de θ dada a base de dados $\mathcal{D}$ é

$$L(\theta|\mathcal{D}) = P(\mathcal{D}|\theta) = \prod_{i=1}^{N} P(d_i|\theta)\,. \tag{5.1}$$

Com as definições e informações mostradas acima, podemos apresentar o teorema que mostra como calcular estimadores por máxima verossimilhança, assim como feito no trabalho de Darwiche (2009).

Teorema 5.3.1. Seja $\mathcal{D}$ uma base de dados completa. A função de verossimilhança 5.1 é maximizada se, e somente se, θ for composto de valores calculados por

$$\hat{\theta}_{x|\mathbf{u}} = \frac{\mathcal{D}\#(x, \mathbf{u})}{\mathcal{D}\#(\mathbf{u})}$$

ou seja,

$$\hat{\theta}^{\text{mv}} = \arg\max_{\theta} L(\theta|\mathcal{D}) \iff \hat{\theta}^{\text{mv}} = \frac{\mathcal{D}\#(x, \mathbf{u})}{\mathcal{D}\#(\mathbf{u})}$$

e chamamos $\hat{\theta}^{\text{mv}}$ de *estimador de máxima verossimilhança* de θ.

Prova (Darwiche (2009), p. 470)**.**

É importante ressaltar que $\hat{\theta}^{\text{mv}}$ é o valor de θ que maximiza a probabilidade condicional de $X = d$ dado θ. Sendo assim, o estimador de máxima verossimilhança é o valor de θ que faz a probabilidade de termos visto nossa amostra a mais alta possível. Entretanto, ele não é, em geral, o valor de θ mais provável dada nossa amostra (DeGroot e Schervish (2012)).

Exemplo 5.3.1 (Darwiche (2009))**.** Suponha que construímos a estrutura da rede bayesiana da figura 5.10, que modela o problema de saber a influência da preocupação com a saúde nos hábitos de fumar e praticar exercícios físicos. As três variáveis do modelo são binárias, com $S = \{V, F\}$, $F = \{V, F\}$ e $E = \{V, F\}$, onde S indica se o indivíduo preocupa-se ou não com a saúde, F representa a relação do indivíduo com o fumo, e E revela se o indivíduo pratica ou não exercícios físicos. Para encontrarmos os parâmetros do modelo usamos o banco de dados apresentado na tabela 5.1. Aplicando a estimação por máxima verossimilhança obtemos os parâmetros mostrados na tabela 5.2.

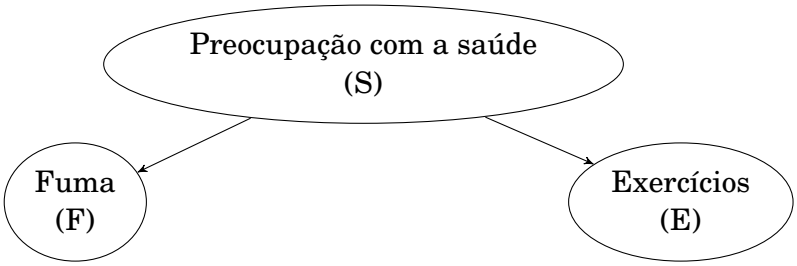


Figura 5.10: Estrutura da rede

Observação	S	F	E
1	V	F	V
2	V	F	V
3	F	V	F
4	F	F	V
5	V	F	F
6	V	F	V
7	F	F	F
8	V	F	V

Observação	S	F	E
9	V	F	V
10	F	F	V
11	V	F	V
12	V	V	V
13	V	F	V
14	V	V	V
15	V	F	V
16	V	F	V

Tabela 5.1: Banco de dados

S	$\hat{\theta}_S^{\text{mv}}$
V	3/4
F	1/4

S	F	$\hat{\theta}_{F\|S}^{\text{mv}}$
V	V	1/6
V	F	5/6
F	V	1/4
F	F	3/4

S	E	$\hat{\theta}_{E\|S}^{\text{mv}}$
V	V	11/12
V	F	1/12
F	V	1/2
F	F	1/2

Tabela 5.2: Estimadores de máxima verossimilhança

5.3.2 Estimação por Aprendizado Sequencial

O método de aprendizado sequencial reduz a estimação dos parâmetros da rede à um problema de inferência. Esse procedimento tem a capacidade de adicionar conhecimento a priori no processo de estimação, por isso é também conhecido como método bayesiano. A atualização sequencial dos parâmetros permite que um especialista da área a ser estudada ajude na elaboração do conhecimento inicial, contribuindo na construção das distribuições a priori do modelo.

Uma das vantagens da estimação por aprendizado sequencial sobre a de máxima verossimilhança, além de conseguir incorporar conhecimentos a priori, é de impedir más modelagens decorrentes de bancos de dados com poucas observações. Imagine que temos um banco de dados pequeno para o problema do exemplo

5.3.1, e que não obtivemos nenhuma observação em que $\{S = V, F = ., E = F\}$, o que nos daria, usando a estimação por máxima verossimilhança, um parâmetro $\hat{\theta}^{\text{mv}}_{E=F|S=V} = 0$, que poderia ser irrealista na visão de um especialista da área, que pode considerar que existam pessoas preocupadas com a saúde mas que, por algum motivo, não se exercitem. Através da estimação bayesiana isso não aconteceria, pois o conhecimento a priori do especialista seria considerado, impedindo que estimássemos esse parâmetro com valor igual a zero.

A abordagem bayesiana considera os parâmetros θ da rede como variáveis aleatórias, por essa razão é comum nesse contexto a representação da rede ser alterada, fazendo-se uso da chamada *meta-rede bayesiana*. Uma meta-rede bayesiana apresenta os parâmetros da rede como vértices pais das variáveis sob sua influência, por exemplo, se temos uma rede $X \to Y$ devemos considerar os parâmetros $\theta_X = P(X)$ e $\theta_{Y|X} = P(Y|X)$, sendo assim, uma meta-rede bayesiana representando esse modelo segue a forma $\theta_X \to X \to Y \leftarrow \theta_{Y|X}$.

Seguindo com o exemplo da rede $X \to Y$, vamos supor que ambas essas variáveis sejam binárias, com $X \in \{x, \bar{x}\}$ e $Y \in \{y, \bar{y}\}$, e que os parâmetros da rede também sejam binários, com $\theta_X \in \{\theta^1_X, \theta^2_X\}$ e $\theta_{Y|X} \in \{\theta^1_{Y|X}, \theta^2_{Y|X}\}$. Podemos caracterizar nosso conhecimento a priori na forma de distribuições incondicionais para os parâmetros θ_X e $\theta_{Y|X}$, por exemplo,

$$\mathbb{P}(\theta_X = \theta^1_X) = p \qquad \mathbb{P}(\theta_X = \theta^2_X) = 1 - p$$

$$\mathbb{P}(\theta_{Y|X} = \theta^1_{Y|X}) = q \qquad \mathbb{P}(\theta_{Y|X} = \theta^2_{Y|X}) = 1 - q\,.$$

Sob a hipótese de meta-independência local para os parâmetros, temos que $\mathbb{P}(\theta_X, \theta_{Y|X}) = \mathbb{P}(\theta_X)\mathbb{P}(\theta_{Y|X})$, sendo que a atualização de nosso conhecimento pode ser feita através da aplicação do teorema de Bayes, que nos diz que

$$\mathbb{P}(\theta_X \,|\, \mathcal{D}) \propto \mathbb{P}(\mathcal{D} \,|\, \theta_X)\mathbb{P}(\theta_X) \quad \text{e} \quad \mathbb{P}(\theta_{Y|X} \,|\, \mathcal{D}) \propto \mathbb{P}(\mathcal{D} \,|\, \theta_{Y|X})\mathbb{P}(\theta_{Y|X}).$$

Sendo possível partir de uma priori para os parâmetros e atualizá-la com novos dados. Após a incorporação da informação, quando já tivermos encontrado, por exemplo, $\mathbb{P}(\theta_X \,|\, \mathcal{D}_1)$, podemos incorporar mais informações a medida que essas cheguem a nós, de modo que $\mathbb{P}(\theta_X|\mathcal{D}_2) \propto \mathbb{P}(\mathcal{D}_2|\theta_X)\mathbb{P}(\theta_X|\mathcal{D}_1)$, seguindo a máxima proposta por Lindley (1972, p. 2), de que a "posteriori de hoje é a priori de amanhã", justificando o nome de *aprendizado sequencial* dado ao método.

Obviamente podemos considerar casos muito mais complexos, em que as variáveis do modelo e os "parâmetros", que aqui também são variáveis aleatórias,

possam assumir vários valores. Na rede $X \rightarrow Y$ poderíamos ter, por exemplo, $\theta_X \in \{\theta_X^1, \theta_X^2 ..., \theta_X^n\}$, com priori dada por $\mathbb{P}(\theta_X = \theta_X^i) = 1/n \quad \forall i = 1, ..., n$, caso não tivéssemos nenhum conhecimento relevante sobre θ_X. Para ilustrar melhor essa grande gama de possibilidades, consideremos o exemplo a seguir, proposto por Cowell et al. (1999).

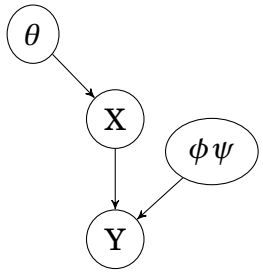


Figura 5.11: Meta-rede bayesiana

Exemplo 5.3.2. Considere a meta-rede bayesiana da figura 5.11, mostrando as variáveis binárias X e Y com espaços de estados dados por $\{x, \bar{x}\}$ e $\{y, \bar{y}\}$, respectivamente, e com parâmetros θ, ϕ e ψ. As probabilidades do modelo são:

$$\begin{aligned} P(X = x \,|\, \theta, \phi, \psi) &= \theta \\ P(Y = y \,|\, X = x, \theta, \phi, \psi) &= \phi \\ P(Y = y \,|\, X = \bar{x}, \theta, \phi, \psi) &= \psi \,. \end{aligned}$$

Sendo que as prioris para os parâmetros seguem distribuições Beta[2] com as seguintes formas:

$$\begin{aligned} p(\theta) &= B(\theta \,|\, 2, 3) \propto \theta(1-\theta)^2 \\ p(\phi) &= B(\phi \,|\, 4, 2) \propto \phi^3(1-\phi) \\ p(\psi) &= B(\psi \,|\, 1, 2) \propto 1-\psi \end{aligned}$$

Com isso, sob a hipótese de meta-independência local nos parâmetros, a priori conjunta é dada por $p(\theta, \phi, \psi) = B(\theta \,|\, 2, 3)\, B(\phi \,|\, 4, 2)\, B(\psi \,|\, 1, 2)$.

Suponha agora que tenhamos observado um dado com a ocorrência $\mathcal{D} = (\bar{x}, y)$, nos permitindo calcular a distribuição a posteriori dos parâmetros através do teorema de Bayes, nos dando que $p(\theta, \phi, \psi \,|\, \mathcal{D}) \propto p(\theta, \phi, \psi)\, P(X = \bar{x}, Y = y \,|\, \theta, \phi, \psi)$. Mas como $P(X = \bar{x}, Y = y \,|\, \theta, \phi, \psi) = (1-\theta)\psi$, chegamos em que

$$\begin{aligned} p(\theta, \phi, \psi \,|\, \mathcal{D}) &\propto (1-\theta)\, \psi\, p(\theta)\, p(\phi)\, p(\psi) \\ &= B(\theta \,|\, 2, 4)\; B(\phi \,|\, 4, 2)\; B(\psi \,|\, 2, 2) \end{aligned}$$

[2] $\text{Beta}(x \mid a, b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}$, onde $B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}\text{dt}$, com $x \in (0,1)$, $a > 0$, $b > 0$.

É importante termos em mente que alguns softwares, como o SAMIAM que será apresentado no capítulo seguinte, exigem que sejam colocados valores pontuais para os parâmetros da rede, não permitindo que sejam usadas suas distribuições de probabilidade. Nesse caso, é possível partir da distribuição desses parâmetros para usarmos valores pontuais, decorrentes de alguma operação, como, por exemplo, o cálculo do valor esperado.

Capítulo 6

Aplicações

Nos capítulos anteriores foram apresentadas várias características de redes probabilísticas e como as redes bayesianas se inserem nesse contexto. Abordamos o problema de selecionar as variáveis do modelo, mostramos algumas formas de como se construir a estrutura da rede e discutimos maneiras de obter seus parâmetros. Agora serão apresentadas, de forma resumida, algumas aplicações de redes bayesianas, na intenção de mostrar a real importância desse tipo de modelo e buscando explicar, sem grandes tecnicidades, alguns métodos usados nessas aplicações.

6.1 Diagnósticos e Previsões

Em vários estudos há questões muito comuns que podem ser resolvidas com a utilização de redes bayesianas. Uma dessas aplicações é a chamada MPE, do inglês *Most Probable Explanation*, ou *explicação mais provável*. Outra muito usada, que não passa de uma generalização da MPE, é a chamada explicação MAP, de *máxima a posteriori*.

A MPE identifica a configuração mais provável para todas as variáveis na rede dada uma certa evidência, e a explicação MAP busca a configuração de um subconjunto das variáveis que maximize a probabilidade a posteriori da evidência. Para formalizarmos a discussão, consideremos as seguintes definições (Mengshoel et al. (2010) e Darwiche (2009)):

Definição 6.1.1. Considere uma rede bayesiana com variáveis $\mathbf{X} = \{X_1, X_2, ..., X_n\}$. Uma *evidência* é uma configuração $\mathbf{e} = \{X_1 = x_1, X_2 = x_2, ..., X_m = x_m\}$ onde $m < n$. Uma *explicação* é definida como $\mathbf{x} = \{X_{m+1} = x_{m+1}, X_{m+2} = x_{m+2}, ..., X_n = x_n\}$.

Definição 6.1.2. Calcular a MPE, ou explicação mais provável, em uma rede bayesiana com evidência $\mathbf{e} = \{X_1 = x_1, X_2 = x_2, ..., X_m = x_m\}$ é o problema de en-

contrar uma explicação $\mathbf{x}^* \in \{X_{m+1} \times X_{m+2} \times ... \times X_n\}$ tal que $P(\mathbf{x}^* | \mathbf{e}) \geq P(\mathbf{y} | \mathbf{e})$, onde $\mathbf{y} \in \{X_{m+1} \times X_{m+2} \times ... \times X_n\}$ é qualquer outra explicação possível na rede.

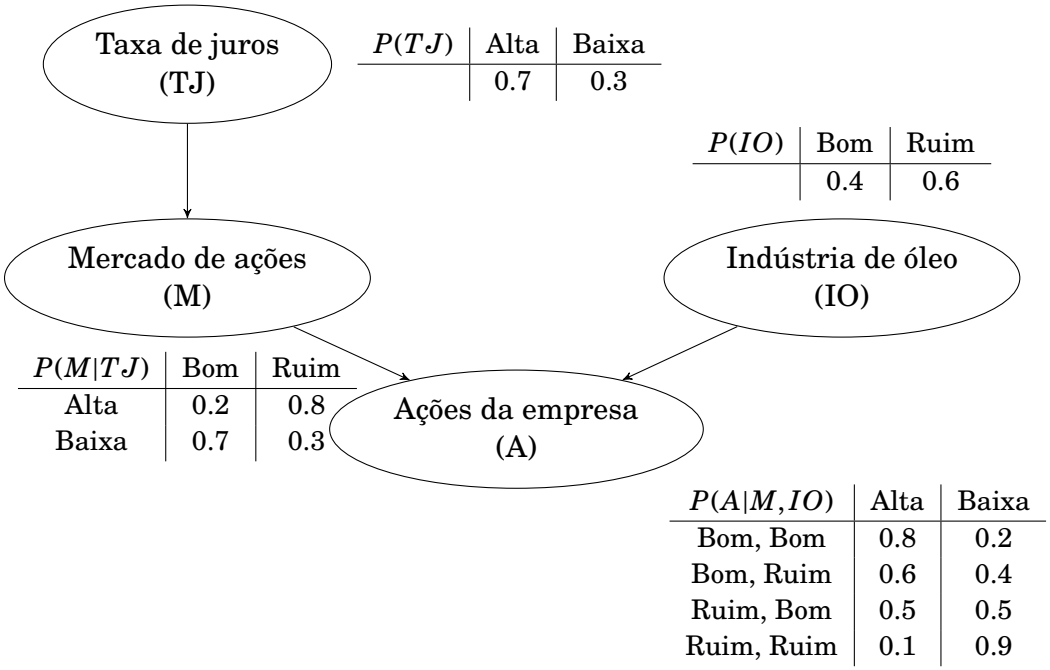
Definição 6.1.3. Calcular a explicação MAP, ou máxima a posteriori, em uma rede bayesiana com evidência $\mathbf{e} = \{X_1 = x_1, X_2 = x_2, ..., X_m = x_m\}$ é o problema de encontrar uma configuração $\mathbf{x}^* \in \{X_{m+1} \times X_{m+2} \times ... \times X_k\}$ com $k \leq n$, tal que $P(\mathbf{x}^* | \mathbf{e}) \geq P(\mathbf{y} | \mathbf{e})$, onde $\mathbf{y} \in \{X_{m+1} \times X_{m+2} \times ... \times X_k\}$ é qualquer outra configuração possível para as variaveis X_i, $\forall i = 1, ..., k$, na rede.

É fácil observar que uma explicação MPE é um caso particular de uma MAP, que ocorre quando o k da definição 6.1.3 é igual a n. Tanto a explicação MPE quanto a MAP podem ser usadas em inúmeros contextos, porém, possuem maior destaque em problemas de diagnóstico e de previsão. Em questões de diagnóstico, essas aplicações podem ser usadas para cálculos de probabilidades do tipo $P(\text{causa} | \text{sintoma})$, onde o sintoma é a evidência e a causa é a explicação buscada. No contexto de previsão, as probabilidades buscadas são do tipo $P(\text{sintoma} | \text{causa})$, onde a causa é a evidência coletada e o sintoma é o efeito futuro a ser previsto.

Para ilustrar as funcionalidades das explicações MAP e MPE citadas acima, tomemos o exemplo dado por Shenoy e Shenoy (1999), mostrado na figura 6.1. Essa rede representa as relações entre as variáveis $TJ \in$ (Alta, Baixa), $M \in$ (Bom, Ruim), $A \in$ (Alta, Baixa) e $IO \in$ (Bom, Ruim), onde TJ é o estado da taxa de juros, M é o estado do mercado de ações, A é o estado das ações de uma determinada empresa de óleo e IO é o estado da indústria de óleo no geral. Podemos, por exemplo, querer diagnosticar o motivo de uma queda nas ações da empresa, onde nossa evidência seria $\{A = \text{Baixa}\}$, ou então, podemos querer prever o estado dessas ações e do mercado caso tenhamos como evidência $\{IO = \text{Bom} \wedge TJ = \text{Alta}\}$.

Uma ferramenta computacional muito útil para esse tipo de análise é o programa SAMIAM (*Sensitivity Analysis, Modeling, Inference And More*)[1], desenvolvido na Universidade da Califórnia, Los Angeles, pelo grupo de Adnan Darwiche. Na figura 6.2 é mostrada a rede descrita acima no formato do SAMIAM, na figura 6.3 é mostrada a explicação MAP de diagnóstico para a evidência $\{A = \text{Baixa}\}$, nos dando que a configuração mais provável para essa informação é $\{IO = \text{Ruim} \wedge M = \text{Ruim} \wedge TJ = \text{Alta}\}$. Para o caso de previsão do estado do mercado e das ações da empresa, dada a evidência $\{IO = \text{Bom} \wedge TJ = \text{Alta}\}$, temos que explicação MAP dá a configuração $\{A = \text{Alta} \wedge M = \text{Ruim}\}$, como mostrado na figura 6.4.

[1] Disponível em: http://reasoning.cs.ucla.edu/samiam/



$P(TJ)$	Alta	Baixa
	0.7	0.3

$P(IO)$	Bom	Ruim
	0.4	0.6

$P(M\|TJ)$	Bom	Ruim
Alta	0.2	0.8
Baixa	0.7	0.3

$P(A\|M,IO)$	Alta	Baixa
Bom, Bom	0.8	0.2
Bom, Ruim	0.6	0.4
Ruim, Bom	0.5	0.5
Ruim, Ruim	0.1	0.9

Figura 6.1: Rede das ações da empresa de óleo

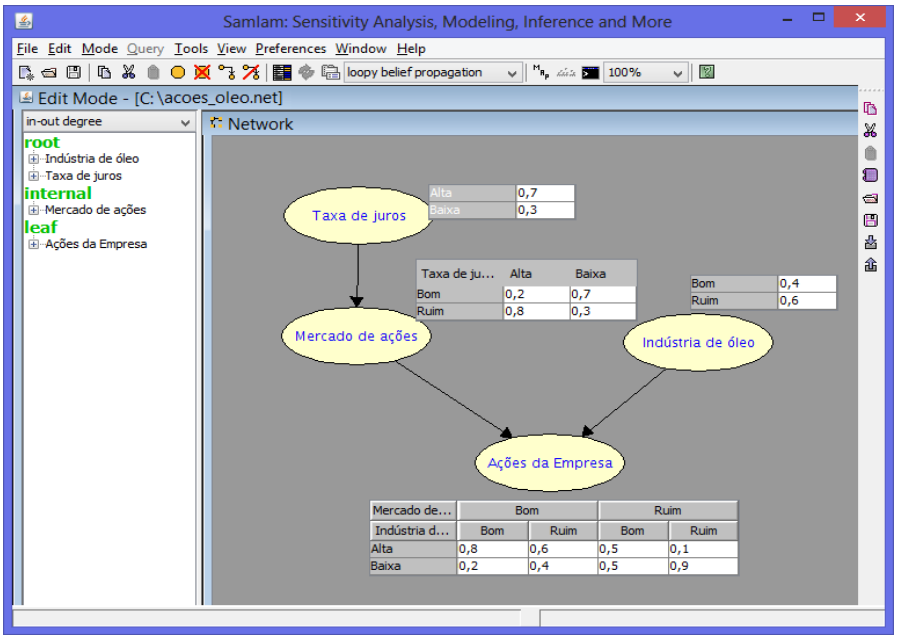


Figura 6.2: Rede da empresa de óleo no programa SAMIAM

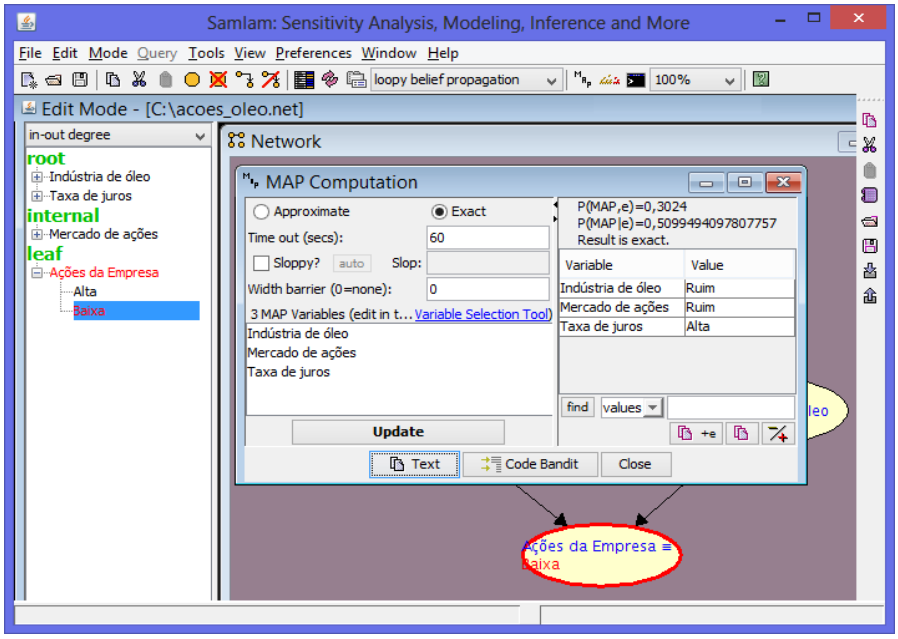


Figura 6.3: Explicação MAP de diagnóstico para $\{A = \text{Baixa}\}$

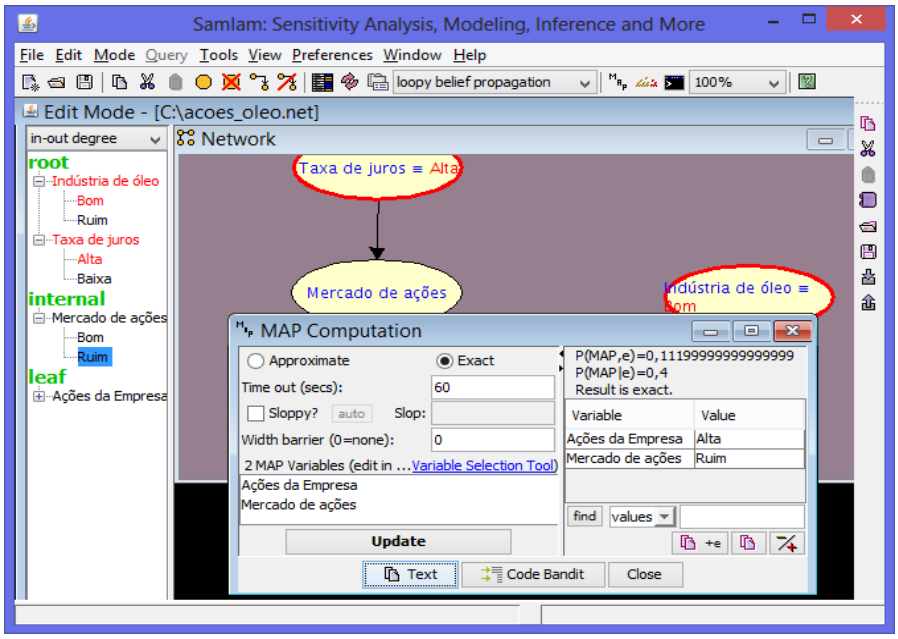


Figura 6.4: Explicação MAP de previsão para $\{IO = \text{Bom} \wedge TJ = \text{Alta}\}$

6.2 Decisões

A informação obtida através de uma rede bayesiana muitas vezes pode ser usada para tomarmos decisões perante o problema estudado. Entretanto, a rede não dá nenhuma "recomendação" sobre qual decisão devemos tomar. No exemplo da seção anterior, em que vimos que a evidência $\{IO = \text{Bom} \wedge TJ = \text{Alta}\}$ tem como mais provável consequência a configuração $\{A = \text{Alta} \wedge M = \text{Ruim}\}$, poderíamos decidir comprar ações da empresa, já que elas estariam em alta. Porém, essa decisão seria embasada somente numa percepção intuitiva sobre o resultado da explicação MAP. Existe uma modelagem que faz com que a rede realmente indique uma decisão a ser tomada. Essa rede é chamada de *diagrama de influência* (Barlow e Pereira (1990)).

Um diagrama de influência contém três tipos de vértices: de incerteza, de decisão e de utilidade. Os vértices de incerteza são os que foram tratados até agora, sendo probabilisticamente dependentes de seus vértices pais. Vértices de decisão representam, como o nome sugere, as decisões a serem tomadas, sendo os valores de seus vértices pais já conhecidos no momento da decisão. Vértices de utilidade representam a ordenação das preferências do decisor em relação às suas possíveis opções. Os vértices de utilidade são deterministicamente dependentes de seus vértices pais. A representação visual desses vértices seguirá o formato proposto por Neapolitan (2003), mostrado na figura 6.5.

Figura 6.5: Representação de vértices de incerteza, decisão e utilidade

Uma *decisão* é um elemento de um conjunto de ações exaustivas e mutuamente exclusivas que o decisor pode tomar. A *utilidade* de um resultado é o valor desse resultado para o decisor. Muitas vezes a utilidade de um resultado depende não somente do valor em si, mas também do risco envolvido na decisão. Se tivermos, por exemplo, duas loterias: na loteria I você ganha \$5.000,00 com 100% de chance, na loteria II você pode ganhar \$7.000,00 com 80% de chance ou \$0,00 com 20% de chance. A loteria I tem lucro médio de \$5.000,00 e a loteria II tem lucro médio maior, de \$5.600,00. Porém, é provável que a maior parte das pessoas escolha a loteria I, mesmo com valor médio menor, pois, além do lucro, também é considerado o risco da decisão. Todavia, para evitar maiores tecnicidades, vamos considerar nos exemplos que virão a seguir que o decisor possui uma fortuna tão grande que

a utilidade de suas decisões será somente baseada na quantia monetária resultante, não sendo necessário para ele considerar o risco envolvido. Para maiores detalhes sobre como incorporar a influência do risco nas decisões, indicamos os trabalhos de Koller e Friedman (2009) e de Bekman e Neto (2009).

Após a identificação das possíveis decisões a serem tomadas e análise da utilidade de cada possível resultado, devemos determinar um critério para tomarmos nossa decisão. Existem diversas formas de se construir esse critério, mas aqui consideraremos o critério da maximização da utilidade esperada, que ficará clara com os exemplos apresentados a seguir, baseados no trabalho de Neapolitan (2003).

Exemplo 6.2.1. Suponha que temos $\$1.000,00$ para investir em uma de duas possíveis aplicações: comprar ações da empresa XPTO, que no momento valem $\$10,00$ cada, ou investir em renda fixa. Através de nossas análises calculamos que as ações da XPTO, ao final do mês, valerão $\$5,00$, $\$10,00$ ou $\$20,00$, com probabilidades $0,25$, $0,25$ e $0,5$, respectivamente. A aplicação de renda fixa terá gerado, também ao final do mês, uma rentabilidade de $0,5\%$ com 100% de chance. Nosso problema de decisão é: em qual das duas aplicações devemos colocar nossos $\$1.000,00$?

Para facilitar a visualização do problema, consideremos o diagrama de influência da figura 6.6, onde $U(d_2, n)$ representa a utilidade de decidirmos investir em renda fixa para qualquer valor de ação da XPTO ao final do mês. A construção desse diagrama de influência considera que nossa decisão não influenciará o preço das ações da XPTO, já que $\$1.000,00$ não deve ter força suficiente para alterar as forças de oferta e demanda do mercado.

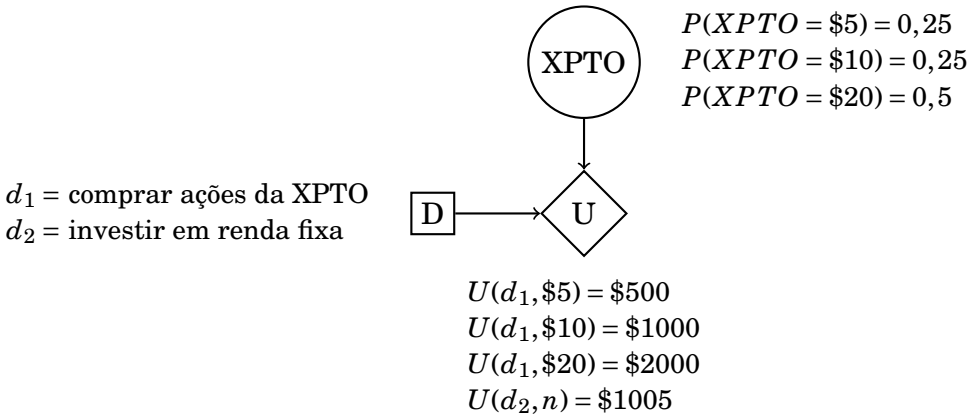


Figura 6.6: Diagrama de influência: investir em ações ou renda fixa?

Com todas essas informações disponíveis, podemos tomar nossa decisão com base nos valores das utilidades esperadas (UE) de cada decisão:

$$\begin{aligned}UE(d_1) &= E(U|d_1)\\ &= P(\$5|d_1)U(d_1,\$5) + P(\$10|d_1)U(d_1,\$10) + P(\$20|d_1)U(d_1,\$20)\\ &= (0,25)(\$500) + (0,25)(\$1000) + (0,5)(\$2000)\\ &= \$1375\end{aligned}$$

$$\begin{aligned}UE(d_2) &= E(U|d_2)\\ &= P(\$5|d_2)U(d_2,\$5) + P(\$10|d_2)U(d_2,\$10) + P(\$20|d_2)U(d_2,\$20)\\ &= (0,25)(\$1005) + (0,25)(\$1005) + (0,5)(\$1005)\\ &= \$1005\end{aligned}$$

Sendo assim, nossa decisão D deve ser aquela que maximiza a utilidade esperada, ou seja, $D = \arg\max\{UE(d_1), UE(d_2)\}$, o que nos leva a decidir por $D = d_1$.

Exemplo 6.2.2. Suponha que somos investidores com um grande capital, e que queremos decidir se compramos 10.000 ações da empresa ABC, com cada ação por $\$10,00$, ou se compramos um contrato de opção de compra por $\$100.000,00$, que nos daria o direito de comprar 50.000 ações da ABC por $\$15,00$ cada ao fim de um mês. Se escolhermos a alternativa de comprar 10.000 ações, vamos impactar o mercado devido ao grande volume da compra, e o preço das ações da ABC subirão. Outro fator a ser considerado é o índice de ações IA, que também pode influenciar o preço da ABC. Acreditamos que, em um mês, o índice de ações IA estará em 10.000 ou 11.000 pontos, e que as ações da ABC custarão $\$5,00$ ou $\$20,00$. O diagrama de influência da figura 6.7 mostra a estrutura do modelo e as probabilidades construídas a partir de nossas crenças.

Para calcularmos a utilidade esperada de cada decisão vamos precisar dos valores de $P(\$5|d_1)$, $P(\$20|d_1)$, $P(\$5|d_2)$ e $P(\$20|d_2)$. O cálculo para $P(\$5|d_1)$ é:

$$\begin{aligned}P(\$5|d_1) &= P(\$5|11.000, d_1)\,P(11.000) + P(\$5|10.000, d_1)\,P(10.000)\\ &= (0,2)(0,6) + (0,5)(0,4)\\ &= 0,32\end{aligned}$$

Os demais cálculos seguem a mesma lógica. Já as utilidades esperadas são:

$$\begin{aligned}UE(d_1) &= E(U|d_1)\\ &= P(\$5|d_1)U(d_1,\$5) + P(\$20|d_1)U(d_1,\$20)\\ &= (0,32)(\$50.000) + (0,68)(\$200.000)\\ &= \$152.000\end{aligned}$$

$$\begin{aligned} UE(d_2) &= E(U|d_2) \\ &= P(\$5|d_2)U(d_2,\$5) + P(\$20|d_2)U(d_2,\$20) \\ &= (0,42)(\$0) + (0,58)(\$250.000) \\ &= \$145.000 \end{aligned}$$

Sendo assim, como $D = \arg\max\{UE(d_1), UE(d_2)\} = d_1$, decidimos por comprar as ações da ABC ou invés de comprarmos o contrato de opção.

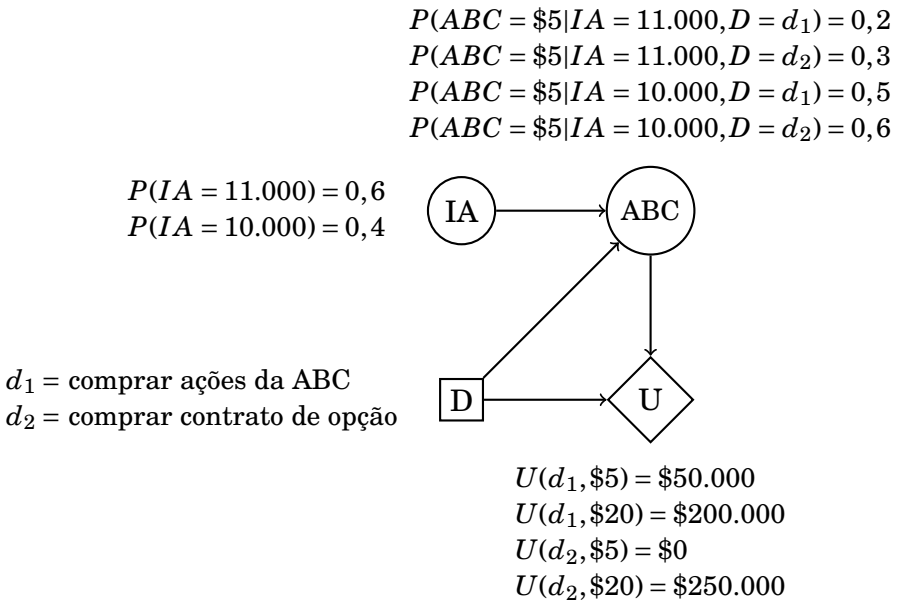


Figura 6.7: Diagrama de influência: investir em ações ou opção de compra?

Capítulo 7

Considerações Finais

Redes probabilísticas são poderosas ferramentas para modelar a interação entre variáveis. As redes bayesianas, em particular, oferecem uma representação gráfica muito intuitiva sobre as relações de independência entre tais variáveis, proporcionando uma formatação simples, mas com grande utilidade para pesquisas e aplicações em diversas áreas.

O processo de construção de uma rede bayesiana pode ser dividido em três etapas: (i) seleção das variáveis de interesse, (ii) construção da estrutura da rede e (iii) estimação dos parâmetros. A primeira etapa, mesmo havendo oferta de métodos de seleção automática, é usualmente feita de forma manual, baseando-se nas percepções subjetivas de um *expert* da área estudada. A segunda etapa pode ser feita de diversas maneiras, sendo possível construir a estrutura da rede manualmente, da qual parte-se dos conhecimentos de um especialista, ou semi-automaticamente, através de algoritmos baseados em restrições ou de algoritmos baseados em escores. A terceira etapa pode ser um pouco controversa, já que os parâmetros da rede podem ser abordados de diferentes formas. A estimação por máxima verossimilhança considera os parâmetros como sendo valores fixos e desconhecidos a serem estimados. No método de aprendizado sequencial, ou bayesiano, os parâmetros são considerados variáveis aleatórias, o que permite a atualização de suas distribuições através do método de inferência bayesiana.

Redes bayesianas podem ser usadas em diversas aplicações, tendo destaque em situações que precisem de diagnósticos, onde são calculadas probabilidades do tipo $P(\text{causa}|\text{sintoma})$, e de cenários em que sejam necessárias previsões, que demandam cálculos de probabilidades do tipo $P(\text{sintoma}|\text{causa})$. Além disso, com a simples adição de variáveis de decisão e de utilidade no modelo, uma rede bayesiana pode se tornar uma ferramenta muito útil para auxiliar na tomada de decisões complexas.

Futuras pesquisas podem ser desenvolvidas em diversos campos citados neste trabalho, havendo ainda possibilidade de muitas melhorias na modelagem de re-

des bayesianas. Ainda há, por exemplo, espaço no desenvolvimento de métodos mais eficientes para a construção de estruturas através de algoritmos baseados em escores. Já na questão do aprendizado sequencial dos parâmetros, há uma grande lacuna a ser preenchida com a aplicação de diferentes métodos de atualização de probabilidades, como, por exemplo, a regra de Jeffrey, citada ainda no início do trabalho.

Referências Bibliográficas

Barlow, R. E. & Pereira, C. A. B. (1990). Conditional Independence and Probabilistic Influence Diagrams. Lecture Notes-Monograph Series, 16, 19-33.

Bekman, O. R. & Neto, P. L. de O. C. (2009). Análise Estatística da Decisão, 2ª Edição, Blucher, São Paulo, SP.

Carnap, R. (1950). Logical foundations of probability. Chicago.

Cobb, B. R., Rumí, R. & Salmerón, A. (2007). Bayesian Network Models with Discrete and Continuous Variables. In: Lucas, P., Gomez, J., Sameron, A. (eds.) Advances in Probabilistic Graphical Models, 81-102.

Corfield, D. & Williamson, J., (eds.) (2001). Foundations of Bayesianism, 1-16, Kluwer Academic Publishers.

Cowell, R. G.; Dawid, A. P.; Lauritzen, S. L. & Spiegelhalter, D. J. (1999). Probabilistic Networks and Expert Systems. Spinger, New York.

Darwiche, A. (2009). Modeling and Reasoning with Bayesian Networks, Cambridge University Press.

Dawid, A. P. (1979). Conditional independence in statistical theory (with discussion). J. R. Stat. Soc. B 41: 1-31.

Dawid, A. P. (2010). Beware of the DAG! In Proceedings of the NIPS 2008 Workshop on Causality. Journal of Machine Learning Research Workshop and Conference Proceedings (D. Janzing, I. Guyon and B. Scholkopf, eds.), 6, 59–86. http://tinyurl.com/33va7tm.

De Finetti, B. (1931). Sul significato soggettivo della probabilità. Fundamenta Mathematicae. 17, 298-329.

DeGroot, M. H. & Schervish, M.J., (2012). Probability and Statistics, 4ª ed., Pearson Education, Boston, MA.

Diaconis, P. & Zabell, S. L. (1982). Updating subjective probability. Journal of the American Statistical Association, 77, 822-830.

Edwards, D. (2000). Introduction to Graphical Modelling, 2ª ed. Springer, New York.

Flesch, I. & Lucas, P. J. F. (2007). Markov equivalence in Bayesian networks. In: Lucas, P., Gomez, J., Sameron, A. (eds.) Advances in Probabilistic Graphical Models, 3–38.

Freedman, D. (2010). Statistical models and causal inference: A dialogue with the social sciences (D. Collier, J. S. Sekhon, & P. B. Stark, Ed.). New York, NY: Cambridge University Press.

George, E. I. (2000). The Variable Selection Problem. Journal of the American Statistical Association, Vol. 95, No. 452, 1304-1308.

Gopnik, A. & Schulz, L. (Eds.). (2007). Causal learning: Psychology, philosophy, and computation. Oxford, U.K.: Oxford University Press.

Hájek, A. (2011). Interpretations of Probability. The Stanford Encyclopedia of Philosophy. Edward N. Zalta (ed.) http://plato.stanford.edu/entries/probability-interpret/.

Jeffrey, R. (1992). Probability and the art of judgement. Cambridge: Cambridge University Press.

Jeffrey, R. (1965). The Logic of Decision. New York: McGraw-Hill.

Jeffreys, H. (1939). Theory of Probability; reprinted in Oxford Classics in the Physical Sciences series, Oxford: Oxford University Press, 1998.

Jensen, F. V. & Nielsen, T. D. (2007). Bayesian Networks and Decision Graphs, 2ª ed. Springer Verlag.

Kadane, J. B., & Lazar, N. A. (2004). Methods and criteria for model selection. Journal of the American Statistical Association, 99, 279–290.

Keynes, J. M. (1921). Treatise on Probability, London: Macmillan & Co.

Kjærulff, U. B. & Madsen A. L. (2013). Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis, 2a ed. Springer.

Koller, D. & Friedman, N. (2009). Probabilistic Graphical Models: Principles and Techniques. MIT Press.

Koski, T. & Noble, J. M. (2009). Bayesian Networks. An Introduction. John Wiley and Sons.

Laplace, P. S. (1814). A Philosophical Essay on Probabilities. Edição em inglês (1951). New York: Dover Publications.

Lauritzen, S. L. (1996). Graphical Models. Clarendon Press, Oxford.

Lindley, D.V. (1972). Bayesian Statistics, A Review. SIAM, Philadelphia.

Loschi, R. & Wechsler, S. (2002). Coherence, Bayes's theorem and posterior distributions. Brazilian Journal of Probability and Statistics, 16, pp. 169-185.

Mahmood, A. (2011). Structure learning of causal bayesian networks: A survey. Technical report TR11-01, Department of Computing Science, University of Alberta, Edmonton, AB, Canada T6G 2E8.

Mengshoel, O. J., Roth, D. & Wilkins, D. C. (2010). Portfolios in stochastic local search: Efficiently computing most probable explanations in Bayesian networks, Journal of Automated Reasoning.

Neapolitan, R. (2003). Learning Bayesian networks. Prentice Hall.

Parmigiani, G. & Inoue, L. (2009). Decision Theory: Principles and Approaches. John Wiley & Sons.

Pearl, J. (2000). Causality: models, reasoning, and inference. Cambridge University Press.

Pearl, J. & Paz, A. (1987). Graphoids: a graph based logic for reasoning about relevancy relations. Advances in Artificial Intelligence - II, (ed. B. D. Boulay, D. Hogg, and L. Steel), pp. 357-63. North-Holland, Amsterdam.

Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Infer- ence. Morgan Kaufmann.

Ramsey, F. P. (1926). Truth and Probability. The Foundations of Mathematics and Other Logical Essays. Publicação de 1931. Cap. VII, 156-198.

Reichenbach, H. (1949). The Theory of Probability, Berkeley: University of California Press.

Robert, C. (2001). The Bayesian Choice: From Decision-Theoretic Motivations to Computational Implementation, 2ª ed. Springer: New York.

Robins, J. & Wasserman, L. (2000). Conditioning, likelihood and concepts: A review of some foundational concepts. Journal of the American Statistical Association, Vol. 95, No. 452, pp. 1340-1346.

Ross, S. M. (1996). Stochastic Process, John Wiley and Sons, Inc.

Roussas, G. G. (1997). A Course in Mathematical Statistics, 2ª ed. Burlington, MA: Harcourt/Academic Press.

Scutari, M. (2010). Learning Bayesian Networks with the bnlearn R Package. Journal of Statistical Software. Volume 35, Issue 3.

Shenoy, C. & Shenoy, P. P. (1999). Bayesian network models of portfolio risk and return. Y. S. Abu-Mostafa, B. LeBaron, A. W. Lo & A. S. Weigend (Eds.), Computational Finance, MIT Press, Cambridge, MA, 87-106.

Spirtes, P., Glymour, C. & Scheines, R. (2000). Causation, Prediction, and Search; 2ª ed. The MIT Press. Cambridge, Massachusetts.

Venn, J. (1876). The Logic of Chance, 2ª ed. London: Macmillan; reprinted, New York: Chelsea Publishing Co., 1962.

Verma, T. & Pearl, J. (1991). Equivalence and Synthesis of Causal Models. Uncertainty in Artificial Intelligence, 6, 255-268.

Von Mises, R. (1928). Wahrscheinlichkeit, Statistik und Wahrheit. Viena.