**Modelos de aprendizado supervisionado
usando métodos kernel, conjuntos fuzzy
e medidas de probabilidade**

Jorge Luis Guevara Díaz

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

Programa: Ciência da Computação
Orientador: Prof. Dr. Roberto Hirata Junior

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da CAPES, CNPq e da FAPESP

São Paulo, 04 Maio de 2015

Modelos de aprendizado supervisionado usando métodos kernel, conjuntos fuzzy e medidas de probabilidade

Esta versão da dissertação/tese contém as correções e alterações sugeridas pela Comissão Julgadora durante a defesa da versão original do trabalho, realizada em 04/05/2015. Uma cópia da versão original está disponível no Instituto de Matemática e Estatística da Universidade de São Paulo.

Comissão Julgadora:

- Prof^a. Dr^a. Roberto Hirata Jr. (orientador) - IME-USP
- Prof. Dr. André Carlos ponde de Leon F. de Carvalho - ICMC-USP
- Prof. Dr. Alejandro César Frery Orgambide - UFAL
- Prof. Dr. Paulo José da Silva e Silva - UNICAMP
- Prof. Dr. Peter Sussner - UNICAMP

Para minha mãe Ester.

Agradecimentos

Primeiro antes que nada, agradeço ao meu orientador e amigo Professor Roberto Hirata Júnior, pela motivação, guia e apoio inestimável nos temas acadêmicos, e a sua preocupação com meu desenvolvimento pessoal e profissional. Agradeço também ao Professor Stéphane Canu, com quem tive o privilégio de trabalhar durante a etapa do meu doutorado sanduíche na França, nos laboratórios do INSA de Rouen.

Agradeço, às diversas instituições que apoiaram o desenvolvimento desta tese. À Capes, pela bolsa de doutorado. À CNPq, pela bolsa de doutorado sanduíche. À Fapesp, pela bolsa de treinamento técnico. Agradeço ao IME e à USP por permitir-me cursar doutorado. Agradeço ao INSA de Rouen por ter-me disponibilizado um excelente ambiente de trabalho e pelo apoio financeiro para assistir a eventos internacionais.

Agradeço mais que todo a minha família toda, ao amor, suporte e motivação da minha esposa Leissi, presente ao meu lado em todas as situações da vida. A o meu filho Christopher, fonte de motivação inesgotável e amor infinito que a vida me deu como presente. À minha irmã Yaque, por seus bons pensamentos, palavras, amor e amizade. Agradeço a meu pai Lucho, meus amigos Orlando, Jihan, Talita, Jesús, David e Leandro pelo seu apoio constante em diversas situações.

Agradeço de maneira especial, com muito amor e carinho, à minha mãe, pessoa lutadora de muita coragem, a qual amo e admiro, pelos sábios ensinamentos que eu tive com ela para minha vida toda. Agradeço ela por ter sido minha maior fonte de inspiração de exemplo de vida.

Finalmente, agradeço lá no céu, a minha madrinha Maria. Obrigado pelo seu amor, e por todas as coisas que você fez por mim.

Resumo

GUEVARA DÍAZ, J. **Modelos de aprendizado supervisionado usando métodos kernel, conjuntos fuzzy e medidas de probabilidade**. 2015. 129 f. Tese (Doutorado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2015.

Esta tese propõe uma metodologia baseada em métodos de kernel, teoria fuzzy e probabilidade para tratar conjuntos de dados cujas observações são conjuntos de pontos. As medidas de probabilidade e os conjuntos fuzzy são usados para modelar essas observações. Posteriormente, graças a kernels definidos sobre medidas de probabilidade, ou em conjuntos fuzzy, é feito o mapeamento implícito dessas medidas de probabilidade, ou desses conjuntos fuzzy, para espaços de Hilbert com kernel reproduzível, onde a análise pode ser feita com algum método kernel. Usando essa metodologia, é possível fazer frente a uma ampla gama de problemas de aprendizado para esses conjuntos de dados. Em particular, a tese apresenta o projeto de modelos de descrição de dados para observações modeladas com medidas de probabilidade. Isso é conseguido graças ao mergulho das medidas de probabilidade nos espaços de Hilbert, e a construção de esferas envolventes mínimas nesses espaços de Hilbert. A tese apresenta como esses modelos podem ser usados como classificadores de uma classe, aplicados na tarefa de detecção de anomalias grupais. No caso que as observações sejam modeladas por conjuntos fuzzy, a tese propõe mapear esses conjuntos fuzzy para os espaços de Hilbert com kernel reproduzível. Isso pode ser feito graças à projeção de novos kernels definidos sobre conjuntos fuzzy. A tese apresenta como esses novos kernels podem ser usados em diversos problemas como classificação, regressão e na definição de distâncias entre conjuntos fuzzy. Em particular, a tese apresenta a aplicação desses kernels em problemas de classificação supervisionada em dados intervalares e teste kernel de duas amostras para dados contendo atributos imprecisos.

Palavras-chave: métodos de kernel, kernel positivo definido, conjunto fuzzy, medidas de probabilidade, máquina de vetor de suporte, descrição de dados usando vetores de suporte.

Abstract

GUEVARA DÍAZ, J. **Supervised machine learning models using kernel methods, probability measures and fuzzy sets**. 2015. 129 f. Tese (Doutorado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2015.

This thesis proposes a methodology based on kernel methods, probability measures and fuzzy sets, to analyze datasets whose individual observations are itself sets of points, instead of individual points. Fuzzy sets and probability measures are used to model observations; and kernel methods to analyze the data. Fuzzy sets are used when the observation contain imprecise, vague or linguistic values. Whereas probability measures are used when the observation is given as a set of multi-dimensional points in a D -dimensional Euclidean space. Using this methodology, it is possible to address a wide range of machine learning problems for such datasets. Particularly, this work presents data description models when observations are modeled by probability measures. Those description models are applied to the group anomaly detection task. This work also proposes a new class of kernels, *the kernels on fuzzy sets*, that are reproducing kernels able to map fuzzy sets to a geometric feature spaces. Those kernels are similarity measures between fuzzy sets. We give from basic definitions to applications of those kernels in machine learning problems as supervised classification and a kernel two-sample test. Potential applications of those kernels include machine learning and pattern recognition tasks over fuzzy data; and computational tasks requiring a similarity measure estimation between fuzzy sets.

Keywords: kernel methods, positive definite kernel, fuzzy set, probability measure, support vector machine, support vector data description.

Sumário

Lista de Abreviaturas	xiii
Lista de Símbolos	xv
Lista de Figuras	xvii
Lista de Tabelas	xix
1 Introdução	1
1.1 Abordagem usada na tese	2
1.1.1 Modelagem das observações usando Medidas de Probabilidade	2
1.1.2 Modelagem das observações usando conjuntos fuzzy	4
1.2 Contribuições	4
1.3 Estrutura da tese	5
2 Teoria fuzzy	9
2.1 Introdução	9
2.2 Normas Triangulares	9
2.3 Conjunto Fuzzy	10
2.4 Sistemas de Regras Fuzzy	13
2.4.1 Sistema Fuzzy de Mamdani	14
3 Métodos de kernel	17
3.1 Introdução	17
3.2 Espaços de Hilbert com Kernel Reproduzível	18
3.2.1 Funções no EHKR	19
3.2.2 Propriedades de fechamento dos kernels	20
3.2.3 Exemplos de kernels	21
3.3 Máquinas de vetores de suporte	21
3.4 Descrição dos dados com vetores de suporte	25
3.5 Discrepância Máxima de Médias nos EHKR	26
3.5.1 Teste de hipótese com a MMD	27
4 Descrição de Dados com Medidas de Suporte	29
4.1 Introdução	29
4.2 Descritor de dados com medidas de suporte no espaço das medidas de probabilidade	35

4.2.1	Formulação usando a desigualdade de Markov	36
4.3	Descrição de dados com medidas de suporte em espaços de Hilbert com kernel reproduzível	40
4.3.1	Mergulho das medidas de probabilidade nos EHKR	40
4.3.2	SMDD no EHKR como problema com restrições de chance	41
4.3.3	SMDD no EHKR como extensão direta do SVDD	46
4.3.4	Modelo SMDD no EHKR usando funções com norma um e kernels invariantes à translação	47
4.4	Relação entre os modelos SMDD	48
4.4.1	Equivalência entre os modelos SMDD	50
4.4.2	Conexão com as Máquinas de vetores de suporte de uma classe	50
4.5	Experimentos	50
4.5.1	Detecção de anomalias grupais pontuais num conjunto de dados cujas observações são distribuições de misturas gaussianas	51
4.5.2	Detecção de anomalias grupais de distribuição num conjunto de dados cujas observações são distribuições de misturas gaussianas	55
4.5.3	Detecção de anomalias grupais em dados astronomicos	57
4.6	Sumário	62
5	Kernels definidos sobre conjuntos fuzzy	63
5.1	Introdução	63
5.1.1	Trabalhos prévios	64
5.2	Kernels definidos sobre conjuntos	64
5.2.1	O kernel de convolução	65
5.2.2	O kernel de produto cruzado	65
5.3	Kernels definidos sobre conjuntos fuzzy	66
5.3.1	O kernel de produto cruzado de conjuntos fuzzy	66
5.3.2	O kernel de convolução fuzzy	68
5.4	O kernel $k_{\cap}$ de interseção de conjuntos fuzzy	69
5.4.1	Medida do suporte dos conjuntos fuzzy	70
5.4.2	O kernel $k_{\cap}$	71
5.5	Kernels em conjuntos fuzzy usando substituição de distância	75
5.5.1	Kernels com substituição de distância	76
5.5.2	Métricas, pseudo-métricas e semi-métricas em conjuntos fuzzy	77
5.6	Aplicações	78
5.6.1	Distâncias de conjuntos fuzzy mediante mergulho no EHKR	78
5.6.2	Engenharia de kernels definidos sobre conjuntos fuzzy	79
5.6.3	Problemas de aprendizado supervisionado	80
5.7	Experimentos	81
5.7.1	Descrição do conjunto de dados de prognóstico de reincidência do câncer de mama	81
5.7.2	Fuzzificação dos atributos do conjunto de dados	82
5.7.3	Teste kernel de duas amostras	84
5.7.4	Resultados	86

5.8	Sumário	88
6	Kernels em Conjuntos Fuzzy baseados no Sistema de Regras Fuzzy Takagi Sugeno Kang	89
6.1	Introdução	89
6.1.1	Contribuições	90
6.2	Sistema Fuzzy Takagi Sugeno Kang	90
6.3	Sistema Fuzzy TKS non-singleton	91
6.4	Kernels baseados em regras fuzzy	93
6.4.1	Kernel PDFC	93
6.4.2	Kernel Takagi Sugeno Kang (TSK)	93
6.5	Kernels TSK non-singleton	94
6.5.1	Kernel gaussiano TSK non-singleton	96
6.6	Kernel TSK non-singleton como relação de equivalencia fuzzy	98
6.7	Experimentos em classificação supervisionada	99
6.7.1	Dados e implementação	99
6.8	Sumário	104
7	Conclusões e Trabalhos Futuros	107
7.1	Resumo das contribuições	108
7.2	Publicações geradas	109
7.2.1	<i>Support Measure Data Description.</i>	110
7.2.2	<i>Positive Definite Kernel Functions on Fuzzy Sets</i>	111
7.2.3	<i>Kernel Functions in Takagi-Sugeno-Kang Fuzzy System with Nonsingleton Fuzzy Input.</i>	112
7.2.4	<i>Support Vector Data Description for Uncertainty Datasets</i>	113
7.2.5	<i>Fuzzy Set Similarity using a Distance-Based Kernel on Fuzzy Sets</i>	114
A	Provas	115
A.1	Prova do Lema 3	115
A.2	Prova do Teorema 5	115
A.3	Prova do Lema 6	116
A.4	Prova do Teorema 8	116
A.5	Prova do Teorema 9	117
A.6	Prova do Corolário 11	118
	Referências Bibliográficas	119

Lista de Abreviaturas

ACC	Acurácia.
AUC	Área abaixo da curva (<i>Area under the curve</i>).
CPD	condicionalmente positivo definido (<i>Conditionally positive definite</i>)
EHKR	Espaço de Hilbert com Kernel Reproduzível.
GMD	Distribuição de misturas Gaussianas (<i>Gaussian Mixture Distribution</i>).
i.d.d	Independente e identicamente distribuídas.
MV-set	Conjunto de volume mínimo (<i>Minimum Volume Set</i>).
MMD	Discrepância máxima de médias (<i>Maximum Mean Discrepancy</i>).
OCSVM	Máquina de vetor de suporte de uma classe (<i>One Class Support Vector Machine</i>).
PCA	Análise de componentes principais (<i>Principal Component Analysis</i>).
PDFC	Classificador fuzzy positivo definido (<i>Positive Definite Fuzzy Classifier</i>).
ROC	Característica Operativa do Receptor (<i>Receiver Operating Characteristic</i>).
SMDD	Descrição de dados com medidas de suporte (<i>Support Measure Data Description</i>).
SVM	Máquina de vetores de suporte (<i>Support Vector Machine</i>).
T-norma	Norma Triangular (<i>Triangular Norm</i>).
TSK	Takagi Sugeno Kang.

Lista de Símbolos

$\mathbb{N}$	Espaço dos números naturais.
$\mathbb{R}$	Espaço dos números reais.
$\mathbb{R}^+$	Espaço dos números reais positivos.
K	Matriz kernel.
s	Conjunto de pontos cujos elementos pertencentes a $\mathbb{R}^D$.
E, Ω	Conjunto não vazio.
D	Dimensão de um vetor.
N	Número de elementos num conjunto de dados.
$(R, \mathbf{c})$	Esfera com raio $R \in \mathbb{R}^+$, e centro $\mathbf{c} \in \mathbb{R}^D$.
λ	Parâmetro de regularização de um kernel.
ξ	Vetor de variáveis de folga pertencente a $\mathbb{R}^D$.
$\mathcal{P}$	Espaço das medidas de probabilidade em $(\mathbb{R}^D, \mathcal{B}(\mathbb{R}^D))$.
$\mathcal{E}$	Medida de probabilidade.
$\mathbb{P} \sim \mathcal{E}$	Medida de probabilidade aleatória distribuída de acordo a $\mathcal{E}$.
$\mathbb{P}$	Medida de probabilidade.
$X \sim \mathbb{P}$	Variável aleatória tomando valores em $\mathbb{R}^D$ e distribuída de acordo a $\mathbb{P}$.
κ	Valores limiar pertencentes a $[0, 1]$.
$tr(A)$	Traço da matriz A .
μ	Média do conjunto de pontos s .
Σ	Matriz de covariância.
α, β, η	Multiplicadores de Lagrange.
α, β	Vetores contendo os multiplicadores de Lagrange α, β , respectivamente.
$\mathcal{H}$	Espaço de Hilbert com kernel reproduzível (EHKR).
k	Kernel reproduzível.
$k(x, \cdot)$	Função de avaliação no ponto $x \in \Omega$, tal que em $k(x, \cdot) \in \mathcal{H}$.
c ou $c(\cdot)$	Centro da esfera no EHKR $\mathcal{H}$.
$\mu_{\mathbb{P}}$	Função média de $\mathbb{P}$ no $\mathcal{H}$.
$\langle \cdot, \cdot \rangle_{\mathcal{H}}$	Produto interno no EHKR $\mathcal{H}$.
$\Sigma^{\mathcal{H}}$	Operador covariância no EHKR $\mathcal{H}$.
$\tilde{k}$	Kernel positivo definido de valor real, definido em $\mathcal{P} \times \mathcal{P}$.
$\ \cdot\ _{\mathcal{H}}$	Norma no EHKR $\mathcal{H}$.
$\mathcal{B}(\mathbb{R}^D)$	σ -álgebra de Borel associada a $\mathbb{R}^D$.
$(\mathbb{R}^D, \mathcal{B}(\mathbb{R}^D))$	Espaço mensurável.
$\delta_{\mathbf{x}}$	Medida de Dirac, definida em algum espaço mensurável.
$k_1 \otimes k_2$	Produto tensorial entre k_1 e k_2 .

T	Norma triangular.
X	Conjunto fuzzy.
$\text{supp}(X)$	Suporte do conjunto fuzzy X .
$\mathcal{F}(\Omega)$	Conjunto dos conjuntos fuzzy definidos em Ω .
$\mathbb{E}_{\mathbb{P}}[X]$	Esperança da variável aleatória X distribuída de acordo a $\mathbb{P}$.
$\mathbb{E}_{X \sim \mathbb{P}}[f(X)]$	Esperança de $f(X)$, onde $X \sim \mathbb{P}$.
$\mathcal{L}$	Função de Lagrange.
μ	Embedding das medidas de probabilidade.
k_{set}	Kernel definido sobre conjuntos.
k_{conv}	Kernel de convolução.
$k_{\times}$	Kernel de produto cruzado de conjuntos fuzzy.
$k_{\star}$	Kernel de convolução fuzzy.
$k_{\cap}$	Kernel de interseção de conjuntos fuzzy.
$\mathcal{S}$	Semi-anel de conjuntos.
k_{tsk}	Kernel Takagi Sugeno Kang.
k_{ntsk}	Kernel Takagi Sugeno Kang non-singleton.

Lista de Figuras

1.1	Esquema ilustrando a abordagem usada na tese.	3
1.2	Esquema ilustrando a estrutura desta tese.	6
2.1	Gráficos de diferentes T-normas.	11
4.1	Detecção de anomalias grupais. A borda verde apresenta dois gráficos de duas observações (grupos de pontos) consideradas não anômalas. Esses dois grupos de pontos foram amostradas a partir de uma distribuição de misturas de Gaussianas (GMD) com três componentes. A borda vermelha apresenta dois grupos anômalos: a parte inferior esquerda apresenta um grupo de pontos amostrado a partir de uma distribuição normal e a parte inferior direita apresenta um grupo de pontos amostrados a partir de uma GMD com cinco componentes. Os dois grupos inferiores são considerados anômalos pois a distribuição deles é diferente dos grupos considerados não anômalos.	31
4.2	Dificuldade de usar somente valores representativos na detecção de anomalias grupais. Os pontos vermelhos e verdes são as médias empíricas dos grupos anômalos e não anômalos, respectivamente, de um conjunto de dados artificial, onde são considerados grupos não anômalos aqueles gerados por uma GMD de três componentes com um conjunto de parâmetros arbitrário, e são considerados grupos anômalos aqueles grupos formados por diversas distribuições normais. A sobreposição das médias empíricas dos grupos anômalos e não anômalos sugerem que os métodos convencionais de DD não sejam adequados para realizar detecção de anomalias para este conjunto de dados. Isso motiva o uso da informação da distribuição dos pontos para esse tipo de tarefa.	32
4.3	Seis esferas envoltentes mínimas para um conjunto de dados cujas observações são distribuições Gaussianas, mostrando o efeito de escolher um valor em particular para κ nas restrições do modelo SMDD. Parte superior: os três modelos compartilham o mesmo valor κ , o qual é incrementado de esquerda para a direita. Parte inferior: os três modelos tem o valor $\kappa = 1$ para suas restrições, excepto as restrições $i = 1$ e $i = 2$, as quais são incrementadas de esquerda para a direita. Decrementar o valor κ para alguma restrição específica, implica que não todas as realizações de $X \sim \mathbb{P}_i$ estarão dentro da descrição.	45
4.4	Conjunto de dados para a detecção de anomalias grupais. As caixas verde e amarela contêm grupos de pontos não anômalos. As caixas vermelha, azul e magenta contem grupos de pontos anômalos.	53

4.5	Resultados experimentais e o gráfico das médias grupais para o experimento de detecção de anomalias grupais. Os boxplots apresentam as estatísticas ACC e AUC. As marcas no eixo x de cada boxplot são os modelos de descrição de dados e no eixo y é a medida de performance. A Figura 4.5d apresenta as médias dos grupos não anômalos vs as médias dos grupos anômalos. Os pontos vermelhos, magenta e azuis são as médias dos grupos anômalos. Os pontos verdes são as médias dos grupos não anômalos.	54
4.6	Resultados experimentais e o gráfico das médias grupais para os dois experimentos na detecção de anomalias grupais.	56
4.7	Imagens obtidas de SDSSIII. Fonte http://www.sdss3.org/	57
4.8	Resultados experimentais e o gráfico das médias grupais para o experimento de detecção de anomalias grupais no conjunto de dados SDSS III dataset.	59
4.9	Resultados experimentais e o gráfico das médias grupais para o experimento de detecção de anomalias grupais no conjunto de dados SDSS III dataset.	60
4.10	Resultados experimentais e o gráfico das médias grupais para o experimento de detecção de anomalias grupais no conjunto de dados SDSS III dataset.	61
5.1	Fuzzificação da variável <i>age</i> , usando conjuntos fuzzy com funções de pertinência trapezoidais	83
5.2	Fuzzificação da variável <i>menopause</i> usando conjuntos fuzzy com funções de pertinência gaussiana, S-shaped e Z-shaped	83
5.3	Fuzzificação da variável <i>tumor-size</i> , usando conjuntos fuzzy com funções de pertinência gaussiana, S-shaped e Z-shaped	84
5.4	Fuzzificação da variável <i>inv-nodes</i> , usando conjuntos fuzzy com funções de pertinência gaussiana, S-shaped e Z-shaped	84
5.5	Nível de significância α vs. O Erro do Tipo II, usando a MMD com cinco diferentes kernels.	87
5.6	Boxplots contendo informação do erro de Tipo I.	87
5.7	Boxplots contendo informação do erro de Tipo II.	88

Lista de Tabelas

2.1	Exemplos de T-normas.	10
3.1	Exemplo de EHKR, com seus respectivos kernels.	20
3.2	Exemplo de funções kernel, onde $\alpha, \beta \in \mathbb{R}^+, d \in \mathbb{N}$	22
4.1	Notação usada neste capítulo.	34
4.2	Resumo da análise das condições KKT.	39
4.3	Notação dos modelos usados no experimento	51
5.1	Formulações de funções Kernel de interseção de conjuntos fuzzy usando T-normas . .	72
5.2	Exemplo de distâncias entre conjuntos fuzzy.	77
5.3	Atributos do conjunto de dados Breast Cancer.	82
5.4	Valores para os atributos <i>node-caps</i> , <i>deg-malig</i> , <i>breast</i> e <i>irradiat</i>	85
5.5	Resultados. O H_0 denota a hipótese nula.	87
6.1	Sumario dos conjuntos de dados de baixa qualidade.	100
6.2	Duas observações do conjunto de dados de baixa qualidade <i>Long-4</i>	100
6.3	Resultados no conjunto de dados <i>100mII-4</i>	102
6.4	Resultados no conjunto de dados <i>Long-4</i>	103
6.5	Resultados no conjunto de dados <i>100mIP-4</i>	103
6.6	Resultados no conjunto de dados <i>Dyslexic-12-4</i>	104

Capítulo 1

Introdução

*If we knew what it was we were doing, it
would not be called research, would it?.*

Albert Einstein

*In the beginning there was nothing,
which exploded.*

Terry Pratchett

Na área de aprendizagem computacional, usualmente, cada elemento ou *observação* é um ponto em $\mathbb{R}^D$. Isto é, um conjunto de dados é uma amostra $\{\mathbf{x}_i\}_{i=1}^N \subset \mathbb{R}^D$, cujos elementos seguem alguma lei de probabilidade, em geral desconhecida. Existe vasta literatura referente a tarefas de aprendizado computacional com esse tipo de conjunto de dados. No entanto, existe um crescente interesse em técnicas de aprendizado computacional para conjuntos de dados onde cada observação é por si, um conjunto de pontos em $\mathbb{R}^D$ (Ben-Tal *et al.*, 2011; Graepel e Herbrich, 2003; Guevara *et al.*, 2013a; Kondor e Jebara, 2003; Muandet e Schoelkopf, 2013; Muandet *et al.*, 2012; Póczos *et al.*, 2012; Shivaswamy *et al.*, 2006; Smola *et al.*, 2007; Viertl, 2011; Xiong *et al.*, 2011a,b; Yang e Gunn, 2007; Zhang, 2005). Esses conjuntos de dados são da forma:

$$\mathcal{T} = \{s_i\}_{i=1}^N, \tag{1.1}$$

onde N é o numero de observações e cada observação s_i é um conjunto de pontos.

Em problemas reais, é possível dizer que existem dois tipos de natureza para as observações s_i , e consequentemente, para os conjuntos de dados do tipo (1.1): a primeira delas é probabilística, ou seja os elementos do conjunto s_i são valores *precisos* que seguem alguma lei de probabilidade. A segunda natureza esta relacionada quando a informação contida em s_i modela dados não precisos, vagos ou linguísticos.

Diremos que um valor é preciso quando aceitamos que o erro de medida desse valor é aceitável para não ser levado em conta na modelagem do problema. Nesse caso é mais importante a análise da distribuição dos elementos da observação s_i para a análise do conjunto de dados ser feita. Exemplos desse tipo de observações são: um conjunto de características de uma imagem num conjunto de imagens (Lowe, 2004); um conjunto de características espaço temporais, (Laptev, 2005); um conjunto descrevendo a invariância de um objeto em particular (Graepel e Herbrich, 2003); um conjunto de

pontos descrevendo um agrupamento de galáxias (Xiong *et al.*, 2011a).

O segundo tipo de natureza para as observações s_i é devido à observações não precisas ou vagas, devidas a flutuações das medições, ou valores em falta, ou ruído, ou préprocessamento dos dados, etc. Citando Viertl (2011), a medição do nível de água de um rio num tempo fixo é um valor impreciso pois a ideia de um valor *verdadeiro* do nível de água não é realista. Alguns exemplos reais de dados imprecisos são: (1) *dados de expressão gênica* - devido à complexidade dos experimentos biológicos, quase todas as causas de imprecisão citadas anteriormente estão presentes, em especial, a quantificação final de um valor de expressão gênica para um gene, depende da estimação dos parâmetros de um modelo estatístico com muitas fontes de variabilidade; (2) *dados médicos* - dados clínicos tendem a ter imprecisão e, em especial, valores faltantes. Por exemplo, no diagnóstico de dislexia, os pacientes, resolvem testes gráficos os quais são avaliados subjetivamente por especialistas, na forma de termos linguísticos, por exemplo, “*perto de 10*”, “*entre 4 e 7*”; (3) *dados climáticos* - por exemplo, temperatura entre 3 e 7 graus, CO₂ entre 394.1 e 394.2 partes por milhão, vento entre 7 e 10 milhas por hora, umidade entre 10 e 16 por cento, nível do mar entre 2.3 e 3.5 metros. Outros exemplos são: um conjunto de valores replicados num processo de medição (Wernisch *et al.*, 2003); um conjunto descrevendo incerteza pontual (Ben-Tal *et al.*, 2011; Shivaswamy *et al.*, 2006; Zhang, 2005); um conjunto descrevendo opiniões subjetivas (Viertl, 2011). Conjuntos de dados provenientes das áreas da meteorologia, economia e bioinformática (Ben-Tal *et al.*, 2011; Bhadra, 2012; Bhadra *et al.*, 2009a,b, 2010), também contém observações desse tipo.

A partir dessa perspectiva, algoritmos de aprendizado que possam ser aplicados em conjuntos de dados do tipo (1.1) terão importância prática devido às potenciais aplicações. No entanto, usualmente, esses conjuntos são processados para obter um valor representativo, levando à perda de informação importante do problema.

Dada a importância teórica e prática da pesquisa em métodos de aprendizado computacional para a análise de desses conjuntos de dados, o objetivo da tese é estudar métodos de kernel, conjuntamente com modelos probabilísticos e modelos da teoria fuzzy que permitam fazer a análise desses conjuntos de dados, respeitando a natureza dos dados.

1.1 Abordagem usada na tese

Nesta tese usamos tanto medidas de probabilidade, quanto conjuntos fuzzy, para modelar as observações dos conjuntos de dados. Além disso, usamos os métodos de kernel para fazer a análise desses dados. A Figura 1.1 apresenta um esquema ilustrando a abordagem proposta.

1.1.1 Modelagem das observações usando Medidas de Probabilidade

Usualmente, em aprendizado computacional, é assumido que os elementos pertencentes ao conjuntos de dados do tipo $\{\mathbf{x}_i\}_{i=1}^N$, $\mathbf{x}_i \in \mathbb{R}^D$, são realizações idênticas e independentemente distribuídas (i.i.d) de uma variável aleatória seguindo uma lei definida por uma medida de probabilidade. No caso dos conjuntos de dados do tipo (1.1), assumiremos que a amostra $\{s_i\}_{i=1}^N$ contém realizações i.i.d. de uma *medida de probabilidade aleatória*, a qual é distribuída de acordo com uma distribuição desconhecida. Ou seja, modelamos cada observação s_i pela medida de probabilidade $\mathbb{P}_i$ definida no espaço mensurável $(\mathbb{R}^D, \mathcal{B}(\mathbb{R}^D))$, onde $\mathcal{B}(\mathbb{R}^D)$ representa os conjuntos de Borel em $\mathbb{R}^D$. Então, (1.1)

Figura 1.1: Esquema ilustrando a abordagem usada na tese.

pode ser escrito como o conjunto de medidas de probabilidade:

$$\{\mathbb{P}_i\}_{i=1}^N. \quad (1.2)$$

Nesse sentido nós fazemos as seguintes suposições:

- os pontos em s_i são gerados de acordo a uma lei de distribuição local dada por $\mathbb{P}_i$;
- cada $\mathbb{P}_i$ é desconhecida, isto é, nada se sabe sobre a forma dessa distribuição.

Representação das medidas de probabilidade num EHKR

Ao invés de fazer a análise diretamente no espaço das medidas de probabilidade, fizemos a análise num Espaço de Hilbert com Kernel Reproduzível (EHKR), mediante o *mergulho* das medidas de probabilidade num EHKR (Berlinet e Thomas-Agnan, 2004; Smola *et al.*, 2007), o qual é feito através do mapeamento $\mathbb{P} \mapsto \mathbb{E}_{\mathbb{P}}[k(X, \cdot)]$, onde k é uma função kernel positiva definida e X é uma variável aleatoriamente distribuída de acordo com $\mathbb{P}$. Os algoritmos de aprendizado são definidos então nas funções $\mathbb{E}_{\mathbb{P}}[k(X, \cdot)]$ que representam as medidas de probabilidade no EHKR. Esta técnica tem sido usada com sucesso em diferentes problemas, por exemplo: redução de dimensionalidade (Fukumizu *et al.*, 2004), medida de dependência de variáveis aleatórias (Gretton *et al.*, 2005a), teste estatístico de duas amostras (Gretton *et al.*, 2012), mergulho de modelos ocultos de Markov no EHKR (Song *et al.*, 2010), regra de Bayes usando kernels (Fukumizu *et al.*, 2010), entre outros (Muandet e Schoelkopf, 2013; Muandet *et al.*, 2012; Sriperumbudur *et al.*, 2010; Zhang *et al.*).

As principais vantagens desta técnica são:

- os dados podem ser modelados sem assumir nenhuma forma para as medidas de probabilidade;
- não precisa de uma estimação de densidade, pois é um método não paramétrico;

- é possível incorporar conhecimento a priori, via uma escolha apropriada do kernel;
- o mergulho com certos tipos de kernels preserva toda a informação da distribuição;
- os kernels usados podem ser aproximados empiricamente.

1.1.2 Modelagem das observações usando conjuntos fuzzy

A teoria dos conjuntos fuzzy (Zadeh, 1965) fornece as ferramentas para modelar e tratar conjuntos de dados do tipo (1.1). Assim, modelamos cada observação s_i como um conjunto fuzzy. Assumimos que a amostra $\{s_i\}_{i=1}^N$, contém realizações i.i.d. de uma *variável aleatória fuzzy*, distribuída de acordo alguma lei de probabilidade. Logo, (1.1) pode ser modelada como o conjunto de dados fuzzy:

$$\{X_i\}_{i=1}^N, \quad (1.3)$$

onde cada observação X_i é um conjunto fuzzy. Nesse sentido fazemos as seguintes suposições:

- os pontos em s_i representam informação vaga, imprecisa, ou fuzzy, de alguma observação cujo valor não pode ser conhecido de maneira precisa;
- os conjuntos fuzzy são projetados arbitrariamente para modelar cada s_i .

Representação dos conjuntos fuzzy em EHKR

O uso de kernels positivos definidos, definidos sobre conjuntos fuzzy garante uma representação implícita dos conjuntos fuzzy como funções num EHKR. As principais vantagens desta técnica são:

- os conjuntos de dados, onde cada observação é uma tupla de termos linguísticos, pode ser facilmente quantizado e analisado num EHKR via a escolha apropriada de um kernel definido sobre conjuntos fuzzy e um algoritmo de análise de dados,
- o mapeamento dos conjuntos fuzzy para o EHKR é implícito, isto é, não é preciso construir explicitamente o EHKR pois todos os cálculos são feitos via um kernel positivo definido.
- é possível definir distâncias entre conjuntos fuzzy usando somente avaliações de kernel,
- conjuntos de dados, onde cada observação tem incerteza e imprecisão, podem ser modelados com conjuntos fuzzy, e analisados com métodos de kernel (Cortes e Vapnik, 1995; Tax e Duin, 2004).

1.2 Contribuições

A tese inclui contribuições teóricas e práticas para as áreas de aprendizado computacional e teoria fuzzy. Essas contribuições são:

1. **Descrição de dados para conjuntos de medidas de probabilidade** $\{\mathbb{P}_i\}_{i=1}^N$: Propomos em (Guevara *et al.*, 2013b, 2014b, 2015b) usar conjuntos de medidas de probabilidades para modelar conjuntos de dados do tipo (1.1). Mediante o mergulho destas medidas de probabilidade para um EHKR, apresentamos alguns métodos de kernel de descrição de dados para esse tipo de dados.

2. **Detecção de anomalias grupais:** Em (Guevara *et al.*, 2014b, 2015b) propomos usar modelos de descrição de dados para os conjuntos de dados do tipo (1.1), na tarefa de detecção de anomalias grupais. Isto é, detectar conjuntos de pontos não usuais, ou anômalos, de outros conjuntos de pontos com comportamento não anômalo.
3. **Mergulho dos conjuntos fuzzy para funções nos espaços de Hilbert com kernel reproduzível:** Em (Guevara *et al.*, 2014a) e no Capítulo 5, propomos representar implicitamente os conjuntos fuzzy como funções nos espaços de Hilbert com kernel reproduzível. Essa representação implícita é feita graças à projeção de funções kernels positivas definidas, de valor real, definidas em conjuntos fuzzy. A importância desta contribuição é que poderão ser adaptados os diversos algoritmos kernel para trabalhar com conjuntos de dados fuzzy. Como por exemplo, distâncias entre conjuntos fuzzy, descrição de dados de conjuntos fuzzy, classificação e regressão de conjuntos fuzzy.
4. **Engenharia de kernels definidos sobre conjuntos fuzzy:** Em (Guevara *et al.*, 2014a) e no Capítulo 5, propomos vários kernels definidos sobre conjuntos fuzzy, alguns deles positivos definidos ou condicionalmente positivos definidos, outros simplesmente simétricos ou kernels indefinidos mais com aplicações práticas. Entre os kernels apresentados encontram-se: o kernel de produto cruzado de conjuntos fuzzy, o kernel de convolução fuzzy, o kernel de interseção de conjuntos fuzzy, o kernel de substituição mediante distâncias entre conjuntos fuzzy Guevara *et al.* (2015a). Além disso, são apresentados kernels derivados a partir das propriedades de fechamento dos kernels. Em (Guevara *et al.*, 2013a) propomos também a partir da extensão do sistema fuzzy TSK para o caso non-singleton, uma nova classe de kernels definidos com conjuntos fuzzy, chamados de kernels TSK non-singleton.
5. **Aplicações em classificação supervisionada e, teste kernel de duas amostras** Em (Guevara *et al.*, 2013a) propomos usar kernels definidos sobre conjuntos fuzzy para a tarefa de classificação supervisionada de conjuntos de dados intervalares. Esses conjuntos de dados contêm atributos imprecisos, vagos, linguísticos e subjetivos, os quais são modelados por intervalos. Além disso, esses conjuntos de dados apresentam atributos com valores em falta. Nos propomos uma metodologia para tratar esses conjuntos de dados usando kernels definidos sobre conjuntos fuzzy e máquinas de vetores de suporte. O Capítulo 5 apresenta uma aplicação em teste de hipótese kernel de duas amostras, para com conjunto de dados com atributos heterogêneos, dos quais alguns dos atributos são modelados com conjuntos fuzzy e onde é usado o kernel de substituição mediante distâncias entre conjuntos fuzzy Guevara *et al.* (2015a).

1.3 Estrutura da tese

O restante da tese encontra-se organizada em seis capítulos. O esquema desses capítulos é apresentado na Figura 1.2. O conteúdo de cada capítulo são sumarizados logo em seguida.

Capítulo 2. Teoria fuzzy

Nesse capítulo, cobriremos o conhecimento necessário em teoria fuzzy para o desenvolvimento de nossa teoria. Isto inclui uma breve introdução as T-normas e conjuntos fuzzy. Posteriormente,

Figura 1.2: Esquema ilustrando a estrutura desta tese.

apresentamos os sistemas de regras fuzzy com o sistema fuzzy de Mamdani como exemplo motivador. Esses conceitos servirão para apresentar em capítulos posteriores, novos kernels definidos sobre conjuntos fuzzy, onde o conceito de operador T-norma é fundamental. Assim como uma extensão do sistema de regras Takagi Sugeno Kang e sua relação com os métodos de kernel.

Capítulo 3. Métodos kernel

Nesse capítulo, apresentamos o conhecimento necessário em métodos de kernel que suportem nossa teoria. Incluímos uma breve introdução aos espaços de Hilbert com kernel reproduzível. Motivamos o uso dos métodos de kernel para a análise de dados. Apresentamos uma serie de kernels positivos definidos, assim como as propriedades de fechamento dos kernels reproduzíveis. Além disso, apresentamos três métodos de kernel: as máquinas de vetores de suporte, a descrição de dados com vetores de suporte e o teste de hipótese de duas amostras usando kernels chamado de discrepância máxima de médias.

Capítulo 4. Descrição de dados com medidas de suporte

Nesse capítulo, discutiremos um novo enfoque para descrição de dados do tipo $\{\mathbb{P}_i\}_{i=1}^N$. Usamos o mergulho de medidas de probabilidade para um espaço de Hilbert com kernel reproduzível. Posteriormente, a descrição de dados será feita nesse espaço de Hilbert, usando uma esfera envolvente mínima para as funções representativas de cada medida de probabilidade nesse espaço de Hilbert. A principal vantagem deste enfoque é que não precisa-se reduzir a informação das observações s_i para um único valor, evitando assim a perda de informação. A descrição somente dependerá de algumas medidas de probabilidade, chamada de medidas de suporte. Além disso, não é assumida forma alguma para as distribuições locais $\mathbb{P}_i$. São apresentados diversos experimentos, entre os quais é estudado o caso de detecção de anomalias grupais em dados astronômicos.

Capítulo 5. Kernels definidos sobre conjuntos fuzzy

Nesse capítulo, propomos novos kernels de valor real, definidos sobre conjuntos fuzzy. Os quais em principio permitirão fazer a análise de conjuntos de dados fuzzy, isto é conjuntos de dados contendo conjuntos fuzzy como elementos. Estes novos kernels, atuarão como medidas de similaridade entre

conjuntos fuzzy, e farão o mapeamento dos conjuntos fuzzy para funções nos espaços de Hilbert com kernel reproduzível. Entre os novos kernels apresentados, temos o kernel de convolução fuzzy, o kernel de produto cruzado de conjuntos fuzzy e o kernel de interseção de conjuntos fuzzy. Além disso, apresentamos a representação funcional desses conjuntos fuzzy nos espaços de Hilbert com kernel reproduzível. Posteriormente, apresentamos como construir kernels definidos sobre conjuntos fuzzy usando substituição de distância entre conjuntos fuzzy. Apresentamos também, as possíveis aplicações destes novos kernels, tais como, definição de distâncias entre conjuntos fuzzy, engenharia de kernels em conjuntos fuzzy, classificação e regressão em conjuntos de dados fuzzy e descrição de conjuntos de dados fuzzy. Finalmente, apresentamos uma aplicação desses kernels no teste de hipótese de duas amostras usando kernels chamado de discrepância máxima de médias para um conjunto de dados de pronóstico de incidência de câncer de mama, o qual contém atributos heterogêneos (precisos e não-precisos).

Capítulo 6. Kernels em conjuntos fuzzy baseados no sistema de regras fuzzy Takagi Sugeno Kang

Nesse capítulo, apresentamos como a partir da extensão do sistema de regras fuzzy Takagi Sugeno Kang para o caso de entradas do tipo non-singleton, é possível induzir uma classe de kernels positivos definidos de valor real, os quais são definidos sobre conjuntos fuzzy. Chamamos a esses kernels de kernel TSK non-singleton, apresentamos as propriedades dele assim como as relações desses kernels com as relações de equivalência fuzzy. Avaliamos experimentalmente esses kernels com conjuntos de dados intervalares, em problemas de classificação supervisionada.

Capítulo 7. Conclusões

Nesse capítulo, resumimos os principais resultados da tese. Também discutiremos possíveis futuros trabalhos.

Capítulo 2

Teoria fuzzy

*Fuzzy logic is not fuzzy. Fuzzy logic is precise.
Basically, fuzzy logic is a precise logic of im-
precision.*

(Zadeh, 2010)

Neste capítulo apresentamos diversos conceitos da teoria fuzzy como conjunto fuzzy, normas triangulares e inferência fuzzy, assim como os sistemas de regras fuzzy.

2.1 Introdução

A teoria fuzzy, introduzida em 1965 por Lotfi A. Zadeh no artigo “*Fuzzy Sets*” (Zadeh, 1965), fornece as ferramentas matemáticas para tratar de forma precisa problemas com informação vaga e informação fornecida subjetivamente, como por exemplo problemas com conjuntos de dados imprecisos. Na atualidade é difícil fazer uma contagem das áreas de conhecimento beneficiadas por essa teoria. Diversas aplicações, tanto em matemática e computação quanto na medicina, na engenharia e na indústria são prova desse fato. Na área de aprendizado computacional, a teoria fuzzy é usada nos métodos de análise de dados em três situações: 1) construindo versões fuzzy dos algoritmos de aprendizado convencionais; 2) mapeando os dados para espaços métricos fuzzy, onde a análise é feita (Diamond *et al.*, 1994); 3) modelando dados não precisos, vagos, subjetivos com conjuntos fuzzy, para posteriormente fazer a análise estatística desses dados fuzzy (Bertoluzza *et al.*, 2002; Viertl, 2011). Alguns exemplos do uso da teoria fuzzy em aprendizado computacional incluem algoritmos de análise de agrupamento fuzzy (Höppner, 1999), aprendizado de sistemas de regras fuzzy (Mendel, 2001), classificadores fuzzy (Olaru e Wehenkel, 2003), ou até inclusive as redes possibilísticas inspiradas nos modelos gráficos (Benferhat e Smaoui, 2007). No que segue, revisamos diversos conceitos que usamos frequentemente nesta tese: normas triangulares, conjuntos fuzzy e sistemas de regras fuzzy.

2.2 Normas Triangulares

As normas triangulares ou T-normas, foram introduzidas no contexto da teoria de probabilidades (Klement *et al.*, 2000; Menger, 1942). Elas são importantes pois elas são amplamente usadas na

definição da interseção de conjuntos fuzzy.

Definição 2.2.1 (T-norma). A norma triangular ou *T-norma* é uma função $T : [0, 1]^2 \rightarrow [0, 1]$, tal que, para todo $x, y, z \in [0, 1]$ satisfaz os seguintes axiomas (Klement *et al.*, 2000; Yu e Zhang, 2008):

T1 comutatividade: $T(x, y) = T(y, x)$;

T2 associatividade: $T(x, T(y, z)) = T(T(x, y), z)$;

T3 monotonicidade: $y \leq z \Rightarrow T(x, y) \leq T(x, z)$;

T4 condição limite $T(x, 1) = x$.

Definição 2.2.2 (Extensão multi-argumento da T-norma). Usando $n \in \mathbb{N}, n \geq 2$, e **T2** da Definição 2.2.1, a extensão multi-argumento $T_n : [0, 1]^n \rightarrow [0, 1]$ da T-norma T é dada pela recorrência:

$$T_2(x_1, x_2) = T(x_1, x_2), \quad n = 2, \quad (2.1)$$

$$T_n(x_1, x_2, \dots, x_n) = T(x_1, T_{n-1}(x_2, x_3, \dots, x_n)), \quad n \geq 3. \quad (2.2)$$

Neste texto usamos indistintamente T para denotar T ou T_n . A Tabela 2.1 contém alguns exemplos de funções que são T-normas.

$T(x, y) = \min(x, y)$	mínimo
$T(x, y) = xy$	produto
$T(x, y) = \max(x + y - 1, 0)$	Łukasiewicz
$T(x, y) = \begin{cases} x, & y = 1 \\ y, & x = 1 \\ 0, & \text{caso contrário} \end{cases}$	Drástico

Tabela 2.1: Exemplos de T-normas.

Os gráficos dessas T-normas são apresentadas na Figura 2.1.

2.3 Conjunto Fuzzy

Existem diversas maneiras de definir e denotar conjuntos fuzzy (Melin e Castillo, 2005; Mendel, 2001; Pedrycz e Gomide, 2007; Zadeh, 1965), aquela que usamos neste texto é baseada na definição dos conjuntos fuzzy como funções dentro do espaços de funções (Aisbett *et al.*, 2010), sendo a simplificação da notação a principal vantagem desta definição.

Figura 2.1: Gráficos de diferentes *T*-normas.

Definição 2.3.1 (Conjunto fuzzy). Seja Ω um conjunto não vazio. Um conjunto fuzzy em Ω é uma função:

$$\begin{aligned} X : \Omega &\rightarrow [0, 1] \\ x &\mapsto X(x). \end{aligned} \quad (2.3)$$

O valor $X(x) \in [0, 1]$ é chamado de *grau de pertinência* do elemento x para o conjunto fuzzy X .

Em seguida, apresentamos uma serie de definições importantes que logo serão usadas no desenvolvimento do trabalho.

Definição 2.3.2 (Suporte de um conjunto fuzzy). O suporte de um conjunto fuzzy é o conjunto:

$$\text{supp}(X) = \{x \in \Omega \mid X(x) > 0\}. \quad (2.4)$$

Definição 2.3.3 (Conjunto fuzzy singleton e non-singleton). O conjunto fuzzy X é singleton se o $\text{supp}(X)$ contem um único elemento. Caso o conjunto $\text{supp}(X)$ possua mais de um elemento, ele é chamado de conjunto fuzzy non-singleton.

Definição 2.3.4 (Conjunto fuzzy normal). Um conjunto fuzzy X é *normal* se existe pelo menos um ponto $x \in \Omega$ satisfazendo $X(x) = 1$. Isto é, X satisfaz:

$$\{x \in \Omega \mid X(x) = 1\} \neq \emptyset. \quad (2.5)$$

Exemplos

- (i) **Conjunto fuzzy singleton normal.** O conjunto fuzzy X , dado por $X(x') = 1$ e $X(x) = 0$ para todo $x \neq x'$, isto é $\text{supp}(X) = x'$, e apenas no ponto x' que $X(x') = 1$, é um conjunto fuzzy singleton normal.
- (ii) **Conjunto fuzzy non-singleton normal.** O conjunto fuzzy definido em $\mathbb{R}$ por: $X(\cdot) = \exp(-\gamma(\cdot - x)^2)$, $\gamma \in \mathbb{R}^+$, é um conjunto fuzzy non-singleton normal.

Definição 2.3.5 (Espaço dos conjuntos fuzzy). O espaço dos conjuntos fuzzy é o conjunto definido por:

$$\mathcal{F}(\Omega) = \{X \mid X \text{ é um conjunto fuzzy em } \Omega\}. \quad (2.6)$$

Definição 2.3.6 (Conjunto interseção fuzzy). Dado dois conjuntos fuzzy $X, Y \in \mathcal{F}(\Omega)$, um conjunto interseção fuzzy é dado por:

$$X \cap Y : \Omega \rightarrow [0, 1] \quad (2.7)$$

$$x \mapsto (X \cap Y)(x) = T(X(x), Y(x)), \quad (2.8)$$

onde T é um operador T-norma.

Em seguida, apresentamos um tipo especial de conjunto fuzzy chamado de *vetor fuzzy*. Usamos vetores fuzzy na construção dos conjuntos de treinamento fuzzy e nos sistemas de regras fuzzy.

Definição 2.3.7 (Vetor fuzzy). Dados os conjuntos fuzzy: $X_1 \in \mathcal{F}(\Omega_1), X_2 \in \mathcal{F}(\Omega_2), \dots, X_D \in \mathcal{F}(\Omega_D)$, $D \in \mathbb{N}$, um vetor fuzzy X é um conjunto fuzzy definido pela função:

$$\begin{aligned} X : \Omega_1 \times \Omega_2 \times \dots \times \Omega_D &\rightarrow [0, 1] \\ (x_1, x_2, \dots, x_D) &\mapsto X(x_1, x_2, \dots, x_D) = T(X_1(x_1), X_2(x_2), \dots, X_D(x_D)). \end{aligned} \quad (2.9)$$

onde T é a extensão multi-argumento do operador T-norma.

Note-se que na definição do vetor fuzzy precisa-se ter definido previamente D conjuntos fuzzy segundo a Definição 2.3.1. Consequentemente, os vetores fuzzy são conjuntos fuzzy definidos no produto cartesiano $\Omega_1 \times \Omega_2 \times \dots \times \Omega_D$. Em seguida, estendemos os conceitos de conjunto fuzzy singleton e non-singleton para o caso dos vetores fuzzy.

Definição 2.3.8 (Vetor fuzzy singleton e non-singleton). Um vetor fuzzy X é *singleton* se o $\text{supp}(X)$ é único ponto $(x_1, x_2, \dots, x_D) \in \Omega_1 \times \Omega_2 \times \dots \times \Omega_D$. Se o suporte contém vários pontos, o vetor fuzzy é chamado de vetor fuzzy non-singleton.

Outro conceito importante na teoria fuzzy é o conceito de relação fuzzy, a qual modela uma relação “vaga” ou “imprecisa” entre objetos, por exemplo, a relação “as partes” $x_1, x_2, \dots, x_{D-1}$ “são elementos de” x_D com um grau de pertinência entre $[0, 1]$.

Definição 2.3.9 (Relação fuzzy). Uma relação fuzzy é um conjunto fuzzy dado por:

$$\begin{aligned} R : \Omega_1 \times \Omega_2 \times \dots \times \Omega_D &\rightarrow [0, 1] \\ (x_1, x_2, \dots, x_D) &\mapsto R(x_1, x_2, \dots, x_D). \end{aligned} \quad (2.10)$$

Dependendo do contexto, uma relação fuzzy pode ser modelada pelo vetor fuzzy segundo a Definição 2.3.7.

2.4 Sistemas de Regras Fuzzy

Os sistemas de regras fuzzy (Zadeh, 1974), chamados também de sistemas de lógica fuzzy, ou sistemas de inferência fuzzy, ou sistemas especialistas fuzzy ou simplesmente de sistemas fuzzy, são aproximadores universais de funções (Kosko, 1994; Wang e Mendel, 1992). Isto é, dada uma função arbitrária f e um número real $\epsilon > 0$, os sistemas de regras fuzzy podem aproximar f com um fator de aproximação ϵ (Kreinovich e Mouzouris, 1996). A característica principal dos sistemas fuzzy é que eles possuem um conjunto de regras fuzzy **If- Then**, nas quais é possível representar o conhecimento de um problema usando termos linguísticos, por exemplo,

If pressão é alta **Then** volume é pequeno.

If velocidade é alta **Then** freiar um pouco.

Nessas regras os termos *pressão*, *volume*, *velocidade* e *freiar* são chamados de *variáveis linguísticas* e os termos *alta*, *pequeno* e *um pouco* são os *valores linguísticos* que essas variáveis podem ter (Zadeh, 1975). Os sistemas fuzzy usam conjuntos fuzzy para representar os valores linguísticos, possibilitando a quantificação e o computo desses valores.

Existem vários tipos de sistemas de regras fuzzy. Os mais conhecidos e usados na literatura são o sistema fuzzy de Mamdani (Mamdani, 1974), e o sistema fuzzy de Takagi Sugeno Kang (Takagi e Sugeno, 1985). Em seguida, descreve-se o sistema fuzzy de Mamdani. O sistema fuzzy de Takagi Sugeno Kang, será descrito oportunamente no Capítulo 6.

2.4.1 Sistema Fuzzy de Mamdani

O sistemas de regras fuzzy de Mamdani (Mamdani, 1974; Mendel, 2001), tem quatro elementos principais: o fuzzificador, as regras fuzzy, o algoritmo de inferência fuzzy e opcionalmente o defuzzificador.

Fuzzificador

Devido a que todas as operações nos sistemas de regras fuzzy são feitas com conjuntos fuzzy, o fuzzificador transforma a *entrada do sistema* num conjunto fuzzy.

Definição 2.4.1. O fuzzificador é um mapeamento do tipo:

$$\Omega \rightarrow \mathcal{F}(\Omega) \quad (2.11)$$

$$x \mapsto X. \quad (2.12)$$

O fuzzificador é chamado de *fuzzificador singleton*, se ele produz um conjunto fuzzy singleton. Por outro lado, se o fuzzificador produz um conjunto fuzzy non-singleton, ele é chamado de *fuzzificador non-singleton* (Mouzouris e Mendel, 1997).

Por exemplo, se a entrada no sistema for $x_1, x_2, \dots, x_D \in \Omega_1 \times \Omega_2 \times \dots \times \Omega_D$, então pode ser usado o vetor fuzzy singleton ou o vetor fuzzy non-singleton segundo a Definição 2.3.8 para representar o conjunto fuzzy proveniente do fuzzificador.

Regras Fuzzy

As regras fuzzy representam o conhecimento de um problema em forma de termos linguísticos. Os sistemas fuzzy de Mamdani possuem um conjunto de regras **If- Then** indexado por $\ell \in \mathcal{L} = \{1, 2, \dots, L\}$. Estes termos linguísticos são representados pelas *variáveis linguísticas* $\tilde{a}_d$, $d \in \mathcal{D} = \{1, 2, \dots, D\}$ contendo os *valores linguísticos* $\tilde{A}_d^\ell$, $\ell \in \mathcal{L}$. Assim, as regras são representadas por:

$$\text{If } \tilde{a}_1 \text{ is } \tilde{A}_1^\ell \text{ and } \dots \text{ and If } \tilde{a}_d \text{ is } \tilde{A}_d^\ell \text{ and } \dots \text{ and If } \tilde{a}_D \text{ is } \tilde{A}_D^\ell \text{ Then } \tilde{c} \text{ is } \tilde{C}^\ell, \quad (2.13)$$

Cada regra fuzzy tem duas partes, o antecedente:

$$\text{If } \tilde{a}_1 \text{ is } \tilde{A}_1^\ell \text{ and } \dots \text{ and If } \tilde{a}_d \text{ is } \tilde{A}_d^\ell \text{ and } \dots \text{ and If } \tilde{a}_D \text{ is } \tilde{A}_D^\ell, \quad (2.14)$$

e o consequente:

$$\tilde{c} \text{ is } \tilde{C}^\ell, \quad (2.15)$$

onde a variável linguística do consequente $\tilde{c}$ pode ter os valores linguísticos C^ℓ (Zadeh, 1975).

Os sistemas fuzzy de Mamdani representam cada termo linguístico: a_d is $\tilde{A}_d^\ell$ em (2.13), pelo conjunto fuzzy: $X_d^\ell \in \mathcal{F}(\Omega_d)$. Isto é, para todo $d \in \mathcal{D}, \ell \in \mathcal{L}$, a regra fuzzy (2.13) pode ser reescrita como:

$$\text{If } X_1^\ell \text{ and } \dots \text{ and If } X_d^\ell \text{ and } \dots \text{ and If } X_D^\ell \text{ Then } X_{D+1}^\ell, \quad (2.16)$$

onde $X_{D+1}^\ell \in \mathcal{F}(\Omega_{D+1})$, é o conjunto fuzzy usado para representar o consequente (2.15).

Para representar o *and* entre os conjuntos fuzzy do antecedente (2.16) é usado o conceito de relação fuzzy (Definição 2.3.9) como é apresentado na seguinte definição.

Definição 2.4.2. O antecedente da regra (2.16) é representado pela relação fuzzy (Definição 2.3.9):

$$\begin{aligned} A^\ell : \Omega_1 \times \Omega_2 \times \dots \times \Omega_D &\rightarrow [0, 1] \\ \mathbf{x} &\mapsto A^\ell(\mathbf{x}) = T(X_1^\ell(x_1), X_2^\ell(x_2), \dots, X_D^\ell(x_D)). \end{aligned} \quad (2.17)$$

para todo $\ell \in \mathcal{L}$, onde $\mathbf{x} = x_1, x_2, \dots, x_D$.

Note que (2.17) satisfaz a definição de vetor fuzzy (Definição 2.3.7), consequentemente, vetores fuzzy são usados para representar os antecedentes das regras.

Finalmente, a regra fuzzy (2.16), é representada pela relação fuzzy:

$$\begin{aligned} R^\ell : \Omega_1 \times \Omega_2 \times \dots \times \Omega_D \times \Omega_{D+1} &\rightarrow [0, 1] \\ (\mathbf{x}, x_{D+1}) &\mapsto R^\ell(\mathbf{x}, x_{D+1}) = I(A^\ell(\mathbf{x}), X_{D+1}^\ell(x_{D+1})), \end{aligned} \quad (2.18)$$

para todo $\ell \in \mathcal{L}$, onde $\mathbf{x} = x_1, x_2, \dots, x_D$ e, I é um operador de implicação fuzzy (Shi *et al.*, 2010).

Algoritmo de inferência fuzzy:

O algoritmo de inferência fuzzy é um algoritmo cujos argumentos são as regras fuzzy e o conjunto fuzzy obtido do fuzzificador. Nos sistemas de regras fuzzy de Mamdani, o algoritmo de inferência usa composição fuzzy (Zadeh, 1996), entre o conjunto fuzzy X proveniente do fuzzificador e cada regra fuzzy R^ℓ , para obter o conjunto $Y^\ell \in \mathcal{F}(\Omega_{D+1})$. Isto é:

$$Y^\ell(x_{D+1}) = \sup_{\mathbf{x} \in \Omega_1 \times \dots \times \Omega_D} T(X(\mathbf{x}), R^\ell(\mathbf{x}, x_{D+1})).$$

Logo, é aplicado um operador de *agregação* de conjuntos fuzzy nesses L conjuntos, por exemplo, o operador T-conorma (Klement *et al.*, 2000), para obter o conjunto fuzzy de saída:

$$Y(.) = \oplus_{\ell=1}^L Y^\ell(x_{D+1}), \quad (2.19)$$

onde $\oplus$ representa o operador T-conorma (Klement *et al.*, 2000).

Defuzzificador:

O defuzzificador é uma parte opcional dos sistemas de regras fuzzy de Mamdani encarregado de converter o conjuntos fuzzy de saída (2.19), em um valor numérico, isto é, o defuzzificador é um

mapeamento:

$$\mathcal{F}(\Omega) \rightarrow \Omega \quad (2.20)$$

$$X \mapsto x. \quad (2.21)$$

Existem muitos tipos de defuzzificadores (Leekwijck e Kerre, 1999), portanto a escolha de algum de eles é baseada na simplicidade computacional e no tipo de aplicação (Driankov *et al.*, 1996; Mendel, 2001, 1995).

Sistema Fuzzy Mamdani com fuzzificador non-singleton (Mouzouris e Mendel, 1997)

A principal característica desse sistema é o fuzzificador non-singleton cujo principal efeito é o cancelamento do ruído e a modelagem das imprecisões e do ruído associado às entradas do sistema (Mouzouris e Mendel, 1997).

Seja X o conjunto fuzzy non-singleton proveniente do fuzzificador. Se o operador T-norma for usado para (2.18), então a saída do sistema fuzzy dada pela Equação (2.19) pode ser escrita como segue:

$$\begin{aligned} Y(.) &= \bigoplus_{\ell=1}^L \sup_{x \in \Omega_1 \times \dots \times \Omega_D} T \left(X(\mathbf{x}), T \left((A^\ell(\mathbf{x}), X_{D+1}^\ell(.)) \right) \right) \\ &\quad \text{Pela extensão multi-argumento do operador T-norma} \\ &= \bigoplus_{\ell=1}^L \sup_{x \in \Omega_1 \times \dots \times \Omega_D} T \left(X(\mathbf{x}), A^\ell(\mathbf{x}), X_{D+1}^\ell(.) \right) \\ &\quad \text{Usando (2.9) e (2.17) para } X \text{ e } A^\ell \\ &= \bigoplus_{\ell=1}^L \sup_{x \in \Omega_1 \times \dots \times \Omega_D} T \left(T \left(X_1(x_1), X_2(x_2), \dots, X_D(x_D) \right), \right. \\ &\quad \left. T \left(X_1^\ell(x_1), X_2^\ell(x_2), \dots, X_D^\ell(x_D) \right), X_{D+1}^\ell(.) \right) \\ &\quad \text{Pela extensão multi-argumento do operador T-norma} \\ &= \bigoplus_{\ell=1}^L \sup_{x \in \Omega_1 \times \dots \times \Omega_D} T \left(X_1(x_1), X_2(x_2), \dots, X_D(x_D), X_1^\ell(x_1), X_2^\ell(x_2), \dots, X_D^\ell(x_D), X_{D+1}^\ell(.) \right) \\ &\quad \text{Pela monotonicidade do operador T-norma} \\ &= \bigoplus_{\ell=1}^L T \left(X_{D+1}^\ell(.), \sup_{x_1 \in \Omega_1} T \left(X_1(x_1), X_1^\ell(x_1) \right), \sup_{x_2 \in \Omega_2} T \left(X_2(x_2), X_2^\ell(x_2) \right), \dots, \right. \\ &\quad \left. \sup_{x_D \in \Omega_D} T \left(X_D(x_D), X_D^\ell(x_D) \right) \right) \end{aligned}$$

O conjunto fuzzy de saída $Y(.)$, pode ser defuzzificado usando algum defuzzificador, como por exemplo aqueles descritos em (Leekwijck e Kerre, 1999). Esse sistema fuzzy tem sido usado com sucesso em diversas aplicações (Chua e Tan, 2011; Hakan Akpolat, 2005; Kim *et al.*, 2004; Mouzouris e Mendel, 1997; Simon e Flowers, 2007) com melhores resultados que o sistema fuzzy Mamdani com fuzzificador singleton (Cara *et al.*, 2011).

Capítulo 3

Métodos de kernel

Simplicity is the ultimate sophistication.

Leonardo Da Vinci

Os métodos de kernel são uma classe de algoritmos de aprendizagem que detectam e exploram padrões complexos nos dados. Qualquer algoritmo de aprendizagem baseado em métodos de kernel possui duas componentes principais: uma *função kernel* que faz o mapeamento dos dados para um *espaço de funções*, e um algoritmo de aprendizagem que trabalha nesse espaço. Uma característica dos métodos de kernel é que a função kernel pode ser usada como medida de similaridade entre observações complexas como, por exemplo, observações na forma de sequências de caracteres, grafos, imagens, medidas de probabilidade, conjuntos, predicados lógicos, entre outros. Neste capítulo, são apresentados os fundamentos necessários para o desenvolvimento do trabalho.

3.1 Introdução

Em aprendizagem computacional, os conjuntos de dados são de grande importância, pois eles são usados para construir modelos que possam dar informação útil para o usuário, por exemplo, realizar predições a partir dos dados disponíveis, realizar seleção das características mais importantes de um conjunto de dados, realizar agrupamento, regressão, inferência, estimação de densidade, classificação, etc (Bishop e Nasrabadi, 2006; Duda *et al.*, 2000). Um enfoque comumente usado em aprendizagem computacional é mapear os dados originais para um *Espaço de Hilbert com Kernel Reproduzível* (EHKR) com uma função chamada de *kernel reproduzível*¹ (Aronszajn, 1950; Berg *et al.*, 1984; Kolmogorov, 1941; Mercer, 1909). O EHKR é um espaço próprio de funções de Hilbert o qual contém às funções representantes dos dados originais. Nesse espaço é onde são usados diversos algoritmos de aprendizado. Os métodos de aprendizagem computacional que trabalham desse jeito são chamados de *métodos de kernel* (Berlinet e Thomas-Agnan, 2004; Schölkopf e Smola, 2002; Shawe-Taylor e Cristianini, 2004; Wahba, 1990), por exemplo, as máquinas de vetores de suporte (Boser *et al.*, 1992; Cortes e Vapnik, 1995; Vapnik, 1995), a descrição de dados usando vetores de suporte (Tax e Duin, 2004), Kernel PCA (Schölkopf e Smola, 2002), processos gaussianos (Rasmussen e Williams, 2006), aprendizagem com múltiplos kernels (Rakotomamonjy *et al.*, 2007).

¹A palavra *kernel* vem da teoria de operadores integrais e não deve ser confundida com o conceito de *kernel* ou *núcleo* de um conjunto fuzzy.

Os métodos kernel são atraentes pelas seguintes razões:

1. o domínio de definição dos dados a serem analisados não tem requisitos adicionais, permitindo a análise de dados não vetoriais como grafos, conjuntos, strings, predicados lógicos, entre outros;
2. em um EHKR, se a função f está próxima, no sentido da norma, da função g , então os valores $f(x)$ estão próximos dos valores $g(x)$;
3. para construir um EHKR somente precisa-se ter uma função kernel $k : E \times E \rightarrow \mathbb{R}$, que seja positiva definida;
4. os cálculos dentro de um EHKR são feitos de maneira implícita usando o fato de que as avaliações da função kernel são iguais a produtos internos num EHKR. Usando a notação $k(., y)$ para denotar o mapeamento $x \mapsto k(x, y)$, com valor fixo y , temos: $k(x, y) = \langle k(., x), k(., y) \rangle_{\mathcal{H}}$, onde $x \in E \mapsto k(., x) \in \mathcal{H}$ é o mapeamento dos dados para o EHKR $\mathcal{H}$, e $k(., x), k(., y) \in \mathcal{H}$ são as funções representativas de $x, y \in E$ no EHKR;
5. $k(x, y)$ é uma medida de similaridade não linear entre os objetos $x, y \in E$, pois o mapeamento $x \in E \mapsto k(x, .) \in \mathcal{H}$ é não linear. Então funções muito simples no EHKR são úteis na análise de dados complexos;
6. os métodos de kernel são modulares, i.e., algoritmos no EHKR são independentes do kernel reproduzível que gera esse espaço, isto é, é possível escolher diferentes tipos de kernel sem ter que fazer mudanças no algoritmo;
7. o *kernel trick* (Schölkopf e Smola, 2002) permite que muitos algoritmos possam ser facilmente *kernelizados*, isto é, substituir produtos internos por funções kernel, isso inclui, algoritmos em problemas de classificação, regressão, agrupamento, redução de dimensionalidade, detecção de anomalias, seleção de características, entre outros.

Na próxima seção, introduziremos formalmente o conceito de kernel reproduzível e EHKR.

3.2 Espaços de Hilbert com Kernel Reproduzível

Um *Espaço de Hilbert com Kernel Reproduzível* (EHKR) é um espaço de funções de Hilbert com a propriedade de que a proximidade de duas funções, no sentido da norma, implica proximidades no seus valores. Na sequência, E denota um conjunto não vazio, $\mathcal{H}$ denota um EHKR de funções de valor real definidas em E . A notação $k(., y)$ quer dizer o mapeamento $x \mapsto k(x, y)$ com o valor y fixo, onde k é uma função de valor real, definida em $E \times E$. O produto interno no EHKR será denotado como $\langle ., . \rangle_{\mathcal{H}}$.

Definição 3.2.1 (Kernel reproduzível). Uma função²

$$\begin{aligned} k : E \times E &\rightarrow \mathbb{R} \\ (x, y) &\mapsto k(x, y), \end{aligned} \tag{3.1}$$

é chamada de *kernel reproduzível* de um espaço de Hilbert $\mathcal{H}$, se e somente se:

²Neste texto considera-se somente kernels de valor real, pois eles são as funções de maior interesse prático.

1. $\forall x \in E, k(., x) \in \mathcal{H}$
2. $\forall x \in E, \forall f \in \mathcal{H} \quad \langle f, k(., x) \rangle_{\mathcal{H}} = f(x)$

A condição (2) é chamada de *propriedade de reprodução*. Consequentemente $\forall (x, y) \in E \times E$:

$$k(x, y) = \langle k(., x), k(., y) \rangle_{\mathcal{H}} \quad (3.2)$$

Definição 3.2.2 (EHKR). Um espaço de Hilbert $\mathcal{H}$, cujos elementos são funções de valor real definidas em E , é um Espaço de Hilbert com Kernel Reprodutível (EHKR) se todos os funcionais de avaliação:

$$e_x : \mathcal{H} \rightarrow \mathbb{R} \quad (3.3)$$

$$f \mapsto e_x(f) = f(x) \quad (3.4)$$

são contínuos em $\mathcal{H}$.

Usando o teorema de representação de Riez (Debnath e Mikusiński, 2005), e a propriedade de reprodução (Definição 3.2.1), temos da Definição de EHKR que, $\forall x \in E, \forall f \in \mathcal{H}$:

$$e_x(f) = f(x) = \langle f, k(., x) \rangle_{\mathcal{H}} \quad (3.5)$$

O teorema de Riez implica que, para todo $x \in E$, a função $k(., x) \in \mathcal{H}$ é o único elemento de $\mathcal{H}$ com essa propriedade. Sendo é possível olhar a avaliação de f em qualquer ponto $x \in \Omega$ como o produto interno dado em (3.5).

O seguinte resultado apresenta a equivalência entre kernels reproduzíveis e funções positivas (Berlinet e Thomas-Agnan, 2004).

Lema 1. *Qualquer kernel reproduzível $k : E \times E \rightarrow \mathbb{R}$ é uma função simétrica positiva definida. Isto é, são funções que satisfazem:*

$$\sum_{i=1}^N \sum_{j=1}^N c_i c_j k(x_i, x_j) \geq 0 \quad (3.6)$$

$\forall N \in \mathbb{N}, \forall \{c_1, \dots, c_N\} \subset \mathbb{R}, \forall \{x_1, \dots, x_N\} \subset E$. O oposto também é verdadeiro.

Desta forma, funções simétricas positivas definidas, as quais chamaremos de *kernels positivos definidos*, são kernels reproduzíveis de algum EHKR. Olhando de outra forma, o espaço gerado por $k(x, .)$, cria um EHKR. Ou seja, para demonstrar que $k(., .)$ é positivo definido, basta demonstrar que k satisfaz (3.2).

3.2.1 Funções no EHKR

Seja o subespaço $\mathcal{H}_0$ de $\mathcal{H}$, definido por:

$$\mathcal{H}_0 = \left\{ f : E \rightarrow \mathbb{R} \mid f(\cdot) = \sum_{i=1}^N c_i k(., x_i), N \in \mathbb{N}, \{c_1, \dots, c_N\} \subset \mathbb{R}, \{x_1, \dots, x_N\} \subset E \right\},$$

com produto interno:

$$\langle f, g \rangle_{\mathcal{H}_0} = \sum_{i=1}^N \sum_{j=1}^M c_i d_j k(x_i, y_j),$$

onde $f(\cdot) = \sum_{i=1}^N c_i k(\cdot, x_i)$ e $g(\cdot) = \sum_{j=1}^M d_j k(\cdot, y_j)$. O seguinte teorema apresenta como o kernel k define um único EHKR e como é o tipo de funções pertencentes a $\mathcal{H}$.

Teorema 2 (Teorema Moore-Aronszajn ([Aronszajn, 1950](#))). *Seja k um kernel positivo definido em $E \times E$. Então, existe um único espaço de Hilbert $\mathcal{H}$ de funções definidas em E com k como seu kernel reproduzível. O subespaço de funções $\mathcal{H}_0$ gerado por combinações lineares das funções $(k(\cdot, x)_{x \in E})$ é denso em $\mathcal{H}$ e $\mathcal{H}$ é o conjunto de funções definidas em E que são limites ponto a ponto das sequências de Cauchy em $\mathcal{H}_0$.*

A Tabela 3.1 apresenta alguns exemplos de EHKR ([Berlinet e Thomas-Agnan, 2004](#)).

EHKR	$\mathcal{H} = \{\phi : E \rightarrow \mathbb{R} \mid \phi(\cdot) = \sum_{i=1}^N \lambda_i e_i(\cdot), \lambda_i \in \mathbb{R},$ $(e_1, e_2, \dots, e_N) \text{ é uma base ortonormal em } \mathcal{H}\}$
kernel	$k(x, y) = \sum_{i=1}^N e_i(x) e_i(y)$
representante de $y \in E$	$k(\cdot, y) = \sum_{i=1}^N e_i(\cdot) e_i(y) \in \mathcal{H}$
propriedade de reprodução	$\forall y \in E \quad \forall \phi \in \mathcal{H}$ $\langle \phi, k(\cdot, y) \rangle_{\mathcal{H}} = \langle \sum_{i=1}^N \lambda_i e_i(\cdot), \sum_{i=1}^N e_i(\cdot) e_i(y) \rangle_{\mathcal{H}} = \phi(y)$
EHKR	$\mathcal{H} = \{(x_i) = (x_1, x_2, \dots, x_i) \mid i \in \mathbb{N}, \sum_{i \in \mathbb{N}} x_i ^2 < \infty\}$
kernel	$k(i, j) = \begin{cases} 1, & i = j \\ 0, & \text{cc} \end{cases}$
representante de $j \in \mathbb{N}$	$k(\cdot, j) = (0, 0, \dots, 0, 1, 0, \dots) \in \mathcal{H}, (1 \text{ no } j\text{-ésimo lugar})$
propriedade de reprodução	$\forall j \in \mathbb{N} \quad \forall (x_i) \in \mathcal{H} \quad \langle (x_i), k(\cdot, j) \rangle_{\mathcal{H}} = x_j$

Tabela 3.1: Exemplo de EHKR, com seus respectivos kernels.

3.2.2 Propriedades de fechamento dos kernels

Os kernels têm propriedades de fechamento muito atraentes. Como consequência algumas operações feitas nos kernels resultam em outros kernels. Essas operações são muito úteis para verificar diretamente se uma função é, ou não, um kernel e adicionalmente, para construir novos kernels a partir de outros ([Shawe-Taylor e Cristianini, 2004](#)).

Sejam k_1, k_2 dois kernels positivos definidos em $E \times E$, $\alpha \in \mathbb{R}^+$, $f : E \rightarrow \mathbb{R}$, k_3 um kernel positivo definido em $\mathbb{R}^D \times \mathbb{R}^D$, Σ uma matriz simétrica de tamanho $N \times N$ positiva definida e $\phi : E \rightarrow \mathbb{R}^D$. As seguintes funções são kernels positivos definidos ([Shawe-Taylor e Cristianini, 2004](#)):

1. $k = k_1 + \alpha$,
2. $k = k_1 + k_2$,
3. $k = \alpha k_1$,
4. $k(x, x') = k_1(x, x')k_2(x, x')$,
5. se $k_1, k_2, \dots$ é uma sequência de kernels, e se $k(x, x') = \lim_{n \rightarrow \infty} k_n(x, x')$ existe para todo x, x' , então, k é um kernel,
6. $k(x, x') = f(x)f(x')$,
7. $k(x, x') = k_3(\phi(x), \phi(x'))$,
8. $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \Sigma \mathbf{x}'$, $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^D$.

Baseado nessas primitivas, é possível construir outros kernels, por exemplo o kernel polinomial definido em $\mathbb{R}^D \times \mathbb{R}^D$:

$$K(\mathbf{x}, \mathbf{x}') = (\alpha \mathbf{x}^\top \mathbf{x}' + \beta)^d, \quad \alpha, \beta \in \mathbb{R}^+, d \in \mathbb{N}^+, \quad (3.7)$$

é um kernel positivo definido.

Outros exemplos de kernels positivos definidos que podem ser construídos a partir das propriedades de fechamento são o kernel exponencial:

$$k(\mathbf{x}, \mathbf{x}') = \sum_{i=0}^{\infty} \frac{\sigma^i}{i!} \langle \mathbf{x}, \mathbf{x}' \rangle^i = \exp(\sigma \langle \mathbf{x}, \mathbf{x}' \rangle) \quad \sigma \in \mathbb{R}^+, \quad (3.8)$$

e o kernel gaussiano:

$$k(\mathbf{x}, \mathbf{x}') = \frac{\exp(2\sigma \langle \mathbf{x}, \mathbf{x}' \rangle)}{\sqrt{\exp(2\sigma \langle \mathbf{x}, \mathbf{x} \rangle) \exp(2\sigma \langle \mathbf{x}', \mathbf{x}' \rangle)}} = \exp(-\sigma \|\mathbf{x} - \mathbf{x}'\|^2). \quad (3.9)$$

3.2.3 Exemplos de kernels

A Tabela 3.2 apresenta alguns exemplos de kernels reproduzíveis ou kernels positivos definidos, definidos em $\mathbb{R}^D \times \mathbb{R}^D$ amplamente usados em aprendizagem computacional. Exemplos mais sofisticados incluem: kernels definidos sobre medidas de probabilidade (Jebara e Kondor, 2003a; Jebara *et al.*, 2004), kernels definidos sobre medidas (Cuturi *et al.*, 2005), kernels definidos sobre strings (Lodhi *et al.*, 2002), e outros kernels definidos em espaços não vetoriais como grafos, conjuntos, predicados lógicos (Gartner, 2008).

3.3 Máquinas de vetores de suporte

A máquinas de vetores de suporte (SVM) (Boser *et al.*, 1992; Cortes e Vapnik, 1995; Vapnik, 1995), são uma classe de algoritmos de aprendizado usadas em problemas de classificação de uma classe, classificação binária e classificação multi-classe, assim como para problemas de regressão, existindo diversas formulações e extensões para outros problemas de aprendizado computacional.

Nome	Definição
Linear	$\langle \mathbf{x}, \mathbf{x}' \rangle$
Polinomial	$(\alpha \mathbf{x}^\top \mathbf{x}' + \beta)^d$
Gaussiano	$\exp(-\sigma \ \mathbf{x} - \mathbf{x}'\ ^2)$
Exponencial	$\exp(\sigma \langle \mathbf{x}, \mathbf{x}' \rangle)$
Laplaciano	$\exp(-\sigma \ \mathbf{x} - \mathbf{x}'\)$
Quadrático racional	$\left(1 + \frac{\ \mathbf{x} - \mathbf{x}'\ ^2}{\alpha \beta^2}\right)^{-\alpha}$
Multi-quadrático	$-\sqrt{\ \mathbf{x} - \mathbf{x}'\ ^2 + \alpha^2}$
Multi-quadrático inverso	$(\sqrt{\ \mathbf{x} - \mathbf{x}'\ ^2 + \alpha^2})^{-1}$
Distância inversa	$1/(\ \mathbf{x} - \mathbf{x}'\ + \sigma)$
Distância inversa quadrática	$1/(\ \mathbf{x} - \mathbf{x}'\ ^2 + \sigma)$
Delta	$\delta(\mathbf{x}, \mathbf{x}') = \begin{cases} 1, & \mathbf{x} = \mathbf{x}' \\ 0, & \text{cc} \end{cases}$

Tabela 3.2: Exemplo de funções kernel, onde $\alpha, \beta \in \mathbb{R}^+, d \in \mathbb{N}$.

Para o caso de classificação binária, isto é, quando é dado um conjunto de dados $\{(x_i, y_i)\}_{i=1}^N$, chamado de conjunto de treinamento, o qual é uma amostra independente e identicamente distribuída (i.i.d.) a partir de um espaço $\Omega \times \mathcal{Y}$, tal que $\mathcal{Y} = \{-1, 1\}$, as SVM “aprendem” uma função $f \in \mathcal{H}$, tal que, dado um conjunto de dados $\{(x_t, y_t)\}_{t=N+1}^T \subset \Omega \times \mathcal{Y}$, chamado de conjunto de teste, diferente ao conjunto de treinamento, f apresenta boa capacidade de generalização, isto é comete pouco erro fazendo as predições $\text{sign}(f(x_t)) = y_t$, para todo $t = N + 1, \dots, T$.

Para aprender f a partir dos conjunto de treinamento $\{(x_i, y_i)\}_{i=1}^N$, as SVM procuram um hiperplano em $\mathcal{H}$ que melhor separe as *funções representantes* do conjunto de treinamento, nesse EHKR, com kernel k . Isto é, um hiperplano que separe o conjunto de pontos $\{k(\cdot, x_i)\}_{i=1}^N$ em dois conjuntos de pontos: $s_1 = \{k(\cdot, x_i)\}_{i=1}^M$, com $y_i = -1$ e $s_2 = \{k(\cdot, x_j)\}_{j=M+1}^N$, com $y_j = 1$. Esse hiperplano é achado usando o conceito da margem.

A teoria das SVM é bem motivada pela teoria de aprendizagem estatística (Vapnik, 1995), a qual cria um laço entre a generalização ou habilidade de f para fazer predições corretas, o risco empírico e a complexidade da classe das funções.

A Margem

A margem γ de um conjunto de pontos $\{k(\cdot, x_i)\}_{i=1}^N \subset \mathcal{H}$ em relação ao hiperplano $\Delta(f, b) = \{s \in \mathcal{H} \mid \langle f, s \rangle_{\mathcal{H}} + b = 0, f \in \mathcal{H}, b \in \mathbb{R}\}$ é o valor:

$$\gamma = \min_{i \in \{1, 2, \dots, N\}} d(k(\cdot, x_i), \Delta(f, b)), \quad (3.10)$$

onde $d(k(\cdot, x_i), \Delta(f, b))$ é a distância do ponto $k(\cdot, x_i)$ para o hiperplano $\Delta(f, b)$. Essa distância é a norma da projeção ortogonal do vetor $k(\cdot, x_i) - s$, $s \in \Delta(f, b)$ no vetor $f/\|f\|_{\mathcal{H}}$, tal que s é um ponto fixo arbitrário, ou seja,

$$d(k(\cdot, x_i), \Delta(f, b)) = \langle k(\cdot, x_i) - s, f/\|f\|_{\mathcal{H}} \rangle_{\mathcal{H}} \quad (3.11)$$

$$= \frac{|\langle k(\cdot, x_i), f \rangle_{\mathcal{H}} + b - \langle f, s \rangle_{\mathcal{H}} - b|}{\|f\|_{\mathcal{H}}} \quad (3.12)$$

$$= \frac{|f(x_i) + b - (\langle f, s \rangle_{\mathcal{H}} + b)|}{\|f\|_{\mathcal{H}}} \quad (3.13)$$

$$= \frac{|f(x_i) + b|}{\|f\|_{\mathcal{H}}} \quad (3.14)$$

onde a terceira linha é devido à propriedade de reprodução dos kernels.

A SVM

A SVM procura então o hiperplano que maximiza o margem γ , sendo formulada pelo seguinte problema de otimização:

Problema 3.3.1.

$$\begin{aligned} & \max_{f \in \mathcal{H}, b \in \mathbb{R}, \gamma \in \mathbb{R}} \quad \gamma \\ & \text{subject to} \quad y_i(f(x_i) + b) \geq \gamma, \quad i = 1, \dots, N \\ & \quad \quad \quad \|f\|_{\mathcal{H}} = 1, \quad i = 1, \dots, N, \end{aligned}$$

onde $f(x_i) = \langle f, k(\cdot, x_i) \rangle_{\mathcal{H}}$, segundo (3.2).

C-SVM

Existem diversas formulações de SVM, como por exemplo, a ν -SVM (Schölkopf e Smola, 2002; Schölkopf *et al.*, 2000), a RHS-SVM (Bennett e Bredensteiner, 2000; Bi e Bennett, 2001), a R-SVM (Bi e Vapnik, 2003) e a C-SVM (Cortes e Vapnik, 1995), as quais permitem que algumas observações do conjunto de treinamento sejam consideradas como erros. Uma comparação dessas formulações pode ser achada em (Bi e Vapnik, 2003).

No caso da C-SVM, chamada assim pois C é um parâmetro de regularização do modelo, é formulada usando o vetor de variáveis de folga $\xi \in \mathbb{R}^N$ com componentes ξ_i , para permitir que alguns elementos do conjunto $\{k(\cdot, x_i)\}_{i=1}^N$ sejam considerados como erros. Então a SVM do Problema 3.3.1 pode ser reformulada usando $C = \lambda/p$, da seguinte forma:

Problema 3.3.2.

$$\begin{aligned} \min_{f \in \mathcal{H}, b \in \mathbb{R}, \xi \in \mathbb{R}^N} \quad & \frac{1}{2} \|f\|_{\mathcal{H}}^2 + \frac{\lambda}{p} \sum_{i=1}^N \xi_i^p \\ \text{subject to} \quad & y_i(f(x_i) + b) \geq 1 - \xi_i, \quad i = 1, \dots, N \\ & \xi_i \geq 0, \quad i = 1, \dots, N, \end{aligned}$$

onde $\lambda \in \mathbb{R}^+$ é o parâmetro de regularização que compensa os erros permitidos entre o conjunto de treinamento e a função a ser procurada e $p \in \mathbb{N}$. No caso de $p = 1$, temos que a função de Lagrange é dada por:

$$\mathcal{L}(f, b, \xi, \alpha, \beta) = \frac{1}{2} \|f\|_{\mathcal{H}}^2 + \lambda \sum_{i=1}^N \xi_i - \sum_{i=1}^n \alpha_i (y_i(f(x_i) + b) - 1 + \xi_i) - \sum_{i=1}^N \beta_i \xi_i, \quad (3.15)$$

sendo a função dual de Lagrange:

$$\inf_{f, b, \xi} \mathcal{L}(f, b, \xi, \alpha, \beta), \quad (3.16)$$

onde $\alpha, \beta \in \mathbb{R}^N$ contêm os multiplicadores de Lagrange α_i, β_i , respectivamente. Como $\nabla_f(\|f\|_{\mathcal{H}}^2) = 2f$, e $\nabla_f(f(x)) = k(\cdot, x_i)$ ³, então diferenciando $\mathcal{L}$ em relação a f, b, ξ , temos:

$$\begin{aligned} \nabla_f \mathcal{L} = 0 & : f - \sum_{i=1}^N \alpha_i y_i k(\cdot, x_i) &= 0 \\ \partial_b \mathcal{L} = 0 & : \sum_{i=1}^N \alpha_i y_i &= 0 \\ \partial_{\xi_i} \mathcal{L} = 0 & : \lambda - \alpha_i - \beta_i &= 0. \end{aligned} \quad (3.17)$$

Substituindo (3.17) em (3.15), o problema dual Lagrangiano é o seguinte problema de otimização quadrática convexa:

³Isso pode ser verificado usando diferenciação funcional. Dado o funcional $J : \mathcal{H} \rightarrow \mathbb{R}$, a derivada funcional $\partial_f J$ é definida como $\partial_f J = \langle \nabla_f J, g \rangle_{\mathcal{H}} = \lim_{\epsilon \rightarrow 0} \frac{J(f+\epsilon g) - J(f)}{\epsilon}$, onde $f, g \in \mathcal{H}$, e $\nabla_f J : \mathcal{H} \rightarrow \mathcal{H}$.

Problema 3.3.3.

$$\begin{aligned}
& \min_{\alpha \in \mathbb{R}^N} \quad \frac{1}{2} \alpha^\top K \alpha - \mathbf{1}^\top \alpha \\
& \text{subject to} \quad \alpha^\top \mathbf{y} = 0. \\
& \quad \lambda \geq \alpha_i \geq 0, \quad i = 1, \dots, N.
\end{aligned}$$

onde $\mathbf{y} \in \mathbb{R}^N$ contem os valores y_i como suas componentes e, $K \in \mathbb{R}^{N \times N}$ tal que $K_{i,j} = k(x_i, x_j)$.

A SVM é interessante pelas seguintes razões:

- a solução é uma combinação linear dos elementos $k(., x_i)$, ou seja, $f = \sum_{i=1}^N \alpha_i y_i k(., x_i)$, e como os pontos corretamente classificados tem $\alpha = 0$, f somente dependerá daqueles pontos com $0 < \alpha \leq C$, chamados de *vetores de suporte*;
- a SVM é independente do kernel, pois K , pode ser construída usando qualquer kernel positivo definido;
- a SVM é formulada como um problema de otimização quadrática convexa, portanto, a solução encontra-se em um ótimo global;
- o custo computacional de trabalhar em espaços de alta dimensionalidade é solucionado pelo mapeamento implícito da função kernel $k(x_i, x_j) = \langle k(., x_i), k(., x_j) \rangle_{\mathcal{H}}$ no espaço de características. O Teorema de Cover ([Cover, 1965](#)) justifica o uso dos kernels, estabelecendo que dado um conjunto de treinamento não linearmente separável, a projeção dele em um espaço de alta dimensionalidade pode se transformar com alta probabilidade, num conjunto de treinamento linearmente separável.

3.4 Descrição dos dados com vetores de suporte

A descrição de dados com vetores de suporte (SVDD) ([Tax e Duin, 2004](#)), é um método kernel amplamente usando na descrição de conjuntos de dados com aplicação em tarefas de classificação de uma classe, detecção de anomalias, detecção de novidades, classificação binária e multi-classe. Dado o conjunto de dados $\{x_i\}_{i=1}^N \subset \Omega$, a SVDD estima a esfera envolvente mínima de raio $R \in \mathbb{R}^+$ e centro $f \in \mathcal{H}$, para as funções representantes $\{k(., x_i)\}_{i=1}^N$ daqueles dados no EHKR $\mathcal{H}$.

A SVDD então é formulada como o seguinte problema de otimização:

Problema 3.4.1.

$$\begin{aligned}
& \min_{f \in \mathcal{H}, R \in \mathbb{R}, \xi \in \mathbb{R}^N} \quad R^2 + \lambda \sum_{i=1}^N \xi_i \\
& \text{subject to} \quad \|k(., x_i) - f\|_{\mathcal{H}}^2 \leq R^2 + \xi_i, \quad i = 1, \dots, N \\
& \quad \xi_i \geq 0, \quad i = 1, \dots, N,
\end{aligned}$$

onde $\lambda \in \mathbb{R}^+$ é um parâmetro de regularização. A função de Lagrange para o Problema 3.4.1 é:

$$\mathcal{L}(f, R, \xi, \alpha, \beta) = R^2 + \lambda \sum_{i=1}^N \xi_i - \sum_{i=1}^n \alpha_i (\|k(., x_i) - f\|_{\mathcal{H}}^2 - R^2 - \xi_i) - \sum_{i=i}^N \beta_i \xi_i, \quad (3.18)$$

sendo a função dual de Lagrange:

$$\inf_{f, R, \xi} \mathcal{L}(f, b, \xi, \alpha, \beta), \quad (3.19)$$

onde $\alpha, \beta \in \mathbb{R}^N$ contém os multiplicadores de Lagrange α_i, β_i , respectivamente. Como $\nabla_f(\|f\|_{\mathcal{H}}^2) = 2f$, e $\nabla_f(f(x)) = k(\cdot, x_i)$, então diferenciando $\mathcal{L}$ em relação a f, R, ξ , temos:

$$\begin{aligned} \nabla_f \mathcal{L} = 0 & : 2f \sum_{i=1}^N \alpha_i - 2 \sum_{i=1}^N \alpha_i k(\cdot, x_i) &= 0 \\ \partial_R \mathcal{L} = 0 & : 1 - \sum_{i=1}^N \alpha_i &= 0 \\ \partial_{\xi_i} \mathcal{L} = 0 & : \lambda - \alpha_i - \beta_i &= 0. \end{aligned} \quad (3.20)$$

Substituindo (3.20) em (3.18), o problema dual Lagrangiano é o seguinte problema de otimização:

Problema 3.4.2.

$$\begin{aligned} \min_{\alpha \in \mathbb{R}^N} \quad & \alpha^\top K \alpha - \alpha^\top \text{diag}(K) \\ \text{subject to} \quad & \alpha^\top \mathbf{1} = 0. \\ & \lambda \geq \alpha_i \geq 0, \quad i = 1, \dots, N, \end{aligned}$$

onde $\mathbf{1} \in \mathbb{R}^N$ contém os valores 1 como suas componentes e, $K \in \mathbb{R}^{N \times N}$ tal que $K_{i,j} = k(x_i, x_j)$. A solução f , pode ser obtida de (3.20) como $f = \sum_{i=1}^N \alpha_i k(\cdot, x_i)$.

3.5 Discrepância Máxima de Médias nos EHKR

A Discrepância Máxima de Médias (MMD) (Borgwardt *et al.*, 2006; Gretton *et al.*, 2006, 2009, 2012) é uma estatística de teste definida da seguinte maneira.

Definição 3.5.1 (MMD). Sejam $f : \Omega \rightarrow \mathbb{R}$ funções pertencentes a $\mathcal{F}$. Sejam X e Y duas variáveis aleatórias definidas em Ω , com suas respectivas medidas de probabilidade $\mathbb{P}$ e $\mathbb{Q}$. A MMD é definida como:

$$\text{MMD}[\mathcal{F}, \mathbb{P}, \mathbb{Q}] = \sup_{f \in \mathcal{F}} (\mathbb{E}_{X \sim \mathbb{P}}[f(X)] - \mathbb{E}_{Y \sim \mathbb{Q}}[f(Y)]), \quad (3.21)$$

onde $\mathbb{E}_{X \sim \mathbb{P}}$ denota a esperança da variável aleatória X distribuída de acordo a $\mathbb{P}$.

Se $\mathcal{F}$ for uma esfera unitária num EHKR, a MMD é definido por:

$$\text{MMD}[\mathcal{F}, \mathbb{P}, \mathbb{Q}] = \sup_{\|f\|_{\mathcal{H}} \leq 1} (\mathbb{E}_{X \sim \mathbb{P}}[f(X)] - \mathbb{E}_{Y \sim \mathbb{Q}}[f(Y)]) \quad (3.22)$$

$$= \sup_{\|f\|_{\mathcal{H}} \leq 1} (\mathbb{E}_{X \sim \mathbb{P}}[\langle k(\cdot, X), f \rangle_{\mathcal{H}}] - \mathbb{E}_{Y \sim \mathbb{Q}}[\langle k(\cdot, Y), f \rangle_{\mathcal{H}}]) \quad (3.23)$$

$$= \sup_{\|f\|_{\mathcal{H}} \leq 1} \langle \mathbb{E}_{X \sim \mathbb{P}}[k(\cdot, X)] - \mathbb{E}_{Y \sim \mathbb{Q}}[k(\cdot, Y)], f \rangle_{\mathcal{H}} \quad (3.24)$$

$$= \|\mathbb{E}_{X \sim \mathbb{P}}[k(\cdot, X)] - \mathbb{E}_{Y \sim \mathbb{Q}}[k(\cdot, Y)]\|_{\mathcal{H}}, \quad (3.25)$$

Logo, usando o kernel k , do EHKR $\mathcal{H}$, teremos:

$$\begin{aligned} \text{MMD}^2[\mathcal{F}, \mathbb{P}, \mathbb{Q}] &= \langle \mathbb{E}_{X \sim \mathbb{P}}[k(\cdot, X)], \mathbb{E}_{X' \sim \mathbb{P}}[k(\cdot, X')] \rangle_{\mathcal{H}} - 2 \langle \mathbb{E}_{X \sim \mathbb{P}}[k(\cdot, X)], \mathbb{E}_{Y \sim \mathbb{Q}}[k(\cdot, Y)] \rangle_{\mathcal{H}} \\ &\quad + \langle \mathbb{E}_{Y \sim \mathbb{Q}}[k(\cdot, Y)], \mathbb{E}_{Y' \sim \mathbb{Q}}[k(\cdot, Y')] \rangle_{\mathcal{H}} \end{aligned} \quad (3.26)$$

$$= \mathbb{E}_{X \sim \mathbb{P}, X' \sim \mathbb{P}} k(X, X') - 2 \mathbb{E}_{X \sim \mathbb{P}, Y \sim \mathbb{Q}} k(X, Y) + \mathbb{E}_{Y \sim \mathbb{Q}, Y' \sim \mathbb{Q}} k(Y, Y'). \quad (3.27)$$

Como sejam observadas amostras do tipo: $s_X = \{x_1, \dots, x_m\}$ e $s_Y = \{y_1, \dots, y_n\}$, distribuídas de acordo com $\mathbb{P}$ e $\mathbb{Q}$, respectivamente, um estimador empírico sem viés, denotado por MMD_u , é dado por (Lema 6 de [Gretton et al. \(2012\)](#)):

$$\text{MMD}_u^2[\mathcal{F}, s_X, s_Y] = \frac{1}{m(m-1)} \sum_{i=1}^m \sum_{j \neq i}^m k(x_i, x_j) - \frac{2}{mn} \sum_{i=1}^m \sum_{j=1}^n k(x_i, y_j) + \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i}^n k(y_i, y_j) \quad (3.28)$$

3.5.1 Teste de hipótese com a MMD

Dadas as amostras i.i.d. $s_X = \{x_1, \dots, x_m\} \sim \mathbb{P}$, e $s_Y = \{y_1, \dots, y_n\} \sim \mathbb{Q}$, o MMD é usado para distinguir entre a hipótese nula $H_0 : \mathbb{P} = \mathbb{Q}$, e a hipótese alternativa: $H_A : \mathbb{P} \neq \mathbb{Q}$. Isso pode ser conseguido comparando o MMD com um limiar ϵ , isto é, se $\text{MMD}[\mathcal{F}, s_X, s_Y] > \epsilon$, a hipótese nula é rejeitada. Caso contrário não existe evidência para rejeitar a hipótese nula. O erro do Tipo I acontece quando H_0 é rejeitada, apesar da hipótese nula ser verdadeira. O erro do Tipo II acontece quando a hipótese nula é aceita apesar das distribuições serem diferentes. O nível de significância do teste α é uma limitante superior nos erros do Tipo I e é usado para determinar o limiar ϵ , o qual pode ser estimado usando vários métodos (Seções 4 e 5 de [Gretton et al. \(2012\)](#)).

Capítulo 4

Descrição de Dados com Medidas de Suporte

Every little bit helps.

Proverb

Neste capítulo, formula-se um modelo de descrição de dados para conjuntos de dados cujas observações são conjuntos de pontos em $\mathbb{R}^D$. Usando medidas de probabilidade para modelar as observações, a descrição do conjunto de dados será feita estimando um conjunto de volume mínimo para aquelas medidas de probabilidade, especificamente, por meio de uma esfera envolvente mínima das funções representantes daquelas medidas de probabilidade, em um Espaço de Hilbert com Kernel Reprodutível (EHKR). A vantagem disso é que não será considerada a "priori" nenhuma definição para as medidas de probabilidade, nem mesmo estimação da densidade. Ao invés disso, será usado um *mergulho* nos EHKR para medidas de probabilidade, o qual será feito com funções kernels positivas definidas, definidas no espaço das medidas de probabilidade. Como resultado, o modelo de descrição de dados é uma função que depende unicamente de algumas medidas de probabilidade: as *medidas de suporte*. Nome dado em analogia aos vetores de suporte [Schölkopf et al. \(2001\)](#). Formulamos três modelos de descrição de dados com medidas de suporte para esses conjuntos de dados: o primeiro é formulado como um problema de otimização com restrições de chance; o segundo é uma extensão direta do método de descrição de dados usando vetores de suporte para o caso das medidas de probabilidade; o terceiro é igual ao segundo, porém definido somente para kernels estacionários e requerendo escalamento dos dados.

4.1 Introdução

Chama-se Descrição de Dados (DD) à tarefa de construir modelos que possam representar ou descrever os elementos de um conjunto de dados, com o objetivo de realizar tarefas de aprendizado computacional como detecção de anomalias, detecção de novidades, agrupamento e classificação de padrões ([Chandola et al., 2009](#); [Garcia et al., 2003](#); [Polonik, 1997](#); [Schölkopf et al., 2001](#); [Scott e Nowak, 2006](#); [Tax e Duin, 2004](#)). A principal ideia dos métodos de DD é assumir que os elementos pertencentes a um conjunto de dados são gerados por uma alguma distribuição desconhe-

cida. Consequentemente, muitos métodos de DD computam alguma informação dessa distribuição a partir do conjunto de dados. Por exemplo, aproximam empiricamente a densidade dessa distribuição, ou algum conjunto de nível dessa densidade, ou extraem informação do suporte da densidade dessa distribuição.

Frequentemente, os métodos de DD trabalham em conjuntos de dados expressos formalmente por: $\{\mathbf{x}_i\}_{i=1}^N$, $\mathbf{x}_i \in \mathbb{R}^D$, onde N é o número de observações. No entanto, existe um interesse crescente em algoritmos de aprendizado computacional para conjuntos de dados cujas observações individuais são grupos, ou conjuntos de pontos em $\mathbb{R}^D$ (Ben-Tal *et al.*, 2011; Graepel e Herbrich, 2003; Guevara *et al.*, 2013a; Kondor e Jebara, 2003; Muandet e Schoelkopf, 2013; Muandet *et al.*, 2012; Póczos *et al.*, 2012; Shivaswamy *et al.*, 2006; Smola *et al.*, 2007; Viertl, 2011; Xiong *et al.*, 2011a,b; Yang e Gunn, 2007; Zhang, 2005). Aqueles conjuntos de dados são conjuntos da forma:

$$\mathcal{T} = \{s_i\}_{i=1}^N, \quad (4.1)$$

onde N indica o número de observações e a observação s_i é um conjunto $\{\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, \dots, \mathbf{x}_{L_i}^{(i)}\}$, com cardinalidade L_i , e cada elemento de s_i é um ponto em $\mathbb{R}^D$. Exemplos práticos de observações tomando a forma de s_i são: um conjunto de características de uma imagem num conjunto de imagens (Lowe, 2004); um conjunto de características espaço-temporal (Laptev, 2005); um conjunto de valores replicados num processo de medição (Wernisch *et al.*, 2003); um conjunto de pontos descrevendo incerteza pontual (Ben-Tal *et al.*, 2011; Shivaswamy *et al.*, 2006; Zhang, 2005); um conjunto de pontos descrevendo opiniões subjetivas (Viertl, 2011); um conjunto de pontos descrevendo a invariância de um objeto em particular (Graepel e Herbrich, 2003); conjuntos de dados provenientes das áreas da meteorologia, economia e bioinformática (Ben-Tal *et al.*, 2011; Bhadra, 2012; Bhadra *et al.*, 2009a,b, 2010), também contém observações desse tipo.

A descrição de dados dos conjuntos do tipo (4.1) é bastante importante devido às possíveis aplicações práticas que ilustramos tomando como exemplo a tarefa de *detecção de anomalias grupais* (Xiong *et al.*, 2011a,b). Detectar anomalias grupais consiste em detectar conjuntos de pontos anômalos, isto é, detectar observações não usuais s_i de (4.1). Cada conjunto anômalo pode ser dado por (Xiong *et al.*, 2011a): 1) *anomalias pontuais*, ou seja, a agregação de pontos anômalos, isto é, todos os elementos em s_i são anômalos; 2) *anomalias de distribuição*, ou seja, a agregação anômala de pontos aparentemente não anômalos, isto é, os elementos em s_i são aparentemente não anômalos, mas a agregação deles é não usual.

Nos métodos clássicos de detecção de anomalias, (detectar pontos anômalos $\mathbf{x}_i \in \mathbb{R}^D$ do conjunto de dados $\{\mathbf{x}_i\}_{i=1}^N$) um ponto é anômalo se ele é diferente em relação a maioria de pontos do conjunto de dados. Por exemplo, um ponto longe da média empírica, ou longe do suporte estimado da densidade da distribuição geradora dos elementos do conjunto de dados. Contudo, quando a tarefa é detectar anomalias grupais, um grupo ou conjunto de pontos é considerado anômalo, se a distribuição local daqueles pontos não é similar às distribuições locais dos grupos não anômalos. Em consequência, a informação fornecida por cada distribuição local de pontos é crucial para a correta descrição de conjuntos de dados segundo (4.1) A Figura 4.1 descreve um caso onde tenta-se descrever um conjunto de dados do tipo (4.1) usando as médias empíricas dos conjuntos s_i . Note que a descrição do conjunto de dados do tipo (4.1) a partir daqueles valores (médias empíricas) não é suficiente para garantir detecção de anomalias grupais pois existe uma sobreposição entre as médias empíricas dos grupos anômalos e dos grupos não anômalos. Além disso, reduzindo a informação

fornecida pelo conjunto s_i a um único valor tornará os métodos convencionais de DD altamente dependentes desse procedimento.

Figura 4.1: Detecção de anomalias grupais. A borda verde apresenta dois gráficos de duas observações (grupos de pontos) consideradas não anômalas. Esses dois grupos de pontos foram amostradas a partir de uma distribuição de misturas de Gaussianas (GMD) com três componentes. A borda vermelha apresenta dois grupos anômalos: a parte inferior esquerda apresenta um grupo de pontos amostrado a partir de uma distribuição normal e a parte inferior direita apresenta um grupo de pontos amostrados a partir de uma GMD com cinco componentes. Os dois grupos inferiores são considerados anômalos pois a distribuição deles é diferente dos grupos considerados não anômalos.

Conjunto de volume mínimo (MV-set) para descrição de dados. Os MV-sets são amplamente usados para descrever conjuntos de dados da forma $\{\mathbf{x}_i\}_{i=1}^N$, $\mathbf{x}_i \in \mathbb{R}^D$, e eles são definidos como conjuntos de pontos que podem ser medidos (Garcia *et al.*, 2003; Polonik, 1997; Schölkopf *et al.*, 2001; Scott e Nowak, 2006). O MV-set é então computado seguindo um critério de otimização, tendo em conta todos os possíveis conjuntos de volume. As classes de conjuntos usadas por os métodos DD vão desde conjuntos convexas (Polonik, 1997), até conjuntos com elementos no EHKR definidos implicitamente via kernels positivos definidos (Muandet e Schoelkopf, 2013; Schölkopf *et al.*, 2001; Tax e Duin, 2004).

Este capítulo descreve como encontrar a descrição de conjuntos de dados definidos por (4.1). Assumimos primeiramente que os elementos em s_i são gerados a partir de uma distribuição local e desconhecida a qual denotamos como $\mathbb{P}_i$. Então, a descrição de (4.1) pode ser formulada como a estimação do MV-set para $\{\mathbb{P}_i\}_{i=1}^N$. Formalmente, assumiremos que os pontos em cada observação: $\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, \dots, \mathbf{x}_{L_i}^{(i)} \in s_i$, são realizações i.i.d¹ de uma variável aleatória X distribuída de acordo uma medida de probabilidade desconhecida $\mathbb{P}_i$ definida no espaço mensurável $(\mathbb{R}^D, \mathcal{B}(\mathbb{R}^D))$, com $\mathcal{B}(\mathbb{R}^D)$

¹Independentes e identicamente distribuídas.

Figura 4.2: Dificuldade de usar somente valores representativos na detecção de anomalias grupais. Os pontos vermelhos e verdes são as médias empíricas dos grupos anômalos e não anômalos, respectivamente, de um conjunto de dados artificial, onde são considerados grupos não anômalos aqueles gerados por uma GMD de três componentes com um conjunto de parâmetros arbitrário, e são considerados grupos anômalos aqueles grupos formados por diversas distribuições normais. A sobreposição das médias empíricas dos grupos anômalos e não anômalos sugerem que os métodos convencionais de DD não sejam adequados para realizar detecção de anomalias para este conjunto de dados. Isso motiva o uso da informação da distribuição dos pontos para esse tipo de tarefa.

denotando a σ -álgebra de Borel associada a $\mathbb{R}^D$. Em seguida, amplia-se a definição de MV-set (Garcia *et al.*, 2003; Polonik, 1997; Schölkopf *et al.*, 2001; Scott e Nowak, 2006) para o caso de medidas de probabilidade.

Definição 4.1.1 (MV-set para medidas de probabilidade). Seja $(\mathcal{P}, \mathcal{A}, \mathcal{E})$ um espaço de probabilidade, onde $\mathcal{P}$ é o espaço de todas as medidas de probabilidade $\mathbb{P}$ definidas em $(\mathbb{R}^D, \mathcal{B}(\mathbb{R}^D))$, $\mathcal{A}$ é uma σ -álgebra de $\mathcal{P}$, e $\mathcal{E}$ é uma medida de probabilidade definida em $(\mathcal{P}, \mathcal{A})$. Um MV-set é o conjunto dado por² ³:

$$G_\alpha^* = \operatorname{argmin}_G \{\rho(G) | \mathcal{E}(G) \geq \alpha, G \in \mathcal{A}\}, \quad (4.2)$$

onde ρ é uma medida de referência definida em $\mathcal{A}$, e $\alpha \in [0, 1]$. O MV-set G_α^* descreve a fração α da concentração de massa de $\mathcal{E}$.

Isto é, dado $\alpha \in [0, 1]$ e uma medida de referência ρ , o MV-set para um conjunto de medidas de probabilidade é calculado como segue:

1. medindo cada possível conjunto de medidas de probabilidade $G \in \mathcal{A}$, mediante $\mathcal{E}(G)$,
2. mantendo todos os conjuntos de medidas de probabilidade $G \in \mathcal{A}$ satisfazendo $\mathcal{E}(G) \geq \alpha$ e,
3. selecionado o conjunto $G_\alpha^* \in \mathcal{A}$, tal que $\rho(G_\alpha^*)$ é o ínfimo $\rho(G)$ sobre todos os conjuntos $G \in \mathcal{A}$ satisfazendo $\mathcal{E}(G) \geq \alpha$.

² $\mathcal{A}$ é por instância a σ -álgebra de Borel em relação a topologia de convergência fraca (Ferguson, 1974; Muandet e Schoelkopf, 2013).

³Assumindo que todas as medidas de probabilidade $\mathbb{P} \in \mathcal{P}$ tem domínio compacto.

Encontrar um MV-set de um conjunto de medidas de probabilidade com esse procedimento é muito geral, no lugar disso, limitamos nossa atenção à classe de conjuntos $\mathcal{A}$ formada por conjuntos de medidas de probabilidade satisfazendo que as realizações das variáveis aleatórias seguindo essas medidas de probabilidade estejam dentro de uma esfera.

Esferas envoltentes nos EHKR. Consideramos o conjunto $\{\mathbb{P}_i\}_{i=1}^N$ como uma amostra i.i.d. de acordo com $\mathcal{E}$. Como cada $\mathbb{P}_i$ é desconhecida, não vai ser assumida nenhuma forma especial para essas medidas, em vez disso, usamos um mapeamento implícito ao espaço de características dado por um kernel positivo definido, definido sobre medidas de probabilidade, i.e., um kernel de valor real definido em $\mathcal{P} \times \mathcal{P}$. Consideramos que a classe $\mathcal{A}$ em (4.2) é implicitamente definida por aquele kernel como o conjunto de esferas envoltentes dos mapeamentos implícitos ou *funções representantes* de $\{\mathbb{P}_i\}_{i=1}^N$ no EHKR. Consequentemente, o MV-set dado por (4.2) é obtido pela mínima esfera envolvente das funções representantes no EHKR.

As funções representantes são dadas pelos mergulhos das medidas de probabilidade no EHKR (Berlinet e Thomas-Agnan, 2004; Guilbart, 1979; Smola *et al.*, 2007; Suquet *et al.*, 1995). Tal mergulho fornece uma maneira para calcular o produto interno em $\mathcal{P}$ usando um kernel definido em $\mathcal{P} \times \mathcal{P}$, sem ter que calcular a densidade das distribuições locais como passo intermediário. Além do mais, uma boa aproximação para o kernel é garantida a partir de uma estimação empírica desse kernel usando (4.1) (Smola *et al.*, 2007). Ou seja, na mesma forma dos métodos de kernel, a descrição de $\{\mathbb{P}_i\}_{i=1}^N$ será uma função que depende somente de alguns pontos $\mathbb{P}_i$, os quais, em analogia aos vetores suporte nos métodos de kernel, serão chamados de *medidas de suporte*. Consequentemente, o método apresentado neste capítulo será chamado como *Descriptor de dados com medidas de suporte*, ou SMDD. Esse método o qual não assume nenhuma forma particular para as densidades das medidas de probabilidade.

Serão apresentados três modelos SMDD neste capítulo. Na Seção 4.2 é apresentado o primeiro modelo como um problema de otimização com restrições de chance (Shapiro *et al.*, 2009; Wallace e Ziemba, 2005) no espaço das medidas de probabilidade. Esses conceitos são logo estendidos para o EHKR na Seção 4.3. O segundo e o terceiro modelo são extensões diretas do modelo SVDD (Tax e Duin, 2004) para o caso de conjuntos de dados, definidos por $\{\mathbb{P}_i\}_{i=1}^N$. O terceiro modelo é igual ao segundo modelo com uma única diferença que ele considera escalamento dos dados e kernels invariantes a translação. Esses modelos SMDD são apresentados na Seção 4.3. A relação entre todos esses modelos é apresentado na Seção 4.4. Mostra-se através de um conjunto de experimentos na Seção 4.5, o comportamento desses modelos na tarefa de detecção de anomalias grupais usando conjuntos de dados artificiais e reais. A Seção 4.6 apresenta um resumo do capítulo.

Notação. Considera-se uma variável aleatória X definida no espaço de probabilidade $(\Omega, \mathcal{F}, \mathbb{P})$ como um mapeamento Borel mensurável: $\Omega \rightarrow \mathbb{R}^D$, satisfazendo $X(\omega) = \omega$, $\forall \omega \in \Omega$, i.e., X é a função identidade. Também, consideramos $\Omega = \mathbb{R}^D$ e $\mathcal{F} = \mathcal{B}(\mathbb{R}^D)$, implicando que, para $B \in \mathcal{B}(\mathbb{R}^D)$, a medida de probabilidade induzida por X definida como $\mathbb{P}_X(B) = \mathbb{P}\{\omega : X(\omega) \in B\}$ é igual à medida de probabilidade $\mathbb{P}(B)$, i.e., $\mathbb{P}_X = \mathbb{P}$. Sempre abreviaremos a expressão $\mathbb{P}\{\omega : a < X(\omega) \leq b\}$ por $\mathbb{P}(a < X \leq b)$. A letra $\mathcal{H}$ denota a EHKR das funções $f : \mathbb{R}^D \rightarrow \mathbb{R}$, com kernel positivo definido $k : \mathbb{R}^D \times \mathbb{R}^D \rightarrow \mathbb{R}$, e norma $\|\cdot\|_{\mathcal{H}}$. Também, denotamos por $k(X_i, \cdot)$, o mapeamento $t \rightarrow k(X_i, t)$, onde $X_i \sim \mathbb{P}_i$. Os produtos internos em $\mathcal{H}$ são denotados como $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. A Tabela 4.1 sumariza a notação usada neste capítulo.

Simbolo	Descrição
$D \in \mathbb{N}$	dimensão
$N \in \mathbb{N}$	número de elementos no conjunto de dados
$R \in \mathbb{R}$	o raio da esfera
$\mathbf{c} \in \mathbb{R}^D$	o centro da esfera
$\lambda \in \mathbb{R}^+$	parâmetro de regularização
$\boldsymbol{\xi} \in \mathbb{R}^D$	vetor de variáveis de folga
$\mathcal{P}$	espaço das medidas de probabilidade
$\mathbb{P} \in \mathcal{P}$	medida de probabilidade
X	variável aleatória tomando valores em $\mathbb{R}^D$
$\kappa \in [0, 1]$	valores limiar
$tr(A)$	traço da matriz A
$\boldsymbol{\mu} \in \mathbb{R}^D$	média de um conjunto de pontos em $\mathbb{R}^D$
Σ	covariância de um conjunto de pontos $\mathbb{R}^D$
$\boldsymbol{\alpha} \in \mathbb{R}^N$	vetor de multiplicadores de Lagrange
$\mathcal{H}$	espaço de Hilbert com kernel reproduzível (EHKR)
k	kernel positivo definido de valor real, definido em $\mathbb{R}^D \times \mathbb{R}^D$
$k(x, \cdot) \in \mathcal{H}$	função de avaliação no ponto x em $\mathcal{H}$
$c \in \mathcal{H}$ ou $c(\cdot) \in \mathcal{H}$	o centro da esfera no EHKR
$\mu_{\mathbb{P}}$	função média de $\mathbb{P}$ no $\mathcal{H}$
$\langle \cdot, \cdot \rangle_{\mathcal{H}}$	produto interno no EHKR $\mathcal{H}$
$\Sigma^{\mathcal{H}}$	operador covariância no EHKR $\mathcal{H}$
$\tilde{k}$	kernel positivo definido de valor real, definido em $\mathcal{P} \times \mathcal{P}$
$\ \cdot\ _{\mathcal{H}}$	norma no EHKR

Tabela 4.1: Notação usada neste capítulo.

4.2 Descritor de dados com medidas de suporte no espaço das medidas de probabilidade

Nesta seção estima-se o MV-set empírico para $\{\mathbb{P}_i\}_{i=1}^N$, como uma esfera envolvente mínima para as realizações das variáveis aleatórias distribuídas de acordo com $\{\mathbb{P}_i\}_{i=1}^N$. Chamamos esse modelo de *SMDD no espaço das medidas de probabilidade*, o qual é apresentado como um problema de otimização com restrições de chance (Shapiro *et al.*, 2009; Wallace e Ziemba, 2005). Logo, usando a desigualdade de Markov (Boucheron *et al.*, 2013), transformaremos o problema de otimização num outro com restrições determinísticas. Todos os resultados desta seção são logo kernelizadas e consequentemente estendidas ao EHKR na Seção 4.3.

Dada uma amostra i.i.d $\{\mathbb{P}_i\}_{i=1}^N$, possível conjunto $G \in \mathcal{A}$ em (4.2) é o conjunto:

$$\hat{G}(R, \mathbf{c}) = \{\mathbb{P}_i \in \mathcal{P} \mid \|X_i - \mathbf{c}\|^2 \leq R^2\}, \quad (4.3)$$

onde $(R, \mathbf{c}) \in \mathbb{R} \times \mathbb{R}^D$ são chamadas de *esferas envolventes* de raio R e centro $\mathbf{c}$, para $\{\mathbb{P}_i\}_{i=1}^N$, e $X_i \sim \mathbb{P}_i$ ⁴. O MV-set empírico é então achado, estimando a esfera envolvente mínima (valores ótimos R e $\mathbf{c}$ seguindo um critério de otimização) usando a amostra $\{\mathbb{P}_i\}_{i=1}^N$. Neste caso, o valor ótimo para o raio será proporcional ao valor α em (4.2). Contudo, (4.3) é uma definição muito conservadora, porque $\mathbb{P}_i$ vai estar em (4.3), somente se todas as possíveis realizações de $X_i \sim \mathbb{P}_i$ estão dentro de $(R, \mathbf{c})$.

Uma formulação mais flexível é dada, permitindo-se que algumas realizações de $X_i \sim \mathbb{P}_i$ não sejam parte do conjunto $\hat{G}$. Isso pode ser feito fixando arbitrariamente alguns valores limiares: $\mathcal{K} = \{\kappa_i\}_{i=1}^N$, $\kappa_i \in [0, 1]$ para controlar que a probabilidade das realizações de $X_i \sim \mathbb{P}_i$ estejam dentro da esfera $(R, \mathbf{c})$. Ou seja, dado o conjunto $\mathcal{K} = \{\kappa_i\}_{i=1}^N$, $\kappa_i \in [0, 1]$, uma formulação mais flexível para uma versão empírica de G em (4.2) é dada por:

$$\hat{G}(\mathcal{K}, R, \mathbf{c}) = \{\mathbb{P}_i \in \mathcal{P} \mid \mathbb{P}_i(\|X_i - \mathbf{c}\|^2 \leq R^2) \geq 1 - \kappa_i\}. \quad (4.4)$$

Como cada medida de probabilidade $\mathbb{P}_i$ está em $\hat{G}(\mathcal{K}, R, \mathbf{c})$, dependendo do valor associado κ_i , é possível observar que se todos os $\kappa_i = 0$, então (4.4) se reduz à (4.3), e se para alguns i , $\kappa_i = 1$, então $\mathbb{P}_i$ está sempre em $\hat{G}(\mathcal{K})$. As medidas de probabilidade não consideradas parte de $\hat{G}(\mathcal{K})$ são aquelas tal que a correspondente função de distribuição $\mathbb{P}_i(\|X_i - \mathbf{c}\|^2 \leq R^2)$ é menor que κ_i .

O MV-set empírico segundo (4.4) pode ser formulado como um problema de otimização com restrições de chance. Dado o conjunto de medidas de probabilidade $\{\mathbb{P}_i\}_{i=1}^N$, e os valores $\{\kappa_i\}_{i=1}^N$, $\kappa_i \in [0, 1]$, esse problema de otimização é o seguinte *programa com restrições de chance*:

$$\begin{aligned} & \min_{\mathbf{c} \in \mathbb{R}^D, R \in \mathbb{R}} R^2 \\ & \text{subject to } \mathbb{P}_i(\|X_i - \mathbf{c}\|^2 \leq R^2) \geq 1 - \kappa_i, \quad i = 1, \dots, N \end{aligned}$$

onde R e $\mathbf{c}$ são o raio e o centro da hiperesfera, respectivamente e, o vetor aleatório $X_i \sim \mathbb{P}_i$ é o parâmetro de incerteza para o modelo com restrições de chance.

Da mesma forma que nos métodos de kernel, introduzimos as varáveis de folga: $\boldsymbol{\xi} = (\xi_1, \xi_2, \dots, \xi_N) \in \mathbb{R}^N$ para permitir que algumas medidas de probabilidade de $\{\mathbb{P}_i\}_{i=1}^N$ não sejam parte da esfera en-

⁴A notação $\sim$ quer dizer *distribuída de acordo a*.

volvente mínima. Então, temos o seguinte programa com restrições de chance:

Problema 4.2.1.

$$\begin{aligned} & \min_{\mathbf{c} \in \mathbb{R}^D, R \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^N} R^2 + \lambda \sum_{i=1}^N \xi_i \\ & \text{subject to} \quad \mathbb{P}_i(\|X_i - \mathbf{c}\|^2 \leq R^2 + \xi_i) \geq 1 - \kappa_i, \quad i = 1, \dots, N \\ & \quad \xi_i \geq 0 \quad i = 1, \dots, N. \end{aligned}$$

onde $\lambda > 0$ é um parâmetro de regularização.

Como os valores κ_i são valores limiares, uma interpretação intuitiva para as restrições de chance do Problema 4.2.1 é que : *a probabilidade de que o vetor aleatório $X_i \sim \mathbb{P}_i$ tenha seus valores fora da esfera $(\sqrt{R^2 + \xi_i}, \mathbf{c})$ é limitado por κ_i* . Equivalentemente, $\mathbb{P}_i(\|X_i - \mathbf{c}\|^2 \leq R^2 + \xi_i) \geq 1 - \kappa_i$ é a função de distribuição da variável aleatória⁵ $Z_i = \|X_i - \mathbf{c}\|^2$ no argumento $R^2 + \xi_i$, isto é, $F_{Z_i}(R^2 + \xi_i) = \mathbb{P}_i(\|X_i - \mathbf{c}\|^2 \leq R^2 + \xi_i)$. Então, o Problema 4.2.1 pode ser reescrito como:

$$\begin{aligned} & \min_{\mathbf{c} \in \mathbb{R}^D, R \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^N} R^2 + \lambda \sum_{i=1}^N \xi_i \\ & \text{subject to} \quad F_{Z_i}(R^2 + \xi_i) \geq 1 - \kappa_i, \quad i = 1, \dots, N \\ & \quad \xi_i \geq 0, \quad i = 1, \dots, N. \end{aligned}$$

Nessa formulação, os valores $1 - \kappa_i$ são *limitantes inferiores* para a função de distribuição F_{Z_i} . Então, todas as medidas $\mathbb{P}_i$, tal que $\xi_i = 0$ e $F_{Z_i}(R) = 1 - \kappa_i$, são *medidas de suporte* (em analogia com os vetores de suporte). Todas as medidas $\mathbb{P}_i$, tal que $\xi_i > 0$ e $F_{Z_i}(R + \xi_i) = 1 - \kappa_i$, são os erros permitidos no conjunto de treinamento. Todas as medidas $\mathbb{P}_i$, tal que $\xi_i = 0$ e $F_{Z_i}(R) > 1 - \kappa_i$, são pontos não críticos por que eles estão dentro do MV-set.

Note que o Problema 4.2.1 é equivalente ao SVDD (Tax e Duin, 2004) quando todas as medidas de probabilidade são as medidas Dirac, i.e., $\mathbb{P}_i = \delta_{\mathbf{x}_i}$, onde $\delta_{\mathbf{x}_i}(X_i) = 1$ se e somente se $X_i = \mathbf{x}_i$ e zero caso contrario (Berlinet e Thomas-Agnan, 2004). Assim, com probabilidade um, a única possível realização de $X_i \sim \delta_{\mathbf{x}_i}$ é $\mathbf{x}_i$. Isso permite eliminar as restrições probabilísticas e formular o problema como o usual SVDD, ou seja, nesse caso, procurar a solução no espaço de entrada é equivalente a procurar a solução no espaço das medidas de probabilidade Dirac.

4.2.1 Formulação usando a desigualdade de Markov

As restrições de chance do Problema 4.2.1 controlam a probabilidade de violação das restrições, permitindo flexibilidade no modelo. No entanto, cada restrição requer que cada possível realização de $X \sim \mathbb{P}_i$, seja levada em conta. Então, é necessário transformar esse problema em outro com restrições determinísticas. Isto pode ser conseguido usando a desigualdade de Markov (Boucheron *et al.*, 2013), a qual para, uma variável aleatória não negativa, $X \sim \mathbb{P}$, e para todo $t > 0$, limita $\mathbb{P}(X \geq t)$ mediante $\mathbb{E}_{\mathbb{P}}[X]/t$.

Cada restrição de chance do Problema 4.2.1 pode ser escrita de forma equivalente como $\mathbb{P}_i(\|X_i - \mathbf{c}\|^2 \geq R^2 + \xi_i) \leq \kappa_i$. Usando a notação $\mathbb{E}_{\mathbb{P}_i}$ para referirmos a $\mathbb{E}_{X \sim \mathbb{P}_i}$, e assumindo que cada medida

⁵A função de distribuição de uma variável aleatória X é a função F_X de $\mathbb{R}$ para $[0, 1]$ definida como $F_X(x) = \mathbb{P}(\{\omega : X(\omega) \leq x\}) \equiv \mathbb{P}(X \leq x)$

de probabilidade $\mathbb{P}_i$ possui esperança: $\boldsymbol{\mu}_i = \mathbb{E}_{\mathbb{P}_i}[X] \in \mathbb{R}^D$, e covariância: $\Sigma_i = \mathbb{E}_{\mathbb{P}_i}[(X - \mathbb{E}_{\mathbb{P}_i}[X])(X - \mathbb{E}_{\mathbb{P}_i}[X])^T] \in \mathbb{R}^{D \times D}$. Como as desigualdades $\|X_i - \mathbf{c}\|^2 \geq 0$ e $(R^2 + \xi_i) \geq 0$ são satisfeitas, a desigualdade de Markov limita cada restrição de chance por:

$$\mathbb{P}_i(\|X_i - \mathbf{c}\|^2 \geq R^2 + \xi_i) \leq \frac{\mathbb{E}_{\mathbb{P}_i}[\|X_i - \mathbf{c}\|^2]}{R^2 + \xi_i}, \quad i = 1, 2, \dots, N,$$

onde $\mathbb{E}_{\mathbb{P}_i}$ denota a esperança de uma variável aleatória distribuída de acordo a $\mathbb{P}_i$.

Lema 3. *Seja $\mathbb{P}$ uma medida de probabilidade com média $\boldsymbol{\mu}$ e matriz de covariância Σ , então para $X \sim \mathbb{P}$*

$$\mathbb{E}_{\mathbb{P}}[\|X - \mathbf{c}\|^2] = \text{tr}(\Sigma) + \|\boldsymbol{\mu} - \mathbf{c}\|^2.$$

Demonstração. A prova encontra-se no Apêndice. $\square$

Aplicando o Lema 3 e a desigualdade de Markov para as restrições de chance do problema 4.2.1 temos:

$$\mathbb{P}_i(\|X_i - \mathbf{c}\|^2 \geq R^2 + \xi_i) \leq \frac{\text{tr}(\Sigma_i) + \|\boldsymbol{\mu}_i - \mathbf{c}\|^2}{R^2 + \xi_i},$$

$\forall i = 1, 2, \dots, N$. Como κ_i é o limitante superior para a restrição de chance i , precisa-se garantir que

$$\frac{\text{tr}(\Sigma_i) + \|\boldsymbol{\mu}_i - \mathbf{c}\|^2}{R^2 + \xi_i} \leq \kappa_i. \quad (4.5)$$

Usando (4.5) e dado $\{\boldsymbol{\mu}_i, \Sigma_i\}_{i=1}^N \subset \mathbb{R}^D \times \mathbb{R}^{D \times D}$ estimado de $\{\mathbb{P}_i\}_{i=1}^N$ e $\{\kappa_i\}_{i=1}^N$, $\kappa_i \in (0, 1]$, a forma determinística do Problema 4.2.1 é o seguinte:

Problema 4.2.2.

$$\begin{aligned} \min_{\mathbf{c} \in \mathbb{R}^D, R \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^N} \quad & R^2 + \lambda \sum_{i=1}^N \xi_i \\ \text{subject to} \quad & \|\boldsymbol{\mu}_i - \mathbf{c}\|^2 \leq (R^2 + \xi_i)\kappa_i - \text{tr}(\Sigma_i), \quad i = 1, \dots, N \\ & \xi_i \geq 0 \quad i = 1, \dots, N, \end{aligned}$$

Problema 4.2.2 é chamado como *SMDD com restrições juntas* se $\kappa_i = \kappa$ para todo $i \in 1, 2, \dots, N$.

Lema 4. *Se $\text{tr}(\Sigma_i) = 0$, e $\kappa_i = 1$, $\forall i$ então, SMDD (Problem 4.2.2) é equivalente ao SVDD (Tax e Duin, 2004) com $\boldsymbol{\mu}_i$ no lugar de $\mathbf{x}_i$.*

Demonstração. Por hipótese, $\text{tr}(\Sigma_i) = 0$, substituindo $\kappa_i = 1, \forall i$ em Problema 4.2.2 obtemos o SVDD (Tax e Duin, 2004) com $\boldsymbol{\mu}_i$ no lugar de $\mathbf{x}_i$. $\square$

Formulação do problema dual

Sejam $\boldsymbol{\alpha}$ e $\boldsymbol{\beta}$ os vetores de multiplicadores de Lagrange com componentes não negativas α_i e β_i , $i = 1, 2, \dots, N$, respectivamente. A função de Lagrange para o Problema 4.2.2 é:

$$\begin{aligned} \mathcal{L}(R, \mathbf{c}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = & R^2 + \lambda \sum_{i=1}^N \xi_i - \sum_{i=1}^N \alpha_i \{(R^2 + \xi_i)\kappa_i \\ & - \|\boldsymbol{\mu}_i - \mathbf{c}\|^2 - \text{tr}(\Sigma_i)\} - \sum_{i=1}^N \beta_i \xi_i \end{aligned} \quad (4.6)$$

As condições Karush-Kuhn-Tucker (KKT) de estabilidade e de complementaridade para esse problema são respectivamente:

$$\left. \begin{aligned} \partial_R \mathcal{L} = 0 & : \quad \sum_{i=1}^N \alpha_i \kappa_i &= 1 \\ \nabla_{\mathbf{c}} \mathcal{L} = 0 & : \quad -2 \sum_{i=1}^N \alpha_i \boldsymbol{\mu}_i + 2 \sum_{i=1}^N \alpha_i \mathbf{c} &= 0 \\ \partial_{\xi_i} \mathcal{L} = 0 & : \quad \lambda - \alpha_i \kappa_i - \beta_i &= 0 \end{aligned} \right\} \quad (4.7)$$

$$\left. \begin{aligned} \alpha_i \{ (R^2 + \xi_i) \kappa_i - \|\boldsymbol{\mu}_i - \mathbf{c}\|^2 - \text{tr}(\Sigma_i) \} &= 0 \\ \beta_i \xi_i &= 0 \end{aligned} \right\} \quad (4.8)$$

Substituindo, as condições KKT em (4.6), obtemos o problema dual, i.e., dado $\{\boldsymbol{\mu}_i, \Sigma_i\}_{i=1}^N \subset \mathbb{R}^D \times \mathbb{R}^{D \times D}$ estimado de $\{\mathbb{P}_i\}_{i=1}^N$ e $\{\kappa_i\}_{i=1}^N$, $\kappa_i \in [0, 1]$, a forma dual do Problema 4.2.2 é

Problema 4.2.3.

$$\begin{aligned} \max_{\boldsymbol{\alpha} \in \mathbb{R}^N} \quad & \sum_{i=1}^N \alpha_i \langle \boldsymbol{\mu}_i, \boldsymbol{\mu}_i \rangle - \frac{\sum_{i,j=1}^N \alpha_i \alpha_j \langle \boldsymbol{\mu}_i, \boldsymbol{\mu}_j \rangle}{\sum_{i=1}^N \alpha_i} \\ & + \sum_{i=1}^N \alpha_i \text{tr}(\Sigma_i) \\ \text{subject to} \quad & 0 \leq \alpha_i \kappa_i \leq \lambda, \quad i = 1, \dots, N \\ & \sum_{i=1}^N \alpha_i \kappa_i = 1. \end{aligned}$$

Teorema de representação e análise das condições KKT

A partir das condições de estabilidade (4.7), o teorema de representação (Schölkopf *et al.*, 2001) para $\mathbf{c}$ é dado por:

$$\mathbf{c} = \frac{\sum_i \alpha_i \boldsymbol{\mu}_i}{\sum_i \alpha_i}, \quad i \in \{i \in \mathcal{I} \mid 0 < \alpha_i \kappa_i \leq \lambda\}, \quad (4.9)$$

onde $\mathcal{I} = \{1, 2, \dots, N\}$.

Analisando as condições de complementaridade (4.8) é possível identificar os seguintes casos para todo $i \in \mathcal{I}$. A Tabela (4.2) apresenta um resumo.

- $\alpha_i = 0, \beta_i > 0 \implies \xi_i = 0$, então $\|\boldsymbol{\mu}_i - \mathbf{c}\|^2 + \text{tr}(\Sigma_i) \leq R^2 \kappa_i$. Todas as realizações $\mathbf{x}'$ de $X_i \sim \mathbb{P}_i$ satisfazendo $\|\mathbf{x}' - \mathbf{c}\|^2 = (\|\boldsymbol{\mu}_i - \mathbf{c}\|^2 + \text{tr}(\Sigma_i)) / \kappa_i$, $\kappa_i \neq 0$ para $i \in \{i \in \mathcal{I} \mid \alpha_i = 0\}$ estarão *dentro* da esfera $(R, \mathbf{c})$ *sem importar o valor de κ_i* . Consequentemente, todas as medidas de probabilidade $\mathbb{P}_i$, $i \in \{i \in \mathcal{I} \mid \alpha_i = 0\}$ são consideradas descritas pela esfera $(R, \mathbf{c})$.
- $\alpha_i > 0, \beta_i = 0 \implies \xi_i > 0$, então $\|\boldsymbol{\mu}_i - \mathbf{c}\|^2 + \text{tr}(\Sigma_i) = (R^2 + \xi_i) \kappa_i$. Todas as realizações $\mathbf{x}'$ de $X_i \sim \mathbb{P}_i$ satisfazendo $\|\mathbf{x}' - \mathbf{c}\|^2 = (\|\boldsymbol{\mu}_i - \mathbf{c}\|^2 + \text{tr}(\Sigma_i)) / \kappa_i$, $\kappa_i \neq 0$ para $i \in \{i \in \mathcal{I} \mid \alpha_i \kappa_i = \lambda\}$ estarão *fora* da esfera $(R, \mathbf{c})$ com probabilidade κ_i . Consequentemente, todas as medidas de probabilidade $\mathbb{P}_i$, $i \in \{i \in \mathcal{I} \mid \alpha_i \kappa_i = \lambda\}$ são consideradas os erros permitidos no conjunto de treinamento $\{\mathbb{P}_i\}_{i=1}^N$.
- $\alpha_i > 0, \beta_i > 0 \implies \xi_i = 0$ e $0 < \alpha_i \kappa_i < \lambda$. Então usando (4.8) o raio pode ser calculado como

$$R^2 = \frac{\|\boldsymbol{\mu}_i - \mathbf{c}\|^2 + \text{tr}(\Sigma_i)}{\kappa_i}, \quad (4.10)$$

onde $i \in \{i \in \mathcal{I} \mid 0 < \alpha_i \kappa_i < \lambda\}$. Note que todas as realizações $\mathbf{x}'$ de $X_i \sim \mathbb{P}_i$ satisfazendo $\|\mathbf{x}' - \mathbf{c}\|^2 = (\|\boldsymbol{\mu}_i - \mathbf{c}\|^2 + \text{tr}(\Sigma_i))/\kappa_i$, $\kappa_i \neq 0$ para $i \in \{i \in \mathcal{I} \mid 0 < \alpha_i \kappa_i < \lambda\}$ estarão *na superfície* da esfera $(R, \mathbf{c})$. Consequentemente, todas as medidas de probabilidade $\mathbb{P}_i$, $i \in \{i \in \mathcal{I} \mid 0 < \alpha_i \kappa_i < \lambda\}$ são consideradas descritas pela esfera $(R, \mathbf{c})$ e são chamadas de *medidas de suporte*.

$\alpha_i = 0$	$\beta_i > 0$	$\xi_i = 0$ $\ \boldsymbol{\mu}_i - \mathbf{c}\ ^2 + \text{tr}(\Sigma_i) \leq R^2 \kappa_i$ $i \in \{i \in \mathcal{I} \mid \alpha_i = 0\}$
----------------	---------------	---

$\alpha_i > 0$	$\beta_i = 0$	$\xi_i > 0$ $\ \boldsymbol{\mu}_i - \mathbf{c}\ ^2 + \text{tr}(\Sigma_i) = (R^2 + \xi_i) \kappa_i$ $i \in \{i \in \mathcal{I} \mid \alpha_i \kappa_i = \lambda\}$
----------------	---------------	---

	$\beta_i > 0$	$\xi_i = 0$ $\ \boldsymbol{\mu}_i - \mathbf{c}\ ^2 + \text{tr}(\Sigma_i) = R^2 \kappa_i$ $i \in \{i \in \mathcal{I} \mid 0 < \alpha_i \kappa_i < \lambda\}$
--	---------------	---

Tabela 4.2: *Resumo da análise das condições KKT.*

O valor do raio dado por (A.1) pode ser obtido diretamente usando a informação dos multiplicadores do Lagrange da função de Lagrange do Problema 4.2.3, tal como descreve-se no seguinte resultado.

Teorema 5. *Seja η o multiplicador de Lagrange da restrição $\sum_{i=1}^N \alpha_i \kappa_i = 1$ da função de Lagrange do Problema 4.2.3, então $R = \sqrt{\eta}$.*

Demonstração. A prova encontra-se no Apêndice. □

Os modelos de otimização com restrições de chance em métodos de kernel tem sido previamente estudado para o caso de dados com incerteza, como por exemplo o caso de dados com ruído aditivo (Zhang, 2005), dados com valores em falta Shivaswamy *et al.* (2006), e dados intervalares (Ben-Tal *et al.*, 2011). Esses modelos assumem uma distribuição local a-priori para modelar as incertezas, e os modelos resultantes são formulados como problemas de otimização cônica de segunda ordem (Lobo *et al.*, 1998). Também é considerada uma aproximação de Taylor no EHKR, para modelar as incertezas nos dados de entrada (Yang e Gunn, 2007). A principal desvantagem desses modelos é no uso de kernels, onde é difícil fazer a kernelização do algoritmo. Outra desvantagem é que alguns desses modelos fazem suposições nos limites dos momentos das restrições de probabilidade ou no suporte das distribuições.

4.3 Descrição de dados com medidas de suporte em espaços de Hilbert com kernel reproduzível

Nesta seção apresentamos três formulações do SMDD no EHKR. Todas elas são formulações de esferas envolventes mínimas no EHKR, as quais correspondem às descrições não lineares do conjunto de treinamento $\{\mathbb{P}_i\}_{i=1}^N$. O primeiro modelo SMDD no EHKR é a kernelização do modelo SMDD apresentado na seção anterior. Este modelo é descrito na Seção 4.3.2. O segundo e o terceiro modelos SMDD no EHKR são extensões diretas do SVDD (Tax e Duin, 2004) para o caso de conjuntos de dados contendo medidas de probabilidade. A diferença entre o segundo e o terceiro modelo é que este último usa escalamento dos dados e kernels invariantes a translação. Aqueles modelos são descritos nas Seções 4.3.3 e 4.3.4, respectivamente. Começamos nossa discussão apresentando os principais fatos dos *mergulhos* das medidas de probabilidade nos EHKR.

4.3.1 Mergulho das medidas de probabilidade nos EHKR

O mergulho das medidas de probabilidade nos espaços de Hilbert (Berlinet e Thomas-Agnan, 2004; Guilbart, 1979; Smola *et al.*, 2007; Suquet *et al.*, 1995), fornece uma forma de representar as medidas de probabilidade $\mathbb{P}_i$ como funções num EHKR. Aquelas funções são comumente chamadas de *funções representantes* ou *funções médias*.

Definição 4.3.1 (função média). Seja $\mathbb{P}$ uma medida de probabilidade e $X \sim \mathbb{P}$. A função média em $\mathcal{H}$ é a função:

$$\begin{aligned} \mu_{\mathbb{P}} : \mathbb{R}^D &\rightarrow \mathbb{R} \\ \mathbf{t} &\mapsto \mu_{\mathbb{P}}(\mathbf{t}) = \mathbb{E}_{\mathbb{P}}[k(X, \mathbf{t})], \end{aligned} \quad (4.11)$$

onde $\mathbb{E}_{\mathbb{P}}[k(X, \mathbf{t})] = \int_{\mathbf{x} \in \mathbb{R}^D} k(\mathbf{x}, \mathbf{t}) d\mathbb{P}(\mathbf{x})$.

Assim, $\mu_{\mathbb{P}}$ é a função representante em $\mathcal{H}$ para $\mathbb{P}$. A condição suficiente para garantir a existência de $\mu_{\mathbb{P}}$ em $\mathcal{H}$ é dada assegurando que $\mu_{\mathbb{P}}(X) = \mathbb{E}_{\mathbb{P}}[k(X, X)] < \infty$, e $k(\cdot, \cdot)$ seja uma função mensurável (Gretton *et al.*, 2012; Smola *et al.*, 2007; Sriperumbudur *et al.*, 2010). Como consequência, a propriedade de reprodução $\langle f, \mu_{\mathbb{P}} \rangle = \langle f, \mathbb{E}_{\mathbb{P}}[k(X, \cdot)] \rangle = \mathbb{E}_{\mathbb{P}}[f(X)]$ é satisfeita para todo $f \in \mathcal{H}$.

O mergulho das medidas de probabilidade $\mathbb{P} \in \mathcal{P}$ para funções médias $\mu_{\mathbb{P}} \in \mathcal{H}$, é dado na seguinte definição.

Definição 4.3.2. O mergulho das medidas de probabilidade $\mathbb{P} \in \mathcal{P}$ em $\mathcal{H}$ é o mapeamento

$$\begin{aligned} \mu : \mathcal{P} &\rightarrow \mathcal{H} \\ \mathbb{P} &\mapsto \mu_{\mathbb{P}} = \mathbb{E}_{\mathbb{P}}[k(X, \cdot)] \end{aligned} \quad (4.12)$$

onde $\mathbb{E}_{\mathbb{P}}[k(X, \cdot)] = \int_{\mathbf{x} \in \mathbb{R}^D} k(\mathbf{x}, \cdot) d\mathbb{P}(\mathbf{x})$.

Escolhendo *kernels característicos* (Fukumizu *et al.*, 2008; Sriperumbudur *et al.*, 2008, 2010) para $k(\cdot, \cdot)$, o mergulho μ é injetivo, isto é, $\langle \mu_{\mathbb{P}}, f \rangle = \langle \mu_{\mathbb{Q}}, f \rangle$ para todo $f \in \mathcal{H}$ implica $\mathbb{P} = \mathbb{Q}$, ou equivalentemente, um kernel positivo definido é característico se $d(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}} = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}$. Note que $\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}$ induz uma métrica d em $\mathcal{P}$. Alguns exemplos de kernels característicos são o kernel gaussiano, Laplaciano, inverso multi-quadrático, kernels B_{2n+1} -splines,

etc (Sriperumbudur *et al.*, 2010). Além disso, um estimador empírico $\mu_{\mathbb{P}}$ a partir da amostra $\{x_i\}_{i=1}^M$ i.i.d. de $\mathbb{P}$ assegura uma boa aproximação para $\mu_{\mathbb{P}}$, i.e., o termo $\|\mu_{\mathbb{P}} - \mu_{emp}\|$, onde μ_{emp} é um estimador empírico para $\mu_{\mathbb{P}}$, é limitado (Smola *et al.*, 2007).

Kernel em medidas de probabilidade

O mapeamento

$$\begin{aligned} \mathcal{P} \times \mathcal{P} &\rightarrow \mathbb{R} \\ (\mathbb{P}, \mathbb{Q}) &\mapsto \langle \mathbb{P}, \mathbb{Q} \rangle_{\mathcal{P}} = \langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}}, \end{aligned} \quad (4.13)$$

define um produto interno em $\mathcal{P}$, onde (usando o Teorema de Fubini Ash e Doleans-Dade (2000)) $\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} = \int_{\mathbf{x} \in \mathbb{R}^D} \int_{\mathbf{x}' \in \mathbb{R}^D} k(\mathbf{x}, \mathbf{x}') d\mathbb{P}(\mathbf{x}) d\mathbb{Q}(\mathbf{x}')$. Consequentemente, o kernel de valor real em $\mathcal{P} \times \mathcal{P}$, definido por:

$$\begin{aligned} \tilde{k}(\mathbb{P}, \mathbb{Q}) &= \langle \mathbb{P}, \mathbb{Q} \rangle_{\mathcal{P}} = \langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \\ &= \int_{\mathbf{x} \in \mathbb{R}^D} \int_{\mathbf{x}' \in \mathbb{R}^D} k(\mathbf{x}, \mathbf{x}') d\mathbb{P}(\mathbf{x}) d\mathbb{Q}(\mathbf{x}'). \end{aligned} \quad (4.14)$$

é positivo definido (Berlinet e Thomas-Agnan, 2004). Note que, $\tilde{k}(\mathbb{P}, \mathbb{Q}) = \mathbb{E}_{\mathbb{P}}[\mathbb{E}_{\mathbb{Q}}[k(X, X')]]$, $X \sim \mathbb{P}$ e $X' \sim \mathbb{Q}$, por virtude da propriedade de reprodução.

O mergulho de medidas para espaços de Hilbert foram introduzidas em Guilbart (1979) e posteriormente em Berlinet e Thomas-Agnan (2004); Suquet *et al.* (1995), e por Smola *et al.* (2007) no caso de medidas de probabilidade. Algumas aplicações em aprendizagem computacional incluem redução de dimensionalidade (Fukumizu *et al.*, 2004), medida de independência de variáveis aleatórias (Gretton *et al.*, 2005a), teste estadístico de duas amostras (Gretton *et al.*, 2012), mergulhos de modelos ocultos de Markov em EHKR (Song *et al.*, 2010), regra de Bayes (Fukumizu *et al.*, 2010), máquinas de vetores de suporte (Muandet e Schoelkopf, 2013; Muandet *et al.*, 2012) entre outros (Sriperumbudur *et al.*, 2008, 2010; Zhang *et al.*). O kernel em medidas de probabilidade pode ser estimado usando (4.1) sem ajustar ou assumir algum modelo probabilístico para as observações. Existem outra classe de kernels em distribuições que assumem modelos probabilísticos para as observações, por exemplo, o kernel de Fisher (Jaakkola *et al.*, 1999), o kernel baseado na divergência Kullback-Leibler (KL) em distribuições (Moreno *et al.*, 2003), o kernel Bhattacharyya (Kondor e Jebara, 2003), e o kernel produto de probabilidades (Jebara e Kondor, 2003b).

4.3.2 SMDD no EHKR como problema com restrições de chance

Aplicando o mergulho das medidas de probabilidade nos espaços de Hilbert, nesta seção generaliza-se o modelo SMDD apresentado na Seção 4.2 para o caso do EHKR. Por exemplo, a desigualdade de Markov e os lemas da Seção 4.2 são estendidos para o caso das funções média e operadores de covariância no EHKR.

Usando o kernel positivo definido de valor real k , definido em $\mathbb{R}^D \times \mathbb{R}^D$, o conjunto $\hat{\mathcal{G}}$ de (4.4) é definido por:

$$\hat{\mathcal{G}}(\mathcal{K}) = \{\mathbb{P}_i \in \mathcal{P} \mid \mathbb{P}_i(\|k(X_i, \cdot) - c(\cdot)\|_{\mathcal{H}}^2 \leq R^2) \geq 1 - \kappa_i\},$$

onde $k(X, \cdot) \in \mathcal{H}$, $X \sim \mathbb{P}$, e o centro e raio da esfera são $c \in \mathcal{H}$, e $R \in \mathbb{R}$, respetivamente. As esferas

envolventes $(R, c(\cdot))$, são definidas agora no EHKR, as quais no espaço de entrada correspondem às descrições não lineares para $\{\mathbb{P}_i\}_{i=1}^N$. A esfera envolvente mínima é então formulada a seguir. Dado o conjunto de medidas de probabilidade $\{\mathbb{P}_i\}_{i=1}^N$ e os valores $\{\kappa_i\}_{i=1}^N$, $\kappa_i \in [0, 1]$, o SMDD no EHKR como um problema de otimização com restrições de chance é o seguinte:

$$\begin{aligned} \min_{c \in \mathcal{H}, R \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^N} \quad & R^2 + \lambda \sum_{i=1}^N \xi_i \\ \text{subject to} \quad & \mathbb{P}_i(\|k(X_i, \cdot) - c(\cdot)\|_{\mathcal{H}}^2 \leq R^2 + \xi_i) \geq 1 - \kappa_i, \quad i = 1, \dots, N \\ & \xi_i \geq 0 \quad i = 1, \dots, N, \end{aligned}$$

onde $\lambda > 0$ é um parâmetro de regularização, $\boldsymbol{\xi} = (\xi_1, \xi_1, \dots, \xi_N)^\top \in \mathbb{R}^N$ é o vetor de variáveis de folga. $\mathcal{H}$ é o EHKR com kernel k e norma $\|\cdot\|_{\mathcal{H}}$. A função $c \in \mathcal{H}$ é o centro da esfera. O valor $R \in \mathbb{R}$ é o raio e $X_i \sim \mathbb{P}_i$ é uma variável aleatório.

Solucionar esse problema com restrições de chance requer que as restrições satisfaçam todas as possíveis realizações de $X_i \sim \mathbb{P}_i$, o qual é difícil de calcular. Em lugar disso, é possível transformar esse problema num problema de otimização com restrições determinísticas, usando as funções médias e a desigualdade de Markov. Usando o mesmo argumento da Seção 4.2.1, a desigualdade de Markov também é satisfeita no EHKR:

$$\mathbb{P}_i(\|k(X_i, \cdot) - c(\cdot)\|_{\mathcal{H}}^2 \geq R^2 + \xi_i) \leq \frac{\mathbb{E}_{\mathbb{P}}[\|k(X_i, \cdot) - c(\cdot)\|_{\mathcal{H}}^2]}{R^2 + \xi_i},$$

para todo $i = 1, 2, \dots, N$.

Traço do operador de covariância

O termo $\mathbb{E}_{\mathbb{P}}[\|k(X_i, \cdot) - c(\cdot)\|_{\mathcal{H}}^2]$ pode ser computado usando o traço do operador de covariância em $\mathcal{H}$ e as funções médias $\mu_{\mathbb{P}}$. O operador de covariância em $\mathcal{H}$ com kernel k é o mapeamento $\Sigma^{\mathcal{H}} : \mathcal{H} \rightarrow \mathcal{H}$, tal que para todo $f, g \in \mathcal{H}$ satisfaz:

$$\langle f, \Sigma^{\mathcal{H}} g \rangle_{\mathcal{H}} = \mathbb{E}_{\mathbb{P}}[f(X)g(X)] - \mathbb{E}_{\mathbb{P}}[f(X)]\mathbb{E}_{\mathbb{P}}[g(X)],$$

devido à propriedade de reprodução⁶. O operador de covariância é então a matriz possivelmente de dimensão infinita (Fukumizu *et al.*, 2004; Gretton *et al.*, 2005b):

$$\Sigma^{\mathcal{H}} = \mathbb{E}_{\mathbb{P}}[k(X, \cdot)k(X, \cdot)^\top] - \mathbb{E}_{\mathbb{P}}[k(X, \cdot)]\mathbb{E}_{\mathbb{P}}[k(X, \cdot)]^\top. \quad (4.15)$$

⁶ $\Sigma^{\mathcal{H}}$ é um operador limitado em um espaço de Hilbert infinito e separável e pode ser representado por uma matriz infinita (Debnath e Mikusiński, 2005).

Assim, o traço de $\Sigma^{\mathcal{H}}$ pode ser obtido como: ⁷

$$\begin{aligned} \text{tr}(\Sigma^{\mathcal{H}}) &= \int_{t \in \mathbb{R}^D} \mathbb{E}_{\mathbb{P}}[k(X, t)k(X, t)^{\top}] \\ &\quad - \mathbb{E}_{\mathbb{P}}[k(X, t)]\mathbb{E}_{\mathbb{P}}[k(X, t)]^{\top} dt \\ &= \mathbb{E}_{\mathbb{P}}[\langle k(X, \cdot)k(X, \cdot) \rangle_{\mathcal{H}}] \\ &\quad - \langle \mathbb{E}_{\mathbb{P}}[k(X, \cdot)], \mathbb{E}_{\mathbb{P}}[k(X, \cdot)] \rangle_{\mathcal{H}} \\ &= \mathbb{E}_{\mathbb{P}}[k(X, X)] - \langle \mu_{\mathbb{P}}, \mu_{\mathbb{P}} \rangle_{\mathcal{H}}, \end{aligned}$$

onde a ultima linha é devida à propriedade de reprodução e Definição 4.3.1. Então, usando (4.14), temos:

$$\text{tr}(\Sigma^{\mathcal{H}}) = \mathbb{E}_{\mathbb{P}}[k(X, X)] - \tilde{k}(\mathbb{P}, \mathbb{P}), \quad (4.16)$$

ou seja, o traço de uma matriz possivelmente de dimensão infinita pode ser computado com avaliações de kernels. Da mesma forma que na Seção 4.2.1, o Lema 3 transforma-se no lema seguinte:

Lema 6.

$$\mathbb{E}_{\mathbb{P}}[\|k(X, \cdot) - c(\cdot)\|_{\mathcal{H}}^2] = \text{tr}(\Sigma^{\mathcal{H}}) + \|\mu_{\mathbb{P}} - c(\cdot)\|_{\mathcal{H}}^2.$$

Demonstração. A prova encontra-se no Apêndice. □

Forma determinística no EHKR

Usando o Lema 6, a desigualdade de Markov e uma análise similar da Seção 4.2.1, dadas as funções média $\{\mu_{\mathbb{P}_i}\}_{i=1}^N$ de $\{\mathbb{P}_i\}_{i=1}^N$ e $\{\kappa_i\}_{i=1}^N$, $\kappa_i \in (0, 1]$, a forma determinística para o SMDD no EHKR é dada por:

Problema 4.3.1.

$$\begin{aligned} \min_{c \in \mathcal{H}, R \in \mathbb{R}, \xi \in \mathbb{R}^N} \quad & R^2 + \lambda \sum_{i=1}^N \xi_i \\ \text{subject to} \quad & \|\mu_{\mathbb{P}_i} - c(\cdot)\|_{\mathcal{H}}^2 \leq (R^2 + \xi_i)\kappa_i - \text{tr}(\Sigma_i^{\mathcal{H}}), \quad i = 1, \dots, N \\ & \xi_i \geq 0 \quad i = 1, \dots, N, \end{aligned}$$

onde $\text{tr}(\Sigma_i^{\mathcal{H}})$ é dado por (4.16).

Formulação do problema dual 4.3.1

Sejam α e β dois vetores de multiplicadores de Lagrange com componentes não negativas α_i e β_i , $i = 1, 2, \dots, N$, respectivamente. A função de Lagrange para o Problema 4.3.1 é dada por:

$$\begin{aligned} \mathcal{L}(R, c(\cdot), \xi, \alpha, \beta) &= R^2 + \lambda \sum_{i=1}^N \xi_i - \sum_{i=1}^N \alpha_i \{(R^2 + \xi_i)\kappa_i \\ &\quad - \|\mu_{\mathbb{P}_i} - c(\cdot)\|_{\mathcal{H}}^2\} - \text{tr}(\Sigma_i^{\mathcal{H}})\} - \sum_{i=1}^N \beta_i \xi_i \end{aligned} \quad (4.17)$$

⁷Note que como $\mu_{\mathbb{P}}(X) < \infty$, segue que $\text{tr}(\Sigma^{\mathcal{H}}) < \infty$.

As condições KKT de estabilidade e de complementaridade para este problema são:

$$\left. \begin{aligned} \partial_R \mathcal{L} = 0 & : \quad \sum_{i=1}^N \alpha_i \kappa_i &= 1 \\ \nabla_{c(\cdot)} \mathcal{L} = 0 & : \quad -2 \sum_{i=1}^N \alpha_i \mu_{\mathbb{P}_i} + 2 \sum_{i=1}^N \alpha_i c(\cdot) &= 0 \\ \partial_{\xi_i} \mathcal{L} = 0 & : \quad \lambda - \alpha_i \kappa_i - \beta_i &= 0 \end{aligned} \right\} \quad (4.18)$$

$$\left. \begin{aligned} \alpha_i \{ (R^2 + \xi_i) \kappa_i - \|\mu_{\mathbb{P}_i} - c(\cdot)\|_{\mathcal{H}}^2 - \text{tr}(\Sigma_i^{\mathcal{H}}) \} &= 0 \\ \beta_i \xi_i &= 0 \end{aligned} \right\} \quad (4.19)$$

Substituindo (4.18) em (4.17), o problema dual é obtido como segue: dada as funções média $\{\mu_{\mathbb{P}_i}\}_{i=1}^N$ de $\{\mathbb{P}_i\}_{i=1}^N$ e $\{\kappa_i\}_{i=1}^N$, $\kappa_i \in [0, 1]$, a forma dual do Problema 4.3.1 é

Problema 4.3.2.

$$\begin{aligned} \max_{\alpha \in \mathbb{R}^N} \quad & \sum_{i=1}^N \alpha_i \langle \mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_i} \rangle_{\mathcal{H}} - \frac{\sum_{i,j=1}^N \alpha_i \alpha_j \langle \mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j} \rangle_{\mathcal{H}}}{\sum_{i=1}^N \alpha_i} \\ & + \sum_{i=1}^N \alpha_i \text{tr}(\Sigma_i^{\mathcal{H}}) \\ \text{subject to} \quad & 0 \leq \alpha_i \kappa_i \leq \lambda, \quad i = 1, \dots, N \\ & \sum_{i=1}^N \alpha_i \kappa_i = 1. \end{aligned}$$

Usando (4.14), $\tilde{k}(\mathbb{P}_i, \mathbb{P}_i) = \langle \mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_i} \rangle_{\mathcal{H}}$, então a função objetivo do Problema 4.3.2 torna-se:

$$\sum_{i=1}^N \alpha_i \tilde{k}(\mathbb{P}_i, \mathbb{P}_i) - \frac{\sum_{i,j=1}^N \alpha_i \alpha_j \tilde{k}(\mathbb{P}_i, \mathbb{P}_j)}{\sum_{i=1}^N \alpha_i} + \sum_{i=1}^N \alpha_i \text{tr}(\Sigma_i^{\mathcal{H}})$$

Teorema de representação e análise das condições KKT

A partir das condições KKT, o teorema de representação é

$$c(\cdot) = \frac{\sum_i \alpha_i \mu_{\mathbb{P}_i}}{\sum_i \alpha_i} = \frac{\sum_i \alpha_i \mathbb{E}_{\mathbb{P}_i}[k(X, \cdot)]}{\sum_i \alpha_i}, \quad (4.20)$$

para todo $i \in \{i \in \mathcal{I} \mid 0 < \alpha_i \kappa_i \leq \lambda\}$, onde $\mathcal{I} = \{1, 2, \dots, N\}$

Analisando as condições estacionárias (4.18) e de complementaridade (4.19) obtemos a mesma informação fornecida na Tabela 4.2, mas com $\mu_{\mathbb{P}_i} \in \mathcal{H}$ ao invés de $\mu_i \in \mathbb{R}^D$, $c \in \mathcal{H}$ ao invés de $c \in \mathbb{R}^D$ e $\Sigma_i^{\mathcal{H}}$ ao invés de Σ_i . Então temos que

- todas as $\mathbb{P}_i$, $i \in \{i \in \mathcal{I} \mid \alpha_i = 0\}$ são descritas por (R, c) , porque $\mu_{\mathbb{P}_i}$ estão dentro de (R, c)
- todas $\mathbb{P}_i$, $i \in \{i \in \mathcal{I} \mid \alpha_i \kappa_i = \lambda\}$ são consideradas os erros permitidos no conjunto de treinamento.
- todas $\mathbb{P}_i$, $i \in \{i \in \mathcal{I} \mid 0 < \alpha_i \kappa_i < \lambda\}$ são as *medidas de suporte*, a partir disso o raio é computado como:

$$R^2 = \frac{\|\mu_{\mathbb{P}_i} - c(\cdot)\|^2 + \text{tr}(\Sigma_i^{\mathcal{H}})}{\kappa_i}, \quad (4.21)$$

para todo $i \in \{i \in \mathcal{I} \mid 0 < \alpha_i \kappa_i < \lambda\}$.

Alternativamente, R pode ser calculado pelo Teorema (5) como $R = \sqrt{\eta}$, onde η é o multiplicador de Lagrange da restrição $\sum_i^N \alpha_i \kappa_i$ da função de Lagrange do Problema 4.3.2. Vale a pena ver que usando o kernel linear: $k(\mathbf{x}, \mathbf{x}') = \langle \mathbf{x}, \mathbf{x}' \rangle$ para implementar (4.14), o Problema 4.3.2 é equivalente ao Problema 4.2.3, porque $\tilde{k}(\mathbb{P}_i, \mathbb{P}_j) = \mathbb{E}_{\mathbb{P}_i}[\mathbb{E}_{\mathbb{P}_j}[\langle X, X' \rangle]] = \langle \boldsymbol{\mu}_i, \boldsymbol{\mu}_j \rangle$.

Relação dos valores κ

Como foi descrito na Seção 4.2, os valores κ_i incrementarão ou decrementarão o raio para cobrir realizações de $X \sim \mathbb{P}_i$. Este comportamento é ilustrado na Figura 4.3, a qual apresenta um conjunto de dados composto por dez distribuições Gaussianas, e seis esferas envoltivas mínimas para esses dados, obtidas solucionando o Problema 4.3.2, com seis diferentes conjuntos de valores para κ . Foi usado o kernel gaussiano $k(\mathbf{x}, \mathbf{x}') = \exp(-\gamma \|\mathbf{x} - \mathbf{x}'\|^2)$ e $\gamma = 1$ para implementar (4.14).

Na parte superior da Figura 4.3, de esquerda para a direita, são mostrados três MV-sets no espaço de entrada, os quais são três esferas envoltivas mínimas no EHKR, construídas solucionando três Problemas SMDD (Problema 4.3.2) com $\kappa_i = 0.8$, $\kappa_i = 0.9$ e $\kappa_i = 1.0$ para todas as restrições i dos três modelos, respectivamente.

Na parte inferior da Figura 4.3, da esquerda para a direita, são mostrados outros três MV-sets. Os três modelos SMDD tem $\kappa_i = 1$ em suas restrições, exceptuando-se $\kappa_1 = \kappa_2 = 0.8$ para o primeiro modelo, $\kappa_1 = \kappa_2 = 0.9$ para o segundo modelo, e $\kappa_1 = \kappa_2 = 1.0$ para o terceiro modelo. Dependendo de valores particulares para κ , as medidas de probabilidade $\mathbb{P}_1$ e $\mathbb{P}_2$, pertencem ao MV-set em algum grau.

Figura 4.3: Seis esferas envoltivas mínimas para um conjunto de dados cujas observações são distribuições Gaussianas, mostrando o efeito de escolher um valor em particular para κ nas restrições do modelo SMDD. Parte superior: os três modelos compartilham o mesmo valor κ , o qual é incrementado de esquerda para a direita. Parte inferior: os três modelos tem o valor $\kappa = 1$ para suas restrições, excepto as restrições $i = 1$ e $i = 2$, as quais são incrementadas de esquerda para a direita. Decrementar o valor κ para alguma restrição específica, implica que não todas as realizações de $X \sim \mathbb{P}_i$ estarão dentro da descrição.

4.3.3 SMDD no EHKR como extensão direta do SVDD

Diferentemente do modelo SMDD apresentado na seção anterior, que usa funções média e operadores de covariância, o modelo SMDD apresentado nesta seção somente usa funções média. Este modelo é uma extensão do SVDD para o caso de conjunto treinamento cujas observações são medidas de probabilidade. Usando as funções médias $\mu_{\mathbb{P}_i}$, uma versão empírica para G de (4.4) é dada por:

$$\hat{G}(R, c) = \{\mathbb{P}_i \in \mathcal{P} \mid \|\mu_{\mathbb{P}_i} - c(\cdot)\|_{\mathcal{H}}^2 \leq R^2\}, \quad (4.22)$$

então, o MV-set empírico $\hat{G}_\alpha^*$ é calculado como a esfera envolvente mínima para as funções média $\{\mu_{\mathbb{P}_i}\}_{i=1}^N$. Isso pode ser formulado a seguir. Dadas as funções médias $\{\mu_{\mathbb{P}_i}\}_{i=1}^N$ de $\{\mathbb{P}_i\}_{i=1}^N$, o modelo SMDD é:

$$\begin{aligned} \min_{c(\cdot) \in \mathcal{H}, R \in \mathbb{R}, \xi \in \mathbb{R}^N} \quad & R^2 + \lambda \sum_{i=1}^N \xi_i \\ \text{subject to} \quad & \|\mu_{\mathbb{P}_i} - c(\cdot)\|_{\mathcal{H}}^2 \leq R^2 + \xi_i, i = 1, \dots, N \\ & \xi_i \geq 0, i = 1, \dots, N. \end{aligned}$$

A função de Lagrange para o problema anterior é:

$$\begin{aligned} \mathcal{L}(R, c(\cdot), \xi, \alpha, \beta) = & R^2 + \lambda \sum_{i=1}^N \xi_i - \sum_{i=1}^N \alpha_i \{(R^2 + \xi_i) \\ & - \|\mu_{\mathbb{P}_i} - c(\cdot)\|_{\mathcal{H}}^2\} - \sum_{i=1}^N \beta_i \xi_i \end{aligned} \quad (4.23)$$

As condições KKT de otimalidade para esse problema são:

$$\left. \begin{aligned} \partial_R \mathcal{L} = 0 & : \sum_{i=1}^N \alpha_i &= 1 \\ \nabla_{c(\cdot)} \mathcal{L} = 0 & : -2 \sum_{i=1}^N \alpha_i \mu_{\mathbb{P}_i} + 2 \sum_{i=1}^N \alpha_i c(\cdot) &= 0 \\ \partial_{\xi_i} \mathcal{L} = 0 & : \lambda - \alpha_i - \beta_i &= 0 \end{aligned} \right\} \quad (4.24)$$

$$\left. \begin{aligned} \alpha_i \{(R^2 + \xi_i) - \|\mu_{\mathbb{P}_i} - c(\cdot)\|_{\mathcal{H}}^2\} &= 0 \\ \beta_i \xi_i &= 0 \end{aligned} \right\} \quad (4.25)$$

Formulação do problema dual

Dadas as funções média $\{\mu_{\mathbb{P}_i}\}_{i=1}^N$ de $\{\mathbb{P}_i\}_{i=1}^N$, a forma dual do problema anterior é dada substituindo (4.24) em (4.23) com segue:

Problema 4.3.3.

$$\begin{aligned} \max_{\alpha \in \mathbb{R}^N} \quad & \sum_{i=1}^N \alpha_i \tilde{k}(\mathbb{P}_i, \mathbb{P}_i) - \sum_{i,j=1}^N \alpha_i \alpha_j \tilde{k}(\mathbb{P}_i, \mathbb{P}_j) \\ \text{subject to} \quad & 0 \leq \alpha_i \leq \lambda, i = 1, \dots, N \\ & \sum_{i=1}^N \alpha_i = 1 \end{aligned}$$

onde foi usada (4.14) para substituir $\langle \mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j} \rangle$ pelo kernel $\tilde{k}(\mathbb{P}_i, \mathbb{P}_j)$. De (4.24) segue que o teorema

de representação é:

$$c(\cdot) = \sum_i \alpha_i \mu_{\mathbb{P}_i}, \quad i \in \{i \in \mathcal{I} \mid 0 < \alpha_i \leq \lambda\}.$$

Analisando (4.25) segue que todas as $\mathbb{P}_i$, $i \in \{i \in \mathcal{I} \mid \alpha_i = 0\}$ são medidas de probabilidade descritas por (R, c) , porque $\mu_{\mathbb{P}_i}$ estão dentro da esfera (R, c) . Todas as $\mathbb{P}_i$, $i \in \{i \in \mathcal{I} \mid \alpha_i = \lambda\}$ são os erros permitidos no conjunto de treinamento $\{\mathbb{P}_i\}_{i=1}^N$. Todas as $\mathbb{P}_i$, $i \in \{i \in \mathcal{I} \mid 0 < \alpha_i < \lambda\}$ são as medidas de suporte.

Se o kernel linear: $k(\mathbf{x}, \mathbf{x}') = \langle \mathbf{x}, \mathbf{x}' \rangle$ for usado em (4.14), então o Problema 4.3.3 é equivalente ao problema dual do SVDD (Tax e Duin, 2004), porque, $\tilde{k}(\mathbb{P}_i, \mathbb{P}_j) = \mathbb{E}_{\mathbb{P}_i}[\mathbb{E}_{\mathbb{P}_j}[\langle X, X' \rangle]]$ vai ser $\langle \mu_i, \mu_j \rangle$.

4.3.4 Modelo SMDD no EHKR usando funções com norma um e kernels invariantes à translação

Os kernels invariantes à translação ou kernels estacionários (Genton, 2002), são kernels da forma $k_I(\mathbf{x}, \mathbf{x}') = f(\mathbf{x} - \mathbf{x}')$, ou seja, tais kernels somente dependem da diferença $\mathbf{x} - \mathbf{x}'$. O kernel k_I é positivo definido se e somente se f pode ser escrito como:

$$f(\mathbf{x} - \mathbf{x}') = \int_{\mathbb{R}^D} \cos(\boldsymbol{\omega}^\top (\mathbf{x} - \mathbf{x}')) dF(\boldsymbol{\omega}),$$

onde F é uma medida positiva.

Os mapeamentos implícitos dos kernels invariantes à translação, são as funções $k_I(\mathbf{x}, \cdot)$ no EHKR que estão sob a superfície de uma esfera, pois eles tem norma constante. Isso pode ser evidenciado notando que os kernels invariantes a translação satisfazem:

$$k_I(\mathbf{x}, \mathbf{x}) = \langle k_I(\mathbf{x}, \cdot), k_I(\mathbf{x}, \cdot) \rangle_{\mathcal{H}} = \epsilon, \quad \forall \mathbf{x} \in \mathbb{R}^D$$

onde ϵ é um valor constante, então imediatamente segue que $\|k_I(\mathbf{x}, \cdot)\|_{\mathcal{H}} = \sqrt{|\epsilon|}$, isto é, as funções $k_I(\mathbf{x}, \cdot)$ estão sob a superfície de uma esfera de raio $\sqrt{|\epsilon|}$.

No entanto, as funções médias $\mu_{\mathbb{P}} = E_{\mathbb{P}}[k_I(X, \cdot)]$, não tem norma constante, porque

$$\|\mu_{\mathbb{P}}\|_{\mathcal{H}} = \|\mathbb{E}_{\mathbb{P}}[k_I(X, \cdot)]\|_{\mathcal{H}} \leq \mathbb{E}_{\mathbb{P}}[\|k_I(X, \cdot)\|_{\mathcal{H}}] = \sqrt{|\epsilon|},$$

pela convexidade de $\|\cdot\|_{\mathcal{H}}$ e a desigualdade de Jensen.

Uma possível solução para prevenir pequenos valores para o raio é escalar as funções média $\mu_{\mathbb{P}}$ para ter norma um e estarem na superfície de alguma esfera (Muandet e Schoelkopf, 2013).

Teorema 7 (Normalização esférica (Muandet e Schoelkopf, 2013)). *Se o kernel $k(\cdot, \cdot)$ é característico e se a amostra é linearmente independente em $\mathcal{H}$, então a normalização esférica:*

$$\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} = \frac{\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}}}{\sqrt{\langle \mu_{\mathbb{P}}, \mu_{\mathbb{P}} \rangle_{\mathcal{H}} \langle \mu_{\mathbb{Q}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}}}}, \quad (4.26)$$

preserva a propriedade injectiva do mapeamento $\mu : \mathcal{P} \rightarrow \mathcal{H}$.

Em outras palavras, o Teorema 7 diz que toda a informação da distribuição $\mathbb{P}$ é preservada após realizar uma normalização esférica nos dados.

Consequentemente, uma versão empírica para o conjunto $\hat{G}$ de (4.4) pode ser definida como:

$$\hat{G}(R, c) = \{\mathbb{P}_i \in \mathcal{P} \mid \|\mu_{\mathbb{P}_i} - c(\cdot)\|_{\mathcal{H}}^2 \leq R^2, \|\mu_{\mathbb{P}}\|_{\mathcal{H}}^2 = 1\}$$

então, o MV-set empírico $\hat{G}_\alpha^*$ é a esfera envolvente mínima das funções média $\{\mu_{\mathbb{P}_i}\}_{i=1}^N$ satisfazendo $\|\mu_{\mathbb{P}_i}\|_{\mathcal{H}}^2 = 1$, tal que o kernel usado em $\mathbb{E}_{\mathbb{P}_i}[k_I(X, \cdot)] = \mu_{\mathbb{P}_i}$ é invariante à translação.

O problema de otimização para esse modelo é similar ao do Problema 4.3.3:

$$\begin{aligned} \max_{\alpha \in \mathbb{R}^N} \quad & \sum_{i=1}^N \alpha_i \tilde{k}(\mathbb{P}_i, \mathbb{P}_i) - \sum_{i,j=1}^N \alpha_i \alpha_j \tilde{k}(\mathbb{P}_i, \mathbb{P}_j) \\ \text{subject to} \quad & 0 \leq \alpha_i \leq \lambda, \quad i = 1, \dots, N \\ & \sum_{i=1}^N \alpha_i = 1, \end{aligned}$$

mas com kernel:

$$\tilde{k}(\mathbb{P}_i, \mathbb{P}_j) = \frac{\tilde{k}(\mathbb{P}_i, \mathbb{P}_j)}{\sqrt{\tilde{k}(\mathbb{P}_i, \mathbb{P}_i) \tilde{k}(\mathbb{P}_j, \mathbb{P}_j)}}, \quad (4.27)$$

o qual é devido ao Teorema 7. Note que $\tilde{k}$ é o kernel segundo (4.14) mas com kernel k_I .

Como $\sum_{i=1}^N \alpha_i \tilde{k}(\mathbb{P}_i, \mathbb{P}_i)$ é constante, o problema anterior pode ser escrito como:

Problema 4.3.4.

$$\begin{aligned} \max_{\alpha \in \mathbb{R}^N} \quad & - \sum_{i,j=1}^N \alpha_i \alpha_j \tilde{k}(\mathbb{P}_i, \mathbb{P}_j) \\ \text{subject to} \quad & 0 \leq \alpha_i \leq \lambda, \quad i = 1, \dots, N \\ & \sum_{i=1}^N \alpha_i = 1. \end{aligned}$$

Essa formulação é similar à formulação dual das Máquinas de Vetores de Suporte de uma classe (OCSVM) (Schölkopf *et al.*, 2001), porém não é diretamente equivalente. Esse ponto será discutido na próxima seção.

4.4 Relação entre os modelos SMDD

Nesta seção descreve-se a relação entre os modelos SMDD. Também discutiremos a equivalência entre os modelos SMDD e as OCSVM (Muandet e Schoelkopf, 2013; Schölkopf *et al.*, 2001). Nesta seção, denotamos o modelo SMDD apresentado na Seção 4.3.2 como M1 (Problemas 4.3.1 e 4.3.2), o SMDD da Seção 4.3.3 como M2 (Problema 4.3.3), e o SMDD da Seção 4.3.4 como M3 (Problema 4.3.4).

Começamos mostrando como M1 pode ser formulado se o restringirmos somente ao caso de restrições conjuntas e uma matriz de covariância compartilhada. Depois, usamos essa formulação para comparar ela com os modelos M1 e M2.

Teorema 8. *A forma primal de M1 (Problema 4.3.1) com restrições conjuntas compartilhando a*

mesma matriz de covariância, i.e, $\kappa_i = \kappa$ e $\Sigma_i^{\mathcal{H}} = \Sigma^{\mathcal{H}}$ para todo $i = 1, 2, \dots, N$ e $\lambda > 0$, pode ser escrita como

Problema 4.4.1.

$$\begin{aligned} \min_{c(\cdot) \in \mathcal{H}, \rho' \in \mathbb{R}, \xi' \in \mathbb{R}^N} \quad & \frac{\|c(\cdot)\|_{\mathcal{H}}^2}{2} - \rho' + \lambda \sum_{i=1}^N \xi'_i \\ \text{subject to} \quad & \langle \mu_{\mathbb{P}_i}, c(\cdot) \rangle_{\mathcal{H}} \geq \rho' - \xi'_i, \quad i = 1, \dots, N \\ & \xi'_i \geq -\frac{\|\mu_{\mathbb{P}_i}\|_{\mathcal{H}}^2}{2}, \quad i = 1, \dots, N. \end{aligned}$$

onde

$$\xi'_i = \frac{1}{2} \kappa \xi_i - \frac{\|\mu_{\mathbb{P}_i}\|_{\mathcal{H}}^2}{2} \quad (4.28)$$

Demonstração. A prova encontra-se no Apêndice. $\square$

Note que o Problema 4.4.1 é uma formulação menos flexível para M1 (Problema 4.3.1), porque essa formulação considera a mesma matriz de covariância local e os mesmos valores κ para todos os pontos. É fácil verificar que, para valores ótimos para $c \in \mathcal{H}$ e ρ' do Problema 4.4.1.⁸, é possível recuperar o raio como:

$$R = \sqrt{(tr(\Sigma) + \|c\|^2 - 2\rho')/\kappa}, \quad (4.29)$$

ou equivalentemente, solucionando o Problema 4.3.1 para $\kappa_i = \kappa$ e $\Sigma_i = \Sigma$ para todo $i = 1, 2, \dots, N$, é possível recuperar ρ' do Problema 4.4.1 como:

$$\rho' = -\frac{1}{2}(R^2\kappa - tr(\Sigma) - \|c\|^2).$$

Teorema 9. Usando o kernel definido para medidas de probabilidade segundo (4.14), o problema dual do Problema 4.4.1 é:

Problema 4.4.2.

$$\begin{aligned} \max_{\alpha \in \mathbb{R}^N} \quad & \frac{1}{2} \sum_{i=1}^N \alpha_i \tilde{k}(\mathbb{P}_i, \mathbb{P}_i) - \frac{1}{2} \sum_{i,j=1}^N \alpha_i \alpha_j \tilde{k}(\mathbb{P}_i, \mathbb{P}_j) \\ \text{subject to} \quad & 0 \leq \alpha_i \leq \lambda, \quad i = 1, \dots, N \\ & \sum_{i=1}^N \alpha_i = 1. \end{aligned}$$

Demonstração. A prova encontra-se no Apêndice. $\square$

Lema 10. Seja η o multiplicador de Lagrange da restrição $\sum_{i=1}^N \alpha_i = 1$ da função de Lagrange do Problema 4.4.2, então $\rho = \eta$.

A partir disso, é possível solucionar o Problema 4.4.2 e aplicar o lema 10 para recuperar ρ , o centro via o teorema de representação dado por (4.24) como $c = \sum_i \alpha_i \mu_{\mathbb{P}_i}$, $i \in \{i | 0 < \alpha_i \leq \lambda\}$, e o raio R usando (4.29).

⁸Veja a demonstração do Teorema 4.4.1 no apêndice.

4.4.1 Equivalência entre os modelos SMDD

M1 vs M2

O Problema 4.4.2 é o dual de M1 com restrições conjuntas e compartilhando a mesma matriz de covariância. Neste cenário, M1 é similar à M2 mas com a diferença num fator de escala de 0.5 na função objetivo do problema dual.

M1 vs M3

Aplicando a normalização esférica nos dados, a função objetivo do Problema 4.4.2, torna-se

$$-0.5 \sum_{i,j=1}^N \alpha_i \alpha_j \tilde{k}(\mathbb{P}_i, \mathbb{P}_j),$$

onde $\tilde{k}$ é o kernel dado por (4.27), porque o outro termo na função objetivo é constante. Em consequência, M1 com restrições conjuntas compartilhando a mesma matriz de covariância e normalização esférica nos dados é equivalente ao modelo M3, com a diferença de um fator de escala de 0.5 na função objetivo do problema dual.

4.4.2 Conexão com as Máquinas de vetores de suporte de uma classe

O método SVDD (Tax e Duin, 2004) e as máquinas de vetores de suporte de uma classe (OCSVM) (Schölkopf *et al.*, 2001), são equivalentes se kernels invariantes à translação são usados nesses modelos (Schölkopf *et al.*, 2001; Tax e Duin, 2004). Embora, o Problema 4.4.1 seja muito similar às OCSVM com medidas de probabilidade, chamadas também de máquinas de medidas de suporte de uma classe (OCSMM) (Muandet e Schoelkopf, 2013; Schölkopf *et al.*, 2001), SMDD não é diretamente equivalente com a OCSVM, pois, mesmo usando um kernel invariante à translação, as normas das funções médias não seriam constantes. No entanto, se fosse realizada normalização esférica nos dados, temos a seguinte equivalência:

Corolário 11. *M2, M3 e OCSMM (Muandet e Schoelkopf, 2013) são equivalentes se é feita normalização esférica no conjunto de treinamento $\{\mathbb{P}_i\}_{i=1}^N$ via (7).*

Demonstração. A prova encontra-se no Apêndice. □

Então é possível deduzir que M1, nesse cenário dado pelo Problema 4.4.2, é equivalente ao OCSMM, com uma diferença no fator de escala de 0.5 na função objetivo do problema dual.

4.5 Experimentos

Nesta seção, testamos os modelos apresentados neste capítulo na tarefa de detecção de anomalias grupais (Muandet e Schoelkopf, 2013; Xiong *et al.*, 2011a,b). Especificamente os modelos SMDD foram testados no caso de detecção de anomalias grupais pontuais na Seção 4.5.1 e, para o caso de detecção de anomalias grupais de distribuição na Seção 4.5.2. Finalmente, na Seção 4.5.3, usam-se dados reais do projeto *Sloan Digital Sky Survey* (SDSS), para detectar grupos anômalos de galáxias. Aqui, usa-se a mesma notação para os modelos SMDD como foi feito na Seção 4.4, i.e., M1 é o SMDD

definido pelo Problema 4.3.2, M2 pelo Problema 4.3.3, e M3 pelo Problema 4.3.4. Também, para fins de comparação, foram usados o OCSMM (Muandet e Schoelkopf, 2013), e o SVDD (Tax e Duin, 2004). As notações para os modelos são mostradas na Tabela 4.3.

Estimação do kernel em medidas de probabilidade e da covariância. O kernel em medidas de probabilidades definido por (4.14) foi estimado usando o estimador empírico:

$$\tilde{k}(\mathbb{P}_i, \mathbb{P}_j) \approx \frac{1}{L_i L_j} \sum_{l=1}^{L_i} \sum_{l'=1}^{L_j} k(\mathbf{x}_l^{(i)}, \mathbf{x}_{l'}^{(j)}), \quad (4.30)$$

no conjunto de treinamento dado por (4.1). Também, o traço do operador de covariância no EHKR dado por (4.16) foi estimado por:

$$\begin{aligned} \text{tr}(\Sigma_i^{\mathcal{H}}) &\approx \frac{1}{L_i - 1} \sum_{l=1}^{L_i} k(\mathbf{x}_l^{(i)}, \mathbf{x}_l^{(i)}) \\ &\quad - \frac{1}{L_i(L_i - 1)} \sum_{l=1}^{L_i} \sum_{l=1}^{L_i} k(\mathbf{x}_l^{(i)}, \mathbf{x}_l^{(i)}). \end{aligned} \quad (4.31)$$

onde k é um kernel positivo definido em $\mathbb{R}^D \times \mathbb{R}^D$.

Modelo	Problema	Seção/Referencia.
M1	4.3.2	4.3.2
M2	4.3.3	4.3.3
M3	4.3.4	4.3.4
OCSMM	-	(Muandet e Schoelkopf, 2013)
SVDD	-	(Tax e Duin, 2004)

Tabela 4.3: Notação dos modelos usados no experimento

Foi usado o pacote de otimização convexa *CVX* (Grant e Boyd, 2014, 2008), para solucionar M1. Para solucionar M2, M3, OCSMM e SVDD foi usado o *SVM and Kernel Methods Matlab Toolbox* (SVM-KM) (Canu et al., 2005).⁹

Nos experimentos, para cada observação s_i de (4.1) foram usados os termos: *grupo* ou conjunto de pontos, indistintamente, i.e., diremos que os conjuntos de treinamento dados por (4.1) tem N grupos ou conjuntos de pontos. Como SVDD não foi originalmente projetado para usar medidas de probabilidade, ele foi treinado usando somente as médias grupais empíricas.

4.5.1 Detecção de anomalias grupais pontuais num conjunto de dados cujas observações são distribuições de misturas gaussianas

A meta da detecção de anomalias grupais é encontrar grupos de pontos com comportamento não usual de conjuntos de dados segundo (4.1). Diferentemente dos métodos de detecção de anoma-

⁹Os arquivos fonte e os conjuntos de dados usados para os experimentos estão disponíveis em <http://www.vision.ime.usp.br/~jorjasso/SMDD.html>.

lias usuais, os grupos anômalos podem estar altamente misturados com os grupos não anômalos, tornando a tarefa de detecção de anomalias grupais um problema difícil.

Em detecção de *anomalias grupais pontuais* (Xiong *et al.*, 2011a), os grupos anômalos estão dados pela agregação individual de pontos anômalos. Para este experimento, foram gerados 50 grupos não anômalos de pontos e 30 grupos de pontos usados como teste, dos quais 20 grupos correspondem a anomalias grupais e os 10 restantes a grupos não anômalos. O procedimento é descrito no paragrafo seguinte. O número de pontos por grupo, para todos os grupos anômalos e não anômalos, foi aleatoriamente escolhido a partir de uma distribuição de Poisson com parâmetro $\beta = 100$.

Os pontos dos grupos não anômalos correspondem a uma amostragem aleatória a partir de uma *distribuição de misturas Gaussianas* ou GMD. Foram considerados *dois tipos* de grupos não anômalos, segundo o procedimento descrito em Muandet e Schoelkopf (2013); Xiong *et al.* (2011b): o primeiro tipo corresponde a grupos de pontos gerados a partir de uma GMD de duas dimensões e com três componentes, com os seguintes parâmetros: pesos das misturas: (0.33, 0.64, 0.03), médias: $(-1.7, -1)$, $(1.7, -1)$, $(0, 2)$, e $0.2 * I_2$ como matriz de covariância compartilhada, onde I_2 denota a matriz identidade de tamanho 2×2 . O segundo tipo corresponde a grupos de pontos gerados a partir de uma GMD com iguais parâmetros mas com pesos das misturas: (0.33, 0.03, 0.64). A probabilidade de escolha de cada tipo de grupo foi: $\pi = (0.48, 0.52)$. A moldura verde na Figura 4.4 apresenta três grupos não anômalos para $\pi = 0.48$ e a moldura amarela apresenta dois grupos não anômalos para $\pi = 0.52$.

Três diferentes tipos de grupos anômalos foram gerados. O primeiro deles foi dado por 10 grupos de pontos correspondentes a uma amostragem aleatória da distribuição normal: $\mathcal{N}((-0.4, 1), I_2)$. A moldura magenta na Figura 4.4 apresenta cinco grupos anômalos deste tipo. O segundo tipo foi dado por 5 grupos de pontos correspondentes a uma amostragem aleatória da GMD com quatro componentes, pesos das misturas: (0.1, 0.08, 0.07, 0.75), médias: $(-1.7, -1)$, $(1.7, -1)$, $(0, 2)$, $(0.6, -1)$, e $0.2 * I_2$ como matriz de covariância compartilhada. A moldura azul na Figura 4.4 apresenta cinco grupos anômalos de este tipo. O terceiro tipo de anomalias grupais foi dada por 5 grupos de pontos amostrados de uma GMD com quatro componentes, pesos: (0.14, 0.1, 0.28, 0.48), médias: $(-1.7, -1)$, $(1.7, -1)$, $(0, 2)$, $(-0.5, 1)$, e $0.2 * I_2$ como matriz de covariância compartilhada. A moldura vermelha da Figura 4.4 apresenta cinco grupos anômalos desse tipo.

Para obter estatísticas confiáveis, foram realizadas 200 execuções, sobre o conjunto de treinamento dado por os 50 grupos não anômalos. Os restantes 10 grupos de pontos do conjunto de teste foram gerados usando o mesmo procedimento usado para gerar os grupos do conjunto de treinamento. As métricas de performance consideradas foram: a área abaixo da curva ROC (AUC), e a acurácia (ACC).

Usualmente em problemas de classificação de uma classe, os dados de treinamento e os dados de teste não têm etiquetas, tornado impossível ter um conjunto de validação para realizar a seleção do modelo, ou seja, ter um conjunto de validação para otimizar os parâmetros dos modelos: o parâmetro de regularização λ e os parâmetros do kernel. Além disso, como o problema de descrição de dados é um problema não trivial, dado que podem existir muitos modelos descrevendo o conjunto de dados. Neste trabalho, considera-se a mesma metodologia usada na literatura, isto é, escolher-se arbitrariamente um valor para o parâmetro de regularização λ do modelo SMDD e, calcular-se os parâmetros do kernel usando alguma heurística nos dados disponíveis, evitando usar assim

Figura 4.4: Conjunto de dados para a detecção de anomalias grupais. As caixas verde e amarela contêm grupos de pontos não anômalos. As caixas vermelha, azul e magenta contêm grupos de pontos anômalos.

o conjunto de treinamento ou de teste para fazer a seleção de modelo (Capítulo 7, pp. 215-219 Scholkopf (2001) e Seção 5 de Muandet e Schoelkopf (2013)).

Foi considerado um parâmetro de regularização $\lambda = 1$, e um kernel em medidas de probabilidade (4.14) implementado pelo kernel gaussiano com parâmetro γ calculado como:

$$\gamma = 1/s(\|\mathbf{x}_k^{(i)} - \mathbf{x}_l^{(j)}\|^2), \quad (4.32)$$

onde i, j são os índices dos grupos, k, l são os índices dos pontos nos grupos i, j , respectivamente e s é o 0.1 quantil da distância euclidiana entre todos os possíveis pares de pontos no conjunto de dados.

As Figuras 4.5a, 4.5b, 4.5c apresentam os resultados em *boxplots* para este experimento. Os *boxplots* (Tukey, 1977) apresentam o AUC, o ACC para os grupos não anômalos, e o ACC para os grupos anômalos respectivamente, usando a informação das 200 execuções do experimento. Este experimento apresenta como os modelos SMDD (M1, M2, e M3) e o OCSMM detectam aquelas anomalias muito bem. Os valores AUC perto do valor um, daqueles modelos, indicam que os modelos SMDD detectam anomalias grupais com poucos falsos positivos e falsos negativos. Por outro lado, o SVDD não pode detectar aquelas anomalias grupais usando somente as médias grupais como conjunto de treinamento.

Para ver porque a detecção de anomalias grupais é um problema difícil, o gráfico das médias de todos os grupos anômalos e não anômalos é apresentado na Figura 4.5d. Os pontos verdes são as médias dos grupos não anômalos. Os pontos vermelho, azul e magenta são as médias dos grupos anômalos correspondendo as molduras vermelha, azul e magenta da Figura 4.4. Porque as médias dos grupos não anômalos estão em sobreposição com as médias dos grupos anômalos, métodos como as OCSVM e SVDD não terão boa performance, porque eles consideram anomalias aos pontos longe da média da descrição de dados.

Figura 4.5: Resultados experimentais e o gráfico das médias grupais para o experimento de detecção de anomalias grupais. Os boxplots apresentam as estatísticas ACC e AUC. As marcas no eixo x de cada boxplot são os modelos de descrição de dados e no eixo y é a medida de performance. A Figura 4.5d apresenta as médias dos grupos não anômalos vs as médias dos grupos anômalos. Os pontos vermelhos, magenta e azuis são as médias dos grupos anômalos. Os pontos verdes são as médias dos grupos não anômalos.

4.5.2 Detecção de anomalias grupais de distribuição num conjunto de dados cujas observações são distribuições de misturas gaussianas

As anomalias grupais de distribuição (Xiong *et al.*, 2011a) são formadas por grupos de pontos que individualmente são não anômalos mas que juntos formam grupos anômalos. Para este experimento, foram gerados 50 grupos não anômalos de pontos para formar o conjunto de treinamento e 15 grupos de pontos anômalos mais 15 grupos de pontos não anômalos para formar o conjunto de teste. O número de pontos por grupo foi estabelecido como o experimento anterior.

Os pontos em cada grupo não anômalo foram amostrados de uma distribuição GMD de duas dimensões e com três componentes. Os parâmetros foram: pesos das misturas: $p = \{1/3, 1/3, 1/3\}$, médias: $(-1.7, 1)$, $(1.7, -1)$, $(0, 2)$, e compartilhando a mesma matriz de covariância: $0.2 * I_2$.

Para gerar os grupos anômalos, primeiramente foi gerado um conjunto de pontos a partir da mesma GMD dos grupos não anômalos, logo, cada ponto gerado $\mathbf{x}_i$ foi rotacionado 45 graus, isto é $\mathbf{x}_i^R = \mathbf{x}_i^\top R$, onde R é uma matriz de rotação de 45 graus. Posteriormente foi estimada a matriz de covariância empírica desses pontos rotacionados. Finalmente os pontos dos grupos anômalos foi gerado com a mesma GMD dos grupos não anômalos, mais com duas de suas matrizes de covariância dadas pelas matrizes de covariância dos pontos rotacionados. Isto é, individualmente os pontos são relativamente não anômalos, mas juntos, como grupo são anômalos.

Como no experimento anterior, também foram usadas as métricas AUC e ACC como medida de performance, o parâmetro de regularização foi $\lambda = 1$, e o kernel entre medidas de probabilidade (4.14) foi implementado com o kernel gaussiano com o seu parâmetro γ computado usando (4.32), mais com s dado pela mediana da distância euclidiana entre todos os possíveis pontos do conjunto de dados.

A Figura 4.6d apresenta as médias grupais. Os pontos verdes indicam as médias dos grupos não anômalos e os pontos vermelhos as médias dos grupos anômalos. Como experimento anterior, encontrar as anomalias grupais é difícil, dado que existe uma sobreposição entres as médias grupais dos grupos não anômalos e as médias grupais dos grupos anômalos.

As Figuras 4.6a, 4.6b e 4.6c apresentam os resultados usando boxplots para mostrar as estatísticas (usando 200 execuções do experimento) do AUC, do ACC para os grupos não anômalos e do ACC para os grupos anômalos respectivamente. Também, para este tipo de anomalias grupais todos os modelos SMDD tiveram um bom desempenho. Como esperado, o desempenho de SVDD foi o pior de todos devido à sobreposição das médias grupais.

Finalmente, para o mesmo conjunto de treinamento, foram consideradas outras anomalias grupais de distribuição, foram geradas dez anomalias grupais a partir de uma GMD com os mesmos parâmetros do conjunto de treinamento, mas com pesos das misturas: $p = \{0.85, 0.08, 0.07\}$ e outras dez anomalias grupais a partir da mesma GMD mas com pesos das misturas: $p = \{0.04, 0.48, 0.48\}$. A Figura 4.6h apresenta as médias dos grupos anômalos (vermelho) e as dos grupos não anômalos (verde). Para esse cenário, é possível notar que os métodos convencionais como a SVDD e a SVM tem bom desempenho baseado somente na informação das médias grupais. As Figuras 4.6e, 4.6f e 4.6g apresentam as estatísticas para o AUC, o ACC para os grupos não anômalos, e o ACC para os grupos anômalos respectivamente. O desempenho de M1 é afectado pois esse modelo usa informação do primeiro e segundo momento para um problema que é solucionado facilmente usando a informação das médias grupais. No entanto, levando em conta a dimensionalidade dos dados, é muito difícil conhecer antecipadamente essa informação.

(a) *AUC.* (b) *ACC Non A.* (c) *ACC Anom.* (d) *Médias grupais* (e) *AUC.* (f) *ACC Non A.* (g) *ACC Anom.* (h) *Médias grupais*

Figura 4.6: Resultados experimentais e o gráfico das médias grupais para os dois experimentos na detecção de anomalias grupais.

n_i galáxias das 7530 galáxias, onde n_i é distribuído de acordo à distribuição Poisson com parâmetro $\beta = 15$. Como as galáxias foram escolhidas aleatoriamente, a agregação delas é um grupo anômalo.

Os grupos anômalos para o segundo, terceiro, quarto e quinto conjunto de teste foram gerados como segue: primeiro, a matriz de covariância das 7530 observações (galáxias) foi empiricamente estimada. Logo, foram selecionadas de forma aleatória três conjuntos de galáxias das 7530 galáxias, cada um dos quais continham n_i galáxias, onde n_i é distribuída de acordo a uma distribuição Poisson com parâmetro $\beta = 15$. Posteriormente, foram computadas as médias empíricas dos três conjuntos.

Usando as três médias e a matriz de covariância empíricas, a GMD com três componentes e pesos das misturas: $p = \{1/3, 1/3, 1/3\}$ foi construída.

Os grupos anômalos para o segundo conjunto de teste foram gerados com essa GMD, com n_i pontos por grupo, onde n_i está distribuída de acordo a uma distribuição Poisson parâmetro $\beta = 15$. Os grupos anômalos de pontos para o terceiro, quarto e quinto conjuntos de teste foram gerados como o segundo conjunto de teste, mas usando as matrizes de covariância $5 * \Sigma$, $10 * \Sigma$, e $100 * \Sigma$, respectivamente.

As Figuras 4.8d, 4.8h, 4.9d, 4.9h, e 4.10d apresentam as médias grupais dos vetores PCA. Os pontos verdes são as médias dos grupos não anômalos e os pontos vermelhos são as médias dos grupos anômalos. Cada figura apresenta quatro sub gráficos. Acima-esquerda: a primeira e a segunda componentes do vetor PCA, acima-direita: a segunda e a terceira componentes, abaixo-esquerda: a terceira e a quarta componentes, abaixo-direita: a quarta e a primeira componentes. As anomalias grupais para este experimento são difíceis de detectar devido à sobreposição dos dados.

Para todos os modelos SMDD, o kernel entre medidas de probabilidade foi implementado com o kernel gaussiano. As estatísticas: AUC e ACC para os grupos anômalos e não anômalos foram obtidas realizando 200 execuções do algoritmo para cada conjunto de teste. As Figuras 4.8a, 4.8b, e 4.8c apresentam o AUC, o ACC para os grupos não anômalos e o ACC para os grupos anômalos para o primeiro conjunto de teste. O parâmetro do kernel foi obtido usando (4.32) mas com s sendo a mediana. Foi considerado um parâmetro de regularização permitindo aproximadamente 30% dos grupos não anômalos serem os erros permitidos no conjunto de treinamento. A performance dos modelos M2 e M3 foi um pouco inferior detectando anomalias que o SVDD para essa escolha de parâmetros. No entanto, a métrica AUC para SVDD mostra que a performance para esse modelo não é melhor que o chance. Por outro lado, o desempenho de M1 e OCSMM é melhor que a linha base e ambos se comportam similarmente detectando as anomalias. Note que o ACC para os grupos é aproximadamente de 70% pela escolha do parâmetro de regularização. O gráfico das médias grupais na Figura 4.8d apresenta a dificuldade do problema.

As Figuras 4.8e, 4.8f, e 4.8g apresentam o AUC, o ACC para os grupos não anômalos e o ACC para os grupos anômalos no segundo conjunto de teste. O parâmetro do kernel gaussiano foi computado usando (4.32). Foi considerado um parâmetro de regularização permitindo aproximadamente 20% dos grupos não anômalos serem os erros permitidos no conjunto de treinamento. A métrica AUC mostra que o desempenho de SVDD é o pior entre os outros modelos, e que a normalização esférica dos dados incrementa o desempenho, tal como pode ser visualizado no valor AUC do modelo M3. Por outro lado, o ACC dos grupos não anômalos é aproximadamente 80% pela escolha de λ . O gráfico das médias grupais é apresentado na Figura 4.8h.

As Figuras 4.9a, 4.9b, e 4.9c apresentam o AUC, o ACC para os grupos não anômalos e o ACC para os grupos anômalos no terceiro conjunto de teste. O parâmetro do kernel gaussiano

(a) *AUC.* (b) *ACC Non A.* (c) *ACC Anom.* (d) *Médias grupais.* (e) *AUC.* (f) *ACC Non A.* (g) *ACC Anom.* (h) *Médias grupais.*

Figura 4.8: Resultados experimentais e o gráfico das médias grupais para o experimento de detecção de anomalias grupais no conjunto de dados SDSS III dataset.

(a) *AUC.* (b) *ACC Non A.* (c) *ACC Anom.* (d) *Médias grupais.* (e) *AUC.* (f) *ACC Non A.* (g) *ACC Anom.* (h) *Médias grupais.*

Figura 4.9: Resultados experimentais e o gráfico das médias grupais para o experimento de detecção de anomalias grupais no conjunto de dados SDSS III dataset.

Figura 4.10: Resultados experimentais e o gráfico das médias grupais para o experimento de detecção de anomalias grupais no conjunto de dados SDSS III dataset.

foi computado usando (4.32) e o parâmetro de regularização foi $\lambda = 1$. O ACC para os grupos anômalos mostra que o desempenho de M2 é ruim detectando anomalias grupais. No entanto, aquela métrica é somente baseada no limiar de 0 para a saída de predição desses modelos (se a saída de predição dos modelos for maior que zero, então o grupo é anômalo, caso contrário o grupo é não anômalo). Por outro lado, a métrica AUC mostra que, para várias escolhas diferentes de valores limiar, o desempenho de todos os modelos é bom, tal como é apresentado na Figura 4.9a. De novo os modelos com pior desempenho são SVDD e M2, e a normalização esférica nos dados tem um efeito positivo, incrementando o valor do AUC perto de um, no modelo M3.

As métricas para quarto conjunto de teste tem similar características que o terceiro conjunto de teste, tal como pode ser visto nas Figuras 4.9e, 4.9f, e 4.9g, onde o AUC, o ACC para os grupos não anômalos e o ACC para os grupos anômalos são apresentados. O parâmetro do kernel e o parâmetro de regularização foi o mesmo que o experimento anterior. Novamente, a normalização esférica tem um efeito positivo na métrica AUC.

Como as médias grupais estão mais espalhadas no quinto conjunto de teste, a métrica AUC mostra que todos o desempenho de todos os modelos é bom, contudo, SVDD é o modelo com o pior desempenho. As figuras 4.10a, 4.10b, e 4.10c apresentam o AUC, o ACC para os grupos não anômalos e o ACC para os grupos anômalos. O parâmetro de regularização foi $\lambda = 1$ e o parâmetro do kernel foi escolhido segundo (4.32) mas com s sendo o 0.9 quantil.

4.6 Sumário

Neste capítulo foi apresentado um método de descrição de dados chamado de SMDD, para conjuntos de dados cujas observações são conjuntos de pontos em $\mathbb{R}^D$. Cada observação foi modelada com uma medida de probabilidade, assim, o SMDD encontra o MV-set para as funções representantes daquelas medidas de probabilidade em um EHKR. A principal vantagem é que não são estimados modelos probabilísticos para cada observação $\mathbb{P}_i$. No lugar disso, todo o processo é feito usando o mergulho das medidas de probabilidade num EHKR, usando kernels definidos sobre medidas de probabilidade. Foram formuladas três modelos SMDD. O primeiro deles usa informação do traço do operador de covariância no EHKR. Aquele modelo é formulado como um programa com restrições de chance, o qual é posteriormente transformado num programa com restrições determinísticas por meio da desigualdade de Markov. O segundo modelo é uma extensão direta do SVDD para o caso das medidas de probabilidade. Esse modelo também usa as funções média e a técnica dos mergulhos das medidas de probabilidade no EHKR. O terceiro modelo é igual ao segundo, mas considera uma normalização esférica dos dados e kernels invariantes a translação. A razão por trás disso é que as funções média geradas usando esse tipo de kernel não tem norma constante no EHKR. Foram comparadas as relações entre os modelos apresentados, mostrando os casos quando eles são equivalentes. Os modelos apresentados foram testados na tarefa de detecção de anomalias grupais. Mostramos empiricamente que o SMDD é uma alternativa para esse tipo de tarefa. Os modelos SMDD realizam detecção de anomalias grupais, descrevendo uma região no EHKR, dado por uma MEB das funções média das medidas de probabilidade usadas para representar os grupos não anômalos. Em consequência, os grupos de pontos, modelados por medidas de probabilidade tal que as respectivas funções média não pertencem ao MEB são consideradas como anomalias.

Capítulo 5

Kernels definidos sobre conjuntos fuzzy

Exploit the tolerance for imprecision, uncertainty and partial truth to achieve tractability, robustness, and low solution cost.

Zadeh.

Na área de aprendizagem computacional, é muito comum transformar os dados a serem analisados para um outro espaço com alguma estrutura adicional que implique em alguma vantagem para a análise a ser feita. Dependendo da estrutura adicional desses espaços, a análise dos dados originais é facilitada. Neste capítulo, propomos uma maneira de analisar conjuntos de dados fuzzy, usando o mergulho desses conjuntos num EHKR. O mergulho é feito usando kernels positivos definidos, os mesmos que são definidos no espaço dos conjuntos fuzzy.

5.1 Introdução

Muitas aplicações de aprendizagem computacional ou reconhecimento de padrões usam conjuntos de dados cujas observações podem ser modeladas por conjuntos fuzzy, i.e., conjuntos de dados da forma $\{X_i\}_{i=1}^N$, onde cada X_i é um conjunto fuzzy ou vetor fuzzy pertencente ao espaço dos conjuntos fuzzy denotado por $\mathcal{F}(\Omega)$ (Bertoluzza *et al.*, 2002; Viertl, 2011) (Veja Capítulo 2 desta tese para as definições e notações usadas neste capítulo). Logo, uma amostra $\{X_i\}_{i=1}^N \subset \mathcal{F}(\Omega)$, pode ser vista como um conjunto contendo as realizações de uma variável aleatória fuzzy, que segue alguma lei de probabilidade (Kwakernaak, 1978; Puri e Ralescu, 1986; Shapiro, 2009). Os conjuntos de dados fuzzy são resultado da modelagem da imprecisão e vaguidade presente nas observações de problemas reais com conjuntos fuzzy. Por exemplo, devido à incerteza causada pelo ruído, ou imprecisão dos instrumentos de medida, os dados de experimentos biológicos são bons candidatos a serem modelados com conjuntos fuzzy. Além disso, conjuntos de dados cujas observações sejam conjuntos de termos linguísticos, palavras, ou intervalos, podem ser modelados com conjuntos fuzzy (Palacios e Alcalá-Fdez, 2012; Palacios *et al.*, 2010, 2011; Sánchez *et al.*, 2009; Wang, 2001; Zadeh, 1994, 1996).

Neste capítulo, propomos fazer a análise dos conjuntos de dados fuzzy usando as funções repre-

sentantes desses conjuntos em um EHKR $\mathcal{H}$ com kernel k , mediante um mapeamento:

$$\phi : \mathcal{F}(\Omega) \rightarrow \mathcal{H} \quad (5.1)$$

$$X \mapsto \phi(X), \quad (5.2)$$

o qual é implicitamente dado graças aos kernels positivo definidos, definidos sobre conjuntos fuzzy, isto é, funções da forma:

$$\mathcal{F}(\Omega) \times \mathcal{F}(\Omega) \rightarrow \mathbb{R} \quad (5.3)$$

$$(X, Y) \mapsto k(X, Y) = \langle \phi(X), \phi(Y) \rangle_{\mathcal{H}}. \quad (5.4)$$

No contexto dos métodos de kernel, kernels definidos sobre conjuntos fuzzy irão permitir usar todas as ferramentas e algoritmos dos métodos de kernel em conjuntos de dados cujas observações sejam conjuntos fuzzy. Esse novo kernel atuará como uma medida de similaridade entre conjuntos fuzzy.

5.1.1 Trabalhos prévios

A literatura reporta o uso de técnicas da teoria fuzzy conjuntamente com kernels positivos definidos para resolver problemas de aprendizagem computacional como, por exemplo, em agrupamento de dados (Huang *et al.*, 2012; Kanzawa *et al.*, 2010; Raza e Rhee, 2012), classificação (Yang *et al.*, 2011), extração de características (Wang *et al.*, 2010) e análise discriminante (Heo e Gader, 2011). Neles as ferramentas da teoria fuzzy são usadas em alguma etapa do algoritmo de aprendizado conjuntamente com os métodos de kernel. Porém, os kernels lá usados não são definidos sobre conjuntos fuzzy, isto é, os dados analisados não são conjuntos de dados fuzzy. Por outro lado, existem trabalhos que estudam a conexão entre kernels positivo definidos, definidos em $\mathbb{R}^D \times \mathbb{R}^D$, e sistemas fuzzy (Chen, 2008; Chen e Wang, 2003; Cheng e Juang, 2011; Jeng e Lee, 1999; Juang *et al.*, 2007; Lin *et al.*, 2006; Lski, 2006; Wee e Wan, 2008; Zhou e Gan, 2007) e ainda, como alguns kernels positivos definidos podem ser vistos como relações de equivalência fuzzy (Liu e Xue, 2012; Moser, 2006a,b). Até onde sabemos, não existe formulação para definir kernels positivos definidos, definidos sobre conjuntos fuzzy. Todos os trabalhos prévios consideram somente kernels definidos em $\mathbb{R}^D \times \mathbb{R}^D$, relacionando conceitos fuzzy no projeto do kernel, ou como algum passo de algum algoritmo. (Heo e Gader, 2011; Huang *et al.*, 2012; Kanzawa *et al.*, 2010; Raza e Rhee, 2012; Wang *et al.*, 2010; Yang *et al.*, 2011)

5.2 Kernels definidos sobre conjuntos

Os kernels definidos sobre conjuntos têm sido amplamente usados em aprendizado computacional (Gärtner *et al.*, 2002; Haussler, 1999). Usamos os conceitos já estabelecidos para estender as definições para o caso de kernels definidos sobre conjuntos fuzzy. Um exemplo de kernel definido sobre conjuntos é o *kernel de produto cruzado* (Gärtner *et al.*, 2002). Esse kernel pode ser derivado a partir do *kernel de convolução* (Haussler, 1999), o qual é também um kernel definido sobre conjuntos. Em seguida, apresentamos a definição do kernel de convolução e a derivação do kernel de produto cruzado a partir dele.

5.2.1 O kernel de convolução

O *kernel de convolução* é um kernel definido sobre conjuntos cujos elementos são estruturas discretas como grafos, arvores ou strings. O kernel de convolução é obtido a partir de *outros* kernels positivo definidos, os quais são multiplicados e somados. O resultado é que o kernel de convolução é positivo definido graças às propriedades de fechamento dos kernels (ver Seção 3.2.2).

Definição 5.2.1 (Kernel de convolução (Gartner, 2008; Haussler, 1999)). Seja o objeto $e \in E$ e seja $\vec{e} = (e_1, e_2, \dots, e_L) \in E_1 \times \dots \times E_L$ uma tupla contendo as partes desse objeto. Dada a relação $R \subseteq (E_1 \times E_2 \times \dots \times E_L) \times E$, definida pela sua *função característica*¹:

$$\begin{aligned} R : E_1 \times E_2 \times \dots \times E_L \times E &\rightarrow \{\text{true}, \text{false}\} \\ (\vec{e}, e) &\mapsto R(\vec{e}, e). \end{aligned} \quad (5.5)$$

e definindo a decomposição $R^{-1}(e) = \{\vec{e} \mid R(\vec{e}, e) = \text{true}\}$. O kernel de convolução é a função de valor real definida em $E \times E$, satisfazendo para todo $e, e' \in E$:

$$k_{conv}(e, e') = \sum_{\vec{e} \in R^{-1}(e), \vec{e}' \in R^{-1}(e')} \prod_{l=1}^L k_l(e_l, e'_l), \quad (5.6)$$

onde k_l é um kernel positivo definido, definido em $E_l \times E_l$.

É importante dizer que a relação R , pode ser definida pela sua *função característica*:

$$\begin{aligned} R : E_1 \times E_2 \times \dots \times E_L \times E &\rightarrow \{\text{true}, \text{false}\} \\ (\vec{e}, e) &\mapsto R(\vec{e}, e). \end{aligned} \quad (5.7)$$

5.2.2 O kernel de produto cruzado

Dado o conjunto não vazio Ω , o *kernel de produto cruzado* pode ser facilmente derivado a partir do kernel de convolução fazendo:

$$E = \{A \mid A \subset \Omega\}, \quad \text{e} \quad E_1 = \Omega,$$

e considerando a relação:

$$R : E_1 \times E \rightarrow \{\text{true}, \text{false}\} \quad (5.8)$$

$$x, A \mapsto R(x, A), \quad (5.9)$$

a qual é verdadeira se $x \in A$. Logo, defina-se:

$$R^{-1}(A) = \{x \mid R(x, A)\}.$$

¹Usaremos a função característica de R na derivação do kernel de produto cruzado e, substituiremos ela com uma relação fuzzy, na definição do kernel de convolução fuzzy.

Como $E_1 = \Omega$, então $L = 1$. Consequentemente, o kernel de convolução para esse caso é o kernel definido em $E \times E = \mathcal{G}(\Omega) \times \mathcal{G}(\Omega)$, dado por:

$$k_{set}(A, A') = \sum_{x \in A, x' \in A'} k(x, x'), \quad (5.10)$$

onde k é um kernel positivo definido, definido em $\Omega \times \Omega$. Esta forma para o kernel de convolução é chamada de *kernel de produto cruzado*.

5.3 Kernels definidos sobre conjuntos fuzzy

Assumindo que os conjuntos fuzzy, $X \in \mathcal{F}(\Omega)$, tem como suporte um conjunto enumerável e finito, nesta seção definimos o kernel de produto cruzado de conjuntos fuzzy e o kernel de convolução fuzzy.

5.3.1 O kernel de produto cruzado de conjuntos fuzzy

Este kernel usa o produto cruzado entre elementos do suporte de dois conjuntos fuzzy, ponderados pela grau de pertinência dos elementos de ambos conjuntos fuzzy. Seja $k : \Omega \times \Omega \rightarrow \mathbb{R}$ um kernel. O *kernel de produto cruzado de conjuntos fuzzy* é a função $k_{\times} : \mathcal{F}(\Omega) \times \mathcal{F}(\Omega) \rightarrow \mathbb{R}$ definida por:

$$k_{\times}(X, Y) = \sum_{x \in \text{supp}(X), y \in \text{supp}(Y)} k(x, y)X(x)Y(y), \quad (5.11)$$

onde $X(x)$ e $Y(y)$ são os graus de pertinência de $x, y \in \Omega$ para os conjuntos fuzzy X, Y .

Se o kernel k for positivo definido, então

$$k_{\times}(X, Y) = \sum_{x \in \text{supp}(X), y \in \text{supp}(Y)} k(x, y)X(x)Y(y) \quad (5.12)$$

$$= \sum_{x \in \text{supp}(X), y \in \text{supp}(Y)} \langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}} X(x)Y(y) \quad (5.13)$$

$$= \sum_{x \in \text{supp}(X), y \in \text{supp}(Y)} \langle X(x)k(\cdot, x), Y(y)k(\cdot, y) \rangle_{\mathcal{H}}, \quad (5.14)$$

onde a segunda linha é devida à propriedade de reprodução dos kernels. Como a soma de kernels positivos definidos é um kernel positivo definido, então o kernel de produto cruzado de conjuntos fuzzy usando o kernel positivo definido k , é também positivo definido (ver as propriedades de fechamento dos kernels na Seção 3.2.2 do Capítulo 3). Note que é possível usar outro tipo de kernels em (5.11), como os kernels condicionalmente definidos ou inclusive outros kernels mais gerais como kernels simétricos que são kernels indefinidos. A representação geométrica para esses casos é garantida (Berg *et al.*, 1984; Haasdonk, 2005; Ong *et al.*, 2004; Scholkopf, 2001).

Generalização e derivação a partir do kernel de convolução

Uma expressão mais geral para o kernel $k_{\times}$, pode ser derivada a partir do kernel de convolução apresentado na Seção 5.2.1. Seja $X \in \mathcal{F}(\Omega)$, $x \in \Omega, X(x) \in [0, 1]$. Defina-se a relação $R \subseteq (\Omega \times$

$[0, 1] \times \mathcal{F}(\Omega)$ com função característica:

$$R : \Omega \times [0, 1] \times \mathcal{F}(\Omega) \rightarrow \{\text{true}, \text{false}\} \quad (5.15)$$

$$x, X(x), X \mapsto R(x, X(x), X) \quad (5.16)$$

a qual é verdadeira se $x, X(x)$ são as partes de X . Defina-se $R^{-1}(X) = \{x, X(x) | R(x, X(x), X)\}$, então $L = 2$. Logo, o kernel de convolução é um kernel definido em $\mathcal{F}(\Omega) \times \mathcal{F}(\Omega)$, satisfazendo para todo $X, Y \subset \mathcal{F}(\Omega)$:

$$k_{conv}(X, Y) = \sum_{\vec{x} \in R^{-1}(X), \vec{y} \in R^{-1}(Y)} \prod_{l=1}^2 k_l(x_l, y_l) \quad (5.17)$$

$$= \sum_{\vec{x} \in R^{-1}(X), \vec{y} \in R^{-1}(Y)} k_1(x, y) k_2(X(x), Y(y)) \quad (5.18)$$

$$\equiv \sum_{x \in \text{supp}(X), y \in \text{supp}(Y)} k_1(x, y) k_2(X(x), Y(y)) \quad (5.19)$$

$$= k_{\times}(X, Y) \quad (5.20)$$

onde $\vec{x} = (x, X(x))$, e os kernels k_1 e k_2 são funções de valor real definidas em $\Omega \times \Omega$ e em $[0, 1] \times [0, 1]$, respectivamente.

Esta forma para o kernel de convolução é uma expressão mais geral para o kernel de produto cruzado de conjuntos fuzzy definida em (5.11), pois se $k_2(X(x), Y(y)) = X(x)Y(y)$ então (5.20) se reduz para (5.11).

Equivalentemente, o kernel de produto cruzado de conjuntos fuzzy (5.20) pode ser escrito como:

$$k_{\times}(X, Y) = \sum_{x \in \text{supp}(X), y \in \text{supp}(Y)} k_1 \otimes k_2((x, X(x)), (y, Y(y))), \quad (5.21)$$

onde o *produto tensorial* $k_1 \otimes k_2 : (\Omega \times [0, 1]) \times (\Omega \times [0, 1]) \rightarrow \mathbb{R}$ é definido como:

$$k_1 \otimes k_2(x, X(x), y, Y(y)) = k_1(x, y) k_2(X(x), Y(y)). \quad (5.22)$$

Lema 12. *Se k_1 e k_2 forem kernels positivos definidos de valor real, então o kernel de produto cruzado de conjuntos fuzzy (Equação (5.20) ou equivalentemente Equação (5.21)) é um kernel positivo definido de valor real.*

Demonstração. Se k_1 é um kernel positivo definido em $\Omega \times \Omega$ e k_2 é um kernel positivo definido em $[0, 1] \times [0, 1]$ então, o produto tensorial $k_1 \otimes k_2((x, X(x)), (y, Y(y))) = k_1(x, y)k_2(X(x), Y(y))$ é um kernel positivo definido em $(\Omega \times [0, 1]) \times (\Omega \times [0, 1])$ (Berg et al., 1984). Logo, pela propriedade de fechamento dos kernels, a soma de kernels positivo definidos é um kernel positivo definido. $\square$

Sejam $\mathcal{H}_1, \mathcal{H}_2$ dois EHKR com kernels k_1, k_2 e produtos internos $\langle \cdot, \cdot \rangle_{\mathcal{H}_1}, \langle \cdot, \cdot \rangle_{\mathcal{H}_2}$, respectivamente. Então, pelo Lema 12 e (5.21), o kernel $k_{\times}$ induz o EHKR $\mathcal{H}$ com kernel $k_1 \otimes k_2$, e com produto interno:

$$\langle v_1 \otimes v_2, w_1 \otimes w_2 \rangle_{\mathcal{H}} = \langle v_1, w_1 \rangle_{\mathcal{H}_1} \langle v_2, w_2 \rangle_{\mathcal{H}_2} \quad (5.23)$$

onde $v_1, w_1 \in \mathcal{H}_1$ e $v_2, w_2 \in \mathcal{H}_2$. Isso implica que:

$$v_1(x)v_2(X(x)) = \left\langle v_1, k_1(., x) \right\rangle_{\mathcal{H}_1} \left\langle v_2, k_2(., X(x)) \right\rangle_{\mathcal{H}_2} = \left\langle v_1 \otimes v_2, k_1(., x) \otimes k_2(., X(x)) \right\rangle_{\mathcal{H}} \quad (5.24)$$

pela propriedade de reprodução.

5.3.2 O kernel de convolução fuzzy

A definição do kernel de convolução (5.6) é baseado numa relação de pertinência das partes $\vec{e} \in E_1 \times E_2 \times \dots \times E_L$, para o elemento $e \in E$, definida pela função característica (5.5). No lugar daquilo, o kernel de convolução fuzzy, usa uma relação fuzzy para modelar a relação de pertinência das partes $\vec{e}$, ao objeto e . Seja a relação fuzzy:

$$\begin{aligned} R : E_1 \times E_2 \times \dots \times E_L \times E &\rightarrow [0, 1] \\ (\vec{e}, e) &\mapsto R(\vec{e}, e), \end{aligned} \quad (5.25)$$

tal que o valor $R(\vec{e}, e) \in [0, 1]$ indica o grau de pertinência das partes $\vec{e}$ para e . Logo, defina-se $R^{-1}(e) = \{(\vec{e}, R(\vec{e}, e)) | R(\vec{e}, e)\}$ (cuja função é *extrair* as partes e o grau de pertinência dessas partes). O kernel de convolução fuzzy é a função de valor real definida em $E \times E$, satisfazendo para todo $e, e' \in \Omega$:

$$k_{\star}(e, e') = \sum_{(\vec{e}, R(\vec{e}, e)) \in R^{-1}(e), (\vec{e}', R(\vec{e}', e')) \in R^{-1}(e')} R(\vec{e}, e) R(\vec{e}', e') \prod_{l=1}^L k_l(e_l, e'_l) \quad (5.26)$$

Lema 13. *Se os kernels $k_1, k_2, \dots, k_L$ forem positivos definidos, então o kernel de convolução fuzzy é um kernel positivo definido.*

Demonstração. Pelas propriedades 2, 3 e 4, de fechamento dos kernels da Seção 3.2.2, o kernel resultante é positivo definido. $\square$

O kernel de convolução fuzzy é um kernel que generaliza o kernel de convolução. A importância dele é que pode ser usado em situações onde o conceito de grau de pertinência das partes para uma estrutura possa ser de utilidade. Apresentamos em seguida exemplos de kernels gerados a partir do kernel de convolução.

Exemplos de kernels de convolução fuzzy

Exemplo 5.3.1 (kernel de convolução fuzzy gaussiano). Suponha que o conjunto de dados é da seguinte forma:

$$\{x_i, \vec{x}_i, R(\vec{x}_i, x_i)\}_{i=1}^N, \quad (5.27)$$

onde $x_i \in \Omega$ é um objeto, $\vec{x}_i = x_1^{(i)}, \dots, x_L^{(i)} \in \Omega_1 \times \dots \times \Omega_L$ são as partes desse objeto e $R(\vec{x}_i, x_i)$ é uma relação fuzzy segundo (5.25) que modela o grau de pertinência das partes $\vec{x}_i$ para a estrutura x_i . Sejam f_l , funções de Ω_l em $\mathbb{R}^D$, isto é, $f_l : \Omega_l \rightarrow \mathbb{R}^D$, $1 \leq l \leq L$ e $D \in \mathbb{N}$, e sejam k_l kernels gaussianos dados por:

$$k_l(x_l^{(i)}, x_l^{(j)}) = \exp(-\gamma_l \|f_l(x_l^{(i)}) - f_l(x_l^{(j)})\|^2), \quad \gamma > 0,$$

onde $0 \leq i, j \leq N$. Então o kernel de convolução fuzzy gaussiano é o kernel de valor real definido em $\Omega \times \Omega$ dado por:

$$k_{\star}(x_i, x_j) = R(\vec{x}_i, x_i)R(\vec{x}_j, x_j) \exp\left(-\sum_l^L \gamma_l \|f_l(x_i^i) - f_l(x_j^j)\|^2\right). \quad (5.28)$$

Exemplo 5.3.2. Sejam $X, Y \in \mathcal{F}(\Omega)$ e a relação fuzzy:

$$\begin{aligned} R : \Omega \times \mathcal{F}(\Omega) &\rightarrow [0, 1] \\ (\vec{x}, X) &\mapsto R(\vec{x}, X) = X(\vec{x}) \end{aligned}$$

Assim, o kernel de convolução fuzzy resultante é um kernel definido em $\mathcal{F}(\Omega) \times \mathcal{F}(\Omega)$:

$$k_{\star}(X, Y) = \sum_{\vec{x}, R(\vec{x}, X) \in R^{-1}(X), \vec{y}, R(\vec{y}, Y) \in R^{-1}(Y)} R(\vec{x}, X)R(\vec{y}, Y)k(\vec{x}, \vec{y}) \quad (5.29)$$

$$= \sum_{\vec{x}, R(\vec{x}, X) \in R^{-1}(X), \vec{y}, R(\vec{y}, Y) \in R^{-1}(Y)} X(\vec{x})Y(\vec{y})k(\vec{x}, \vec{y}). \quad (5.30)$$

Esse kernel é o kernel de produto cruzado entre conjuntos fuzzy definido em (5.11).

Generalização do kernel de convolução fuzzy

O kernel de convolução fuzzy apresentado, pode ser facilmente estendido usando um kernel de valor real k_R definido em $[0, 1] \times [0, 1]$ para:

$$k_{\star}(e, e') = \sum_{(\vec{e}, R(\vec{e}, e)) \in R^{-1}(e), (\vec{e}', R(\vec{e}', e')) \in R^{-1}(e')} k_R(R(\vec{e}, e), R(\vec{e}', e')) \prod_{l=1}^L k_l(e_l, e'_l). \quad (5.31)$$

5.4 O kernel $k_{\cap}$ de interseção de conjuntos fuzzy

O *kernel de interseção de conjuntos* (Gartner, 2008) é definido tal que dados dois conjuntos não vazios, a avaliação desses conjuntos pelo kernel é um valor de similaridade dado pela *medida* do conjunto interseção desses conjuntos. A definição desse kernel é baseada no conceito de *semi-anel de conjuntos* e uma *medida* cujo domínio de definição é justamente esse semi-anel de conjuntos.

Definição 5.4.1 (Semi-anel de conjuntos). Um semi-anel de conjuntos, $\mathcal{S} \subseteq \Omega$, é um subconjunto do conjunto potência $\mathcal{P}(\Omega)$, satisfazendo:

- 1 $\phi \in \mathcal{S}$, ϕ denota o conjunto vazio,
- 2 $A, B \in \mathcal{S}, \implies A \cap B \in \mathcal{S}$,
- 3 para todo $A, A_1 \in \mathcal{S}$ tal que $A_1 \subseteq A$, existe uma sequência de conjuntos par a par disjuntos $A_2, A_3, \dots, A_N \in \mathcal{S}$, que satisfaz:

$$A = \bigcup_{i=1}^N A_i.$$

A condição 3 é chamada de *decomposição finita de A*.

Definição 5.4.2 (Medida definida em semi-anel de conjuntos). Seja $\mathcal{S}$ um semi-anel de conjuntos e seja $\rho : \mathcal{S} \rightarrow [0, \infty]$ uma função satisfazendo:

$$1 \quad \rho(\phi) = 0;$$

$$2 \quad \text{Dado } A \in \mathcal{S} \text{ com decomposição finita } \bigcup_{i=1}^N A_i, \quad \rho(A) = \rho(\bigcup_{i=1}^N A_i) = \sum_{i=1}^N \rho(A_i).$$

usando o *teorema de extensão de Carathéodory* (Ash e Doleans-Dade, 2000), dizemos que ρ é uma medida definida em $\sigma(\mathcal{S})$, onde $\sigma(\mathcal{S})$ é a menor σ -álgebra incluindo $\mathcal{S}$.

O kernel de interseção de conjuntos. Usando esses dois conceitos de semi-anel de conjuntos e de medida definida em semi-anel de conjuntos, o kernel de interseção de conjuntos, é a função $k : \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}$ definida por:

$$k(A, A') = \rho(A \cap A'), \quad (5.32)$$

a qual é um kernel positivo definido (Gartner, 2008).

5.4.1 Medida do suporte dos conjuntos fuzzy

Usamos o conceito de semi-anel de conjuntos para definir conjuntos fuzzy cujo suporte pertence a um semi-anel de conjuntos. Nesse sentido, definimos $\mathcal{F}_{\mathcal{S}}(\Omega)$ como o conjunto de todos os conjuntos fuzzy cujo suporte pertence a $\mathcal{S}$, ou seja,

$$\mathcal{F}_{\mathcal{S}}(\Omega) = \{X \subset \Omega \mid \text{supp}(X) \in \mathcal{S}\},$$

então, pela definição do semi-anel de conjuntos, qualquer conjunto fuzzy em $\mathcal{F}_{\mathcal{S}}(\Omega)$, possui suporte com decomposição finita. Ou seja, se $\mathcal{A} = \{A_1, A_2, \dots, A_M\} \subseteq \mathcal{S}$, for um sistema finito de conjuntos par a par disjuntos, então o suporte de cada conjunto fuzzy em $\mathcal{F}_{\mathcal{S}}(\Omega)$, é a união de elementos de $\mathcal{A}$. Isto é, o suporte do conjunto fuzzy $X \in \mathcal{F}_{\mathcal{S}}(\Omega)$ pode ser escrito como:

$$\text{supp}(X) = \bigcup_{A \in \mathcal{B} \subseteq \mathcal{A}} A,$$

então, será possível quantificar a decomposição finita do suporte de um conjunto fuzzy usando a medida definida em semi-anel de conjuntos da Definição 5.4.2. Para isso, introduzimos a seguir a função indicadora:

$$\begin{aligned} \mathbf{1}_B : \mathcal{S} &\rightarrow \{0, 1\} \\ A &\mapsto \mathbf{1}_B(A) = \begin{cases} 1, & A \subseteq B, \\ 0, & \text{cc} \end{cases}, \end{aligned} \quad (5.33)$$

onde $B \subseteq \mathcal{S}$. A função indicadora será de utilidade na formulação de medida do suporte dos conjuntos fuzzy em $\mathcal{F}_{\mathcal{S}}(\Omega)$.

Definição 5.4.3 (Medida do suporte de um conjunto fuzzy). Dado $X \in \mathcal{F}_{\mathcal{S}}(\Omega)$, tal que

$$\text{supp}(X) = \bigcup_{A \in \mathcal{B} \subseteq \mathcal{A}} A,$$

onde $\mathcal{A}$ é um sistema finito de conjuntos par a par disjuntos. Seja $\mathbf{1}_{\text{supp}(X)}$ segundo (5.33), então, a medida do suporte de um conjunto fuzzy é

$$\rho(\text{supp}(X)) = \sum_{A \in \mathcal{B} \subseteq \mathcal{A}} \rho(A) \quad (5.34)$$

$$= \sum_{A \in \mathcal{A}} \rho(A) \mathbf{1}_{\text{supp}(X)}(A) \quad (5.35)$$

onde o somatório é devido ao fato que, segundo a Definição 5.4.2, ρ é uma medida aditiva.

Por exemplo, usando as funções $\mathbf{1}_B(\cdot)$ e $\rho(\cdot)$, e o sistema finito de conjuntos par a par disjuntos $\mathcal{A} = \{A_1, A_2, \dots, A_M\} \subseteq \mathcal{S}$, é possível medir o suporte da interseção de dois conjuntos fuzzy $X, Y \in \mathcal{F}_\mathcal{S}(\Omega)$ por:

$$\rho(\text{supp}(X \cap Y)) = \sum_{A \in \mathcal{A}} \rho(A) \mathbf{1}_{\text{supp}(X)}(A) \mathbf{1}_{\text{supp}(Y)}(A), \quad (5.36)$$

Apresentamos em seguida o conceito de contribuição fuzzy, o qual permitirá quantificar a contribuição fuzzy dos elementos do semi-anel que definem o suporte dos conjuntos fuzzy.

Definição 5.4.4 (Contribuição fuzzy). Seja $X \in \mathcal{F}_\mathcal{S}(\Omega)$, então a contribuição fuzzy de um conjunto $A \in \mathcal{S}$ é o valor:

$$\sum_{x \in A} X(x). \quad (5.37)$$

Abusando da notação, denotamos a contribuição fuzzy de $A \in \mathcal{S}$, como $X(A) = \sum_{x \in A} X(x)$.

5.4.2 O kernel $k_\cap$

O kernel de interseção fuzzy $k_\cap$, é uma medida de similaridade entre dois conjuntos fuzzy pertencentes a $\mathcal{F}_\mathcal{S}(\Omega)$. O kernel $k_\cap$ é baseado em dois conceitos principais: o primeiro deles é a medida da decomposição finita do conjunto suporte da interseção de dois conjuntos fuzzy (ver (5.36)) e, o segundo conceito é a contribuição fuzzy dessa decomposição finita.

Definição 5.4.5 (kernel de interseção de conjuntos fuzzy). Sejam X, Y dois conjuntos fuzzy em $\mathcal{F}_\mathcal{S}(\Omega)$, O kernel de interseção de conjuntos fuzzy é a função de valor real $k_\cap : \mathcal{F}_\mathcal{S}(\Omega) \times \mathcal{F}_\mathcal{S}(\Omega) \rightarrow \mathbb{R}$, tal que:

$$k_\cap(X, Y) = \sum_{A \in \mathcal{A}} (X \cap Y)(A) \rho(A) \mathbf{1}_{\text{supp}(X)}(A) \mathbf{1}_{\text{supp}(Y)}(A), \quad (5.38)$$

onde $(X \cap Y)(A)$ é a contribuição fuzzy do conjunto $A \in \mathcal{A}$.

O kernel $k_\cap$ pondera a medida de cada elemento da decomposição finita do suporte de um conjunto fuzzy em $\mathcal{F}_\mathcal{S}(\Omega)$ (ver (5.36)) com a contribuição fuzzy desse elemento $X \cap Y(A)$.

Introduzindo $\mathcal{C}_{X,Y} = \{A \in \mathcal{A} | \mathbf{1}_{\text{supp}(X)}(A) \mathbf{1}_{\text{supp}(Y)}(A) = 1\}$, (5.38) pode ser escrita como:

$$k_\cap(X, Y) = \sum_{A \in \mathcal{C}_{X,Y}} X \cap Y(A) \rho(A). \quad (5.39)$$

Usando os conceitos de T-norma e contribuição fuzzy é possível rescrever $k_{\cap}$ como:

$$k_{\cap}(X, Y) = \sum_{A \in \mathcal{C}_{X,Y}} X \cap Y(A) \rho(A) \quad (5.40)$$

Pela definição de contribuição fuzzy (Definição 5.4.4)

$$= \sum_{A \in \mathcal{C}_{X,Y}} \left(\sum_{x \in A} X \cap Y(x) \right) \rho(A)$$

Usando a T-norma

$$= \sum_{A \in \mathcal{C}_{X,Y}} \left(\sum_{x \in A} T(X(x), Y(x)) \right) \rho(A) \quad (5.41)$$

Exemplo 5.4.1 (Kernel de interseção fuzzy e T-normas). A Tabela 5.1 apresenta os kernels derivados usando as diferentes T-normas da Tabela 2.1 em (5.41).

Kernel $k_{\cap}$	T-norma
$k_{\cap_min}(X, Y) = \sum_{A \in \mathcal{C}_{X,Y}} \sum_{x \in A} \min(X(x), Y(x)) \rho(A)$	mínimo
$k_{\cap_pro}(X, Y) = \sum_{A \in \mathcal{C}_{X,Y}} \sum_{x \in A} X(x)Y(x) \rho(A)$	produto
$k_{\cap_Luk}(X, Y) = \sum_{A \in \mathcal{C}_{X,Y}} \sum_{x \in A} \max(X(x) + Y(x) - 1, 0) \rho(A)$	Łukasiewicz
$k_{\cap_Dra}(X, Y) = \sum_{A \in \mathcal{C}_{X,Y}} \sum_{x \in A} Z(X(x), Y(x)) \rho(A)^a$	Drástico

^aA função Z é a T-norma Drástico da Tabela 2.1

Tabela 5.1: Formulações de funções Kernel de interseção de conjuntos fuzzy usando T-normas

Exemplo 5.4.2. Se a σ -álgebra onde é definida a medida ρ é uma álgebra de Borel, (5.40) pode ser escrita como:

$$k_{\cap}(X, Y) = \sum_{A \in \mathcal{C}_{X,Y}} \int_{x \in A} X \cap Y(x) d\rho(x) \equiv \int_{x \in \Omega} X \cap Y(x) d\rho(x), \quad (5.42)$$

como exemplo temos:

$$k_{\cap}(X, Y) = \int_{\mathbb{R}^D} T(X(x), Y(x)) d\mathbb{P}(x), \quad (5.43)$$

se for usada a medida de probabilidade. E se forem usadas T-normas específicas como as funções de mínimo e produto, então (5.43) pode ser escrita como:

$$k_{\cap_min}(X, Y) = \int_{\mathbb{R}^D} \min(X(x), Y(x)) d\mathbb{P}(x) \quad (5.44)$$

$$k_{\cap_pro}(X, Y) = \int_{\mathbb{R}^D} X(x)Y(x) d\mathbb{P}(x) \quad (5.45)$$

As seguintes proposições mostram que os kernels de interseção fuzzy $k_{\cap_min}$ e $k_{\cap_pro}$ são kernels reproduzíveis de algum EHKR.

Proposição 14. *o kernel $k_{\cap_min}(X, Y)$ é positivo definido.*

Demonstração. Primeiramente, defina-se a função $\mathbf{1}_{[0,a]}$ como:

$$\begin{aligned} \mathbf{1}_{[0,a]} : [0, 1] &\rightarrow \{0, 1\} \\ t &\mapsto \mathbf{1}_{[0,a]}(t) = \begin{cases} 1, & t \in [0, a], \quad a \in [0, 1] \\ 0, & \text{cc} \end{cases} \end{aligned}$$

então a função $\min : [0, 1] \times [0, 1] \rightarrow [0, 1]$ pode ser escrita como, $\forall a, b \in [0, 1]$

$$\begin{aligned} \min(a, b) &= \int_{[0,1]} \mathbf{1}_{[0,a]}(t) \mathbf{1}_{[0,b]}(t) d\lambda(t) \\ &= \langle \mathbf{1}_{[0,a]}(\cdot), \mathbf{1}_{[0,b]}(\cdot) \rangle_{\mathcal{H}} \end{aligned}$$

onde $\mathcal{H} = L^2([0, 1], \lambda)$, i.e., o espaço das funções quadrado integráveis no intervalo $[0, 1]$ em relação a medida de Lebesgue λ . Pelo Lema (1) e Equação (3.2), segue que $\min$ é positivo definido. Isto é, fixando $x \in \Omega$ temos que $\forall N \in \mathbb{N}$, $\forall c_i, c_j \in \mathbb{R}$, $\forall X_i, X_j \in \mathcal{F}_{\mathcal{S}}(\Omega)$:

$$\sum_{i=1}^N \sum_{j=1}^N c_i c_j \min(X_i(x), X_j(x)) \geq 0,$$

Assim, definindo $\mathcal{C}(X_i, X_j) = \{A \in \mathcal{A} | \mathbf{1}_{\text{supp}(X_i)}(A) \mathbf{1}_{\text{supp}(X_j)}(A) = 1\}$, segue que:

$$\begin{aligned} &\sum_{i=1}^N \sum_{j=1}^N c_i c_j \sum_{A \in \mathcal{C}(X_i, X_j)} \sum_{x \in A} \min(X_i(x), X_j(x)) \rho(A) \\ &= \sum_{A \in \mathcal{C}(X_i, X_j)} \sum_{x \in A} \left(\sum_{i=1}^N \sum_{j=1}^N c_i c_j \min(X_i(x), X_j(x)) \right) \rho(A) \\ &\geq 0 \end{aligned}$$

$\forall N \in \mathbb{N}$, $\forall c_i, c_j \in \mathbb{R}$, $\forall X_i, X_j \in \mathcal{F}_{\mathcal{S}}(\Omega)$, $\forall x \in \Omega$. Então $k_{\cap_min}$ é positivo definido. $\square$

Proposição 15. *o kernel $k_{\cap_pro}(X, Y)$ é positivo definido.*

Demonstração. Note que fixando $x \in \Omega$, é valido que $\forall N \in \mathbb{N}$, $\forall c_i, c_j \in \mathbb{R}$, $\forall X_i, X_j \in \mathcal{F}_{\mathcal{S}}(\Omega)$:

$$\sum_{i=1}^N \sum_{j=1}^N c_i c_j X_i(x) X_j(x) = \left(\sum_{i=1}^N c_i X_i(x) \right)^2 \geq 0. \quad (5.46)$$

Assim, definindo $\mathcal{C}(X_i, X_j) = \{A \in \mathcal{A} | \mathbf{1}_{\text{supp}(X_i)}(A) \mathbf{1}_{\text{supp}(X_j)}(A) = 1\}$, segue que:

$$\begin{aligned} & \sum_{i=1}^N \sum_{j=1}^N c_i c_j \sum_{A \in \mathcal{C}(X_i, X_j)} \sum_{x \in A} X_i(x) X_j(x) \rho(A_l) \\ &= \sum_{A \in \mathcal{C}(X_i, X_j)} \sum_{x \in A} \left(\sum_{i=1}^N \sum_{j=1}^N c_i c_j X_i(x) X_j(x) \right) \rho(A_l) \\ &\geq 0 \end{aligned}$$

$\forall N \in \mathbb{N}, \forall c_i, c_j \in \mathbb{R}, \forall X_i, X_j \in \mathcal{F}_S(\Omega), \forall x \in \Omega$. Então $k_{\cap_pro}$ é positivo definido. $\square$

Usando esses resultados (Proposições 14 e 15) é possível intuir que qualquer função positiva definida que seja norma triangular (ver Seção 2.2.1) possa ser usada para definir kernels de interseção de conjuntos fuzzy, que sejam positivos definidos. Mostramos esse resultado no seguinte teorema.

Teorema 16 (kernel positivo de interseção de conjuntos fuzzy). *Se a norma triangular $T(.,.)$ é uma função positiva definida, então o kernel da Equação (5.38) é positivo definido.*

Demonstração. Sejam os conjuntos fuzzy $X, Y \in \mathcal{F}_S(\Omega)$. Logo, se a norma triangular $T(.,.)$ é uma função positiva definida, então pela propriedade de reprodução dos kernels temos:

$$T(X(x), Y(x)) = \left\langle T(X(x), .), T(Y(x), .) \right\rangle_{\mathcal{H}},$$

onde $T(X(x), .), T(Y(x), .) \in \mathcal{H}$.

Assim, $\forall N \in \mathbb{N}, \forall c_i, c_j \in \mathbb{R}, \forall X_i, X_j \in \mathcal{F}_S(\Omega), \forall x \in \Omega$,

$$\begin{aligned} & \sum_{i=1}^N \sum_{j=1}^N c_i c_j k(X_i, X_j) \\ &= \sum_{i=1}^N \sum_{j=1}^N c_i c_j \sum_{A \in \mathcal{A}} X_i \cap X_j(A) \rho(A) \mathbf{1}_{\text{supp}(X_i)}(A) \mathbf{1}_{\text{supp}(X_j)}(A). \end{aligned}$$

Usando $\mathcal{C}(X_i, X_j) = \{A \in \mathcal{A} | \mathbf{1}_{\text{supp}(X_i)}(A) \mathbf{1}_{\text{supp}(X_j)}(A) = 1\}$, temos:

$$= \sum_{i=1}^N \sum_{j=1}^N c_i c_j \sum_{A \in \mathcal{C}(X_i, X_j)} X_i \cap X_j(A) \rho(A).$$

Usando (5.37) para $X_i \cap X_j(A)$, temos:

$$= \sum_{i=1}^N \sum_{j=1}^N c_i c_j \sum_{A \in \mathcal{C}(X_i, X_j)} \sum_{x \in A} X_i \cap X_j(x) \rho(A).$$

Usando o operador T-norma, temos

$$\begin{aligned}
 &= \sum_{i=1}^N \sum_{j=1}^N c_i c_j \sum_{A \in \mathcal{C}(X_i, X_j)} \sum_{x \in A} T(X_i(x), X_j(x)) \rho(A) \\
 &= \sum_{A \in \mathcal{C}(X_i, X_j)} \sum_{x \in A} \sum_{i=1}^N \sum_{j=1}^N c_i c_j T(X_i(x), X_j(x)) \rho(A).
 \end{aligned}$$

Usando a hipótese de que $T(., .)$ é positiva definida, temos:

$$\begin{aligned}
 &= \sum_{A \in \mathcal{C}(X_i, X_j)} \sum_{x \in A} \sum_{i=1}^N \sum_{j=1}^N c_i c_j \langle T(X_i(x), .), T(X_j(x), .) \rangle_{\mathcal{H}} \rho(A) \\
 &= \sum_{A \in \mathcal{C}(X_i, X_j)} \sum_{x \in A} \left\| \sum_{i=1}^N c_i T(X_i(x), .) \right\|_{\mathcal{H}}^2 \rho(A) \in \mathbb{R}^+.
 \end{aligned}$$

□

5.5 Kernels em conjuntos fuzzy usando substituição de distância

Nesta seção exploramos como o uso de distâncias pode ser usado para construir kernels definidos sobre conjuntos fuzzy. Usaremos o termo distância para referirmos indistintamente aos conceitos de métrica, pseudo-métrica e semi-métrica.

Definição 5.5.1 (Métrica, pseudo-métrica e semi-métrica). Uma métrica é uma função: $d : \Omega \times \Omega \rightarrow \mathbb{R}_0^+$, satisfazendo:

1. $d(x, x') \geq 0, \forall x, x' \in \Omega$ (não negativa)
2. $d(x, x') = 0$ se e somente se $x = x'$ (diagonal zero)
3. $d(x, x') = d(x', x)$ (simetria)
4. $d(x, z) \leq d(x, y) + d(y, z)$ (desigualdade triangular)

A função d é chamada de *pseudo-métrica* se ela satisfaz os axiomas 1, 3 e 4, e adicionalmente satisfaz $d(x, x) = 0, \forall x \in \Omega$ e $d(x, x') = 0$ para alguns valores $x \neq x'$. A função d é uma *semi-métrica* se ela satisfaz os axiomas 1, 2 e 3.

O produto interno $\langle x, x' \rangle$, pode ser visto como uma medida de *similaridade* entre $x, x' \in \mathbb{R}^D$. Os kernels positivos definidos, então, são generalizações não lineares dessa medida de similaridade, pois eles satisfazem:

$$k(x, x') = \langle k(., x), k(., x') \rangle_{\mathcal{H}}, \quad k(., x), k(., x') \in \mathcal{H}.$$

Assim, é possível dizer que uma generalização não linear da medida de *dissimilaridade*: $\|x - x'\|^2$, (distância), é dada pela expressão:

$$\|k(., x) - k(., x')\|_{\mathcal{H}}^2 = k(x, x) - 2k(x, x') + k(x', x'). \quad (5.47)$$

Não somente os kernels positivos definidos podem ser usados para generalizar o conceito de distância, como também podem ser usados os kernels condicionalmente positivos definidos (CPD), isto é, funções kernels que além de satisfazer o Lema 1 (Capítulo 3), satisfazem $\sum_i c_i = 0, \forall c_i \in \mathbb{R}$. Se k' for um kernel CPD, então o kernel (Proposição 3 em (Scholkopf, 2001) e Lema 2.1 em (Berg *et al.*, 1984)):

$$k(x, x') = \frac{1}{2}(k'(x, x') - k'(x, x_0) - k'(x_0, x') + k'(x_0, x_0))$$

é positivo definido. Logo a expressão (Proposição 3.2 em (Berg *et al.*, 1984)):

$$\|k(\cdot, x) - k(\cdot, x')\|_{\mathcal{H}}^2 = -k'(x, x') + 1/2(k'(x, x) + k'(x', x')), \quad (5.48)$$

é uma métrica ou semi-métrica dependendo da escolha do kernel CPD k' . Além do mais, kernels simétricos em geral, os quais não são necessariamente positivos definidos ou CPD, podem ser usados em (5.47) e (5.48), mais eles podem violar a definição de métrica, isto é, ou não satisfazer a desigualdade triangular ou ser negativos. Porém, (5.47) e (5.48) com kernel simétricos serão simétricas e com zero na diagonal. Os kernels simétricos podem ser representados usando formas bilineares simétricas em espaços lineais de funções (Scholkopf, 2001)². Esses kernels, tem sido usados exitosamente em problemas de aprendizado computacional com resultados satisfatórios (Haasdonk, 2005). Inclusive kernels indefinidos não simétricos³, baseados em distâncias, tem sido amplamente usados com bons resultados (Haasdonk e Keysers, 2002; Moreno *et al.*, 2003; Noma, 2002). Os espaços pseudo-euclidianos (Haasdonk, 2005; Pekalska *et al.*, 2002) e os espaços de Krein (Ong *et al.*, 2004) têm sido usados para dar uma interpretação geométrica para as SVM com kernels indefinidos. Nesses espaços o produto interno não é necessariamente positivo definido e a norma não define métrica.

5.5.1 Kernels com substituição de distância

Os kernels com substituição de distância (Haasdonk e Bahlmann, 2004), usam a função d da Definição 5.5.1, reuquirindo que ela seja pelo minimo não negativa: $d(x, x') \geq 0, \forall x, x' \in \Omega$, com diagonal zero: $d(x, x') = 0$ para $x = x'$, e simétrica: $d(x, x') = d(x', x)$. Logo, como

$$\|(x - x_0) - (x' - x_0)\|^2 = \|(x - x_0)\|^2 - 2\langle(x - x_0), (x' - x_0)\rangle + \|(x' - x_0)\|^2,$$

então, segue que:

$$\langle(x - x_0), (x' - x_0)\rangle = \frac{1}{2}(\|(x - x_0)\|^2 + \|(x' - x_0)\|^2 - \|x - x'\|^2) \quad (5.49)$$

onde $x_0 \in \Omega$ é um origem escolhido a priori. Logo, denotando $\langle(x - x_0), (x' - x_0)\rangle$ como $\langle x, x' \rangle_d^{x_0}$, onde d é a função distância para ser substituída em (5.49), temos:

$$\langle x, x' \rangle_d^{x_0} = \frac{1}{2}(d(x, x_0)^2 + d(x', x_0)^2 - d(x, x')^2) \quad (5.50)$$

²Se k for um kernel simétrico, então ele satisfaz $k(x, x') = Q(k(\cdot, x), k(\cdot, x'))$, e ele tem a propriedade de reprodução $Q(k(\cdot, x), f) = f(x)$, nesse espaço linear de funções. Veja Proposição 6 em (Scholkopf, 2001).

³Se a matriz kernel $K = (k(x_i, x_j))_{i,j=1}^N$ for gerada por um kernel indefinido, então existem os vetores $\mathbf{c}, \mathbf{c}' \in \mathbb{R}^D$ satisfazendo $\mathbf{c}^\top K \mathbf{c} > 0$ e $\mathbf{c}'^\top K \mathbf{c}' < 0$.

Logo, as seguintes expressões são alguns kernels de substituição de distância:

$$1) \langle x, x' \rangle_d^{x_0}, \quad 2) (\alpha + \gamma \langle x, x' \rangle_d^{x_0})^\beta, \quad 3) -d(x, x')^p, \quad 4) \exp(-\gamma d(x, x')^2) \quad (5.51)$$

para $\alpha, \gamma \in \mathbb{R}^+$, $\beta \in \mathbb{N}$, $p \in [0, 2]$. Se a função d for uma métrica, Os kernels 1, 2 e 4 de (5.51) serão positivos definidos e o kernel 3 de (5.51) será CPD. No caso de d não for uma métrica, por exemplo não satisfaz a desigualdade triangular (semi-métrica) ou d for pseudo-métrica, isto é, existe $d(x, x') = 0$ para $x \neq x'$. Temos que os kernels resultantes são kernels simétricos, isto é, eles podem ser interpretados geometricamente em espaços pseudo-euclidianos ou espaços de Krein, e poderiam ser usados satisfatoriamente na prática (Chapelle *et al.*, 1999; Haasdonk e Keysers, 2002; Moreno *et al.*, 2003).

5.5.2 Métricas, pseudo-métricas e semi-métricas em conjuntos fuzzy

Diversas métricas, pseudo-métricas e semi-métricas tem sido propostas para conjuntos fuzzy (Bloch, 1999; Chaudhur e Rosenfeld, 1996; Diamond *et al.*, 1994; Rosenfeld, 1985). Essas métricas podem ser usadas no contexto de kernels por substituição de distância. Por exemplo, se X e X' são conjuntos fuzzy, a Tabela 5.2 apresenta algumas expressões que poderiam ser usadas para a criação de kernels de substituição de distâncias:

$d(X, X')$	Descrição
$\left[\int_{x \in \Omega} X(x) - X'(x) ^p \right]^{1/p}$	pseudo-métrica, $p \in \mathbb{N}$
$\sup_{x \in \Omega} X(x) - X'(x) $	métrica
$\frac{\sum_{x \in \Omega} X(x) - X'(x) }{\sum_{x \in \Omega} X(x) + X'(x) }$	métrica, caso discreto
$ E(X) - E(X') $	pseudo-métrica,
$E(X) = -K \sum_{x \in \Omega} [X(x) \log X(x) + (1 - X(x)) \log(1 - X(x))]$	
$\sum_{x \in \Omega} [D_x(X, X') + D_x(X', X)]$	semi-métrica,
$D_x(X, X') = X(x) \log \frac{X(x)}{X'(x)} + (1 - X(x)) \log \frac{(1 - X(x))}{(1 - X'(x))}$	
$1 - \frac{\sum_{x \in \Omega} \min[X(x), X'(x)]}{\sum_{x \in \Omega} \max[X(x), X'(x)]}$	semi-métrica
$1 - \sum_{x \in \Omega} \frac{\min[X(x), X'(x)]}{\max[X(x), X'(x)]}$	semi-métrica
$1 - \max_{x \in \Omega} \min[X(x), X'(x)]$	pseudo-métrica
$\int_0^1 D(X_\alpha, X'_\alpha) d\alpha$	X_α é o conjunto α -cut, D é qualquer métrica entre conjuntos
$\frac{\sum_{x, x' \in \Omega} d(x, x') \min[X(x), X'(x')]}{\sum_{x, x' \in \Omega} \min[X(x), X'(x')]}$	métrica

Tabela 5.2: Exemplo de distâncias entre conjuntos fuzzy.

Logo, qualquer uma das distâncias apresentadas na Tabela 5.2 poderiam ser usadas em (5.50) ou em (5.51) para obter kernels definidos sobre conjuntos fuzzy. Por exemplo, as seguintes pseudo-

métricas poderiam ser usadas em kernels com substituição de distância.

Lema 17. *Seja $\Omega \subseteq \mathbb{R}^D$, tal que cada conjunto fuzzy em $\mathcal{F}(\Omega)$ é identificado pelos parâmetros: $(a_1, \dots, a_D)^\top \in \mathbb{R}^D$ e $(b_1, \dots, b_D)^\top \in \mathbb{R}^D$. Logo, para todo $X, X' \in \mathcal{F}(\Omega)$ as funções:*

$$d_1(X, X') = \frac{\prod_{j=1}^D (a_j - a'_j)^2}{\sum_{j=1}^D b_j^2 + \sum_{j=1}^D (b'_j)^2}$$

$$d_2(X, X') = \frac{\prod_{j=1}^D (a_j - a'_j)^2}{\prod_{j=1}^D (b_j^2 + (b'_j)^2)}$$

tem zero diagonal e são pseudo-métricas em $\mathcal{F}(\Omega)$.

Demonstração. d_1 e d_2 são simétricas por construção. Elas são pseudo-métricas pois elas podem ser zero para distintos valores dos parâmetros b . $\square$

Então os kernels, de substituição com essas medidas serão kernels simétricos. Por exemplo, os seguintes kernels definidos sobre conjuntos fuzzy, por construção são kernels simétricos com substituição de distância:

$$k_\gamma(X, X') = \exp(-\gamma d_1(X, X')^2), \gamma \in \mathbb{R}^+, \quad (5.52)$$

$$k_\gamma(X, X') = \exp(-\gamma d_2(X, X')^2), \gamma \in \mathbb{R}^+, \quad (5.53)$$

em geral, substituindo qualquer função d da Tabela 5.2.

5.6 Aplicações

Em seguida, são mostradas diversas áreas de aplicação em aprendizado computacional onde o conceito de kernels definidos sobre conjuntos fuzzy pode ser de utilidade. Esboçamos um apanhado de diversos algoritmos de aprendizado baseados no mapeamento de conjuntos fuzzy aos EHKR, de acordo com os kernels apresentados. Essas aplicações incluem definição de distâncias entre conjuntos fuzzy, engenharia de kernels e problemas de classificação supervisionada

5.6.1 Distâncias de conjuntos fuzzy mediante mergulho no EHKR

Usando os kernels apresentados neste capítulo é possível definir distâncias entre conjuntos fuzzy de maneira implícita. Dado que as funções kernel fazem o mapeamento dos conjuntos fuzzy para funções em um EHKR usando um mapeamento não linear da forma:

$$\begin{aligned} k(., .) : \mathcal{F}(\Omega) &\rightarrow \mathcal{H} \\ X &\mapsto k(., X), \end{aligned}$$

então é possível definir a função $D : \mathcal{F}(\Omega) \times \mathcal{F}(\Omega) \rightarrow \mathbb{R}$, por

$$D(X, Y) \stackrel{\text{def}}{=} \|k(., X) - k(., Y)\|_{\mathcal{H}}. \quad (5.54)$$

Assim, é possível definir métricas entre conjuntos fuzzy, de forma análoga às Equações (5.47) e (5.48). Ou seja, se o kernel definido sobre conjuntos fuzzy for positivo definido, temos:

$$D(X, Y)^2 = k(X, X) - 2k(X, Y) + k(Y, Y),$$

ou se o kernel definido sobre conjuntos fuzzy for CPD, temos:

$$D(X, Y)^2 = -k(X, Y) + \frac{1}{2}(k(X, X) + k(Y, Y)),$$

5.6.2 Engenharia de kernels definidos sobre conjuntos fuzzy

Usando as propriedades de fechamento dos kernels é simples criar novos kernels positivo definidos, definidos sobre conjuntos fuzzy, a partir das funções kernel apresentadas neste capítulo. Por exemplo, definindo $k_1, \dots, k_m, \dots, k_M$ como kernels positivo definidos, definidos sobre conjuntos fuzzy, a função de valor real f definida no espaço dos conjuntos fuzzy, e a função ϕ definida em conjuntos fuzzy, tal que ela é uma função de valor $\mathbb{R}^D$, temos que as propriedades 1,2,3,4,5,6 e 7 da Seção 3.2.2 podem ser aplicadas diretamente para construir novos kernels positivo definidos, definidos sobre conjuntos fuzzy. Por exemplo,

$$k_{mkl}(X, X') = \sum_{m=1}^M \alpha_m k_m(X, X'), \text{ com } \alpha_m \in \mathbb{R}^+, \quad (5.55)$$

é um kernel positivo definido definido sobre conjuntos fuzzy. Esse kernel pode ser usado no contexto de *aprendizado de kernels múltiplos* (Rakotomamonjy *et al.*, 2008) em conjuntos de dados com atributos compostos, nos quais alguns deles sejam conjuntos fuzzy.

Kernels positivos definidos, definidos sobre conjuntos fuzzy

Alguns kernels positivo definidos, definidos sobre conjuntos fuzzy, são:

- *Kernel polinomial fuzzy*. Sejam $\alpha \geq 0, \beta \in \mathbb{N}$ e k um kernel positivo definido, definido sobre conjuntos fuzzy, então o kernel

$$\begin{aligned} k_{pol}(X, Y) &= (\langle k(., X), k(., Y) \rangle_{\mathcal{H}} + \alpha)^\beta \\ &= (k(X, Y) + \alpha)^\beta, \end{aligned}$$

é positivo definido. Veja (5.51).

- *Kernel gaussiano fuzzy*. Seja $\gamma > 0$, então o kernel

$$\begin{aligned} k_{gauss}(X, Y) &= \exp(-\gamma \|k(., X) - k(., Y)\|_{\mathcal{H}}^2) \\ &= \exp(-\gamma D(X, Y)), \end{aligned}$$

é positivo definido se D for uma métrica (Haasdonk e Bahlmann, 2004). Veja por (5.51).

- *Kernel quadrático racional fuzzy.* Sejam $\alpha, \beta > 0$, então o kernel

$$\begin{aligned} k_{ratio}(X, Y) &= \left(1 + \frac{\|k(\cdot, X) - k(\cdot, Y)\|_{\mathcal{H}}^2}{\alpha\beta^2}\right)^{-\alpha} \\ &= \left(1 + \frac{D(X, Y)}{\alpha\beta^2}\right)^{-\alpha}, \end{aligned}$$

é positivo definido, se D for uma métrica.

Kernels condicionalmente positivos definidos, definidos sobre conjuntos fuzzy

Alguns exemplos de kernels CPD (Veja Seção 5.5.1), se D for uma métrica, são:

- *Kernel multi-quádrico fuzzy.* Seja $\alpha \in \mathbb{R}$, então o kernel

$$\begin{aligned} k_{multi}(X, Y) &= -\sqrt{\|\phi(X) - \phi(Y)\|_{\mathcal{H}}^2 + \alpha^2} \\ &= -\sqrt{D(X, Y) + \alpha^2}, \end{aligned}$$

é um kernel condicionalmente positivo definido.

- *Kernel multi-quádrico inverso fuzzy.* Seja $\alpha \in \mathbb{R}$, então o kernel

$$\begin{aligned} k_{invmult}(X, Y) &= (\sqrt{\|\phi(X) - \phi(Y)\|_{\mathcal{H}}^2 + \alpha^2})^{-1} \\ &= (\sqrt{D(X, Y) + \alpha^2})^{-1}, \end{aligned}$$

é um kernel condicionalmente positivo definido.

A partir de kernels condicionalmente positivo definidos é possível construir kernels positivo definidos, pois um kernel k é condicionalmente positivo definido se e somente se $\exp(\gamma k)$ for positivo definido para todo $\gamma > 0$ (Schölkopf e Smola, 2002).

5.6.3 Problemas de aprendizado supervisionado

De posse de um conjunto de dados fuzzy $\{X_i\}_{i=1}^N$, definindo a matriz kernel K , tal que $K = k(X_i, X_j)$, onde k é um kernel positivo definido, definido sobre conjuntos fuzzy, os kernels apresentados neste capítulo são úteis em diversas formulações de classificação como por exemplo, classificação binária, regressão, classificação de uma classe, etc. Porém não podemos afirmar que esses kernels serão sempre bem sucedidos. Nas seções seguintes, apresentamos algumas possíveis usos desses kernels e alguns experimentos para ilustrar a ideia.

Classificação binária

Se o conjunto de treinamento for $\{(x_i, y_i)\}_{i=1}^N \subset \mathcal{X} \times \{-1, 1\}$, tal que x_i é um valor impreciso, então, é possível usar conjuntos fuzzy para modelar esses dados, resultando no conjunto de treinamento $\{(X_i, y_i)\}_{i=1}^N \subset \mathcal{F}(\Omega) \times \{-1, 1\}$. Nesse contexto, é possível incluir a natureza dos dados, como o erro associado aos dados imprecisos, via a matriz kernel K , em diversos algoritmos de classificação binária como a SVM (apresentada no Problema 3.3.3), análise de discriminante de Fisher com kernels (Scholkopf e Mullert, 1999), entre outros.

Análise de regressão

Dado o conjunto de treinamento $\{(X_i, y_i)\}_{i=1}^N \subset \mathcal{F}(\Omega) \times \mathbb{R}$, a análise de regressão pode ser estendida para incluir imprecisão nos dados de entrada, modelada pelos conjuntos fuzzy X_i . Nesse sentido o foco da análise de regressão para dados fuzzy é estimar a função de regressão $f : \mathcal{F}(\Omega) \rightarrow \mathbb{R}$, a qual modela a dependência entre X_i e y_i . Ou seja, se $\{y_i\}_{i=1}^N$ forem i.d.d realizações da variável aleatória Y , e $\{X_i\}_{i=1}^N$ forem i.d.d realizações de variável aleatória fuzzy X , então $f(X_i) = \mathbb{E}[Y|X = X_i]$. A vantagem de usar conjuntos fuzzy é que, não somente será considerada a incerteza na saída dada por Y , como também a imprecisão na entrada modelada por X_i . Por exemplo, pode ser usada a matriz kernel K nas máquinas de vetores de suporte para regressão propostas em Smola e Schölkopf (2004), ou podem ser usados os *processos gaussianos* para regressão (Rasmussen e Williams, 2006), simplesmente colocando a matriz kernel gerada pelo kernel definido sobre conjuntos fuzzy, como a matriz de covariância do processo gaussiano.

Descrição de conjuntos de dados fuzzy

Se os elementos da amostra $\{X_i\}_{i=1}^N \subset \mathcal{F}(\Omega)$ forem consideradas realizações de uma variável aleatória fuzzy X tal que essa variável aleatória segue uma lei probabilística descrita por alguma medida de probabilidade $\mathbb{P}$. Então, é possível criar métodos de kernel de descrição de dados para conjuntos fuzzy. Essa descrição pode ser interpretada como a estimação de um conjunto de nível da densidade de $\mathbb{P}$ ou como uma estimativa do conjunto suporte da densidade de $\mathbb{P}$. Para isso, pode ser usada a matriz kernel K gerada pelas funções kernel definidas em conjuntos fuzzy em métodos como as *máquinas de vetores de suporte de uma classe* (Schölkopf et al., 2001) ou no *descriptor de dados usando vetores de suporte* (Tax e Duin, 2004).

Em geral, vários métodos de kernel podem ser estendidos para o caso de dados modelados por conjuntos fuzzy.

5.7 Experimentos

Nesta seção, apresentamos uma aplicação dos kernels definidos sobre conjuntos fuzzy, no prognóstico de reincidência do câncer de mama. Para isso usamos a técnica de Discrepância Máxima das Médias (MMD), nos EHKR, descrita na Seção 3.5.

5.7.1 Descrição do conjunto de dados de prognóstico de reincidência do câncer de mama

Pacientes submetidos a procedimentos cirúrgicos de câncer de mama possuem uma chance de 30% de reincidência da doença após cinco anos. O conjunto de dados *Breast Cancer* Lichman (2013)⁴, o qual obtido do Instituto de Oncologia do Centro Médico Universitário em Ljubljana, Yugoslavia⁵, contém dados de 286 pacientes, das quais 201 pacientes não apresentaram reincidência de câncer de mama após cinco anos, e as outras 85 pacientes apresentaram incidência de câncer de mama nesse período.

⁴Disponível em <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer>

⁵University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia.

O conjunto de dados contém doze atributos, os quais são apresentados na Tabela 5.3. É possível observar que os atributos *age*, *menopause*, *tumor-size* e *inv-node* contém valores imprecisos ou vagos, definidos por intervalos.

Atributo	intervalo de valores
class	no-recurrence-events, recurrence-events
age	10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99.
menopause	lt40, ge40, premeno.
tumor-size	0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59.
inv-nodes	0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39.
node-caps	yes, no.
deg-malig	1, 2, 3.
breast	left, right.
breast-quad	left-up, left-low, right-up, right-low, central.
irradiat	yes, no.

Tabela 5.3: Atributos do conjunto de dados Breast Cancer.

5.7.2 Fuzzificação dos atributos do conjunto de dados

Neste experimento, foram construídos arbitrariamente conjuntos fuzzy para representar os valores dos atributos *age*, *menopause*, *tumor-size* e *inv-node*. Para o atributo *age* foram usados conjuntos fuzzy com funções de pertinência trapezoidais, isto é:

$$X(x; a, b, c, d) = \begin{cases} 0, & x \leq a \\ \frac{x-a}{b-a}, & a \leq x \leq b \\ 1, & b \leq x \leq c \\ \frac{d-x}{d-c}, & c \leq x \leq d \\ 0, & d \leq x \end{cases} \quad (5.56)$$

Segundo o intervalo de valores desse atributo, foi usada a seguinte regra arbitrária para a definição dos parâmetros a, b, c, d . Denotando por l e r o menor e o maior valor de cada intervalo (ver intervalo de valores para o atributo *age* na tabela 5.3), fazemos, $a = l - 5$, $b = l$, $c = r$, $d = r + 5$. A razão de usar o valor cinco é por que o início (ou final) da função de pertinência trapezoidal irá estar no meio do intervalo anterior ou posterior. A Figura 5.1 mostra os conjuntos fuzzy trapezoidais usados para representar o intervalo de valores da variável *age*.

No caso da variável *menopause*, foi usado um conjunto fuzzy com função de pertinência *Z-shaped*

Figura 5.1: Fuzzificação da variável age, usando conjuntos fuzzy com funções de pertinência trapezoidais

para representar o valor *lq40* (menor do que 40 anos de idade), isto é:

$$X(x; a, b) = \begin{cases} 1, & x \leq a \\ 1 - 2\left(\frac{x-a}{b-a}\right)^2, & a \leq x \leq \frac{a+b}{2} \\ 2\left(\frac{x-a}{b-a}\right)^2, & \frac{a+b}{2} \leq x \leq b \\ 0, & x \geq b, \end{cases} \quad (5.57)$$

onde os parâmetros a e b assumem os valores de 40 e 45, respectivamente. A Figura 5.2 apresenta este conjunto fuzzy como uma curva de color vermelho. Foi usado também um conjunto fuzzy com função de pertinência *S-shaped* para representar o valor *ge40* (maior do que 40 anos de idade), isto é:

$$X(x; a, b) = \begin{cases} 0, & x \leq a \\ 2\left(\frac{x-a}{b-a}\right)^2, & a \leq x \leq \frac{a+b}{2} \\ 1 - 2\left(\frac{x-a}{b-a}\right)^2, & \frac{a+b}{2} \leq x \leq b \\ 1, & x \geq b, \end{cases} \quad (5.58)$$

onde os parâmetros a e b tomaram os valores de 35 e 40, respectivamente. A Figura 5.2 apresenta este conjunto fuzzy como uma curva de cor azul. Finalmente foi usado um conjunto fuzzy com função de pertinência Gaussiana para representar o valor *premeno*, isto é

$$X(x; \sigma, c) = \exp\left(-\frac{(x-c)^2}{2\sigma^2}\right) \quad (5.59)$$

onde o valor c foi estabelecido a ser 45 e o valor σ é dado pela expressão $(50 - 40)/(2\sqrt{2\log 2})$ ⁶. A Figura 5.2 apresenta este conjunto fuzzy como uma curva de cor amarelo.

Figura 5.2: Fuzzificação da variável menopause usando conjuntos fuzzy com funções de pertinência gaussiana, *S-shaped* e *Z-shaped*

⁶Esse valor é conhecido como *Largura à média altura* ou *full width at half maximum*.

Foram usados conjuntos fuzzy com funções de pertinência *Z-shaped* com $a = 0$ e $b = 9$, e *S-shaped* com $a = 50$ e $b = 59$ para representar os valores nos extremos da variável *tumor-size*. Também foram usados conjuntos fuzzy com funções de pertinência Gaussiana para representar os demais valores usando para eles os parâmetros $c = (l + r)/2$ e $\sigma = (r - l)/(2\sqrt{2\log 2})$. A figura 5.3 apresenta os conjuntos fuzzy usados para representar os valores deste atributo.

Figura 5.3: Fuzzificação da variável *tumor-size*, usando conjuntos fuzzy com funções de pertinência gaussiana, *S-shaped* e *Z-shaped*

Finalmente foram usados conjuntos fuzzy com funções de pertinência *Z-shaped* com $a = 0$ e $b = 5$, e *S-shaped* com $a = 33$ e $b = 39$ para representar os valores extremos da variável *inv-nodo*. Os demais valores foram representados com conjuntos fuzzy com funções de pertinência Gaussiana com parâmetros $c = (l + r)/2$ e $\sigma = (r - l)/(2\sqrt{2\log 2})$. A Figura 5.4 apresenta os conjuntos fuzzy usados para representar os valores deste atributo.

Figura 5.4: Fuzzificação da variável *inv-nodes*, usando conjuntos fuzzy com funções de pertinência gaussiana, *S-shaped* e *Z-shaped*

A Tabela 5.4, apresenta os valores usados para os outros atributos.

Foi construída também, uma versão *crisp* do conjunto de dados. Para isso, no caso dos atributos *age*, *tumor-size* e *inv-nodes* foram considerados os valores médios dos intervalos, isto é, $(l + r)/2$. No caso dos demais atributos, os valores são os mesmos da Tabela 5.4.

5.7.3 Teste kernel de duas amostras

Neste experimento, exploramos se a MMD (ver Seção 3.5), é capaz de distinguir entre amostras que pertencem a pacientes com reincidência de câncer de mama e amostras de pacientes que não apresentam essa reincidência (hipótese alternativa), usando alguns dos kernels apresentados neste capítulo. Para isso, o nível de significância do teste MMD foi estabelecido em $\alpha = 0.05$. Logo, o teste foi aplicado 100 vezes sobre 25 amostras escolhidas aleatoriamente de pacientes apresentando reincidência de câncer de mama e 25 amostras de pacientes que não apresentavam essa reincidência. Assim, foi contado o número de vezes em que o teste MMD aceita ou rejeita a hipótese alternativa. Para obter estatísticas de média e variância, o procedimento foi repetido 250 vezes.

atributo	valores	valores discretos
class	no-recurrence-events	1
	recurrence-events	-1
node-caps	yes	1
	no	-1
deg-malig	1	-1
	2	0
	3	1
breast	left	1
	right	-1
breast-quad	left-up	-1
	left-low	-0.5
	right-up	0
	right-low	0.5
	central	1
irradiat	yes	1
	no	-1

Tabela 5.4: Valores para os atributos node-caps, deg-malig, breast e irradiat.

Logo, exploramos se o teste MMD pode determinar quando duas amostras pertencem à mesma distribuição (hipótese nula), usando alguns dos kernels apresentados neste capítulo. Para isso, o teste foi aplicado 100 vezes sobre 25 pares de amostras pertencentes à mesma classe. Para obter as estatísticas de média e variância, o procedimento foi repetido 250 vezes.

No caso da versão *crisp* do conjunto de dados, foram usados o kernel linear, denotado como k_{lin} , e o kernel gaussiano, denotado como k_{RBF} . O parâmetro para este ultimo foi computado usando a heurística da mediana, ou seja, como o inverso da mediana de todas as possíveis distâncias entre os elementos do conjunto de dados. No caso do conjunto de dados fuzzificado, como ele contém dados heterogêneos, ou seja fuzzy e crisp, (ver Tabela 5.3), foi usado uma combinação de kernels definidos sobre conjuntos fuzzy e kernels convencionais como o kernel linear e o kernel gaussiano. Foi usada a somatória de quatro kernels de interseção de conjuntos fuzzy⁷ (um kernel fuzzy por cada atributo que contem conjuntos fuzzy, ver atributos *age*, *menopause*, *tumor-size* e *inv-nodes* na Tabela 5.3), com um kernel linear definido em todos os demais atributos. Este kernel foi denotado como $k_{\cap} + k_{lin}$, note que este kernel não tem seleção de parâmetro.

Foi usado também o somatório de quatro kernels de interseção fuzzy com um kernel gaussiano definido nos atributos *crisp*. Este kernel foi chamado de $k_{\cap} + k_{RBF}$. O parâmetro deste kernel foi obtido pela heurística da mediana, considerando somente as dimensões tipo *crisp*.

Finalmente foi usado o kernel de substituição de distância entre conjuntos fuzzy, onde a métrica:⁸

$$D(X, X') = \frac{\sum_{x \in \Omega} |X(x) - X'(x)|}{\sum_{x \in \Omega} |X(x) + X'(x)|},$$

foi usada no cálculo de $K_D(X, X') = \exp(-\lambda D(X, X')^2)$. O parâmetro deste kernel foi computado pela a heurística da mediana usando a versão *crisp* do conjunto de dados e considerando somente dimensões com valores dados por conjuntos fuzzy. Foi somado a esse kernel, o kernel gaussiano definido nos atributos do tipo *crisp* e com parâmetro computado com a heurística da mediana considerando somente as dimensões do tipo *crisp*. Este kernel foi denotado como $K_D + k_{RBF}$.

5.7.4 Resultados

A Figura 5.5, apresenta o erro do Tipo II versus o nível de significância do teste α . É possível observar que o kernel $k_D + k_{RBF}$ é o kernel que apresenta menor erro do tipo II para distintos valores α .

Os resultados para $\alpha = 0.05$, são apresentados na Tabela 5.5. É possível observar que todos os testes MMD tem pouco erro do Tipo I (a condição de que as amostras pertencem à mesma distribuição), os diferentes testes MMD apresentam erros abaixo de 3%. Já no caso do erro do Tipo II, (a condição de que as amostras pertencem a distribuições diferentes), somente o teste MMD usando o kernel com substituição de distância fuzzy apresenta erro menor com relação aos outros testes.

Finalmente, as Figuras 5.6 e 5.7, apresentam os erros do Tipo I e Tipo II usando boxplots.

⁷A implementação foi feita com a função mínimo como T-norma

⁸Ver Tabela 5.2.

Figura 5.5: Nível de significância α vs. O Erro do Tipo II, usando a MMD com cinco diferentes kernels.

Condição	H_0	k_{lin}	k_{RBF}	$k_{\cap} + k_{lin}$	$k_{\cap} + k_{RBF}$	$k_D + k_{RBF}$
Igual	aceita	97.63	97.98	97.38	97.51	97.92
Igual	rejeitada	02.37	02.02	02.62	02.49	02.08
Diferente	aceita	73.79	73.68	92.78	93.85	61.09
Diferente	rejeitada	26.21	26.32	07.22	06.15	38.91

Tabela 5.5: Resultados. O H_0 denota a hipótese nula.

Figura 5.6: Boxplots contendo informação do erro de Tipo I.

Figura 5.7: Boxplots contendo informação do erro de Tipo II.

5.8 Sumário

O conjunto dos conjuntos fuzzy não formam um espaço vetorial. Porém, é possível analisar esses dados mapeando eles para um EHKKR. O mapeamento é feito implicitamente via um kernel positivo definido, definido sobre conjuntos fuzzy. Foram estendidos a partir de kernels definidos sobre conjuntos, o kernel de produto cruzado de conjuntos fuzzy e o kernel de convolução fuzzy. Também, usando os conceitos de semi-anel de conjuntos e medida definida nesse semi-anel, foi apresentado o kernel de interseção de conjuntos fuzzy. Foi também descrito como construir novos kernels definidos sobre conjuntos fuzzy a partir da substituição de distâncias entre conjuntos fuzzy. Apresentamos potenciais aplicações práticas desses kernels em distâncias entre conjuntos fuzzy, problemas de classificação, regressão, descrição de conjuntos de dados fuzzy, e engenharia de novos kernels definidos sobre conjuntos fuzzy a partir dos kernels apresentados. Isto é, os kernels definidos sobre conjuntos fuzzy podem ser usados modularmente em vários métodos de kernel como as máquinas de vetores de suporte, a descrição de dados usando vetores de suporte, regressão com vetores de suporte, processos gaussianos, kernel PCA, e assim por diante. Para isso somente requerem uma matriz Gram positiva definida induzida por algum kernel positivo definido, definido sobre conjuntos fuzzy. Finalmente, foi apresentado um experimento usando o MMD para determinar se os kernels apresentados podem ser utilizados em testes de estatísticos de duas amostras.

Capítulo 6

Kernels em Conjuntos Fuzzy baseados no Sistema de Regras Fuzzy Takagi Sugeno Kang

As complexity rises, precise statements lose meaning and meaningful statements lose precision.

To what degree is something true or false?

Zadeh.

Neste capítulo, apresentamos kernels positivos definidos, definidos sobre conjuntos fuzzy, derivados a partir da interação entre as regras de um sistema fuzzy do tipo Takagi Sugeno Kang (TSK) [Takagi e Sugeno \(1985\)](#), e as entradas associadas a esse tipo de sistema fuzzy. Como as entradas ao sistema fuzzy serão modeladas como conjuntos fuzzy do tipo non-singleton, a derivação dos kernels será feita usando uma extensão do sistema fuzzy TSK para processar esse tipo de entradas.

6.1 Introdução

A literatura em aprendizado computacional usando conceitos da teoria fuzzy e kernels positivo definidos para a análise de dados (não necessariamente dados imprecisos e vagos modelados com conjuntos fuzzy), apresenta duas principais vertentes. Na primeira delas, o uso de conceitos de teoria fuzzy não implicam a criação de kernels positivos definidos. Em outras palavras, técnicas específicas da teoria fuzzy e kernels positivos definidos são usados de maneira conjunta para solucionar diversos problemas de aprendizado ([Heo e Gader, 2011](#); [Huang et al., 2012](#); [Kanzawa et al., 2010](#); [Raza e Rhee, 2012](#); [Wang et al., 2010](#); [Yang et al., 2011](#)). Na segunda vertente, o uso da teoria fuzzy induz kernels positivos definidos, isto é, diversos kernels positivo definidos podem ser induzidos a partir de conceitos da teoria fuzzy. A pesquisa neste grupo pode ser subdividida em:

(1) Kernels e regras fuzzy. Usando operadores T-norma (Capítulo 2), a interação entre sistemas de regras fuzzy e suas entradas pode ser vista como a avaliação de um kernel positivo definido (Chen, 2008; Chen e Wang, 2003). Diversos trabalhos usam essa propriedade para treinar classificadores fuzzy usando máquinas de vetores de suporte (Cheng e Juang, 2011; Jeng e Lee, 1999; Juang *et al.*, 2007; Lin *et al.*, 2006; Wee e Wan, 2008; Zhou e Gan, 2007).

(2) Kernels e funções base de sistemas fuzzy. O produto interno entre vetores cujos componentes são avaliações de funções base de sistemas fuzzy (Wang e Mendel, 1992), definem um kernel positivo definido. Essas funções base são obtidas a partir de operadores de implicação fuzzy (Lski, 2006).

(3) Kernels e relações de equivalência fuzzy. Kernels positivo definidos, tomando valores no intervalo unitário com valor constante na diagonal, são relações de equivalência fuzzy ou vice versa, com algumas hipóteses (Moser, 2006a,b). O trabalho Liu e Xue (2012) apresenta como o conhecimento a priori de algum problema pode ser incluído nos kernels como relações de equivalência fuzzy.

6.1.1 Contribuições

Até onde conhecemos, os trabalhos prévios relacionando a iteração das regras fuzzy e as entradas no sistema fuzzy com avaliações de kernels positivo definidos foram feitos somente no contexto de entradas no sistema fuzzy modeladas por conjuntos fuzzy singleton. Porém, os conjuntos fuzzy non-singleton são os mais adequados para modelar dados não precisos, além disso, os conjuntos de dados fuzzy, implicitamente, contém conjuntos non-singleton como elementos. Assim levamos em consideração esse tipo de conjuntos na análise dos sistemas fuzzy Takagi Sugeno Kang. As contribuições deste capítulo são as seguintes:

- Extensão dos sistemas fuzzy TSK para o caso cujas entradas são conjuntos fuzzy non-singleton.
- Apresentamos como a interação das regras do sistema fuzzy TSK non-singleton e as entradas dadas por conjuntos non-singleton são avaliações de kernels definidos sobre conjuntos fuzzy.
- Mostramos como o kernel TSK non-singleton pode ser visto como uma relação de equivalência fuzzy.

6.2 Sistema Fuzzy Takagi Sugeno Kang

Os sistemas TSK (Takagi e Sugeno, 1985), usam os antecedentes das regras (Definição 2.4.2, Capítulo 2) para dividir o espaço de entrada do sistema em regiões. Assim, cada regra define uma região e, o comportamento do sistema nessa região é descrito usando uma função de valor real. Diferentemente do sistema de Mamdani, a saída do sistema fuzzy TSK é um valor real e não um conjunto fuzzy.

Os sistemas fuzzy TSK possuem as seguintes características:

- consideram o fuzzificador singleton (Definição 2.4.1) de forma implícita;

- os antecedentes das regras são da forma da Equação (2.14), portanto podem ser representados por relações fuzzy segundo a Definição 2.4.2;
- Os consequentes das regras são modeladas por funções de valor real;
- consideram a saída do sistema como um valor real, portanto não precisam usar um defuzzificador.

Definido os conjuntos de índices $\mathcal{D} = \{1, \dots, d, \dots, D\}$ e $\mathcal{L} = \{1, \dots, \ell, \dots, L\}$, temos que as regras dos sistemas fuzzy TSK são da forma:

$$\underbrace{\text{If } X_1^\ell \text{ and } \dots \text{ and If } X_d^\ell \text{ and } \dots \text{ and If } X_D^\ell}_{\text{antecedente}} \quad \underbrace{\text{Then } g^\ell(\mathbf{x})}_{\text{consequente}}, \quad (6.1)$$

para todo $d \in \mathcal{D}$, $\ell \in \mathcal{L}$, onde $X_d^\ell \in \mathcal{F}(\Omega_d)$, e g^ℓ é a função de valor real, definida em $\Omega_1 \times \Omega_2 \times \dots \times \Omega_D$, que define o comportamento do sistema fuzzy TSK na região definida pelo antecedente da regra fuzzy ℓ .

Usando uma relação fuzzy para representar o antecedente da regras (Definição 2.4.2, Capítulo 2) conjuntamente com a função g^ℓ que modela o consequente, as regras dos sistemas fuzzy TSK podem ser representadas por funções, como apresentamos a continuação.

Definição 6.2.1. A regra $\ell \in \mathcal{L}$ do sistema fuzzy TSK (6.1) é representada pela função:

$$\begin{aligned} \Omega_1 \times \Omega_2 \times \dots \times \Omega_D &\rightarrow \mathbb{R} \\ \mathbf{x} &\mapsto A^\ell(\mathbf{x})g^\ell(\mathbf{x}) = T\left(X_1^\ell(x_1), X_2^\ell(x_2), \dots, X_D^\ell(x_D)\right)g^\ell(\mathbf{x}), \end{aligned} \quad (6.2)$$

onde A^ℓ é o conjunto fuzzy da Definição 2.4.2 que representa o antecedente de (6.1), g^ℓ é a função de valor real definida em $\Omega_1 \times \Omega_2 \times \dots \times \Omega_D$, que representa o consequente da regra, e $\mathbf{x} = x_1, x_2, \dots, x_D$.

Desta esta forma, usando a definição anterior, o sistema fuzzy TSK define funções de valor real, definidas em $\Omega_1 \times \Omega_2 \times \dots \times \Omega_D$ do tipo:

$$f_s(\mathbf{x}) = \frac{\sum_{\ell=1}^L A^\ell(\mathbf{x})g^\ell(\mathbf{x})}{\sum_{\ell=1}^L A^\ell(\mathbf{x})}. \quad (6.3)$$

6.3 Sistema Fuzzy TKS non-singleton

Nesta seção, apresentamos uma extensão do sistema fuzzy TSK, chamada de sistema fuzzy TSK non-singleton ou *Sistema fuzzy TSK-NS*, para o caso de entradas modeladas com conjuntos fuzzy non-singleton. Essa nova formulação inclui como caso particular o sistema fuzzy TSK da Seção (6.2), o qual processa entradas modeladas com conjuntos singleton (Mendel, 2001).

Definição 6.3.1 (Sistema Fuzzy TSK-NS). O sistema fuzzy TSK-NS, possui L regras da forma (6.1), onde cada uma delas define funções como as da Definição (6.2.1). A entrada ao sistema é dada pela fuzzificação non-singleton de $\mathbf{x}' \in \Omega_1 \times \dots \times \Omega_D$ para o conjunto fuzzy X . O sistema

fuzzy TSK-NS define um mapeamento de valor real definido em $\mathcal{F}(\Omega_1 \times \Omega_2 \times \cdots \times \Omega_D)$, do tipo:

$$f_{ns}(X; \mathbf{x}'') = \frac{\sum_{\ell=1}^L \sup_{\mathbf{x} \in \Omega_1 \times \cdots \times \Omega_D} (T(X(\mathbf{x}), A^\ell(\mathbf{x}))) g^\ell(\mathbf{x}'')}{\sum_{\ell=1}^L \sup_{\mathbf{x} \in \Omega_1 \times \cdots \times \Omega_D} T(X(\mathbf{x}), A^\ell(\mathbf{x}))}, \quad (6.4)$$

onde $\mathcal{F}(\Omega_1 \times \cdots \times \Omega_D)$ é a família de conjuntos fuzzy definidas em $\Omega_1 \times \cdots \times \Omega_D$. O valor $\mathbf{x}'' \in \text{supp}(X)$ é o parâmetro de f_{ns} .

O seguinte esquema ilustra o processo de fuzzificação não singleton e a saída do sistema fuzzy TSK-NS.

$$\mathbf{x}' \xrightarrow{\text{Fuzzificador non-singleton}} X \xrightarrow{\text{Sistema fuzzy TSK-NS}} f_{ns}(X; \mathbf{x}'')$$

Note que o valor $\mathbf{x}''$ pode ser qualquer ponto no suporte de X . Por exemplo $\mathbf{x}'' = \mathbf{x}_{sup}$, onde $\mathbf{x}_{sup}$ é o valor onde $\sup_{\mathbf{x} \in \Omega_1 \times \cdots \times \Omega_D} T(X(\mathbf{x}), A^\ell(\mathbf{x}))$ é atingido. Se $\mathbf{x}'' = \mathbf{x}'$, então, f_{ns} é uma função.

Devido a que o fuzzificador transforma $\mathbf{x}'$ em X , é possível dizer que o sistema fuzzy TSK-NS define equivalentemente os mapeamentos de valor real $\tilde{f}_{ns}$, definidos em $\Omega_1 \times \cdots \times \Omega_D$ por:

$$\tilde{f}_{ns}(\mathbf{x}'; \mathbf{x}'') = \frac{\sum_{\ell=1}^L \sup_{\mathbf{x} \in \Omega_1 \times \cdots \times \Omega_D} (T(X(\mathbf{x}), A^\ell(\mathbf{x}))) g^\ell(\mathbf{x}'')}{\sum_{\ell=1}^L \sup_{\mathbf{x} \in \Omega_1 \times \cdots \times \Omega_D} T(X(\mathbf{x}), A^\ell(\mathbf{x}))}, \quad \mathbf{x}'' \in \text{supp}(X), \quad (6.5)$$

note que $\mathbf{x}'$ não aparece explicitamente no lado direito de (6.5), no lugar disso aparece X como representante de $\mathbf{x}'$. Assim podemos dizer que $f_{ns} \equiv \tilde{f}_{ns}$.

O sistema fuzzy TSK-NS pode ser reduzido para o sistema fuzzy TSK da Seção 6.2, como é apresentado no seguinte lema.

Lema 18. *Se o fuzzificador do sistema fuzzy TSK produz conjuntos singleton normal, então (6.5) é equivalente a (6.3).*

Demonstração. Seja $\mathbf{x}' \in \Omega_1 \times \cdots \times \Omega_D$ a entrada no sistema fuzzy TSK, e X , o conjunto fuzzy singleton, resultado da fuzzificação singleton de $\mathbf{x}'$. Então, X satisfaz: i) $\text{supp}(X) = \{\mathbf{x}'\}$, e ii) $X(\mathbf{x}') = 1$, pois X é um conjunto fuzzy normal. Então,

$$\sup_{\mathbf{x} \in \Omega_1 \times \cdots \times \Omega_D} T(X(\mathbf{x}), A^\ell(\mathbf{x})) = T(X(\mathbf{x}_{sup}), A^\ell(\mathbf{x}_{sup})).$$

Como X é singleton e normal, então $\mathbf{x}_{sup} = \mathbf{x}'$. Portanto:

$$\begin{aligned} \sup_{\mathbf{x} \in \Omega_1 \times \cdots \times \Omega_D} T(X(\mathbf{x}), A^\ell(\mathbf{x})) &= T(X(\mathbf{x}_{sup}), A^\ell(\mathbf{x}_{sup})) \\ &= T(X(\mathbf{x}'), A^\ell(\mathbf{x}')) \\ &= T(1, A^\ell(\mathbf{x}')) \\ &= A^\ell(\mathbf{x}'). \end{aligned} \quad (6.6)$$

Logo, como $\mathbf{x}'' \in \text{supp}(X)$, e $\text{supp}(X) = \{\mathbf{x}'\}$, implica $\mathbf{x}'' = \mathbf{x}'$. Finalmente, substituindo (6.6) e $\mathbf{x}'' = \mathbf{x}'$ em (6.4), resulta em (6.3) com $\mathbf{x}$ no lugar de $\mathbf{x}'$. $\square$

6.4 Kernels baseados em regras fuzzy

Nesta seção apresentamos como construir kernels baseados na iteração das entradas no sistema fuzzy e as regras fuzzy do sistema. Começamos com uma revisão de do kernel PDFC (Chen e Wang, 2003) e estenderemos o kernel apresentado em (Cheng e Juang, 2011; Juang *et al.*, 2007) para o caso mas geral. Posteriormente, definimos um kernel definido sobre conjuntos fuzzy, o qual é resultado da análise da iteração das entradas e o sistema fuzzy TSK-NS da Seção 6.3. Baseado nos fatos apresentados, mostramos como os sistemas fuzzy representam funções no EHKR.

6.4.1 Kernel PDFC

O kernel PDFC (Chen e Wang, 2003) é um kernel de valor real, definido em $\mathbb{R}^D \times \mathbb{R}^D$, o qual pode ser deduzido a partir da representação dos antecedentes das regras dos sistema fuzzy dos sistemas Mamdani e TSK segundo a Definição 2.4.2.

Lema 19 (PDFC kernel Chen e Wang (2003)). *Se para todo $d \in \mathcal{D}$ e $\ell \in \mathcal{L}$, cada um dos conjuntos fuzzy X_d^ℓ , do antecedente da regra (2.16) foram gerados transladando as funções positivas definidas: $u_d : \mathbb{R} \rightarrow [0, 1]$, $d \in \mathcal{D}$, definidas por:*

$$u_d(x) = \begin{cases} u_d(-x) & \text{if } x \neq 0 \\ 1 & \text{if } x = 0, \end{cases}, x \in \mathbb{R}, \quad (6.7)$$

e se $\mathbf{x} = (x_1, x_2, \dots, x_D) \in \mathbb{R}^D$, for uma entrada no sistema fuzzy. Se $\mathbf{x}^\ell = (x_1^\ell, x_2^\ell, \dots, x_D^\ell) \in \mathbb{R}^D$, forem parâmetros de deslocamento dos conjuntos fuzzy do antecedente da regra, tais que $X_d^\ell(x_d) = u_d(x_d - x_d^\ell)$, e se a T -norma usada em (2.17) for a função produto, então o kernel: $k_{pdfc} : \mathbb{R}^D \times \mathbb{R}^D \rightarrow \mathbb{R}$, definido por:

$$\begin{aligned} k_{pdfc}(\mathbf{x}, \mathbf{x}^\ell) &= \prod_{d=1}^D u_d(x_d - x_d^\ell) \\ &= \prod_{d=1}^D X_d^\ell(x_d) \\ &= A^\ell(\mathbf{x}), \end{aligned} \quad (6.8)$$

será positivo definido;

A prova pode ser encontrada em Chen e Wang (2003) (Teorema 3.11).

6.4.2 Kernel Takagi Sugeno Kang (TSK)

Nos trabalhos Cheng e Juang (2011); Juang *et al.* (2007), observou-se que se todos os conjuntos fuzzy dos antecedentes das regras de um sistema fuzzy TSK forem funções gaussianas, e se for usada a função produto como T -norma e, se o consequente fosse a função $g^\ell(\mathbf{x}) = \langle \mathbf{x}, \mathbf{x}^\ell \rangle$, onde $\mathbf{x}^\ell$ é o parâmetro de g^ℓ , então, as funções definidas pelas regras do sistema fuzzy TSK (Definição 6.2.1), serão kernels positivos definidos, isto é, $\forall \ell \in \mathcal{L}$

$$\begin{aligned} A^\ell(\mathbf{x})g^\ell(\mathbf{x}) &= \prod_{d=1}^D \exp(-\gamma(x_d - x_d^\ell)^2) \langle \mathbf{x}, \mathbf{x}^\ell \rangle \\ &= \exp(-\gamma \|\mathbf{x} - \mathbf{x}^\ell\|^2) \langle \mathbf{x}, \mathbf{x}^\ell \rangle \\ &= k_{tsk}(\mathbf{x}, \mathbf{x}^\ell), \end{aligned}$$

onde $\gamma \in \mathbb{R}^+$, $\mathbf{x} \in \mathbb{R}^D$ representa a entrada do sistema fuzzy TSK, $\mathbf{x}^\ell \in \mathbb{R}^D$ são parâmetros dos conjuntos fuzzy do antecedente das regras e, $k_{tsk} : \mathbb{R}^D \times \mathbb{R}^D \rightarrow \mathbb{R}$, é o kernel positivo definido, definido pelas regras dos sistemas fuzzy TSK.

O seguinte Lema generaliza esse resultado.

Lema 20 (Kernels TSK). *Seja $\mathbf{x} \in \mathbb{R}^D$ a entrada do sistema fuzzy TSK e $\mathbf{x}^\ell \in \mathbb{R}^D$ os parâmetros dos conjuntos fuzzy do antecedente das regras. Se A^ℓ e g^ℓ forem kernels positivo definidos, isto é $A^\ell(\mathbf{x}) = k_1(\mathbf{x}, \mathbf{x}^\ell)$, $g^\ell(\mathbf{x}) = k_2(\mathbf{x}, \mathbf{x}^\ell)$, onde k_1 e k_2 são kernels positivos definidos, então a função definida pelas regras de um sistema fuzzy TSK segundo a Definição 6.2.1, é um kernel positivo definido.*

Demonstração. O produto de dois kernels positivo definidos é positivo definido, segundo as propriedades de fechamento dos kernels do Capítulo 3. $\square$

Definição de funções em EHKR pelos sistemas fuzzy TSK

O uso de kernels positivos definidos nos sistemas fuzzy TSK faz que as funções (6.3) definidas pelos sistemas fuzzy TSK pertençam a um EHKR $\mathcal{H}$ de funções $f : \mathbb{R}^D \rightarrow \mathbb{R}$, com kernel k_{tsk} ou kernel k_{pdfc} , segundo seja o kernel considerado. No caso do kernel k_{tsk} , basta substituir $A^\ell(\mathbf{x})g^\ell(\mathbf{x})$ por $k_{tsk}(\mathbf{x}, \mathbf{x}^\ell)$ no numerador de (6.3), para que (6.3) possa ser escrita como:

$$f(\mathbf{x}) = \sum_{\ell=1}^L \alpha k_{tsk}(\mathbf{x}, \mathbf{x}^\ell) \in \mathcal{H}, \quad \alpha = \frac{1}{\sum_{\ell=1}^L A^\ell(\mathbf{x})}.$$

Se for considerado o kernel k_{pdfc} , então substituindo $A^\ell(\mathbf{x})$ por $k_{pdfc}(\mathbf{x}, \mathbf{x}^\ell)$ em (6.3), temos que (6.3) pode ser escrita como:

$$f(\mathbf{x}) = \sum_{\ell=1}^L \alpha_\ell k_{pdfc}(\mathbf{x}, \mathbf{x}^\ell) \in \mathcal{H}, \quad \alpha_\ell = \frac{g^\ell(\mathbf{x})}{\sum_{\ell=1}^L A^\ell(\mathbf{x})}.$$

A importância desse fato esta na possibilidade de modelar problemas usando termos linguísticos num sistema fuzzy, conjuntamente com os métodos de kernel para analisar os dados desse problema.

6.5 Kernels TSK non-singleton

Nesta seção consideramos funções kernels induzidas pelos sistemas fuzzy TSK-NS apresentados na Seção 6.3. Estes kernels, diferentemente dos kernels apresentados na Seção 6.4, tem como domínio de definição o espaço dos conjuntos fuzzy, isto é, são funções kernels de valor real, definidos em $\mathcal{F}(\Omega) \times \mathcal{F}(\Omega)$ onde $\Omega = \Omega_1 \times \dots \times \Omega_D$. Por simplicidade na notação, nesta seção sempre simplificamos $\Omega_1 \times \dots \times \Omega_D$ pelo simbolo Ω .

Nos sistemas fuzzy TSK-NS (Definição 6.3.1) a interação entre o conjunto fuzzy non-singleton de entrada X e cada regra A^ℓ é computada como:

$$\sup_{\mathbf{x} \in \Omega} \left(T(X(\mathbf{x}), A^\ell(\mathbf{x})) \right)$$

baseado nesse fato, apresentamos o seguinte kernel em conjuntos fuzzy.

Definição 6.5.1 (Kernel TSK non-singleton). O kernel TSK non-singleton é uma função $\mathcal{F}(\Omega) \times \mathcal{F}(\Omega) \rightarrow [0, 1]$ definida por:

$$k_{ntsk}(X, X') = \sup_{\mathbf{x} \in \Omega} \left(T(X(\mathbf{x}), X'(\mathbf{x})) \right). \quad (6.9)$$

Como a função T-norma é usada para implementar a interseção de conjuntos fuzzy, então podemos dizer que k_{ntsk} é um kernel baseado na interseção dos conjuntos fuzzy X, X' . i.e., $k_{ntsk}(X, X') = \sup_{\mathbf{x} \in \Omega} X \cap X'(\mathbf{x})$.

Lema 21. Se $\mathcal{F}$ contém somente conjuntos fuzzy normais, então k_{ntsk} é positivo definido.

Demonstração. Denotamos k_{ntsk} por k para simplificar a notação. Seja I um conjunto arbitrário de índices. Pela propriedade comutativa da T-norma, k é simétrico. Note também que:

$$\sum_{i,j \in I} c_i c_j k(X_i, X_j) = \sum_{i \in I} c_i^2 k(X_i, X_i) + 2 \sum_{i > j, i, j \in I} c_i c_j k(X_i, X_j).$$

Como $\sup_{x \in \Omega} T(X_i(x), X_i(x)) = 1, \forall i \in I$, então,

$$\sum_{i,j \in I} c_i c_j k(X_i, X_j) = \sum_{i \in I} c_i^2 + 2 \sum_{i > j, i, j \in I} c_i c_j k(X_i, X_j).$$

Usando $(\sum_{i \in I} c_i)^2 = \sum_{i \in I} c_i^2 + 2 \sum_{i > j, i, j \in I} c_i c_j \geq 0$, e pelo fato que $k(X_i, X_j) \in [0, 1]$, temos que:

a) Se $k(X_i, X_j) = 0, \forall i, j \in I, i > j$ e, portanto,

$$\sum_{i,j \in I} c_i c_j k(X_i, X_j) = \sum_{i \in I} c_i^2 \geq 0.$$

b) Se $k(X_i, X_j) = 1, \forall i, j \in I, i > j$ e, portanto,

$$\begin{aligned} \sum_{i,j \in I} c_i c_j k(X_i, X_j) &= \sum_{i \in I} c_i^2 + 2 \sum_{i,j \in I, i \neq j, i > j} c_i c_j \\ &= \left(\sum_{i \in I} c_i \right)^2 \geq 0. \end{aligned}$$

Assim, k é positivo definido. □

Lema 22. O kernel k_{ntsk} pode ser escrito como:

$$k_{ntsk}(X, X') = T \left(\sup_{x_1 \in \Omega_1} T(X_1(x_1), X'_1(x_1)), \dots, \sup_{x_D \in \Omega_D} T(X_D(x_D), X'_D(x_D)) \right)$$

Demonstração.

$$\begin{aligned}
 k_{ntsk}(X, X') &= \sup_{\mathbf{x} \in \Omega} (T(X(\mathbf{x}), X'(\mathbf{x}))) \\
 &= \sup_{\mathbf{x} \in \Omega} (T(X_1(x_D), X'_1(x_1), \dots, X_D(x_D), X'_D(x_D))) \\
 &= \sup_{\mathbf{x} \in \Omega} \left(T \left(T(X_1(x_D), X'_1(x_1)), \dots, T(X_D(x_D), X'_D(x_D)) \right) \right) \\
 &= T \left(\sup_{x_1 \in \Omega_1} T(X_1(x_1), X'_1(x_1)), \dots, \sup_{x_D \in \Omega_D} T(X_D(x_D), X'_D(x_D)) \right)
 \end{aligned}$$

Onde a segunda e terceira linha são graças à extensão multi-argumento e as propriedades associativa e comutativa da T-norma. $\square$

Considerando cada consequente das regras dos sistemas fuzzy TSK como kernels positivo definidos, é possível definir um novo kernel definido em $\mathcal{F}(\Omega) \times \Omega \times \mathcal{F}(\Omega) \times \Omega$, como apresentamos em seguida.

Definição 6.5.2 (Kernel TSK non-singleton composto). O kernel TSK non-singleton composto é a função: $k_{ntsk} \otimes k : (\mathcal{F}(\Omega) \times \Omega) \times (\mathcal{F}(\Omega) \times \Omega) \rightarrow [0, 1]$, definida por:

$$k_{ntsk} \otimes k((X, \mathbf{x}^I), (X^\ell, \boldsymbol{\theta}^\ell)) = \sup_{\mathbf{x} \in \Omega} (T(X(\mathbf{x}), X^\ell(\mathbf{x}))) g^\ell(\mathbf{x}^I; \boldsymbol{\theta}^\ell) \quad (6.10)$$

$$= k_{ntsk}(X, X^\ell) k(\mathbf{x}^I, \boldsymbol{\theta}^\ell), \quad (6.11)$$

onde, $\mathbf{x}^I$ representa a entrada no sistema fuzzy TSK-NS, X é o conjunto fuzzy non-singleton resultado da fuzzificação de $\mathbf{x}^I$, O mapeamento $g^\ell(\cdot, \boldsymbol{\theta}) : \Omega \rightarrow \mathbb{R}$ é a função representando o consequente do sistema fuzzy TSK-NS da Definição 6.3.1 parametrizada por $\boldsymbol{\theta}^\ell \in \Omega$. O conjunto fuzzy X^ℓ é o vetor fuzzy que representa o antecedente das regras fuzzy, e $k : \Omega \times \Omega$ é um kernel positivo definido de valor real.

Lema 23. O kernel $k_{ntsk} \otimes k$ é positivo definido se k_{ntsk} e k forem kernels positivos definidos

Demonstração. O produto de dois kernels positivo definidos é positivo definido. $\square$

6.5.1 Kernel gaussiano TSK non-singleton

Neste seção, exploramos um kernel TSK non-singleton induzido quando todos os conjuntos fuzzy dos antecedentes das regras e os provenientes do fuzzificador forem funções Gaussianas. Usando a definição de vetor fuzzy (Definição 2.3.7) para os elementos de $\mathcal{F}(\Omega)$, define-se $\forall d \in \mathcal{D}$:

$$X_d(x_d) = \exp \left(-\frac{1}{2} \frac{(x_d - m_d)^2}{\sigma_d^2} \right),$$

tal que $X_d \in \mathcal{F}(\mathbb{R})$, $x_d \in \mathbb{R}$, e onde os valores $m_d \in \mathbb{R}$ e $\sigma_d \in \mathbb{R}^+$, são os parâmetros do conjunto fuzzy X_d . Logo, se para todo $X \in \mathcal{F}(\Omega)$ fosse usada a função produto como T-norma, segundo a Definição 2.3.7, então,

$$X(\mathbf{x}) = \prod_{d=1}^D \exp \left(-\frac{1}{2} \frac{(x_d - m_d)^2}{\sigma_d^2} \right).$$

Então, o kernel k_{ntfs} da Definição 6.5.1, definido para os elementos de $\mathcal{F}(\Omega)$ é dado por:

$$\begin{aligned}
k_{ntsk}(X, X') &= \sup_{\mathbf{x} \in \mathbb{R}^D} (X(\mathbf{x})X'(\mathbf{x})) \\
&= \sup_{\mathbf{x} \in \mathbb{R}^D} \left(\prod_{d=1}^D \exp \left(-\frac{1}{2} \frac{(x_d - m_d)^2}{\sigma_d^2} \right) \prod_{d=1}^D \exp \left(-\frac{1}{2} \frac{(x_d - m'_d)^2}{(\sigma'_d)^2} \right) \right) \\
&= \sup_{\mathbf{x} \in \mathbb{R}^D} \left(\prod_{d=1}^D \exp \left(-\frac{1}{2} \frac{(x_d - m_d)^2}{\sigma_d^2} - \frac{1}{2} \frac{(x_d - m'_d)^2}{(\sigma'_d)^2} \right) \right) \\
&= \prod_{d=1}^D \sup_{x_d \in \mathbb{R}} \left(\exp \left(-\frac{1}{2} \frac{(x_d - m_d)^2}{\sigma_d^2} - \frac{1}{2} \frac{(x_d - m'_d)^2}{(\sigma'_d)^2} \right) \right) \\
&= \prod_{d=1}^D \exp \left(-\frac{1}{2} \frac{(x_d^{sup} - m_d)^2}{\sigma_d^2} - \frac{1}{2} \frac{(x_d^{sup} - m'_d)^2}{(\sigma'_d)^2} \right). \tag{6.12}
\end{aligned}$$

Onde a quarta linha é graças ao Lema 22. Logo, o valor x_d^{sup} pode ser encontrado derivando e igualando a zero a seguinte expressão.

$$h(x_d) = -\frac{1}{2} \frac{(x_d - m_d)^2}{\sigma_d^2} - \frac{1}{2} \frac{(x_d - m'_d)^2}{(\sigma'_d)^2},$$

ou seja,

$$\frac{d}{dh(x_d)} \exp(h(x_d)) = \frac{d}{dh(x_d)} \exp(h(x_d)) \frac{d}{dx_d} h(x_d) = \exp(h(x_d)) \frac{d}{dx_d} h(x_d) = 0,$$

como a função $\exp(h(x_d)) \neq 0, \forall x_d \in \mathbb{R}$, então $\frac{d}{dx_d} h(x_d) = 0$, logo calculando o ponto $x_d^{sup} \in \mathbb{R}$ que satisfaz $\frac{d}{dx_d} h(x_d) = 0$, temos que:

$$x_d^{sup} = \frac{m_d(\sigma'_d)^2 + m'_d \sigma_d^2}{\sigma_d^2 + (\sigma'_d)^2}, \tag{6.13}$$

sustituindo x_d^{sup} em (6.12) temos que o kernel TSK non-singleton usando funções gaussianas é dado por:

$$k_{ntsk}(X, X') = \prod_{d=1}^D \exp \left(-\frac{1}{2} \frac{(m_d - m'_d)^2}{\sigma_d^2 + (\sigma'_d)^2} \right), \tag{6.14}$$

o qual pelo Lema 21, é positivo definido.

Kernel gaussiano TSK non-singleton com seleção de parâmetro

O kernel k_{ntsk} é um kernel sem parâmetro. É possível estender esse kernel para uma versão com parâmetro:

$$k_{ntsk}^\gamma(X, X') = \prod_{d=1}^D \exp \left(-\frac{1}{2} \frac{(m_d - m'_d)^2}{\sigma_d^2 + (\sigma'_d)^2 + \gamma} \right), \tag{6.15}$$

onde $\gamma \in \mathbb{R}^+$. Esse kernel é positivo definido pois trocando as variáveis $\sigma_d'' = \sigma_d^2 + \gamma$ em (6.15) obtém-se (6.14).

Kernel gaussiano TSK non-singleton composto

Variando as definições para as funções k_{ntsk} e g^ℓ , do kernel non-singleton composto da Definição 6.5.2, obtém-se os seguintes kernels:

- se k_{ntsk} for o kernel (6.14) e $g^\ell(\mathbf{x}^I, \boldsymbol{\theta}^\ell) = \langle \mathbf{x}^I, \boldsymbol{\theta}^\ell \rangle$, então (6.10) é o kernel:

$$k_{ntsk} \otimes k((X, \mathbf{x}^I), (X^\ell, \boldsymbol{\theta}^\ell)) = \prod_{d=1}^D \exp\left(-\frac{1}{2} \frac{(m_d - m_d^\ell)^2}{\sigma_d^2 + (\sigma_d^\ell)^2}\right) \langle \mathbf{x}^I, \boldsymbol{\theta}^\ell \rangle, \quad (6.16)$$

- se k_{ntsk} for o kernel (6.15) e g^ℓ fora a expressão $\langle \mathbf{x}^I, \boldsymbol{\theta}^\ell \rangle$, tal que $\boldsymbol{\theta}^\ell = x_1^{sup}, \dots, x_d^{sup}, \dots, x_D^{sup}$, onde cada x_d^{sup} é dado por (6.13), então (6.10) é o kernel:

$$k_{ntsk} \otimes k((X, \mathbf{x}^I), (X^\ell, \boldsymbol{\theta}^\ell)) = \prod_{d=1}^D \exp\left(-\frac{1}{2} \frac{(m_d - m_d^\ell)^2}{\sigma_d^2 + (\sigma_d^\ell)^2 + \gamma}\right) \sum_{d=1}^D x_d^I \frac{m_d(\sigma_d^\ell)^2 + m_d^\ell \sigma_d^2}{\sigma_d^2 + (\sigma_d^\ell)^2}, \quad (6.17)$$

onde $x_1^I, \dots, x_d^I, \dots, x_D^I$ são as componentes de $\mathbf{x}^I$,

- se k_{ntsk} for o kernel (6.14) e g^ℓ for o kernel gaussiano, então (6.10) é o kernel:

$$k_{ntsk} \otimes k((X, \mathbf{x}^I), (X^\ell, \boldsymbol{\theta}^\ell)) = \prod_{d=1}^D \exp\left(-\frac{1}{2} \frac{(m_d - m_d^\ell)^2}{\sigma_d^2 + (\sigma_d^\ell)^2}\right) \exp\left(-\gamma \frac{1}{2} \|\mathbf{x}^I - \boldsymbol{\theta}^\ell\|^2\right). \quad (6.18)$$

Assim, existem varias possibilidades para construir kernels a partir de (6.10). Isto pode ser possível variando as definições dos conjuntos fuzzy e as T-normas para o kernel k_{ntsk} , e a definição para a função g^ℓ .

6.6 Kernel TSK non-singleton como relação de equivalencia fuzzy

Nesta seção apresentamos duas revisões de resultados conhecidos. O Colorário 6 em Moser (2006a) e o Teorema 9 em Moser (2006b). Mostramos como eles estão relacionados ao kernel TSK non-singleton apresentado na seção anterior. O primeiro resultado prova que todo kernel positivo definido, com valores no intervalo unitário e com constante um na diagonal é uma relação de equivalência fuzzy referente a uma T-norma. O segundo resultado, apresenta que esses kernels podem ser vistos como fórmulas de lógica fuzzy usadas para representar regras fuzzy. Mostraremos como o kernel TSK non-singleton satisfaz esses requisitos.

Definição 6.6.1 (Relação de Equivalência Fuzzy). A função $\mathcal{E} : \mathcal{X} \times \mathcal{X} \rightarrow [0, 1]$ é chamada de relação de equivalência fuzzy referente à T-norma T se

1. $\forall x \in \mathcal{X}, \mathcal{E}(x, x) = 1$;
2. $\forall x, y \in \mathcal{X}, \mathcal{E}(x, y) = \mathcal{E}(y, x)$;
3. $\forall x, y, z \in \mathcal{X}, T(\mathcal{E}(x, y), \mathcal{E}(y, z)) \leq \mathcal{E}(x, z)$.

O valor $\mathcal{E}(x, y)$ pode ser interpretado como “ x é igual à y ”. A condição 3 é chamada de *transitividade-T*, e pode ser interpretada como a sentença: “Se x e y são similares, e y e z são similares então x e z também são similares.” (Moser, 2006a).

Lema 24 (Kernels são ao menos transitividade- T_{cos} (Moser, 2006a)). *Seja $\mathcal{X}$ um conjunto não vazio. Seja $k : \mathcal{X} \times \mathcal{X} \rightarrow [0, 1]$ um kernel positivo definido tal que $\forall x \in \mathcal{X} : k(x, x) = 1$. Então $\forall x, y, z \in \mathcal{X}$, o kernel k satisfaz a transitividade- T_{cos} :*

$$T_{cos}(k(x, y), k(y, z)) \leq k(x, z), \quad (6.19)$$

onde

$$T_{cos}(a, b) = \max(ab - \sqrt{1 - a^2}\sqrt{1 - b^2}, 0), \quad (6.20)$$

é uma T -norma arquimediana e é a maior T -norma com esta propriedade (Moser, 2006a).

Lema 25 (Kernels como fórmula de lógica fuzzy para regras fuzzy (Moser, 2006b)). *Seja $\mathcal{X}$ um conjunto não vazio. Seja $k : \mathcal{X} \times \mathcal{X} \rightarrow [0, 1]$ um kernel positivo definido tal que $\forall x \in \mathcal{X} : k(x, x) = 1$. Então $\forall x, y, z \in \mathcal{X}$, existe uma família de conjuntos fuzzy $\mu_{i \in I} : \mathcal{X} \rightarrow [0, 1]$, onde I é um conjunto não vazio, tal que*

$$\forall x, y \in \mathcal{X} : k(x, y) = \inf_{i \in I} \overrightarrow{T}_M(\mu_i(x), \mu_i(y)), \quad (6.21)$$

onde $\overrightarrow{T}_M = \min(\overrightarrow{T}(x, y), \overrightarrow{T}(y, x))$ é um operador de bi-implicação e $\overrightarrow{T}(x, y) = \sup\{t \in [0, 1] | T(x, t) \leq y\}$, é a função de implicação gerada a partir da T -norma T (Moser, 2006b).

Lema 26. *Se forem usados conjuntos fuzzy normais, então o kernel TSK non-singleton satisfaz a transitividade- T_{cos} (Lemma (24)) e admite a representação dada pelo Lema (25).*

Demonstração. Por construção, o kernel TSK non-singleton com conjuntos normais é um kernel positivo definido tal que $k_{ntsk}(X, X) = 1$, $\forall X \in \mathcal{F}(\Omega)$: tomando valores no intervalo $[0, 1]$. Pelo Lema (24) k_{ntsk} satisfaz a transitividade- T_{cos} e pelo Lema (25), k_{ntsk} admite a representação da Equação (6.21). $\square$

6.7 Experimentos em classificação supervisionada

Apresentamos a seguir, experimentos em classificação supervisionada, usando quatro conjuntos de dados intervalares. Os experimentos foram feitos usando uma SVM (Seção 3.3 do Capítulo 3) com os kernels definidos sobre conjuntos fuzzy.

6.7.1 Dados e implementação

Foram usados quatro conjuntos de dados intervalares, chamados de de *baixa qualidade*, do repositório de dados KEEL (Alcalá-Fdez et al., 2011). *Dados de baixa qualidade* (Palacios et al., 2009), refere-se aos dados com informação vaga do valor correspondente às observações de um conjunto de dados. Essas observações são dadas como tuplas de intervalos, Além disso, alguns desses conjuntos de dados contêm observações com valores faltantes. A Tabela 6.1 apresenta um resumo desses conjuntos de dados. Palacios et al. (2009, 2011). A Tabela 6.2 apresenta duas observações do conjunto de dados *Long-4*, onde cada observação tem quatro atributos x_1, x_2, x_3, x_4 e duas classes 0 e 1. Note que a segunda observação pertence a duas classes no mesmo tempo.

Conjuntos de dados	#Observações	# Classes	Dimensão	Valores faltantes
Long-4	25	2	4	Não
100mlI-4	52	2	4	Não
100mlP-4	52	2	4	Não
dyslexic-12-4	65	4	12	Sim

Tabela 6.1: Sumario dos conjuntos de dados de baixa qualidade.

x_1	x_2	x_3	x_4	y
[8.7, 10.1]	[45, 47]	[2, 2.15]	[5, 5.1]	{1}
[9.5, 10]	[60, 64]	[2.21, 2.23]	[5.33, 5.4]	{0, 1}

Tabela 6.2: Duas observações do conjunto de dados de baixa qualidade Long-4.

Escalamento e fuzzificação

Dado que algumas obervações desses conjuntos de dados pertencem às duas classes ao mesmo tempo e dado que o propósito deste experimento não é tratar esses casos, então essas observações foram removidas. Além disso, todos os intervalos dos conjuntos de dados foram escalados ao intervalo unitário.

Cada atributo foi modelado com conjuntos fuzzy da seguinte forma: para cada atributo (intervalo): $[\underline{x}_d, \bar{x}_d]$, $\underline{x}_d \leq \bar{x}_d$, $\underline{x}_d, \bar{x}_d \in \mathbb{R}$, de todos os conjuntos de dados de baixa qualidade, foi construído o conjunto fuzzy $X_d : \mathbb{R} \rightarrow [0, 1]$:

$$X_d(.) = \exp \left(-\frac{1}{2} \frac{|\cdot - m_d|^2}{\sigma_d^2} \right), \quad (6.22)$$

onde $m_d \in \mathbb{R}$ e $\sigma_d \in \mathbb{R}^+$ são seus parâmetros definidos como:

$$m_d = \frac{(\underline{x}_d + \bar{x}_d)}{2}, \quad (6.23)$$

$$\sigma_d = \frac{|\underline{x}_d - \bar{x}_d|}{2\sqrt{2 \ln 2}}, \quad (6.24)$$

a medida para o parâmetro σ é chamada de *largura à meia altura*.

Isso torna o conjunto de dados de baixa qualidade:

$$\{[\underline{x}_1^{(i)}, \bar{x}_1^{(i)}], [\underline{x}_2^{(i)}, \bar{x}_2^{(i)}], \dots, [\underline{x}_D^{(i)}, \bar{x}_D^{(i)}]\}_{i=1}^N,$$

no conjunto de dados fuzzy:

$$\{(X_1^{(i)}, X_2^{(i)}, \dots, X_D^{(i)})\}_{i=1}^N.$$

No caso de valores em falta, foi adoptada a heurística de construir um intervalo com limites nos valores mínimo e máximo do atributo.

Seleção de modelo e validação cruzada

Foi feita validação cruzada 10-fold, para cada ponto da grade (C, γ) formada pelos parâmetros C da SVM com valores em $\{2^{-1}, \dots, 2^{14}\}$, e o parâmetro kernel γ com valores em $\{2^4, \dots, 2^{-10}\}$. Para o 10-fold, foram usadas as 10 partições disponibilizadas no repositório KEEL (Alcalá-Fdez *et al.*, 2011) para esses conjuntos de dados. Como trata-se de conjuntos de dados com poucas amostras, foi reportado o par (C, γ) com a melhor acurácia de validação cruzada para cada um dos métodos comparados. Nos acreditamos que uma comparação baseada no erro de validação cruzada refletirá a performance relativa dos diferentes kernels.

Nos experimentos foi usado o kernel gaussiano k_γ^{rbf} como base de comparação, o qual foi definido nos valores médios dos intervalos, isto é:

$$k_\gamma^{\text{rbf}}(\mathbf{m}, \mathbf{m}') = \exp\left(-\gamma \frac{1}{2} \|\mathbf{m} - \mathbf{m}'\|^2\right), \quad (6.25)$$

tal que, $\mathbf{m} = (m_1, m_2, \dots, m_d, \dots, m_D)$, onde m_d é dado pela Equação (6.23).

Resultados no conjunto de dados 100mII-4 usando kernels derivados de sistemas fuzzy

O conjunto de dados 100mII-4, apresenta dados de desempenho de atletas para classificar quando uma marca de tempo de corrida de 100 metros rasos de corrida pode ser, ou não, atingida satisfatoriamente. As observações são na forma de com tuplas de 4 valores intervalares contendo:

- informação da proporção entre peso e altura,
- o tempo de reação,
- a velocidade de início ou velocidade aos 20 metros e
- a velocidade máxima ou velocidade aos 40 metros.

As medidas foram feitas por três observadores diferentes.

Foram usados três kernels definidos sobre conjuntos fuzzy derivados dos sistemas fuzzy TSK-NS, apresentados anteriormente. A Tabela 6.3 apresenta os kernels usados, o melhor parâmetro de regularização C e o melhor parâmetro γ obtidos na etapa de seleção de modelo. É apresentada também, a respetiva acurácia de validação cruzada (Acc.) e o número médio de vetores de suporte (SV's) da validação cruzada 10-fold para esse par (C, γ) . O kernel gaussiano é denotado como *RBF*, o kernel TSK non-singleton é denotado como TSK-NS

Foi feita seleção de parâmetro (C, γ) para todos os kernels com exceção do kernel da ultima linha da Tabela 6.3, dado pela Equação (6.16). Para esse kernel foi feita somente seleção de modelo e validação cruzada usando o parâmetro C .

Os resultados indicam que os kernels definidos sobre conjuntos fuzzy derivados dos sistemas fuzzy TSK apresentam melhor desempenho em termos de acurácia que o kernel gaussiano. Porém usam um numero maior na média de vetores de suporte.

Kernel	Definição	Tipo	SVs	C	γ	Acc.
$k_{\gamma}^{\text{rbf}}(\mathbf{m}, \mathbf{m}')$	Equação (6.25)	RBF	11.8	2^{15}	2^{-4}	96
$k_{ntsk}^{\gamma}(X, X')$	Equação (6.15)	TSK-NS	31	2^1	2^{-5}	98
$k_{ntsk} \otimes k((X, \mathbf{m}), (X^l, \mathbf{m}'))$	Equação (6.18)	TSK-NS	37.1	2^2	2^{-4}	98
$k_{ntsk} \otimes k((X, \mathbf{m}), (X^l, \mathbf{m}'))$	Equação (6.16)*	TSK-NS	41.5	2^0	-	98

* é um kernel sem seleção de parâmetro.

Tabela 6.3: Resultados no conjunto de dados 100mlI-4.

Resultados no conjunto de dados *Long-4* usando kernels com substituição de distância

O conjunto de dados *Long-4*, apresenta dados de desempenho de atletas, para classificar se um atleta consegue melhorar um certo limiar em salto em distância. Ele foi gerado pelas opiniões subjetivas de um árbitro, onde cada observação é representada por uma tupla de 4 valores intervalares:

- a proporção entre peso e altura,
- a máxima velocidade aos 40 metros,
- o número de flexões de abdominais por minuto e
- uma prova de alongamento.

Dois dos atributos são valores linguísticos que posteriormente são convertidos para intervalos e, os outros dois atributos intervalares são resultados de uma medida.

A Tabela 6.4 apresenta os kernels usados, os resultados para o kernel TSK non-singleton são apresentados na segunda linha. Os kernels da terceira e quarta linha são kernels definidos sobre conjuntos fuzzy com substituição de pseudo-métricas, isto é, são kernels simétricos (Capítulo 5). A Tabela 6.4 reporta o melhor par (C, γ) da validação cruzada 10-fold conjuntamente com a acurácia (Acc.) e o número de vetores de suporte médio (SV's) para esse par. Neste experimento, o kernel TSK non-singleton tem um desempenho comparável com o kernel gaussiano. Da mesma forma, o kernel em conjuntos fuzzy com substituição de distância da última linha da Tabela 6.4 tem acurácia igual que o kernel gaussiano, mas com, menor número médio de vetores de suporte. Por outro lado, o kernel em conjuntos fuzzy com substituição de distância da terceira linha da Tabela 6.4 apresenta melhor acurácia e menor número de suportes do que os outros kernels.

Resultados no conjunto de dados *100mlP-4* e *Dyslexic-12-4* usando SVM com múltiplos kernels

O conjunto de dados *100mlP-4* tem as mesmas características que o conjunto de dados 100mlI-4, com a diferença que as medidas foram obtidas pela opinião de um treinador em termos de valores linguísticos como:0 “o tempo de reação é lento”. O conjunto de dados *Dyslexic-12-4* apresenta dados

Kernel	Definição	Tipo	SVs	C	γ_0	Acc.
$k_{\gamma}^{\text{rbf}}(\mathbf{m}, \mathbf{m}')$	Equação (6.25)	RBF	15.1	2^0	2^{-3}	73.33
$k_{ntsk}^{\gamma}(X, X')$	Equação (6.15)	TSK-NS	15.1	2^0	2^2	73.33
$k_{\gamma}(X, X')$	Equação (5.52)	subs. distância	13.2	2^1	2^{-3}	85.00
$k_{\gamma}(X, X')$	Equação (5.53)	subs. distância	9.8	2^{11}	2^{-7}	73.33

Tabela 6.4: Resultados no conjunto de dados Long-4

Kernel	Definição	Tipo	SVs	C	γ	Acc.
$k_{\gamma}^{\text{rbf}}(\mathbf{m}, \mathbf{m}')$	Eq. (6.25)	RBF	28.6	2^3	2^2	88
$k_{ntsk}^{\gamma}(X, X')$	Eq. (6.15)	TSK-NS	18.9	2^4	2^{-2}	88
$k_{ntsk} \otimes k((X, \mathbf{m}), (X^l, \mathbf{m}'))$	Eq. (6.16)*	TSK-NS	42.3	2^3	-	82.67
$\beta k^{\text{eq}(6.15)}(X, X') + (1 - \beta) k_{\gamma}^{\text{rbf}}(\mathbf{m}, \mathbf{m}')$		MK	37.2	2^4	2^{-3}	88
$\beta k^{\text{eq}(6.18)}((X, \mathbf{m}), (X', \mathbf{m}')) + (1 - \beta) k_{\gamma}^{\text{rbf}}(\mathbf{m}, \mathbf{m}')$		MK	37.3	2^4	2^{-3}	88
$\beta k_{\gamma'}^{(5.52)}(X, X') + (1 - \beta) k_{\gamma}^{\text{rbf}}(\mathbf{m}, \mathbf{m}')$	$\gamma' = 1$	MK	26.4	2^{16}	2^4	88
$\beta k_{\gamma'}^{\text{eq}(5.53)}(X, X') + (1 - \beta) k_{\gamma}^{\text{rbf}}(\mathbf{m}, \mathbf{m}')$	$\gamma' = 1$	MK	26	2^2	2^8	88

* é um kernel sem seleção de parâmetro.

Tabela 6.5: Resultados no conjunto de dados 100mlP-4.

Kernel	Definição	Tipo	SVs	C	γ	Acc.
$k_{\gamma}^{\text{rbf}}(\mathbf{m}, \mathbf{m}')$	Eq. (6.25)	RBF	33	2^5	2^{-8}	36
$k_{ntsk}^{\gamma}(X, X')$	Eq. (6.15)	TSK-NS	33.3	2^3	2^5	36
$k_{ntsk} \otimes k((X, \mathbf{m}), (X', \mathbf{m}'))$	Eq. (6.16)*	TSK-NS	45	2^1	-	38.45
$\beta k^{\text{eq}(6.15)}(X, X') + (1 - \beta)k_{\gamma}^{\text{rbf}}(\mathbf{m}, \mathbf{m}')$		MK	38.1	2^1	2^{-4}	36
$\beta k^{\text{eq}(6.18)}((X, \mathbf{m}), (X', \mathbf{m}')) + (1 - \beta)k_{\gamma}^{\text{rbf}}(\mathbf{m}, \mathbf{m}')$		MK	38.1	2^1	2^{-4}	36
$\beta k_{\gamma'}^{\text{eq}(5.52)}(X, X') + (1 - \beta)k_{\gamma}^{\text{rbf}}(\mathbf{m}, \mathbf{m}')$	$\gamma' = 1$	MK	33.8	2^2	2^{-4}	36
$\beta k_{\gamma'}^{\text{eq}(5.53)}(X, X') + (1 - \beta)k_{\gamma}^{\text{rbf}}(\mathbf{m}, \mathbf{m}')$	$\gamma' = 1$	MK	30.6	2^{16}	2^{-6}	44.10

* é um kernel sem seleção de parâmetro.

Tabela 6.6: Resultados no conjunto de dados Dyslexic-12-4.

para predizer se um paciente tem, ou não, dislexia. Os dados foram obtidos por avaliações subjetivas de especialistas sobre diversos testes gráficos usados para diagnosticar dislexia. Esse conjunto de dados tem doze atributos e quatro classes: {dyslexia, no dyslexia, controle, outros problemas}. Além disso, ele contém valores ausentes.

Neste experimento, foram usadas combinações de kernels, denotadas como MK, nas Tabelas 6.5 e 6.6. Foram combinados os kernels TSK non-singleton com o kernel gaussiano, e os kernels em conjuntos fuzzy com substituição de distâncias com o kernel gaussiano. A combinação foi do tipo $\beta k_1 + (1 - \beta)k_2$, com $\beta = 0.25$ (definido arbitrariamente). Os resultados para esses kernels estão na quarta e quinta linha das Tabelas 6.5 e 6.6 para o caso de combinação dos kernels TSK non-singleton com o kernel gaussiano, e na sexta e sétima linha das Tabelas 6.5 e 6.6 no caso de combinação dos kernels em conjuntos fuzzy com substituição de distâncias com o kernel gaussiano.

Para o conjunto de dados *100mlP-4*, todos os resultados são comparáveis em termos de acurácia, com exceção do kernel da terceira linha da Tabela 6.5. Note que alguns kernels em conjuntos fuzzy apresentam menor número de vetores de suporte, em média, que o kernel gaussiano. No caso do conjunto de dados *Dyslexic-12-4*, os kernels em conjuntos fuzzy apresentam igual o melhor desempenho em termos de acurácia que o kernel gaussiano. Particularmente, a combinação dos kernels em conjuntos fuzzy com substituição de distância com o kernel gaussiano, melhora muito a acurácia de classificação para estes dados com menor número médio de vetores de suporte.

6.8 Sumário

Neste capítulo, foi apresentada como a iteração entre as regras dos sistemas fuzzy TSK e a entrada do sistema modelada com conjuntos fuzzy non-singleton induzem kernels positivos definidos, definidos sobre conjuntos fuzzy. Para isso, foi apresentada uma nova formulação de sistema fuzzy TSK que leva em conta as entradas do sistema modeladas por conjuntos fuzzy non-singleton. Foram

apresentadas diversas formulações de kernels positivos definidos baseados nesses conceitos. Também foi apresentada a relação existente entre relações de equivalência fuzzy e os kernels apresentados. Além disso, foram apresentados experimentos em problemas de classificação em dados intervalares, onde foi usada uma SVM conjuntamente com os kernels apresentados. Os resultados obtidos foram satisfatórios.

Capítulo 7

Conclusões e Trabalhos Futuros

Everything must be made as simple as possible. But not simpler.

Albert Einstein

Neste trabalho, investigamos como as medidas de probabilidade e a teoria fuzzy conjuntamente com os métodos de kernel, podem ser usadas para tratar conjuntos de dados do tipo (1.1). Dependendo da natureza das observações s_i em (1.1), elas podem ser modeladas por medidas de probabilidade ou por conjuntos fuzzy. No caso das observações serem modeladas por medidas de probabilidade, nós assumimos que a distribuição de pontos de cada observação s_i segue alguma lei de probabilidade desconhecida, codificada pela medida de probabilidade. Graças aos kernels definidos sobre medidas de probabilidade e o mergulho dessas medidas de probabilidade para funções média nos espaços de Hilbert, é possível fazer aprendizado desse tipo de dados. Nós apresentamos como é possível criar modelos de descrição de dados para conjuntos de dados do tipo (1.1) usando esse enfoque, e como aplicar os modelos propostos na tarefa de detecção de anomalias grupais. No caso das observações serem modeladas por conjuntos fuzzy, nós assumimos que cada observação codifica a natureza não precisa ou vaga de problemas reais. Isso permite que possam ser projetados conjuntos fuzzy para modelar essa informação, de forma arbitrária. Graças à projeção de kernels definidos sobre conjuntos fuzzy, é possível analisar conjuntos de dados com observações modeladas usando conjuntos fuzzy, nos espaços de Hilbert com kernel reproduzível. Nos apresentamos a teoria por trás dos kernels definidos sobre conjuntos fuzzy, assim como uma aplicação em classificação supervisionada de padrões em conjuntos de dados intervalares. Além dos problemas apresentados neste trabalho, a metodologia usada pode ser útil para solucionar diversos problemas de aprendizado computacional para esse tipo de dados, como por exemplo:

- *Projeção de classificadores sobre conjuntos de medidas de probabilidade.* Isso pode ser alcançado estendendo o modelo de descrição de dados com medidas de suporte apresentado no Capítulo 4, para o caso de classificação binária e multi-classe, considerando a descrição de cada classe de forma independente usando esferas envolventes mínimas. É possível também estender outros algoritmos conhecidos para esses casos, considerando o mergulho das medidas de probabilidade para os espaços de Hilbert. Por exemplo, o caso de regressão de conjuntos de dados de medidas de probabilidade poderá ser feito considerando as máquinas de vetores

de suporte ou os processos gaussianos com kernels definidos sobre medidas de probabilidade. O caso de agrupamento de dados também pode ser considerado, tendo em conta possíveis extensões dos algoritmos usuais.

- *Distâncias entre conjuntos fuzzy.* O conceito de distâncias entre conjuntos fuzzy tem sido amplamente explorado na comunidade fuzzy, com diversas aplicações práticas (Bloch, 1999; Chaudhur e Rosenfeld, 1996; Chaudhuri e Rosenfeld, 1999; Chen *et al.*, 1995; Rosenfeld, 1985; Xuecheng, 1992; Zwick *et al.*, 1987). A pesquisa referente às distâncias entre conjuntos fuzzy pode ser feita usando mapeamentos dos conjuntos fuzzy para um EHKR, onde pode ser usada uma métrica entre as funções representantes desses conjuntos fuzzy. Isso pode ser conseguido usando os kernels apresentados nos Capítulos 5 e 6.
- *Descrição de dados fuzzy.* Assumindo que os dados fuzzy são realizações independentes e identicamente distribuídas de uma variável aleatória fuzzy seguindo uma distribuição de probabilidade, é possível aproximar o suporte e os conjuntos de nível daquela distribuição, usando os kernels definidos sobre conjuntos fuzzy. Potenciais aplicações estão nas áreas de detecção de anomalias e novidades, e classificação de uma classe em conjuntos de dados fuzzy.
- *Regressão para conjuntos de dados fuzzy.* Os kernels definidos sobre conjuntos fuzzy podem ser usados na tarefa de regressão de conjuntos fuzzy. O problema pode ser formulado via máquinas de vetores de suporte ou com processos gaussianos. Potenciais aplicações será na predição de valores, em conjuntos de dados reais que possam ser modelados como conjuntos fuzzy.
- *Mergulho de medidas fuzzy para os EHKR.* Outra pesquisa a ser abordada será estudar a teoria e possíveis aplicações práticas dos mergulhos de outras medidas que não sejam medidas de probabilidade (por exemplo medidas fuzzy) para os EHKR. Usando a teoria por trás dos mergulhos de medidas apresentadas em Berlinet e Thomas-Agnan (2004), permitirá pesquisar tais casos.
- *Modelos híbridos probabilísticos e fuzzy.* Considerando a não-precisão dos dados, modelada por conjuntos fuzzy, e considerando as observações s_i , como conjuntos de conjuntos fuzzy, será possível combinar os dois enfoques. Isso terá alta aplicação prática em problemas que modelem as observações como agrupamento de pontos não precisos. Para isso, uma metodologia a seguir, é estender os mergulhos das medidas de probabilidade definidas em $(\mathbb{R}^D, \mathcal{B}(\mathbb{R}^D))$ para o caso das distribuições das variáveis aleatórias fuzzy.

7.1 Resumo das contribuições

A tese apresenta contribuições originais para as áreas de aprendizado computacional e teoria fuzzy, a saber:

- No Capítulo 4, introduzimos três modelos de descrição de dados para conjuntos de dados modelados com medidas de probabilidade. Os três modelos não assumem nenhuma forma para as medidas de probabilidade. Fazendo com que o passo de estimação de densidade não seja necessário. Ao invés disso, são usados os mergulhos dessas medidas de probabilidade para o

EHKR, onde a descrição é computada por meio de esferas envolventes mínimas. Os modelos apresentados são métodos de kernel que usam um subconjunto das observações do conjunto de dados, para representar a função objetivo. Como essas observações são medidas de probabilidade, nós chamamos aquelas de medidas suporte. Os modelos são apresentados como problemas de otimização, sendo o primeiro deles formulado como problema de otimização com restrições de chance. Esses modelos de descrição de dados poderão ser usados em problemas de aprendizado, como classificação, regressão, agrupamento de dados, entre outros. Na Seção 4.5, foi proposto usar o modelo de descrição de dados com medidas de suporte para a tarefa de detecção de anomalias grupais, as quais são caracterizadas por agrupamentos de pontos anômalos, ou agrupamentos não usuais de pontos não anômalos. Os modelos propostos apresentaram um bom desempenho em problemas com conjuntos de dados artificiais e reais. A aplicação prática abordada foi a detecção de grupos anômalos de galáxias no conjunto de dados SDSSIII.

- No Capítulo 5, foram apresentados os kernels definidos sobre conjuntos fuzzy. A teoria por trás desses kernels foi motivada pelos kernels definidos sobre conjuntos. Entre os kernels apresentados encontram-se, o kernel de convolução fuzzy, o kernel de produto cruzado entre conjuntos fuzzy, o kernel de interseção de conjuntos fuzzy, e os kernels em conjuntos fuzzy usando substituição de distância. Esses kernels permitem o mergulho dos conjuntos fuzzy para funções nos EHKR, permitindo assim usar os métodos de kernel para fazer aprendizado em conjuntos de dados fuzzy usando esses kernels. Na Seção 5.7 foi apresentado um experimento em teste de hipótese usando a MMD com os kernels apresentados sobre um conjunto de dados de predição de recorrência de câncer de mama, onde o kernel com substituição de distância, definidos sobre conjuntos fuzzy apresentou um bom desempenho.
- No Capítulo 6, foi apresentada uma nova classe de kernels em conjuntos fuzzy induzidos pelos sistemas de regras fuzzy TSK. Esses kernels são o resultado da interação das regras de uma nova generalização do sistema de regras fuzzy TSK para o caso non-singleton. Foram apresentados alguns kernels, baseados em funções de pertinência Gaussianas. Essa nova classe de kernels foi chamada de kernels TSK non-singleton. Na Seção 5.6, foram apresentadas as aplicações práticas dos kernels definidos sobre conjuntos fuzzy, como o cálculo de distâncias entre conjuntos fuzzy, a criação de novos kernels em conjuntos fuzzy a partir dos apresentados e os problemas nos quais eles poderão ser de utilidade. Na Seção 6.7, foram apresentados experimentos em classificação supervisionada para dados intervalares, onde os kernels apresentados mostraram um bom desempenho.

7.2 Publicações geradas

Parte do conteúdo do Capítulo 5, especificamente o kernel de interseção entre conjuntos fuzzy apresentado na Seção 5.4, foi publicado na conferência internacional FUZZ-IEEE 2014 (*IEEE International Conference on Fuzzy Systems*), principal conferência internacional da área da teoria fuzzy. O kernel em conjuntos fuzzy com substituição de distâncias foi apresentado em Guevara *et al.* (2015a) como parte do capítulo do livro *Handbook of Fuzzy Sets Comparison - Theory, Algorithms and Applications*

Pretende-se submeter o conteúdo total do Capítulo 5 para o periódico *IEEE Transactions on Fuzzy Systems*, o qual é o principal periódico dos sistemas fuzzy.

O conteúdo do Capítulo 5.6, foi apresentado na conferência internacional FUZZ-IEEE 2013. O conteúdo apresentado na tese, é uma versão revisada e estendida daquele artigo.

O modelo de descrição de dados com medidas de suporte, apresentado no Capítulo 4 foi apresentado em Guevara *et al.* (2015b) no congresso ACM SIGKDD (*Conference on Knowledge Discovery and Data Mining*), uma das principais conferencias em mineração de dados. A Seção 4.2, foi apresentada em formato poster na escola de aprendizado computacional *MLSS-2013* (The machine learning summer school).

A continuação descrevemos os principais artigos publicados pelo autor deste projeto, os quais estão relacionados com esta mesma linha de pesquisa. Nos apresentamos o resumo dos artigos e as principais contribuições de cada um deles.

7.2.1 *Support Measure Data Description.*

Submetido a *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Periódico com Conceito Capes A, e conceito A*, pela Computing Research and Education Association. Uma versão menor titulada como *Support Measure Data Description for Group Anomaly Detection* foi apresentada no congresso ACM SIGKDD Guevara *et al.* (2015b) ¹.

Resumo

O foco deste trabalho é o aprendizado de um modelo de descrição de dados para conjuntos de dados contendo elementos ou observações dadas por conjuntos de pontos em $\mathbb{R}^D$. Modelando cada observação como uma medida de probabilidade, a descrição desses conjuntos de dados é estimada computando o conjunto de volume mínimo para aquelas medidas de probabilidade, usando uma esfera envolvente mínima para as funções representantes das medidas de probabilidade num Espaço de Hilbert com Kernel Reproduzível ou EHKR. A vantagem disso é que não é considerada nenhuma forma particular para as medidas de probabilidade, no lugar de aquilo, é usado o embedding daquelas medidas num EHKR dado por um kernel positivo definido, definido no espaço das medidas de probabilidade. Como resultado, o modelo de descrição de dados é uma função que somente depende em algumas medidas de probabilidade chamadas de *medidas de suporte*. No trabalho apresentamos três modelos de descrição de dados com medidas de suporte para aqueles conjuntos de dados. O primeiro modelo é formulado como um problema de otimização com restrições de chance. O segundo modelo é uma extensão do método chamado *descrição de dados usando vetores de suporte* para o caso de medidas de probabilidade. O terceiro modelo é igual ao segundo modelo, mas definido somente para funções kernels invariantes a traslação, requerendo este modelo um escalamento prévio dos dados. Os modelos apresentados são validados na tarefa de detecção de anomalias grupais, com conjuntos de dados artificiais e reais.

Contribuições

As principais contribuições são:

¹Conceito Capes A, e conceito A, pela Computing Research and Education Association.

1. *Descrição de dados usando medidas de suporte como problema de otimização com restrições de chance.* Este problema é formulado primeiramente no espaço de medidas de probabilidade, e logo apos o método é kernelizado, usando o embedding de medidas de probabilidade em EKHR.
2. *A extensão do método de descrição de dados usando vetores de suporte para o caso de medidas de probabilidade.* Usando o conceito de embeddings de medidas de probabilidade para EHKR, foram projetados dois modelos de descrição de dados adicionais, ambos modelos são formulados como problemas de otimização quadrática.
3. *Deteção de anomalias grupais em dados astronômicos* Foi feito um estudo de aplicação do método no caso de detecção de anomalias grupais, usando dados artificiais e usando um conjunto de dados reais de grupos de galaxias do projeto *The Sloan Digital Sky Survey*² (SDSS). Este conjunto de dados contem informação do universo. A ideia do experimento é usar o conjunto de dados para detectar grupos anômalos de galaxias, o conjunto de dados contem aproximadamente 7×10^5 galaxias, onde cada galaxia é representada como um vetor de características de 4000 dimensões representado a informação espectral. Os resultados obtidos, provam que o método é trabalha bem para esse tipo de problemas.

7.2.2 Positive Definite Kernel Functions on Fuzzy Sets

Artigo apresentado, em "2014, IEEE International Conference on Fuzzy Sets".³

Resumo do Trabalho

O mapeamento de dados não vetoriais em espaços vectoriais é muito comum em aprendizado computacional, com o objetivo de realizar tarefas de classificação, regressão ou agrupamento de dados. Os conjuntos de dados fuzzy, ou conjuntos cujas observações são conjuntos fuzzy, são exemplos de dados não vetoriais e, tem sido formulados diversos algoritmos de reconhecimento de padrões para analisar eles no espaço formado pelo conjunto dos conjuntos fuzzy. Sem embargo, o analise dos conjuntos de dados fuzzy nesse espaço tem a limitação de não ser um espaço vectorial. Nos propomos, o mapeamento dos dados fuzzy para um Espaço de Hilbert com Kernel Reproduzível Este embedding é possível usando funções kernel positivas definidas em conjuntos fuzzy. Como resultado, apresentamos a formulação de funções kernel de valor real, definidas no espaço dos conjuntos fuzzy, em particular, usando operadores T-norma, definimos o kernel baseado na interseção de conjuntos fuzzy e o kernel fuzzy de produto cruz. As potenciais aplicações destes kernels incluem problemas de aprendizado computacional onde os dados são representados como conjuntos fuzzy, como o kernel é independente do algoritmo, estes kernels podem ser facilmente usados em algoritmos como as máquinas de vetores de suporte, kernel PCA, descrição de dados com medidas de suporte, entre outros.

Contribuições

As principais contribuições são:

²<http://www.sdss3.org/>

³Conceito Capes A, e conceito A, pela Computing Research and Education Association.

1. Definição do kernel sobre conjuntos fuzzy.
2. Definição do kernel de interseção de conjuntos fuzzy usando medidas aditivas.
3. Uma análise do kernel apresentado em [Guevara *et al.* \(2013a\)](#)
4. Definição de vários kernels sobre conjuntos fuzzy.

7.2.3 *Kernel Functions in Takagi-Sugeno-Kang Fuzzy System with Nonsingleton Fuzzy Input.*

Artigo apresentado na “2013 IEEE International Conference on Fuzzy Systems”, FUZZ ⁴

Resumo do Trabalho

Os algoritmos para classificação supervisionada de padrões, na maioria dos casos, não consideram dados com imprecisão, por exemplo, coleções de intervalos, amostras dadas por um conjunto de histogramas, lista de valores, conjuntos fuzzy, entre outros, para modelar as observações. A teoria fuzzy é uma escolha natural para modelar a imprecisão e os métodos de kernel estão o estado da arte em algoritmos de aprendizado computacional. Alguns trabalhos anteriores descrevem a ligação entre as duas áreas: a interação das regras fuzzy dos sistemas fuzzy do tipo Takagi-Sugeno-Kang com as entradas do sistema definem um kernel positivo definido [Chen e Wang \(2003\)](#). Por outro lado a pesquisa atual em conjuntos fuzzy mostra que o uso de entradas nonsingleton no sistema fuzzy tem vantagens no modelamento da imprecisão nos dados. Este trabalho descreve a relação existente entre kernels do tipo positivo definido e sistemas fuzzy do tipo Takagi-Sugeno-Kang, mas com entradas do tipo nonsingleton. Como resultado, foi proposto um novo sistema fuzzy Takagi-Sugeno-Kang com entradas do tipo nonsingleton, foi provado que este tipo de sistema fuzzy induz uma família de kernels positivos definidos chamados de nonsingleton TSK kernels, finalmente foi proposto um procedimento para formular kernels para dados com intervalos. Os experimentos realizados com dados reais mostraram um melhor desempenho que os métodos do estado da arte.

Contribuições

As principais contribuições do trabalho são:

1. *Sistema fuzzy do tipo Takagi-Sugeno-Kang com entradas do tipo nonsingleton.* Foi apresentada uma extensão dos sistemas fuzzy do tipo Takagi-Sugeno-Kang para o caso de entradas do tipo nonsingleton.
2. *Kernels positivo definidos induzidos pelo sistema Takagi-Sugeno-Kang com entradas do tipo nonsingleton.* Foi provado que a interação entre regras fuzzy do sistema Takagi-Sugeno-Kang com entradas do tipo nonsingleton e suas entradas induzem uma família de kernels positivos definidos.
3. *Kernels para dados intervalares.* Usando os kernels induzidos e o conceito de kernels por substituição de distâncias, foi proposta uma maneira geral para criar kernels positivos definidos para dados intervalares.

⁴Conceito Capes A, e conceito A, pela Computing Research and Education Association.

4. *Experimentos em dados reais.* Usando máquinas de vetores de suporte, foram testados os kernels propostos com conjunto de dados com dados faltantes e valores intervalares. Foram usados quatro conjunto de dados intervalares, chamados de baixa qualidade, do repositório KEEL [35]. O termo baixa qualidade [Cadenas et al. \(2011\)](#) é usado para se referir a dados com incerteza no valor dos atributos. Os conjuntos de dados usados foram:

Long-4 É um conjunto de dados de desempenho de atletismo com quatro atributos. Um técnico definiu os dois primeiros atributos em forma de valores linguísticos ou intervalos, os outros dois atributos são medidos três vezes produzindo incerteza, sendo representada esta informação como intervalos.

100mII-4 É um conjunto de dados de desempenho de atletismo usado para classificar se uma marca de 100 metros de corrida é conseguida. As características são dadas pelo peso e altura do atleta, tempo de reação, velocidade de saída, velocidade nos 40 metros. As medições foram feitas por diferentes pessoas.

100mIP-4 Medidas foram obtidas mediante uma avaliação subjetiva de uma pessoa em termos linguísticos.

Dyslexic-12-4 É um conjunto de dados para avaliar os graus de dislexia em crianças, o conjunto de dados contém valores faltantes com doze atributos e quatro classes.

7.2.4 *Support Vector Data Description for Uncertainty Datasets*

Poster apresentado na escola de aprendizado computacional: The Machine Learning Summer School MLSS 2013⁵.

Resumo do trabalho

Este trabalho descreve um método para estimar o suporte da distribuição dos dados [Schölkopf et al. \(2001\)](#) levando-se em conta dados com incertezas modelados com medidas de probabilidade. O método usada foi a descrição dos dados usando vetores de suporte [Tax e Duin \(2004\)](#). Resultados sobre conjunto de dados artificiais foram apresentados.

Em aprendizado computacional, assume-se geralmente que os dados são gerados a partir de uma distribuição de probabilidade desconhecida. Na literatura encontram-se vários métodos para estimar essa distribuição. Um método alternativo é estimar o suporte da distribuição, isto é, o conjunto de dados para os quais a medida de probabilidade tem valor maior ou igual a um valor dado. Há dois métodos bastante conhecidas e usados: as máquinas de vetores de suporte para apenas uma classe [Schölkopf et al. \(2001\)](#) e a descrição dos dados usando vetores de suporte [Tax e Duin \(2004\)](#). Estes dois métodos são equivalentes se o kernel usado for um kernel invariante a translação. Os algoritmos que usam essas técnicas, geralmente, treinam o classificador com apenas uma classe, com o propósito de construir um modelo de classe única. Isso é útil para prever amostras que pertencem a essa classe, ou apontar amostras anômalas, isto é, que não são reconhecidas dentro do suporte da distribuição estimada. O ponto fraco desses métodos é que não consideram o caso de conjuntos de dados com incertezas e imprecisões. Neste trabalho, foi considerado o modelamento desses conjuntos de dados usando medidas de probabilidade por cada amostra e a estimação do suporte da distribuição que gera esses dados.

⁵<http://www.mlss.cc/>

Descrição da escola.

A “machine learning summer school” *MLSS* é uma das escolas mais importantes da área de aprendizado computacional. Esta edição teve lugar no instituto Max Planck para Sistemas Inteligentes em Tübingen Alemanha, de 25 de agosto a 8 de setembro, de 2013.

Contribuições

As principais contribuições foram:

1. A formulação do problema usando programação de restrições com chance.
2. A formulação determinística desse problema usando o método de momentos de desigualdade.
3. A formulação do dual do problema e o análise das condições de otimalidade.

7.2.5 *Fuzzy Set Similarity usinga Distance-Based Kernel on Fuzzy Sets*

Publicado em [Guevara *et al.* \(2015a\)](#) como capítulo do livro *Handbook of Fuzzy Sets Comparison - Theory, Algorithms and Applications*

Resumo

O capítulo apresenta como construir kernels definidos sob conjuntos fuzzy usando distancias definidas em conjuntos fuzzy. O capítulo apresenta um experimento em teste de duas amostras kernels, com dados heterogêneos, onde um subconjunto das variáveis contém conjuntos fuzzy como valores.

Contribuições

As principais contribuições são:

1. *Kernels definidos sob conjuntos fuzzy usando distancias entre conjuntos fuzzy.* Propomos uma nova forma para criar kernels definidos sob conjuntos fuzzy baseados no conceito de substituição de distancias sob conjuntos fuzzy, na definição do kernel.
2. *Distancias entre conjuntos fuzzy induzidas por kernels definidos sob conjuntos fuzzy.* Mostramos como os kernels definidos em conjuntos fuzzy induzem distancias entre conjuntos fuzzy.
3. *Teste kernel de duas amostras usando um conjunto de dados contendo variáveis fuzzy* Foi feito um estudo de aplicação na tarefa de teste kernel de duas amostras sob um conjunto de dados heterogêneo, o qual contém um subconjunto de variáveis fuzzy.

Apêndice A

Provas

A.1 Prova do Lema 3

Seja $X = (X_1, \dots, X_j, \dots, X_D)^\top$ e $\mathbf{c} = (c_1, \dots, c_j, \dots, c_D)^\top$, então

$$\begin{aligned} E[\|X - \mathbf{c}\|^2] &= E[X^\top X] - 2E[X^\top \mathbf{c}] + \|\mathbf{c}\|^2 \\ &\quad \text{pela definição de covariância} \\ &= \sum_j^D (\text{cov}(X_j, X_j) - \mathbb{E}[X_j]\mathbb{E}[X_j]) - 2 \sum_{j=1}^D E[X_j]c_j + \|\mathbf{c}\|^2 \\ &= \sum_j^D (\Sigma)_{jj} + \boldsymbol{\mu}^\top \boldsymbol{\mu} - 2\boldsymbol{\mu}^\top \mathbf{c} + \|\mathbf{c}\|^2 \\ &= \text{tr}(\Sigma) + \|\boldsymbol{\mu} - \mathbf{c}\|^2 \end{aligned}$$

Alternativamente, usando a esperança da forma quadrática $\mathbb{E}[X^\top \mathbf{A} X] = \boldsymbol{\mu}^\top \mathbf{A} \boldsymbol{\mu} + \text{tr}(\Sigma)$, para a matriz $\mathbf{A}$ de tamanho $N \times N$, e substituindo $\mathbf{A}$ pela matriz identidade $\mathbf{I}$ temos:

$$\begin{aligned} \mathbb{E}[\|X - \mathbf{c}\|^2] &= \mathbb{E}[X^\top X] - 2\mathbb{E}[X^\top \mathbf{c}] + \|\mathbf{c}\|^2 \\ &= \mathbb{E}[X^\top \mathbf{I} X] - 2\boldsymbol{\mu}^\top \mathbf{c} + \|\mathbf{c}\|^2 \\ &= \boldsymbol{\mu}^\top \boldsymbol{\mu} + \text{tr}(\Sigma) - 2\boldsymbol{\mu}^\top \mathbf{c} + \|\mathbf{c}\|^2 \\ &= \text{tr}(\Sigma) + \|\boldsymbol{\mu} - \mathbf{c}\|^2 \end{aligned}$$

A.2 Prova do Teorema 5

A função de Lagrange do Problema 4.2.3 é

$$\mathcal{L}(\alpha, \eta, \nu) = - \sum_{i=1}^N \alpha_i \langle \boldsymbol{\mu}_i, \boldsymbol{\mu}_i \rangle + \frac{\sum_{i,j=1}^N \alpha_i \alpha_j \langle \boldsymbol{\mu}_i, \boldsymbol{\mu}_j \rangle}{\sum_{i=1}^N \alpha_i} - \sum_{i=1}^N \alpha_i \text{tr}(\Sigma_i) - \eta \left(\sum_{i=1}^N \alpha_i \kappa_i - 1 \right) - \sum_{i=1}^N \nu_i (\alpha_i \kappa_i - \lambda)$$

onde

$$\begin{aligned} \partial_{\alpha_i} \mathcal{L} = & \frac{(\sum_{j=1}^N \alpha_j) 2 \langle \boldsymbol{\mu}_i, \sum_{j=1}^N \alpha_j \boldsymbol{\mu}_j \rangle}{(\sum_{j=1}^N \alpha_j)^2} - \frac{\sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j \langle \boldsymbol{\mu}_i, \boldsymbol{\mu}_j \rangle}{(\sum_{j=1}^N \alpha_j)^2} - \|\boldsymbol{\mu}_i\|^2 - \text{tr}(\Sigma_i) - \eta \kappa_i \\ & - \nu(\kappa_i - c) = 0 \end{aligned}$$

mas pela condição de complementaridade de (4.2.3) segue que $\nu_i = 0$ implica $\alpha_i > 0$, então

$$\partial_{\alpha_i} \mathcal{L} = 2 \langle \boldsymbol{\mu}_i, \mathbf{c} \rangle - \|\mathbf{c}\|^2 - \|\boldsymbol{\mu}_i\|^2 - \text{tr}(\Sigma_i) - \eta \kappa_i = 0.$$

Analisando as condições KKT do Problema 4.2.3, temos que $\alpha_i > 0, \beta_i > 0 \implies \xi_i = 0$ e $0 < \alpha_i \kappa_i < \lambda$. A partir disto e a condição de complementaridade das KKT, o raio pode ser encontrado como

$$R^2 = \frac{\|\boldsymbol{\mu}_i - \mathbf{c}\|^2 + \text{tr}(\Sigma_i)}{\kappa_i}, \quad (\text{A.1})$$

onde $i \in \{i \in \mathcal{I} \mid 0 < \alpha_i \kappa_i < \lambda\}$. Assim, segue que $\eta = R^2$, então $R = \sqrt{\eta}$.

A.3 Prova do Lema 6

$$\begin{aligned} \mathbb{E}_{\mathbb{P}}[\|k(X, \cdot) - c(\cdot)\|^2] &= \mathbb{E}_{\mathbb{P}}[\langle k(X, \cdot), k(X, \cdot) \rangle_{\mathcal{H}}] - 2 \langle \mu_{\mathbb{P}}, c(\cdot) \rangle_{\mathcal{H}} + \|c(\cdot)\|_{\mathcal{H}}^2 \\ &= \text{tr}(\Sigma^{\mathcal{H}}) + \|\mu_{\mathbb{P}}\|_{\mathcal{H}}^2 - 2 \langle \mu_{\mathbb{P}}, c(\cdot) \rangle_{\mathcal{H}} + \|c(\cdot)\|_{\mathcal{H}}^2 \\ &= \text{tr}(\Sigma^{\mathcal{H}}) + \|\mu_{\mathbb{P}} - c(\cdot)\|_{\mathcal{H}}^2 \end{aligned}$$

A.4 Prova do Teorema 8

Trocando as variáveis no Problema 4.3.1 por $-\rho = \frac{1}{2}(R^2 \kappa - \text{tr}(\Sigma^{\mathcal{H}}) - \|c(\cdot)\|_{\mathcal{H}}^2)$, implica $R^2 = (\text{tr}(\Sigma^{\mathcal{H}}) + \|c(\cdot)\|_{\mathcal{H}}^2 - 2\rho)/k$. Logo o Problema 4.3.1 pode ser escrito como:

$$\begin{aligned} \min_{c(\cdot) \in \mathcal{H}, R \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^N} \quad & \frac{\text{tr}(\Sigma^{\mathcal{H}}) + \|c(\cdot)\|_{\mathcal{H}}^2 - 2\rho}{k} + \lambda \sum_{i=1}^N \xi_i \\ \text{subject to} \quad & \langle \mu_{\mathbb{P}_i}, c(\cdot) \rangle_{\mathcal{H}} \geq \rho - \frac{1}{2}(\kappa \xi_i - \|\mu_{\mathbb{P}_i}\|_{\mathcal{H}}^2), \\ & \xi_i \geq 0, \end{aligned}$$

para todo $i = 1, \dots, N$. Fazendo $\xi'_i = \frac{1}{2}(\kappa \xi_i - \|\mu_{\mathbb{P}_i}\|_{\mathcal{H}}^2)$ implica $\xi_i = \frac{2\xi'_i + \|\mu_{\mathbb{P}_i}\|_{\mathcal{H}}^2}{k}$ e multiplicando a função objetivo por $\frac{k}{2}$ resulta:

$$\begin{aligned} \min_{c(\cdot) \in \mathcal{H}, R \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^N} \quad & \frac{1}{2}(\text{tr}(\Sigma^{\mathcal{H}}) + \|c(\cdot)\|_{\mathcal{H}}^2 - 2\rho) \\ & + \frac{k}{2} \lambda \sum_{i=1}^N \frac{2\xi'_i + \|\mu_{\mathbb{P}_i}\|_{\mathcal{H}}^2}{k} \\ \text{subject to} \quad & \langle \mu_{\mathbb{P}_i}, c(\cdot) \rangle_{\mathcal{H}} \geq \rho - \xi'_i, \quad i = 1, \dots, N \\ & \frac{2\xi'_i + \|\mu_{\mathbb{P}_i}\|_{\mathcal{H}}^2}{k} \geq 0, \quad i = 1, \dots, N. \end{aligned}$$

Isso pode ser simplificado como

$$\begin{aligned}
 & \min_{c(\cdot) \in \mathcal{H}, R \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^N} \quad \frac{1}{2}(\text{tr}(\Sigma^{\mathcal{H}}) + \|c(\cdot)\|_{\mathcal{H}}^2 - 2\rho) + \\
 & \quad \lambda \sum_{i=1}^N \xi'_i + \frac{1}{2} \lambda \sum_{i=1}^N \|\mu_{\mathbb{P}_i}\|_{\mathcal{H}}^2 \\
 & \text{subject to} \quad \langle \mu_{\mathbb{P}_i}, c(\cdot) \rangle_{\mathcal{H}} \geq \rho - \xi'_i, \quad i = 1, \dots, N \\
 & \quad \xi'_i \geq -\frac{\|\mu_{\mathbb{P}_i}\|_{\mathcal{H}}^2}{2}, \quad i = 1, \dots, N.
 \end{aligned}$$

Apagando o término constante, temos

$$\begin{aligned}
 & \min_{c(\cdot) \in \mathcal{H}, R \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^N} \quad \frac{\|c(\cdot)\|_{\mathcal{H}}^2}{2} - \rho + \lambda \sum_{i=1}^N \xi'_i \\
 & \text{subject to} \quad \langle \mu_{\mathbb{P}_i}, c(\cdot) \rangle_{\mathcal{H}} \geq \rho - \xi'_i, \quad i = 1, \dots, N \\
 & \quad \xi'_i \geq -\frac{\|\mu_{\mathbb{P}_i}\|_{\mathcal{H}}^2}{2}, \quad i = 1, \dots, N.
 \end{aligned}$$

A.5 Prova do Teorema 9

A partir de (4.28), se:

$$\xi'' = \frac{1}{2}\kappa\xi \implies \xi'' = \xi' + \frac{\|\mu_{\mathbb{P}_i}\|_{\mathcal{H}}^2}{2},$$

então, o Problema 4.4.1 pode ser escrito como:

$$\begin{aligned}
 & \min_{c(\cdot) \in \mathcal{H}, \rho' \in \mathbb{R}, \boldsymbol{\xi}'' \in \mathbb{R}^N} \quad \frac{\|c(\cdot)\|_{\mathcal{H}}^2}{2} - \rho' + \lambda \sum_{i=1}^N (\xi''_i - \frac{\|\mu_{\mathbb{P}_i}\|_{\mathcal{H}}^2}{2}) \\
 & \text{subject to} \quad \langle \mu_{\mathbb{P}_i}, c(\cdot) \rangle_{\mathcal{H}} \geq \rho' - \xi''_i + \frac{\|\mu_{\mathbb{P}_i}\|_{\mathcal{H}}^2}{2}, \\
 & \quad \xi''_i \geq 0,
 \end{aligned}$$

para todo $i = 1, \dots, N$. A função de Lagrange para o problema anterior é

$$\begin{aligned}
 \mathcal{L}(c(\cdot), \rho, \boldsymbol{\xi}, \boldsymbol{\alpha}, -\boldsymbol{\beta}) &= \frac{\|c(\cdot)\|_{\mathcal{H}}^2}{2} - \rho' + \lambda \sum_{i=1}^N (\xi''_i - \frac{\|\mu_{\mathbb{P}_i}\|_{\mathcal{H}}^2}{2}) \\
 &\quad - \sum_{i=1}^N \alpha_i \{ \langle \mu_{\mathbb{P}_i}, c(\cdot) \rangle_{\mathcal{H}} - \rho' + \xi''_i - \frac{\|\mu_{\mathbb{P}_i}\|_{\mathcal{H}}^2}{2} \} \\
 &\quad - \sum_{i=1}^N \beta_i \xi''_i
 \end{aligned} \tag{A.2}$$

As condições KKT para este problema são:

$$\left. \begin{array}{lll} \partial_\rho \mathcal{L} = 0 & : & \sum_{i=1}^N \alpha_i \\ \nabla_{c(\cdot)} \mathcal{L} = 0 & : & c(\cdot) - \sum_{i=1}^N \alpha_i \mu_{\mathbb{P}_i} \\ \partial_{\xi_i''} \mathcal{L} = 0 & : & \lambda - \alpha_i - \beta_i \end{array} \right\} = \begin{array}{l} 1 \\ 0 \\ 0 \end{array} \quad (\text{A.3})$$

$$\left. \begin{array}{ll} \alpha_i \{ \langle \mu_{\mathbb{P}_i}, c(\cdot) \rangle_{\mathcal{H}} - \rho' + \xi_i'' - \frac{\|\mu_{\mathbb{P}_i}\|_{\mathcal{H}}^2}{2} \} & = 0 \\ \beta_i \xi_i'' & = 0 \end{array} \right\} \quad (\text{A.4})$$

Substituindo (A.3) em (A.2) resulta em $\frac{1}{2} \sum_{i=1}^N (\alpha_i - \lambda) \tilde{k}(\mathbb{P}_i, \mathbb{P}_i) - \frac{1}{2} \sum_{i,j=1}^N \alpha_i \alpha_j \tilde{k}(\mathbb{P}_i, \mathbb{P}_j)$, mas $-\frac{\lambda}{2} \sum_{i=1}^N \tilde{k}(\mathbb{P}_i, \mathbb{P}_i)$ é constante, então, a forma dual do problema anterior é:

$$\begin{aligned} \max_{\alpha \in \mathbb{R}^N} \quad & \frac{1}{2} \sum_{i=1}^N \alpha_i \tilde{k}(\mathbb{P}_i, \mathbb{P}_i) - \frac{1}{2} \sum_{i,j=1}^N \alpha_i \alpha_j \tilde{k}(\mathbb{P}_i, \mathbb{P}_j) \\ \text{subject to} \quad & 0 \leq \alpha_i \leq \lambda, \ i = 1, \dots, N \\ & \sum_{i=1}^N \alpha_i = 1, \end{aligned}$$

A.6 Prova do Corolário 11

Após de uma normalização esférica $\|\mu_{\mathbb{P}_i}\|^2 = 1$ é satisfeito. Então, se um kernel invariante a translação fosse usado, implica que o valor: $\sum_{i=1}^N \alpha_i \tilde{k}(\mathbb{P}_i, \mathbb{P}_i)$ é constante, consequentemente aqueles problemas são equivalentes.

Referências Bibliográficas

- Aisbett et al. (2010)** J. Aisbett, J.T. Rickard e D.G. Morgenthaler. Type-2 fuzzy sets as functions on spaces. *Fuzzy Systems, IEEE Transactions on*, 18(4):841–844. ISSN 1063-6706. doi: 10.1109/TFUZZ.2010.2046176. Citado na pág. [10](#)
- Alcalá-Fdez et al. (2011)** Jesús Alcalá-Fdez, Alberto Fernández, Julián Luengo, Joaquín Derrac e Salvador García. Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. *Multiple-Valued Logic and Soft Computing*, 17(2-3):255–287. Citado na pág. [99](#), [101](#)
- Aronszajn (1950)** N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68. Citado na pág. [17](#), [20](#)
- Ash e Doleans-Dade (2000)** Robert B Ash e Catherine Doleans-Dade. *Probability and measure theory*. Academic Press. ISBN 978-0120652020. Citado na pág. [41](#), [70](#)
- Ben-Tal et al. (2011)** Aharon Ben-Tal, Sahely Bhadra, Chiranjib Bhattacharyya e J Saketha Nath. Chance constrained uncertain classification via robust optimization. *Mathematical programming*, 127(1):145–173. Citado na pág. [1](#), [2](#), [30](#), [39](#)
- Benferhat e Smaoui (2007)** Salem Benferhat e Salma Smaoui. Hybrid possibilistic networks. *International journal of approximate reasoning*, 44(3):224–243. Citado na pág. [9](#)
- Bennett e Bredensteiner (2000)** Kristin P Bennett e Erin J Bredensteiner. Duality and geometry in svm classifiers. Em *ICML*, páginas 57–64. Citado na pág. [24](#)
- Berg et al. (1984)** C. Berg, J. P. R. Christensen e P. Ressel. *Harmonic Analysis on Semigroups*. Springer, Berlin. ISBN 978-1-4612-1128-0. Citado na pág. [17](#), [66](#), [67](#), [76](#)
- Berlinet e Thomas-Agnan (2004)** Alain Berlinet e Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*, volume 3. Kluwer Academic Boston. ISBN 978-1-4613-4792-7. Citado na pág. [3](#), [17](#), [19](#), [20](#), [33](#), [36](#), [40](#), [41](#), [108](#)
- Bertoluzza et al. (2002)** Carlo Bertoluzza, Maria-Ángeles Gil e Dan A Ralescu. *Statistical modeling, analysis and management of fuzzy data*. Physica-Verlag. ISBN 978-3-7908-1800-0. Citado na pág. [9](#), [63](#)
- Bhadra (2012)** Sahely Bhadra. *Learning Robust Support Vector Machine Classifiers with Uncertain Observations*. Tese de Doutorado. Citado na pág. [2](#), [30](#)
- Bhadra et al. (2009a)** Sahely Bhadra, Chiranjib Bhattacharyya, Nagasuma R Chandra e I Saira Mian. A linear programming approach for estimating the structure of a sparse linear genetic network from transcript profiling data. *Algorithms for Molecular Biology*, 4(1):5. Citado na pág. [2](#), [30](#)
- Bhadra et al. (2009b)** Sahely Bhadra, J Nath, Aharon Ben-Tal e Chiranjib Bhattacharyya. Interval data classification under partial information: A chance-constraint approach. *Advances in Knowledge Discovery and Data Mining*, páginas 208–219. Citado na pág. [2](#), [30](#)

- Bhadra et al. (2010)** Sahely Bhadra, Sourangshu Bhattacharya, Chiranjib Bhattacharyya e Aharon Ben-Tal. Robust formulations for handling uncertainty in kernel matrices. Em *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, páginas 71–78. Citado na pág. [2](#), [30](#)
- Bi e Bennett (2001)** Jinbo Bi e Kristin P Bennett. Duality, geometry, and support vector regression. Em *NIPS*, páginas 593–600. Citado na pág. [24](#)
- Bi e Vapnik (2003)** Jinbo Bi e Vladimir N. Vapnik. Learning with rigorous support vector machines. Em *Proceedings of the 16th Annual Conference on Learning Theory*, páginas 35–42. AAAI Press. Citado na pág. [24](#)
- Bishop e Nasrabadi (2006)** Christopher M Bishop e Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 1. springer New York. ISBN 9780387310732. Citado na pág. [17](#)
- Bloch (1999)** Isabelle Bloch. On fuzzy distances and their use in image processing under imprecision. *Pattern Recognition*, 32(11):1873–1895. Citado na pág. [77](#), [108](#)
- Borgwardt et al. (2006)** Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf e Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57. Citado na pág. [26](#)
- Boser et al. (1992)** Bernhard E. Boser, Isabelle M. Guyon e Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. Em *Proceedings of the fifth annual workshop on Computational learning theory, COLT '92*, páginas 144–152, New York, NY, USA. ACM. ISBN 0-89791-497-X. doi: 10.1145/130385.130401. Citado na pág. [17](#), [21](#)
- Boucheron et al. (2013)** Stéphane Boucheron, Gábor Lugosi e Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press. ISBN 978-0199535255. Citado na pág. [35](#), [36](#)
- Cadenas et al. (2011)** J.M. Cadenas, M.C. Garrido, R. Martinez e P.P. Bonissone. Towards the learning from low quality data in a fuzzy random forest ensemble. Em *Fuzzy Systems (FUZZ), 2011 IEEE International Conference on*, páginas 2897–2904. doi: 10.1109/FUZZY.2011.6007512. Citado na pág. [113](#)
- Canu et al. (2005)** S. Canu, Y. Grandvalet, V. Guigue e A. Rakotomamonjy. Svm and kernel methods matlab toolbox. Perception Systemes et Information, INSA de Rouen, Rouen, France, 2005. Citado na pág. [51](#)
- Cara et al. (2011)** A.B. Cara, I. Rojas, H. Pomares, C. Wagner e H. Hagrais. On comparing non-singleton type-1 and singleton type-2 fuzzy controllers for a nonlinear servo system. Em *Advances in Type-2 Fuzzy Logic Systems (T2FUZZ), 2011 IEEE Symposium on*, páginas 126–133. doi: 10.1109/T2FUZZ.2011.5949560. Citado na pág. [16](#)
- Chandola et al. (2009)** Varun Chandola, Arindam Banerjee e Vipin Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3):15:1–15:58. ISSN 0360-0300. Citado na pág. [29](#)
- Chapelle et al. (1999)** Olivier Chapelle, Patrick Haffner e Vladimir N Vapnik. Support vector machines for histogram-based image classification. *Neural Networks, IEEE Transactions on*, 10(5):1055–1064. Citado na pág. [77](#)
- Chaudhur e Rosenfeld (1996)** BB Chaudhur e Azriel Rosenfeld. On a metric distance between fuzzy sets. *Pattern Recognition Letters*, 17(11):1157–1160. Citado na pág. [77](#), [108](#)
- Chaudhuri e Rosenfeld (1999)** Bidyut Baran Chaudhuri e Azriel Rosenfeld. A modified hausdorff distance between fuzzy sets. *Information Sciences*, 118(1):159–171. Citado na pág. [108](#)

- Chen et al. (1995)** Shyi-Ming Chen, Ming-Shiow Yeh e Pei-Yung Hsiao. A comparison of similarity measures of fuzzy values. *Fuzzy sets and systems*, 72(1):79–89. Citado na pág. 108
- Chen (2008)** Yixin Chen. Support vector machines and fuzzy systems. Em Oded Maimon e Lior Rokach, editors, *Soft Computing for Knowledge Discovery and Data Mining*, páginas 205–223. Springer US. ISBN 978-0-387-69934-9. doi: 10.1007/978-0-387-69935-6_9. Citado na pág. 64, 90
- Chen e Wang (2003)** Yixin Chen e J.Z. Wang. Support vector learning for fuzzy rule-based classification systems. *IEEE Transactions on Fuzzy Systems*, 11(6):716 – 728. ISSN 1063-6706. doi: 10.1109/TFUZZ.2003.819843. Citado na pág. 64, 90, 93, 112
- Cheng e Juang (2011)** Wei-Yuan Cheng e Chia-Feng Juang. An incremental support vector machine-trained ts-type fuzzy system for online classification problems. *Fuzzy Sets Syst.*, 163: 24–44. ISSN 0165-0114. doi: <http://dx.doi.org/10.1016/j.fss.2010.08.006>. Citado na pág. 64, 90, 93
- Chua e Tan (2011)** Teck Wee Chua e Woei Wan Tan. Non-singleton genetic fuzzy logic system for arrhythmias classification. *Engineering Applications of Artificial Intelligence*, 24(2):251–259. Citado na pág. 16
- Cortes e Vapnik (1995)** Corinna Cortes e Vladimir Vapnik. Support-vector networks. *Mach. Learn.*, 20(3):273–297. ISSN 0885-6125. doi: 10.1023/A:1022627411411. URL <http://dx.doi.org/10.1023/A:1022627411411>. Citado na pág. 4, 17, 21, 24
- Cover (1965)** T. M. Cover. Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition. *Electronic Computers, IEEE Transactions on*, EC-14 (3):326–334. Citado na pág. 25
- Cuturi et al. (2005)** Marco Cuturi, Kenji Fukumizu e Jean-Philippe Vert. Semigroup kernels on measures. Em *Journal of Machine Learning Research*, páginas 1169–1198. Citado na pág. 21
- Debnath e Mikusiński (2005)** L. Debnath e P. Mikusiński. *Hilbert Spaces with Applications*. Elsevier Academic Press. ISBN 9780122084386. Citado na pág. 19, 42
- Diamond et al. (1994)** Phil Diamond, Peter E Kloeden, Peter Eris Kloeden, Australia Mathematician e Peter Eris Kloeden. *Metric spaces of fuzzy sets: theory and applications*. World Scientific. ISBN 978-981-02-1731-0. Citado na pág. 9, 77
- Driankov et al. (1996)** Dimitar Driankov, Hans Hellendoorn e Michael Reinfrank. *An introduction to fuzzy control (2. Aufl.)*. Springer. ISBN 978-3-540-60691-8. Citado na pág. 16
- Duda et al. (2000)** Richard O. Duda, Peter E. Hart e David G. Stork. *Pattern Classification (2Nd Edition)*. Wiley-Interscience. ISBN 0471056693. Citado na pág. 17
- Ferguson (1974)** Thomas S Ferguson. Prior distributions on spaces of probability measures. *The Annals of Statistics*, páginas 615–629. Citado na pág. 32
- Fukumizu et al. (2004)** Kenji Fukumizu, Francis R Bach e Michael I Jordan. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *The Journal of Machine Learning Research*, 5:73–99. Citado na pág. 3, 41, 42
- Fukumizu et al. (2008)** Kenji Fukumizu, Arthur Gretton, Xiaohai Sun e Bernhard Schölkopf. Kernel measures of conditional dependence. Em J.C. Platt, D. Koller, Y. Singer e S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, páginas 489–496. MIT Press, Cambridge, MA. Citado na pág. 40
- Fukumizu et al. (2010)** Kenji Fukumizu, Le Song e Arthur Gretton. Kernel bayes' rule. *arXiv preprint arXiv:1009.5736*. Citado na pág. 3, 41

- Garcia et al. (2003)** Javier Nuñez Garcia, Zoltan Kutalik, Kwang-Hyun Cho e Olaf Wolkenhauer. Level sets and minimum volume sets of probability density functions. *International Journal of Approximate Reasoning*, 34(1):25 – 47. ISSN 0888-613X. Citado na pág. 29, 31, 32
- Gartner (2008)** Thomas Gartner. *Kernels for structured data*, volume 72. World Scientific. ISBN 978-3-540-00567-4. Citado na pág. 21, 65, 69, 70
- Gärtner et al. (2002)** Thomas Gärtner, Peter A Flach, Adam Kowalczyk e Alex J Smola. Multi-instance kernels. Em *ICML*, volume 2, páginas 179–186. Citado na pág. 64
- Genton (2002)** Marc G. Genton. Classes of kernels for machine learning: A statistics perspective. *J. Mach. Learn. Res.*, 2:299–312. ISSN 1532-4435. Citado na pág. 47
- Graepel e Herbrich (2003)** Thore Graepel e Ralf Herbrich. Invariant pattern recognition by semidefinite programming machines. Em *Advances in Neural Information Processing Systems 16*, página 2004. MIT Press. Citado na pág. 1, 30
- Grant e Boyd (2014)** Michael Grant e Stephen Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. <http://cvxr.com/cvx>, Março 2014. Citado na pág. 51
- Grant e Boyd (2008)** Michael Grant e Stephen Boyd. Graph implementations for nonsmooth convex programs. Em V. Blondel, S. Boyd e H. Kimura, editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, páginas 95–110. Springer-Verlag Limited. Citado na pág. 51
- Gretton et al. (2005a)** Arthur Gretton, Ralf Herbrich, Alexander Smola, Olivier Bousquet e Bernhard Schölkopf. Kernel methods for measuring independence. *J. Mach. Learn. Res.*, 6: 2075–2129. ISSN 1532-4435. Citado na pág. 3, 41
- Gretton et al. (2005b)** Arthur Gretton, Ralf Herbrich, Alexander Smola, Olivier Bousquet e Bernhard Schölkopf. Kernel methods for measuring independence. *The Journal of Machine Learning Research*, 6:2075–2129. Citado na pág. 42
- Gretton et al. (2006)** Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf e Alex J Smola. A kernel method for the two-sample-problem. Em *Advances in neural information processing systems*, páginas 513–520. Citado na pág. 26
- Gretton et al. (2009)** Arthur Gretton, Kenji Fukumizu, Zaid Harchaoui e Bharath K Sriperumbudur. A fast, consistent kernel two-sample test. Em *Advances in neural information processing systems*, páginas 673–681. Citado na pág. 26
- Gretton et al. (2012)** Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf e Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13: 723–773. Citado na pág. 3, 26, 27, 40, 41
- Guevara et al. (2014a)** J. Guevara, R. Hirata e S. Canu. Positive definite kernel functions on fuzzy sets. Em *Fuzzy Systems (FUZZ-IEEE), 2014 IEEE International Conference on*, páginas 439–446. doi: 10.1109/FUZZ-IEEE.2014.6891628. Citado na pág. 5
- Guevara et al. (2015a)** J. Guevara, R. Hirata e S. Canu. *Fuzzy Set Similarity using a Distance-Based Kernel on Fuzzy Sets*. Citado na pág. 5, 109
- Guevara et al. (2013a)** Jorge Guevara, Roberto Hirata e Stephane Canu. Kernel functions in takagi-sugeno-kang fuzzy system with nonsingleton fuzzy input. Em *Fuzzy Systems (FUZZ), 2013 IEEE International Conference on*, páginas 1–8. doi: 10.1109/FUZZ-IEEE.2013.6622409. Citado na pág. 1, 5, 30, 112

- Guevara et al. (2013b)** Jorge Guevara, Roberto Hirata e Stephane Canu. Support vector data description for uncertainty datasets, 2013b. Poster presented at The Machine Learning Summer School 26 August to 6 September 2013 at the Max Planck Institute for Intelligent Systems, Tübingen, Germany. Citado na pág. 4
- Guevara et al. (2014b)** Jorge Guevara, e Roberto Canu, Stephane Hirata. Support measure data description. submitted to pattern analysis and machine intelligence, Institute of Mathematics and Statistics. Citado na pág. 4, 5
- Guevara et al. (2015b)** Jorge Guevara, Roberto Canu, Stephane Hirata, J. Guevara, R. Hirata e S. Canu. Support measure data description for group anomaly detection. páginas 24–33. Citado na pág. 4, 5, 110
- Guilbart (1979)** C Guilbart. Produits scalaires sur l'espace des mesures. Em *Annales de l'institut Henri Poincaré (B) Probabilités et Statistiques*, volume 15, páginas 333–354. Gauthier-Villars. Citado na pág. 33, 40, 41
- Haasdonk e Bahlmann (2004)** B. Haasdonk e C. Bahlmann. Learning with distance substitution kernels. *Pattern Recognition*, páginas 220–227. Citado na pág. 76, 79
- Haasdonk (2005)** Bernard Haasdonk. Feature space interpretation of svms with indefinite kernels. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(4):482–492. Citado na pág. 66, 76
- Haasdonk e Keysers (2002)** Bernard Haasdonk e Daniel Keysers. Tangent distance kernels for support vector machines. Em *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 2, páginas 864–868. IEEE. Citado na pág. 76, 77
- Hakan Akpolat (2005)** Z. Hakan Akpolat. Non-singleton fuzzy logic control of a dc motor. *Journal of Applied Sciences*, 5:887–891. Citado na pág. 16
- Haussler (1999)** David Haussler. Convolution kernels on discrete structures. Relatório técnico, Technical report, Department of Computer Science, University of California at Santa Cruz. Citado na pág. 64, 65
- Heo e Gader (2011)** Gyeongyong Heo e Paul Gader. Robust kernel discriminant analysis using fuzzy memberships. *Pattern Recogn.*, 44(3):716–723. ISSN 0031-3203. doi: 10.1016/j.patcog.2010.10.007. Citado na pág. 64, 89
- Höppner (1999)** Frank Höppner. *Fuzzy cluster analysis: methods for classification, data analysis and image recognition*. John Wiley & Sons. ISBN 978-0-471-98864-9. Citado na pág. 9
- Huang et al. (2012)** Hsin-Chien Huang, Yung-Yu Chuang e Chu-Song Chen. Multiple kernel fuzzy clustering. *Fuzzy Systems, IEEE Transactions on*, 20(1):120–134. ISSN 1063-6706. doi: 10.1109/TFUZZ.2011.2170175. Citado na pág. 64, 89
- Jaakkola et al. (1999)** Tommi Jaakkola, David Haussler et al. Exploiting generative models in discriminative classifiers. *Advances in neural information processing systems*, páginas 487–493. Citado na pág. 41
- Jebara e Kondor (2003a)** Tony Jebara e Risi Kondor. Bhattacharyya and expected likelihood kernels. Em *In Conference on Learning Theory*. press. Citado na pág. 21
- Jebara e Kondor (2003b)** Tony Jebara e Risi Kondor. Bhattacharyya and expected likelihood kernels. Em *Learning Theory and Kernel Machines*, páginas 57–71. Springer. Citado na pág. 41
- Jebara et al. (2004)** Tony Jebara, Risi Kondor e Andrew Howard. Probability product kernels. *The Journal of Machine Learning Research*, 5:819–844. Citado na pág. 21

- Jeng e Lee (1999)** Jin-Tsong Jeng e Tsu-Tain Lee. Support vector machines for the fuzzy neural networks. Em *Systems, Man, and Cybernetics, 1999. IEEE SMC '99 Conference Proceedings. 1999 IEEE International Conference on*, volume 6, páginas 115–120 vol.6. doi: 10.1109/ICSMC.1999.816469. Citado na pág. 64, 90
- Juang et al. (2007)** Chia-Feng Juang, Shih-Hsuan Chiu e Shu-Wew Chang. A self-organizing ts-type fuzzy network with support vector learning and its application to classification problems. *IEEE Transactions on Fuzzy Systems*, 15(5):998–1008. ISSN 1063-6706. doi: 10.1109/TFUZZ.2007.894980. Citado na pág. 64, 90, 93
- Kanzawa et al. (2010)** Y. Kanzawa, Y. Endo e S. Miyamoto. On kernel fuzzy c-means for data with tolerance using explicit mapping for kernel data analysis. Em *Fuzzy Systems (FUZZ), 2010 IEEE International Conference on*, páginas 1–6. IEEE. Citado na pág. 64, 89
- Kim et al. (2004)** Dongwon Kim, Sung-hoe Huh e Gwi-tae Park. Modeling corrupted time series data via nonsingleton fuzzy logic system 2 nonsingleton fuzzy logic system. *Time*, páginas 1298–1303. Citado na pág. 16
- Klement et al. (2000)** Erich Peter Klement, Radko Mesiar e Endre Pap. *Triangular Norms*. Springer, 1 ed. ISBN 978-90-481-5507-1. Citado na pág. 9, 10, 15
- Kolmogorov (1941)** A. N. Kolmogorov. Stationary sequences in Hilbert space. *Bull. Math. Univ. Moscow*, 2(6). Citado na pág. 17
- Kondor e Jebara (2003)** Risi Kondor e Tony Jebara. A kernel between sets of vectors. Em *ICML*, páginas 361–368. Citado na pág. 1, 30, 41
- Kosko (1994)** B. Kosko. Fuzzy systems as universal approximators. *Computers, IEEE Transactions on*, 43(11):1329–1333. ISSN 0018-9340. doi: 10.1109/12.324566. Citado na pág. 13
- Kreinovich e Mouzouris (1996)** V. Kreinovich e G. C. Mouzouris. Fuzzy rule based modeling as a universal approximation tool, 1996. Citado na pág. 13
- Kwakernaak (1978)** Huibert Kwakernaak. Fuzzy random variables—i. definitions and theorems. *Information Sciences*, 15(1):1–29. Citado na pág. 63
- Laptev (2005)** Ivan Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123. Citado na pág. 1, 30
- Leekwijck e Kerre (1999)** Werner Van Leekwijck e E E Kerre. Defuzzification: criteria and classification. *Fuzzy Sets and Systems*, 108(2):159–178. Citado na pág. 16
- Lichman (2013)** M. Lichman. UCI machine learning repository, 2013. URL <http://archive.ics.uci.edu/ml>. Citado na pág. 81
- Lin et al. (2006)** Chin-Teng Lin, Chang-Mao Yeh, Sheng-Fu Liang, Jen-Feng Chung e N. Kumar. Support-vector-based fuzzy neural network for pattern classification. *IEEE Transactions on Fuzzy Systems*, 14(1):31–41. ISSN 1063-6706. doi: 10.1109/TFUZZ.2005.861604. Citado na pág. 64, 90
- Liu e Xue (2012)** Fengqiu Liu e Xiaoping Xue. Design of natural classification kernels using prior knowledge. *Trans. Fuz Sys.*, 20(1):135–152. ISSN 1063-6706. doi: 10.1109/TFUZZ.2011.2170428. Citado na pág. 64, 90
- Lobo et al. (1998)** Miguel Sousa Lobo, Lieven Vandenberghe, Stephen Boyd e Hervé Lebrete. Applications of second-order cone programming. *Linear algebra and its applications*, 284(1):193–228. Citado na pág. 39

- Lodhi et al. (2002)** Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini e Chris Watkins. Text classification using string kernels. *The Journal of Machine Learning Research*, 2: 419–444. Citado na pág. [21](#)
- Lowe (2004)** David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110. Citado na pág. [1](#), [30](#)
- Lski (2006)** Jacek M. Lski. On support vector regression machines with linguistic interpretation of the kernel matrix. *Fuzzy Sets Syst.*, 157(8):1092–1113. ISSN 0165-0114. doi: 10.1016/j.fss.2005.09.012. Citado na pág. [64](#), [90](#)
- Mamdani (1974)** E.H. Mamdani. Application of fuzzy algorithms for control of simple dynamic plant. *Electrical Engineers, Proceedings of the Institution of*, 121(12):1585–1588. ISSN 0020-3270. doi: 10.1049/piee.1974.0328. Citado na pág. [14](#)
- Melin e Castillo (2005)** Patricia Melin e Oscar Castillo. *Hybrid Intelligent Systems for Pattern Recognition Using Soft Computing: An Evolutionary Approach for Neural Networks and Fuzzy Systems (Studies in Fuzziness and Soft Computing)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA. ISBN 3540241213. Citado na pág. [10](#)
- Mendel (2001)** J. M. Mendel. *Uncertain rule-based fuzzy logic system: introduction and new directions*. Prentice–Hall PTR. ISBN 978-0130409690. Citado na pág. [9](#), [10](#), [14](#), [16](#), [91](#)
- Mendel (1995)** J.M. Mendel. Fuzzy logic systems for engineering: a tutorial. *Proceedings of the IEEE*, 83(3):345–377. ISSN 0018-9219. doi: 10.1109/5.364485. Citado na pág. [16](#)
- Menger (1942)** K Menger. Statistical metrics. *Proceedings of the National Academy of Sciences of the United States of America*, 28. Citado na pág. [9](#)
- Mercer (1909)** J. Mercer. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 209:pp. 415–446. ISSN 02643952. Citado na pág. [17](#)
- Moreno et al. (2003)** Pedro J Moreno, Purdy P Ho e Nuno Vasconcelos. A kullback-leibler divergence based kernel for svm classification in multimedia applications. Em *Advances in neural information processing systems*, página None. Citado na pág. [41](#), [76](#), [77](#)
- Moser (2006a)** Bernhard Moser. On the t-transitivity of kernels. *Fuzzy Sets and Systems*, 157(13):1787–1796. ISSN 0165-0114. doi: 10.1016/j.fss.2006.01.007. Citado na pág. [64](#), [90](#), [98](#), [99](#)
- Moser (2006b)** Bernhard Moser. On representing and generating kernels by fuzzy equivalence relations. *J. Mach. Learn. Res.*, 7:2603–2620. ISSN 1532-4435. Citado na pág. [64](#), [90](#), [98](#), [99](#)
- Mouzouris e Mendel (1997)** G.C. Mouzouris e J.M. Mendel. Nonsingleton fuzzy logic systems: theory and application. *Fuzzy Systems, IEEE Transactions on*, 5(1):56–71. ISSN 1063-6706. doi: 10.1109/91.554447. Citado na pág. [14](#), [16](#)
- Muandet e Schoelkopf (2013)** Krikamol Muandet e Bernhard Schoelkopf. One-class support measure machines for group anomaly detection. Em *Proceedings of the Twenty-Ninth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-13)*, páginas 449–458, Corvallis, Oregon. AUAI Press. Citado na pág. [1](#), [3](#), [30](#), [31](#), [32](#), [41](#), [47](#), [48](#), [50](#), [51](#), [52](#), [53](#), [57](#)
- Muandet et al. (2012)** Krikamol Muandet, Kenji Fukumizu, Francesco Dinuzzo e Bernhard Schölkopf. Learning from distributions via support measure machines. Em P. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Bottou e K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, páginas 10–18. Citado na pág. [1](#), [3](#), [30](#), [41](#)

- Noma (2002)** Hiroshi Shimodaira Ken-ichi Noma. Dynamic time-alignment kernel in support vector machine. *Advances in neural information processing systems*, 14:921. Citado na pág. 76
- Olaru e Wehenkel (2003)** Cristina Olaru e Louis Wehenkel. A complete fuzzy decision tree technique. *Fuzzy sets and systems*, 138(2):221–254. Citado na pág. 9
- Ong et al. (2004)** C.S. Ong, X. Mary, S. Canu e A.J. Smola. Learning with non-positive kernels. Em *Proceedings of the twenty-first international conference on Machine learning*, página 81. ACM. Citado na pág. 66, 76
- Palacios e Alcalá-Fdez (2012)** A. Palacios e J. Alcalá-Fdez. Mining fuzzy association rules from low quality data. *Soft Computing - A Fusion of Foundations, Methodologies and Applications*, páginas 0–0. Citado na pág. 63
- Palacios et al. (2010)** A. Palacios, L. Sánchez e I. Couso. Diagnosis of dyslexia with low quality data with genetic fuzzy systems. *International Journal of Approximate Reasoning*, 51(8):993–1009. Citado na pág. 63
- Palacios et al. (2009)** A.M. Palacios, L. Sánchez e I. Couso. Extending a simple genetic cooperative-competitive learning fuzzy classifier to low quality datasets. *Evolutionary Intelligence*, 2(1-2):73–84. Citado na pág. 99
- Palacios et al. (2011)** Ana M. Palacios, Luciano Sánchez e Inés Couso. Future performance modeling in athletics with low quality data-based genetic fuzzy systems. *Multiple-Valued Logic and Soft Computing*, 17(2-3):207–228. Citado na pág. 63, 99
- Pedrycz e Gomide (2007)** Witold Pedrycz e Fernando A. C. Gomide. *Fuzzy Systems Engineering - Toward Human-Centric Computing*. Wiley. ISBN 978-0-471-78857-7. Citado na pág. 10
- Pekalska et al. (2002)** Elzbieta Pekalska, Pavel Paclik e Robert PW Duin. A generalized kernel approach to dissimilarity-based classification. *The Journal of Machine Learning Research*, 2: 175–211. Citado na pág. 76
- Póczos et al. (2012)** Barnabás Póczos, Liang Xiong e Jeff G. Schneider. Nonparametric divergence estimation with applications to machine learning on distributions. *CoRR*, abs/1202.3758. Citado na pág. 1, 30, 57
- Polonik (1997)** Wolfgang Polonik. Minimum volume sets and generalized quantile processes. *Stochastic Processes and their Applications*, 69(1):1 – 24. ISSN 0304-4149. Citado na pág. 29, 31, 32
- Puri e Ralescu (1986)** Madan L Puri e Dan A Ralescu. Fuzzy random variables. *Journal of mathematical analysis and applications*, 114(2):409–422. Citado na pág. 63
- Rakotomamonjy et al. (2007)** A. Rakotomamonjy, F. Bach, S. Canu e Y. Grandvalet. More efficiency in multiple kernel learning. Em *Proceedings of the 24th international conference on Machine learning*, páginas 775–782. ACM. Citado na pág. 17
- Rakotomamonjy et al. (2008)** A. Rakotomamonjy, F. Bach, S. Canu e Y. Grandvalet. Simplemkl. *Journal of Machine Learning Research*, 9:2491–2521. Citado na pág. 79
- Rasmussen e Williams (2006)** Carl E. Rasmussen e Christopher Williams. *Gaussian Processes for Machine Learning*. MIT Press. ISBN 0-262-18253-X. Citado na pág. 17, 81
- Raza e Rhee (2012)** M.A. Raza e F.C.-H. Rhee. Interval type-2 approach to kernel possibilistic c-means clustering. Em *Fuzzy Systems (FUZZ-IEEE), 2012 IEEE International Conference on*, páginas 1 –7. doi: 10.1109/FUZZ-IEEE.2012.6251233. Citado na pág. 64, 89
- Rosenfeld (1985)** Azriel Rosenfeld. Distances between fuzzy sets. *Pattern Recognition Letters*, 3 (4):229–233. Citado na pág. 77, 108

- Sánchez et al. (2009)** L. Sánchez, I. Couso e J. Casillas. Genetic learning of fuzzy rules based on low quality data. *Fuzzy Sets and Systems*, 160(17):2524–2552. Citado na pág. 63
- Scholkopf (2001)** Bernhard Scholkopf. The kernel trick for distances. *Advances in neural information processing systems*, páginas 301–307. Citado na pág. 53, 66, 76
- Schölkopf e Smola (2002)** Bernhard Schölkopf e Alexander J. Smola. *Learning with kernels : support vector machines, regularization, optimization, and beyond*. Adaptive computation and machine learning. MIT Press. ISBN 9780585477596. Citado na pág. 17, 18, 24, 80
- Schölkopf et al. (2000)** Bernhard Schölkopf, Alex J Smola, Robert C Williamson e Peter L Bartlett. New support vector algorithms. *Neural computation*, 12(5):1207–1245. Citado na pág. 24
- Schölkopf et al. (2001)** Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola e Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471. Citado na pág. 29, 31, 32, 48, 50, 81, 113
- Scholkopft e Mullert (1999)** Bernhard Scholkopft e Klaus-Robert Mullert. Fisher discriminant analysis with kernels. *Neural networks for signal processing IX*. Citado na pág. 80
- Schölkopf et al. (2001)** Bernhard Schölkopf, Ralf Herbrich e AlexJ. Smola. A generalized representer theorem. Em David Helmbold e Bob Williamson, editors, *Computational Learning Theory*, volume 2111 of *Lecture Notes in Computer Science*, páginas 416–426. Springer Berlin Heidelberg. ISBN 978-3-540-42343-0. Citado na pág. 38
- Scott e Nowak (2006)** Clayton Scott e Robert D. Nowak. Learning minimum volume sets. *Journal of Machine Learning Research*, 7:665–704. Citado na pág. 29, 31, 32
- Shapiro et al. (2009)** Alexander Shapiro, Darinka Dentcheva e Andrzej P Ruszczyński. *Lectures on stochastic programming: modeling and theory*, volume 9. SIAM. ISBN 9780898716870. Citado na pág. 33, 35
- Shapiro (2009)** Arnold F Shapiro. Fuzzy random variables. *Insurance: Mathematics and Economics*, 44(2):307–314. Citado na pág. 63
- Shawe-Taylor e Cristianini (2004)** John Shawe-Taylor e Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, New York, NY, USA. ISBN 0521813972. Citado na pág. 17, 20
- Shi et al. (2010)** Y. Shi, B. Van Gasse, D. Ruan e E.E. Kerre. On dependencies and independencies of fuzzy implication axioms. *Fuzzy Sets and Systems*, 161(10):1388 – 1405. ISSN 0165-0114. doi: 10.1016/j.fss.2009.12.003. Citado na pág. 15
- Shivaswamy et al. (2006)** Pannagadatta K. Shivaswamy, Chiranjib Bhattacharyya e Alexander J. Smola. Second order cone programming approaches for handling missing and uncertain data. *J. Mach. Learn. Res.*, 7:1283–1314. ISSN 1532-4435. Citado na pág. 1, 2, 30, 39
- Simon e Flowers (2007)** Andr Simon e George T Flowers. Non-singleton fuzzy sets for disturbance attenuation. *International Journal*, (September):36849–36849. Citado na pág. 16
- Smola et al. (2007)** Alex Smola, Arthur Gretton, Le Song e Bernhard Schölkopf. A hilbert space embedding for distributions. Em *Algorithmic Learning Theory*, páginas 13–31. Springer. Citado na pág. 1, 3, 30, 33, 40, 41
- Smola e Schölkopf (2004)** Alex J Smola e Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222. Citado na pág. 81

- Song et al. (2010)** Le Song, Byron Boots, Sajid M Siddiqi, Geoffrey J Gordon e Alex J Smola. Hilbert space embeddings of hidden markov models. Em *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, páginas 991–998. Citado na pág. 3, 41
- Sriperumbudur et al. (2008)** Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Gert Lanckriet e Bernhard Schölkopf. Injective hilbert space embeddings of probability measures. Em *In COLT*. Citado na pág. 40, 41
- Sriperumbudur et al. (2010)** Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf e Gert RG Lanckriet. Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research*, 99:1517–1561. Citado na pág. 3, 40, 41
- Suquet et al. (1995)** Charles Suquet et al. Distances euclidiennes sur les mesures signees et applications a des theoremes de berry-esseen. *Bulletin of the Belgian Mathematical Society Simon Stevin*, 2(2):161–182. Citado na pág. 33, 40, 41
- Takagi e Sugeno (1985)** T Takagi e M Sugeno. Fuzzy identification of systems and its applications to modeling and control. *IEEE Transactions On Systems Man And Cybernetics*, 15(1):116–132. Citado na pág. 14, 89, 90
- Tax e Duin (2004)** David MJ Tax e Robert PW Duin. Support vector data description. *Machine learning*, 54(1):45–66. Citado na pág. 4, 17, 25, 29, 31, 33, 36, 37, 40, 47, 50, 51, 81, 113
- Tukey (1977)** John W. Tukey. *Exploratory Data Analysis*. Addison-Wesley. ISBN 978-0201076165. Citado na pág. 53
- Vapnik (1995)** Vladimir N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA. ISBN 0-387-94559-8. Citado na pág. 17, 21, 23
- Viertl (2011)** R. Viertl. *Statistical Methods for Fuzzy Data*. Wiley Series in Probability and Statistics. Wiley. ISBN 9780470699454. Citado na pág. 1, 2, 9, 30, 63
- Wahba (1990)** Grace Wahba. *Spline models for observational data*. Number 59. Siam. ISBN 978-0-898712-44-5. Citado na pág. 17
- Wallace e Ziemba (2005)** Stein W Wallace e William T Ziemba. *Applications of stochastic programming*. Siam. ISBN 978-0898715552. Citado na pág. 33, 35
- Wang et al. (2010)** Jianguo Wang, Jizhao Hua e Jianbo Guo. Fuzzy maximum scatter discriminant analysis with kernel methods. Em *Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference on*, volume 2, páginas 560–564. doi: 10.1109/FSKD.2010.5569484. Citado na pág. 64, 89
- Wang e Mendel (1992)** L.-X. Wang e J.M. Mendel. Fuzzy basis functions, universal approximation, and orthogonal least-squares learning. *Neural Networks, IEEE Transactions on*, 3(5):807–814. ISSN 1045-9227. doi: 10.1109/72.159070. Citado na pág. 13, 90
- Wang (2001)** Paul P Wang. *Computing with words*. John Wiley & Sons, Inc. ISBN 978-3-642-27472-5. Citado na pág. 63
- Wee e Wan (2008)** Chua Teck Wee e Tan Woei Wan. Efsvmrobust-fcm: Evolutionary fuzzy rule-based support vector machines classifier with fcm clustering. Em *Fuzzy Systems, 2008. FUZZ-IEEE 2008. (IEEE World Congress on Computational Intelligence). IEEE International Conference on*, páginas 606–612. doi: 10.1109/FUZZY.2008.4630431. Citado na pág. 64, 90
- Wernisch et al. (2003)** Lorenz Wernisch, Sharon L. Kendall, Shamit Soneji, Andreas Wietzorrek, Tanya Parish, Jason Hinds, Philip D. Butcher e Neil G. Stoker. Analysis of whole-genome microarray replicates using mixed models. *Bioinformatics*, 19(1):53–61. Citado na pág. 2, 30

- Xiong et al. (2011a)** Liang Xiong, Barnabás Póczos e Jeff G. Schneider. Group anomaly detection using flexible genre models. Em *NIPS*, páginas 1071–1079. Citado na pág. [1](#), [2](#), [30](#), [50](#), [52](#), [55](#)
- Xiong et al. (2011b)** Liang Xiong, Barnabás Póczos, Jeff G. Schneider, Andrew J. Connolly e Jake VanderPlas. Hierarchical probabilistic models for group anomaly detection. Em *AISTATS*, páginas 789–797. Citado na pág. [1](#), [30](#), [50](#), [52](#), [57](#)
- Xuecheng (1992)** Liu Xuecheng. Entropy, distance measure and similarity measure of fuzzy sets and their relations. *Fuzzy sets and systems*, 52(3):305–318. Citado na pág. [108](#)
- Yang e Gunn (2007)** Jianqiang Yang e Steve Gunn. Exploiting uncertain data in support vector classification. Em Bruno Apolloni, RobertJ. Howlett e Lakhmi Jain, editors, *Knowledge-Based Intelligent Information and Engineering Systems*, volume 4694 of *Lecture Notes in Computer Science*, páginas 148–155. Springer Berlin Heidelberg. ISBN 978-3-540-74828-1. Citado na pág. [1](#), [30](#), [39](#)
- Yang et al. (2011)** Xiaowei Yang, Guangquan Zhang, Jie Lu e Jun Ma. A kernel fuzzy c-means clustering-based fuzzy support vector machine algorithm for classification problems with outliers or noises. *Fuzzy Systems, IEEE Transactions on*, 19(1):105–115. ISSN 1063-6706. doi: 10.1109/TFUZZ.2010.2087382. Citado na pág. [64](#), [89](#)
- Yu e Zhang (2008)** Fusheng Yu e Mingxin Zhang. Generalized triangular norms based product and similarity of fuzzy sets. Em *FSKD (1)'08*, páginas 286–290. Citado na pág. [10](#)
- Zadeh (1994)** L.A. Zadeh. Soft computing and fuzzy logic. *Software, IEEE*, 11(6):48–56. ISSN 0740-7459. doi: 10.1109/52.329401. Citado na pág. [63](#)
- Zadeh (1965)** Lotfi A. Zadeh. Fuzzy sets. *Information and Control*, 8(3):338–353. Citado na pág. [4](#), [9](#), [10](#)
- Zadeh (1975)** Lotfi A. Zadeh. The concept of a linguistic variable and its application to approximate reasoning - ii. *Inf. Sci.*, 8(4):301–357. Citado na pág. [13](#), [14](#)
- Zadeh (1974)** Lotfi A. Zadeh. Fuzzy logic and its application to approximate reasoning. Em *World Computer Congress*, páginas 591–594. Citado na pág. [13](#)
- Zadeh (2010)** Lotfi A. Zadeh. A summary and update of fuzzy logic. Em *IEEE International Conference on Granular Computing*, páginas 42–44. doi: 10.1109/GrC.2010.144. Citado na pág. [9](#)
- Zadeh (1996)** Lotfi A. Zadeh. Fuzzy logic = computing with words. *Fuzzy Systems, IEEE Transactions on*, 4(2):103–111. Citado na pág. [15](#), [63](#)
- Zhang (2005)** Jinbo Bi Tong Zhang. Support vector classification with input data uncertainty. Em *Advances in Neural Information Processing Systems 17: Proceedings of the 2004 Conference*, volume 17, página 161. MIT Press. Citado na pág. [1](#), [2](#), [30](#), [39](#)
- Zhang et al. ()** Kun Zhang, Bernhard Schölkopf, Krikamol Muandet e Zhikun Wang. Domain adaptation under target and conditional shift. Citado na pág. [3](#), [41](#)
- Zhou e Gan (2007)** Shang-Ming Zhou e J.Q. Gan. Constructing l2-svm-based fuzzy classifiers in high-dimensional space with automatic model selection and fuzzy rule ranking. *IEEE Transactions on Fuzzy Systems*, 15(3):398–409. ISSN 1063-6706. doi: 10.1109/TFUZZ.2006.882464. Citado na pág. [64](#), [90](#)
- Zwick et al. (1987)** Rami Zwick, Edward Carlstein e David V Budescu. Measures of similarity among fuzzy concepts: A comparative analysis. *International Journal of Approximate Reasoning*, 1(2):221–242. Citado na pág. [108](#)