UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE CIÊNCIAS

DEPARTAMENTO DE ESTATÍSTICA E MATEMÁTICA APLICADA

PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM E MÉTODOS QUANTITATIVOS

MESTRADO ACADÊMICO EM MODELAGEM E MÉTODOS QUANTITATIVOS

DOUGLAS CHIELLE

UMA PROPOSTA DE MÁQUINA DE VETOR-SUPORTE NEBULOSA COM OPÇÃO DE REJEIÇÃO

FORTALEZA

2019

DOUGLAS CHIELLE

UMA PROPOSTA DE MÁQUINA DE VETOR-SUPORTE NEBULOSA COM OPÇÃO DE REJEIÇÃO

Dissertação apresentada ao Curso de Mestrado Acadêmico em Modelagem e Métodos Quantitativos do Programa de Pós-Graduação em Modelagem e Métodos Quantitativos do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Modelagem e Métodos Quantitativos. Área de Concentração: Interdisciplinar

Orientador: Prof. Dr. Ricardo Coelho Silva

Co-Orientador: Prof. Dr. Guilherme de Alencar Barreto

FORTALEZA

2019

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

C461p Chielle, Douglas.
Uma Proposta de Máquina de Vetor-Suporte Nebulosa com Opção de Rejeição / Douglas Chielle. – 2019.
71 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Centro de Ciências, Programa de Pós-Graduação em Modelagem e Métodos Quantitativos, Fortaleza, 2019.
Orientação: Prof. Dr. Ricardo Coelho Silva.
Coorientação: Prof. Dr. Guilherme de Alencar Barreto.

1. Classificação. 2. Máquinas de Vetores-Suporte. 3. Lógica Nebulosa. I. Título.

CDD 510

DOUGLAS CHIELLE

UMA PROPOSTA DE MÁQUINA DE VETOR-SUPORTE NEBULOSA COM OPÇÃO DE REJEIÇÃO

Dissertação apresentada ao Curso de Mestrado Acadêmico em Modelagem e Métodos Quantitativos do Programa de Pós-Graduação em Modelagem e Métodos Quantitativos do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Modelagem e Métodos Quantitativos. Área de Concentração: Interdisciplinar

Aprovada em: 21 de Junho de 2019

BANCA EXAMINADORA

Prof. Dr. Ricardo Coelho Silva (Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. Guilherme de Alencar Barreto (Co-Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. Tiberius de Oliveira e Bonates
Universidade Federal do Ceará (UFC)

Prof. Dr. Ajalmar Rêgo da Rocha Neto
Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE)

À minha mãe (*in memorian*).

AGRADECIMENTOS

Primeiramente, agradeço minha mãe, por ter sempre se esforçado para dar o melhor para seus filhos. Apesar de não estar mais presente entre nós, é graças a ela que estou aqui.

Ao meu pai e irmãos, obrigado pelo apoio e parceria. Um agradecimento especial ao Eduardo, pela ajuda e conselhos no desenvolvimento deste trabalho.

À minha esposa, Raisa, pelo seu amor, carinho e companheirismo. Obrigado por estar sempre presente, me incentivando e fazendo acreditar que chegaria ao final desta etapa.

Ao meu orientador, Prof. Dr. Ricardo Coelho, meus sinceros agradecimentos por sua serenidade, paciência e por todo o conhecimento compartilhado.

Ao meu co-orientador, Prof. Dr. Guilherme de Alencar Barreto, por sua sabedoria e contribuições dedicados à este trabalho.

Aos professores Dr. Ajalmar Rêgo da Rocha Neto e Dr. Tiberius de Oliveira e Bonates, obrigado pelas excelentes sugestões dadas no exame de qualificação.

Também agradeço aos colegas, funcionários e professores do MMQ. Obrigado pela prontidão e ensinamentos.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

E aqueles que foram vistos dançando foram julgados insanos por aqueles que não podiam escutar a música. (Friedrich Nietzsche)

RESUMO

O objetivo em problemas de classificação de padrões é atribuir um elemento de um conjunto de dados a uma dentre diversas classes. Devido à clara fundamentação matemática e em geral boa capacidade de generalização, métodos de *kernel* baseados em máquinas de vetores-suporte (SVMs) têm sido aplicados com sucesso em problemas de classificação. Para construir um classificador SVM busca-se pela superfície de decisão que melhor separe os elementos de classes diferentes, com as mais largas margens possíveis. A partir dessa superfície constrói-se uma função de decisão, que é utilizada para classificar novos elementos. Embora classificadores SVM tenham sido inicialmente propostos para tratar conjuntos linearmente separáveis, a fim de lidar com conjuntos de dados mais complexos, um parâmetro de regularização e variáveis de folga foram adicionadas à formulação original. Um problema dessa abordagem é que, além de ser necessário calibrar um parâmetro extra, há um aumento na quantidade de vetores-suporte usados para construir a função de decisão. Quanto maior o número de vetores-suporte, maior o custo computacional para classificar novos padrões. Nesta dissertação, introduz-se um classificador SVM que utiliza a lógica nebulosa para lidar com incertezas nos conjuntos de dados. Com esta abordagem, os problemas oriundos da introdução de um parâmetro regularizador são evitados. Além disso, o classificador resultante da proposta atribui valores de pertinência aos elementos do conjunto de dados, permitindo que uma classe de rejeição seja introduzida na formulação sem maiores dificuldades. As duas versões do modelo proposto, com e sem opção de rejeição, foram testadas em diversos conjuntos de dados oriundos da área médica e comparadas a outras formulações de classificadores SVM. A versão sem opção de rejeição apresentou taxas de acerto similares àquelas do SVM com margem flexível, com a vantagem de necessitar de uma quantidade consideravelmente menor de vetores-suporte. A versão com classe de rejeição também apresentou resultados promissores.

Palavras-chave: Classificação. Máquinas de vetores-suporte. Lógica nebulosa.

ABSTRACT

The goal of pattern classification is to assign an element of a data set to one out of many available classes. Due to a precise mathematical foundation and excellent generalization performance, kernel methods based on support vector machines (SVMs) have been successfully applied to pattern classification problems. To build an SVM classifier, we search for the best decision surface that separates the elements of the different classes from each other, keeping the largest margins possible. From this surface, we build a decision function, which is used to classify new elements. Despite being proposed initially to handle linearly separable data sets, to deal with more complex data sets, a regularization parameter and slack variables were introduced into the original SVM formulation. A drawback of this approach is that, in addition to the need for calibrating an additional parameter, there is an increase in the number of support vectors required to build the decision function. The higher the number of support vectors, the higher the computational cost for classifying new incoming patterns. In this thesis, we introduce a novel SVM formulation based on fuzzy logic to handle uncertainties in the data set. Using this approach, drawbacks resulting from the introduction of a regularization parameter are avoided. Additionally, the resulting classifier assigns membership values to the elements of the data set, allowing the introduction of a rejection class into the proposed formulation without further difficulties. The two versions of the proposed model, with and without rejection class, were evaluated on several benchmarking data sets originating from the biomedical research and their performances were compared to those from other SVM formulations. The version without rejection class presented recognition rates comparable to those from soft margin SVM, with the advantage of using considerable less amount of support vectors. The version with rejection class also presented promising results.

Palavras-chave: Pattern classification. Support-vector machines. Fuzzy logic.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

FSVM	*Fuzzy Support Vector-Machine*
HSO	Hiperplano Separador Ótimo
KKT	Karush-Kuhn-Tucker
PPQ	Problema de Programação Quadrática
SVM	Máquina de Vetores-Suporte, tradução livre de *Support-Vector Machine*
SVM-HM	Máquina de Vetores-Suporte com Margem Rígida, tradução livre de *Hard Margin Support-Vector Machine*
SVM-PA	Máquina de Vetores-Suporte com Abordagem Paramétrica, tradução livre de *Support-Vector Machine with Parametric Approach*
SVM-PA-RO	Máquina de Vetores-Suporte com Abordagem Paramétrica e Opção de Rejeição, tradução livre de *Support-Vector Machine with Parametric Approach and Rejection Option*
SVM-SM	Máquina de Vetores-Suporte com Margem Flexível, tradução livre de *Soft Margin Support-Vector Machine*
WCS-FSVM	*FSVM with minimum within-class scatter*

SUMÁRIO

1 INTRODUÇÃO

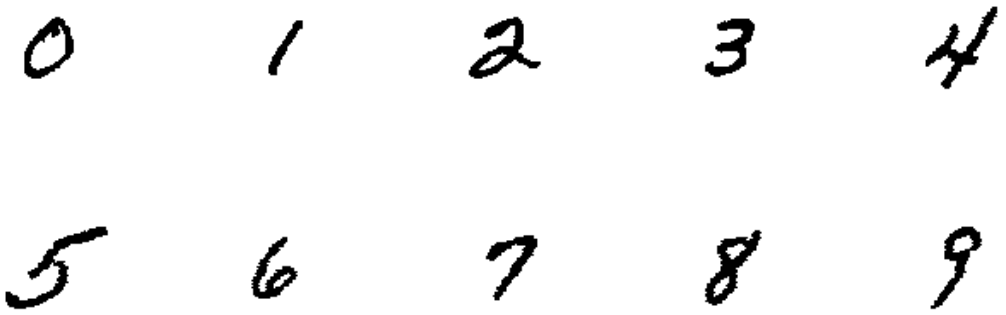
O volume de informação gerado a cada segundo nos dias atuais tornou o trabalho de análise dessa informação um desafio, inviável de ser realizado sem o auxílio de computadores. Ademais, inúmeras técnicas e modelos são desenvolvidos para esse fim, os quais são objetos de estudos na área de pesquisa sobre análise de dados.

Denomina-se de aprendizado de máquina uma área da análise de dados que busca automatizar a construção de modelos analíticos. É um ramo da inteligência artificial baseado na ideia de que os sistemas podem aprender usando dados observados, identificar padrões e tomar decisões com mínima intervenção humana.

Algoritmos de aprendizado de máquina podem ser separados, em um primeiro momento, em duas categorias: algoritmos de aprendizagem supervisionada e algoritmos de aprendizagem não supervisionada. Nos algoritmos supervisionados, o conjunto de dados de treino é composto de vetores de características e seus rótulos (a classe a qual cada vetor pertence). Problemas onde o objetivo é atribuir a cada vetor de entrada uma dentre diversas classes são chamados de problemas de **classificação**. Caso o objetivo seja obter uma ou mais variáveis contínuas, então o problema é chamado de **regressão**. Já nos algoritmos não supervisionados, o conjunto de dados é composto apenas pelos vetores de características, de tal maneira que os elementos não possuem rótulos. Nessa categoria de algoritmos, o objetivo pode ser encontrar grupos de vetores com características similares, o que é conhecido como **agrupamento** (*clustering*); determinar a distribuição dos dados no espaço dos vetores de características, o que é chamado de **estimação de densidade**; ou ainda projetar o conjunto de dados de um espaço de alta dimensão para um espaço com duas ou três dimensões, para que seja possível visualizar os dados, processo que é conhecido como **redução de dimensionalidade** (BISHOP, 2011).

Tome, como exemplo, o problema de classificação/reconhecimento de números manuscritos. Na Figura 1 são apresentados alguns exemplos de números extraídos do conjunto de dados NIST (GROTHER, 1995). Cada número corresponde a uma imagem de 28×28 pixels. O objetivo é construir um classificador capaz de tomar uma imagem como entrada e retornar a classe a qual ela pertence. Neste exemplo, as classes são os números inteiros de 0 a 9. Para isso, utiliza-se um conjunto de dados de treino (fase de aprendizagem), composto por diversas imagens de cada um dos números e pelas classes dessas imagens.

Figura 1 – Exemplos de números manuscritos contidos no conjunto de dados NIST.



Fonte: elaborada pelo autor, a partir de imagens dos dígitos do conjunto de dados NIST.

O resultado da execução de um algoritmo de classificação pode ser representado por uma função $f(x)$, chamada de função de decisão. Essa função é determinada durante a fase de treinamento, ou fase de aprendizagem, com base nos dados de treino. Tal função tem como entrada um vetor de características, x, e como saída a previsão da classe desse vetor, y. Então, essa função é utilizada para classificar novos vetores de características, os quais compõem o conjunto de teste. A habilidade de classificar corretamente esses novos vetores, o que é um dos objetivos principais em problemas de classificação de padrões, é conhecida como **generalização**.

Também existem algoritmos de classificação cuja função de decisão pode, além de atribuir uma classe a um determinado vetor de características, abster-se de tomar uma decisão (CHOW, 1957; CHOW, 1970; PLATT *et al.*, 1999; FUMERA; ROLI, 2002; BARTLETT; WEGKAMP, 2008; GRANDVALET *et al.*, 2009). Essa opção de se abster, conhecida como **opção de rejeição**, é muito importante na prática, principalmente em problemas cujos custos de se cometer erros são elevados (CHOW, 1970; WEBB, 2003).

Problemas de classificação estão cada vez mais presentes no nosso dia a dia como por exemplo, *Smartphones* usam detecção facial ou leitura da impressão digital para desbloquear o aparelho, lojas de vendas online usam informações de compras anteriores para sugerir novos produtos, e-mails são classificados dentre diversas categorias (promoções, compras, viagens, *spam*, etc). Há na literatura diversos algoritmos de classificação, tais como árvores de decisão (QUINLAN, 1986), redes neurais (HAYKIN, 1994), funções discriminantes (MIKA *et al.*, 1999) e Máquina de Vetores-Suporte, tradução livre de *Support-Vector Machine* (SVM) (BOSER *et al.*, 1992; CORTES; VAPNIK, 1995).

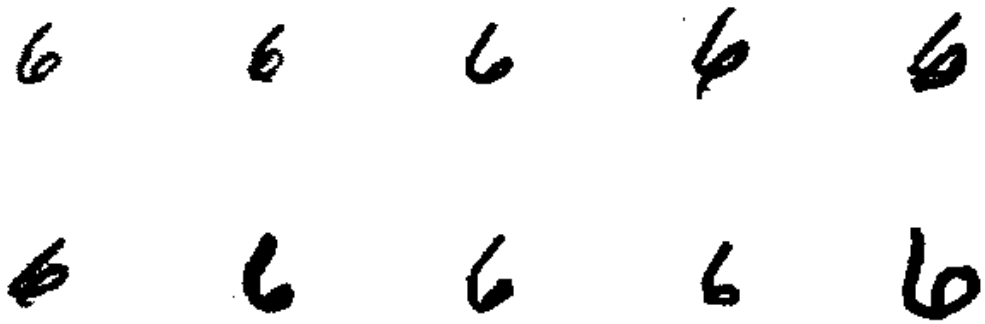
Dentre os algoritmos de classificação, SVMs têm tido grande destaque desde seu

surgimento na década de 90. Esse tipo de classificador vem sendo aplicado com sucesso em diversos problemas, tais como reconhecimento facial (OSUNA *et al.*, 1997; DÉNIZ *et al.*, 2003; HEISELE *et al.*, 2001), classificação de imagens (CHAPELLE *et al.*, 1999; FOODY; MATHUR, 2006), categorização de textos (JOACHIMS, 1998; DUMAIS *et al.*, 1998) e bioinformática (BYVATOV; SCHNEIDER, 2003). Além de classificação, SVMs também podem ser utilizadas para regressão. Para esses casos, o método é denominado *Support-Vector Regression* (BURGES, 1998).

Resumidamente, dado um conjunto de dados binário (com apenas duas classes) uma SVM realiza um mapeamento, de forma implícita, dos dados de entrada para um espaço de dimensão elevada, chamado de espaço de características. Nesse espaço ela então procura pelo hiperplano que maximiza a margem de separação das duas classes, sendo que margem de separação é definida como a soma das distâncias do hiperplano até o elemento mais próximo de cada classe.

Dados oriundos de problemas reais podem possuir incertezas. No problema do reconhecimento de dígitos, essas incertezas podem ser variações na escrita dos números, uma vez que cada pessoa escreve de forma única. Por exemplo, na Figura 2 são apresentadas algumas imagens do número 6 contidas no conjunto de dados NIST. Em problemas de reconhecimento facial (ou de imagens em geral) as fotos podem estar embaçadas, em ângulos diversos, com diferentes ambientes ao fundo, etc. Ademais, dados obtidos através de medições também terão incertezas, uma vez que nenhuma forma de medição está livre de imprecisões e dados incertos.

Figura 2 – Alguns exemplos do número 6, extraídos do conjuntos de dados NIST.



Fonte: elaborada pelo autor, a partir de imagens dos dígitos do conjunto de dados NIST.

Uma forma de lidar com incertezas é através do uso da lógica nebulosa. Nela, a cada

elemento e subconjunto de um conjunto universo é atribuído uma pertinência do elemento ao subconjunto, onde a pertinência é um valor real no intervalo $[0,1]$. A interpretação da pertinência é feita com o uso de variáveis linguísticas. Por exemplo, vamos usar a lógica nebulosa para definir um conjunto nebuloso das pessoas adultas. Pode-se então associar um número real à idade de cada pessoa da população e assim construir o conjunto universo $U = [0,122]$[1]. Para caracterizar o subconjunto nebuloso, A, dos adultos definimos uma função para representar a pertinência de um elemento, x, a esse subconjunto. Essa função pode ser definida como

$$\chi_{\tilde{A}}(x) = \begin{cases} \frac{x}{30}, & \text{se } x \in [0,30) \\ 1, & \text{se } x \in [30,59] \\ \frac{59}{x}, & \text{se } x \in (59,122] \end{cases} .$$

Note que todos os elementos do conjunto universo possuem uma pertinência ao subconjunto nebuloso dos adultos. Para esse exemplo, quanto maior a pertinência de alguma pessoa "mais adulto" ela seria considerada. Uma grande vantagem dessa abordagem é que abre-se a possibilidade de se tomar decisões menos bruscas. Ao invés de falar que uma pessoa de 29 anos não é adulta, como seria na teoria de conjuntos clássica, na teoria de conjuntos nebulosos podemos dizer que essa pessoa é um adulto com pertinência $29/30$ - ou seja, essa pessoa poderia ser considerada um "quase adulto".

As primeiras aplicações da lógica nebulosa foram em problemas de sistemas de controle, na década de 70. Hoje em dia, além de sistemas de controle, ela tem sido aplicada com sucesso em diversas áreas, tais como processamento de imagens, engenharia eletrotécnica e automação industrial (SINGH *et al.*, 2013).

1.1 Objetivos

1.1.1 Objetivo Geral

O trabalho apresentado nesta dissertação compreende o desenvolvimento de dois classificadores, baseados em máquinas de vetores-suporte. O primeiro classificador generaliza a primeira formulação do SVM com o uso da lógica nebulosa para lidar com incertezas em conjuntos de dados. O segundo classificador introduz a estratégia de opção de rejeição ao primeiro classificador proposto.

[1] O intervalo é fechado em 122 devido a pessoa mais velha já documentada da história ter vivido 122 anos (WHITNEY, 1997).

1.1.2 Objetivos Específicos

- Avaliar a acurácia dos classificadores propostos em conjuntos de dados reais;
- Avaliar a quantidade de vetores-suporte encontrada pelos classificadores propostos em conjuntos de dados reais;
- Comparar os classificadores propostos com outros classificadores baseados em máquinas de vetores-suporte, com e sem opção de rejeição.

1.2 Estrutura do texto

O restante deste trabalho está organizado da seguinte forma: no Capítulo 2 é apresentada uma fundamentação matemática dos classificadores SVM, assim como a formulação matemática de métodos que introduzem uma opção de rejeição a esses classificadores. Neste capítulo ainda é feita uma revisão bibliográfica de trabalhos que desenvolveram classificadores SVM nebulosos. No Capítulo 3 é apresentado o desenvolvimento do método proposto neste trabalho, além de ser discutido como uma região de rejeição pode ser obtida a partir da saída desse método. No Capítulo 4 são apresentados resultados da aplicação do classificador proposto e de outros classificadores SVM, com e sem opção de rejeição, em quatro conjuntos de dados oriundos da área médica. Por fim, o Capítulo 5 traz as considerações finais sobre o método proposto neste trabalho.

2 REVISÃO BIBLIOGRÁFICA

Este capitulo apresenta uma revisão bibliográfica sobre classificadores baseados em SVM, assim como formulações desses classificadores.

Conceitos básicos e a primeira formulação de um classificador SVM, proposta para conjuntos linearmente separáveis, são apresentados na Seção 2.1. Uma generalização desse método, a qual permite a construção de classificadores SVM para qualquer conjunto de dados, é apresentada na Seção 2.2. A Seção 2.3 trata sobre funções *kernel*, as quais podem ser utilizadas para a construção de superfícies de decisão não-lineares. Na Seção 2.4 introduz-se o conceito de opção de rejeição e são apresentados classificadores SVM com essa característica. Por fim, na Seção 2.5 é feita uma discussão sobre métodos que utilizam a lógica nebulosa em classificadores SVM, chamados de *Fuzzy Support Vector-Machine* (FSVM).

2.1 Máquina de vetores-suporte com margem rígida

Em 1964, Chervonenkis e Vapnik criaram o primeiro método para a construção de um Hiperplano Separador Ótimo (HSO) (VAPNIK, 2006). Para entender o que é um HSO, primeiro são necessárias algumas definições.

Definição 2.1 *O **hiperplano** é um conceito matemático que generaliza a noção de reta ou plano para várias dimensões. No espaço $\mathbb{R}^n$, um hiperplano é definido pelo conjunto de pontos que satisfaz a equação*

$$w^\intercal x + b = 0,$$

sendo que $w \in \mathbb{R}^n$ é o vetor normal ao hiperplano, $b \in \mathbb{R}$ e $x \in \mathbb{R}^n$.

Definição 2.2 *Seja $T = \{(x_i, y_i) \mid i = 1, \ldots, \ell\}$ um conjunto de dados tal que $x_i \in \mathbb{R}^p$ é o i-ésimo vetor de atributos, $y_i \in \{-1, 1\}$ representa a classe de x_i e p a é quantidade de atributos de x_i. Esse conjunto é dito ser **linearmente separável** se existem $w \in \mathbb{R}^p$ e $b \in \mathbb{R}$ tais que*

$$w^\intercal x_i + b \geq 1, \quad \text{para } y_i = 1 \tag{2.1}$$

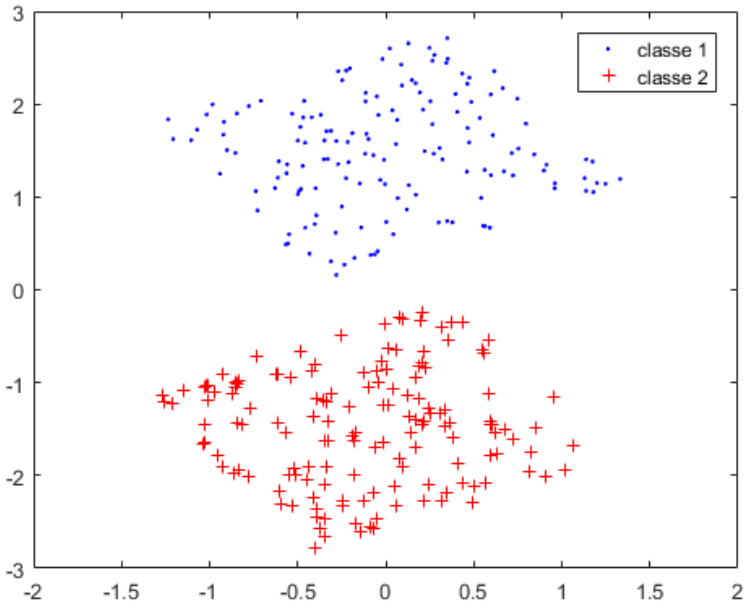
e

$$w^\intercal x_i + b \leq -1, \quad \text{para } y_i = -1 \tag{2.2}$$

são satisfeitas para $i = 1, \ldots, \ell$. O conjunto das Inequações (2.1) e (2.2) garante que os elementos das classes positiva podem ser separados perfeitamente dos elementos da classe negativa.

Um exemplo de um conjunto de dados linearmente separável pode ser observado na Figura 3.

Figura 3 – Exemplo de um conjunto de dados linearmente separável.



Fonte: elaborada pelo autor.

Definição 2.3 *Considere o conjunto de dados T e um hiperplano H que satisfaz as Inequações* (2.1) *e* (2.2)*. A* ***margem de separação*** *é definida como a soma das distâncias de H até os elementos mais próximos de cada classe. Sejam d_+ a menor distância entre H e um elemento da classe positiva e d_- a menor distância entre H e um elemento da classe negativa. A margem ρ é então*

$$\rho(w,\, b) = d_+ + d_-$$

$$\rho(w,\, b) = \min_{\{x:\, y=1\}} \frac{|w^\intercal x + b|}{\|w\|} + \min_{\{x:\, y=-1\}} \frac{|w^\intercal x + b|}{\|w\|},$$

sendo que $|\cdot|$ representa o valor absoluto de um número real e $\|\cdot\|$ representa a norma euclidiana de um vetor. Como é possível garantir, através de um escalonamento de w e b, a existência de pontos, em ambas as classes, que satisfazem a igualdade nas Inequações (2.1) *e* (2.2)*, tem-se que*

$$\min_{\{x:\, y=1\}} \frac{|w^\intercal x + b|}{\|w\|} = \frac{1}{\|w\|}$$

e

$$\min_{\{x:\ y=-1\}} \frac{|w^\intercal x + b|}{\|w\|} = \frac{1}{\|w\|}.$$

Portanto,

$$\rho(w,\ b) = \frac{2}{\|w\|} = \frac{2}{w^\intercal w}. \tag{2.3}$$

Definidos os conceitos de hiperplano, de conjunto linearmente separável e de margem de separação, pode-se agora definir o que Chervonenkis e Vapnik chamaram de HSO.

Definição 2.4 *Dado um conjunto de dados binário e linearmente separável, o **HSO** é o hiperplano que separa os dados com a maior margem possível.*

Portanto, para encontrar o HSO deve-se encontrar o hiperplano que maximiza a margem de separação. Da Equação (2.3), percebe-se que maximizar a margem é equivalente a minimizar a norma de w. Ademais, notando que as Inequações (2.1) e (2.2) podem ser unificadas na seguinte inequação

$$y_i(w^\intercal x_i + b) \geq 1, \quad i = 1, \ldots, \ell. \tag{2.4}$$

Assim, o HSO pode ser encontrado através da solução do seguinte Problema de Programação Quadrática (PPQ):

$$\begin{aligned} &\min_{(w,b)} \quad \frac{1}{2} w^\intercal w \\ &\text{s.a.} \quad y_i(w^\intercal x_i + b) \geq 1, \qquad i = 1, 2, \ldots, \ell. \end{aligned} \tag{2.5}$$

O Problema (2.5) pode ser resolvido pelo método dos multiplicadores de Lagrange. Para isso, constrói-se a função Lagrangeana descrita a seguir

$$L(w, b, \alpha) = \frac{1}{2} w^\intercal w - \sum_{i=1}^{\ell} \alpha_i \left[y_i(w^\intercal x_i + b) - 1 \right], \tag{2.6}$$

tal que $\alpha^\intercal = [\alpha_1, \ldots, \alpha_\ell]$ é o vetor dos multiplicadores de Lagrange não-negativos, sendo α_i correspondente à i-ésima inequação de (2.5). Vale ressaltar que α^0 está associado ao ponto viável $(w_0,\ b_0)$ ótimo do Problema primal (2.5).

Note que o Problema (2.5) é um problema de programação quadrática sendo a Hessiana uma matriz semi-definida positiva, de forma que o mesmo ponto $(w_0,\ b_0,\ \alpha^0)$ otimiza os problemas primal e dual. A solução do Problema (2.5) é determinada pelo ponto que satisfaz

o conjunto de restrições, minimiza a função L em relação às variáveis do problema primal e maximiza L em relação aos multiplicadores de Lagrange.

Para encontrar o ponto que minimiza L em relação às variáveis w e b, calcula-se a derivada de L em relação a essas variáveis e iguala-se essas derivadas a 0:

$$\left.\frac{\partial L(w,b,\alpha)}{\partial w}\right|_{w=w_0} = w_0 - \sum_{i=1}^{\ell} \alpha_i y_i x_i = 0, \tag{2.7}$$

$$\left.\frac{\partial L(w,b,\alpha)}{\partial b}\right|_{b=b_0} = \sum_{i=1}^{\ell} y_i \alpha_i = 0. \tag{2.8}$$

Das Equações (2.7) e (2.8), encontra-se que:

$$w_0 = \sum_{i=1}^{\ell} \alpha_i y_i x_i \tag{2.9}$$

e

$$\sum_{i=1}^{\ell} y_i \alpha_i = 0. \tag{2.10}$$

Essas equações podem ser usadas para simplificar a função Lagrangeana apresentada na Equação (2.6). Note que

$$L(w,b,\alpha) = \frac{1}{2} w_0^{\intercal} w_0 - \sum_{i=1}^{\ell} \alpha_i \left[y_i (w_0^{\intercal} x_i + b_0) - 1 \right]. \tag{2.11}$$

$$L(w,b,\alpha) = \frac{1}{2} w_0^{\intercal} w_0 - w_0^{\intercal} \left(\sum_{i=1}^{\ell} \alpha_i y_i x_i \right) - b_0 \sum_{i=1}^{\ell} \alpha_i y_i + \sum_{i=1}^{\ell} \alpha_i. \tag{2.12}$$

Da Equação (2.10), percebe-se que o terceiro termo da equação acima é zero. Além disso, de acordo com a Equação (2.9), o segundo termo pode ser reescrito como $w_0^{\intercal} w_0$. Logo,

$$L(w,b,\alpha) = \frac{1}{2} w_0^{\intercal} w_0 - w_0^{\intercal} w_0 + \sum_{i=1}^{\ell} \alpha_i \tag{2.13}$$

$$L(w,b,\alpha) = -\frac{1}{2} w_0^{\intercal} w_0 + \sum_{i=1}^{\ell} \alpha_i \tag{2.14}$$

$$L(w,b,\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} \alpha_i \alpha_j y_i y_j x_i^{\intercal} x_j, \tag{2.15}$$

obtém-se uma função que depende somente dos multiplicadores de Lagrange α e dos produtos internos entre os vetores de entrada. Assim, a formulação dual do Problema (2.5) pode ser

modelada da seguinte forma:

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} \alpha_i \alpha_j y_i y_j x_i^{\intercal} x_j \\ \text{s.a.} \quad & \sum_{i=1}^{\ell} y_i \alpha_i = 0 \\ & \alpha_i \geq 0, \qquad i = 1, 2, \ldots, \ell. \end{aligned} \tag{2.16}$$

A solução que maximiza o Problema (2.16) será o vetor de multiplicadores de Lagrange associado ao ponto que minimiza o Problema (2.5).

As condições de Karush-Kuhn-Tucker (KKT) (KARUSH, 1939; KUHN; TUCKER, 1951) desempenham um papel fundamental em problemas de otimização. De acordo com tais condições, no ponto ótimo (w_0, b_0, α^0), cada componente do multiplicador de Lagrange, α_i^0, e sua restrição correspondente estão conectados pela equação

$$\alpha_i^0 \left[y_i (w_0^{\intercal} x_i + b_0) - 1 \right] = 0, \qquad i = 1, 2, \ldots, \ell, \tag{2.17}$$

a qual é conhecida como Condição de Complementaridade. Note que, na Equação (2.17), α_i^0 pode ser diferente de zero somente quando

$$y_i (w_0^{\intercal} x_i + b_0) = 1. \tag{2.18}$$

Os vetores x_i que satisfazem a Equação (2.18) são chamados de vetores-suporte.

Além disso, de acordo com as Equações (2.9) e (2.17), o vetor solução w_0 pode ser escrito como uma combinação linear dos vetores-suporte, *i.e.*,

$$w_0 = \sum_{i \in SV} \alpha_i^0 y_i x_i. \tag{2.19}$$

sendo que SV representa o conjunto com os índices dos vetores-suporte. Isso quer dizer que para a classificação de novos vetores de atributos necessita-se apenas dos vetores-suporte. Ou seja, uma vez que esses vetores tenham sido encontrados, todos os demais elementos do conjunto de dados de treinamento podem ser descartados. Essa classificação é feita de acordo com a seguinte função de decisão:

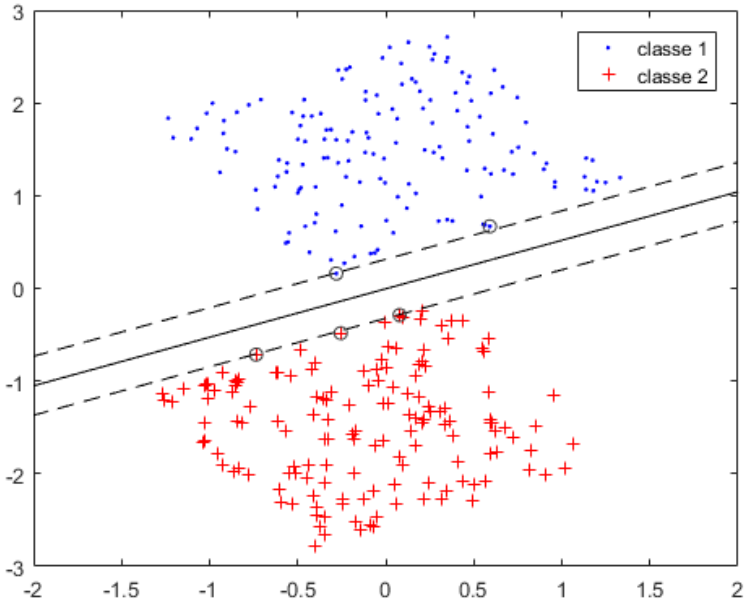
$$f(x) = sign\left(\sum_{i \in SV} \alpha_i^0 y_i x_i^{\intercal} x + b_0 \right), \tag{2.20}$$

sendo que

$$sign(x) = \begin{cases} -1, & \text{se } x < 0 \\ 1, & \text{se } x \geq 0 \end{cases}.$$

O método descrito acima é conhecido como Máquina de Vetores-Suporte com Margem Rígida, tradução livre de *Hard Margin Support-Vector Machine* (SVM-HM). Na Figura 4 pode-se visualizar o hiperplano separador ótimo encontrado pelo SVM-HM, para o conjunto de dados exposto na Figura 3.

Figura 4 – Hiperplano separador ótimo (linha contínua) encontrado pelo SVM-HM para o conjunto de dados exposto na Figura 3. Os vetores-suporte estão destacados com círculos cinzas. As linhas tracejadas representam as margens do HSO.



Fonte: elaborada pelo autor.

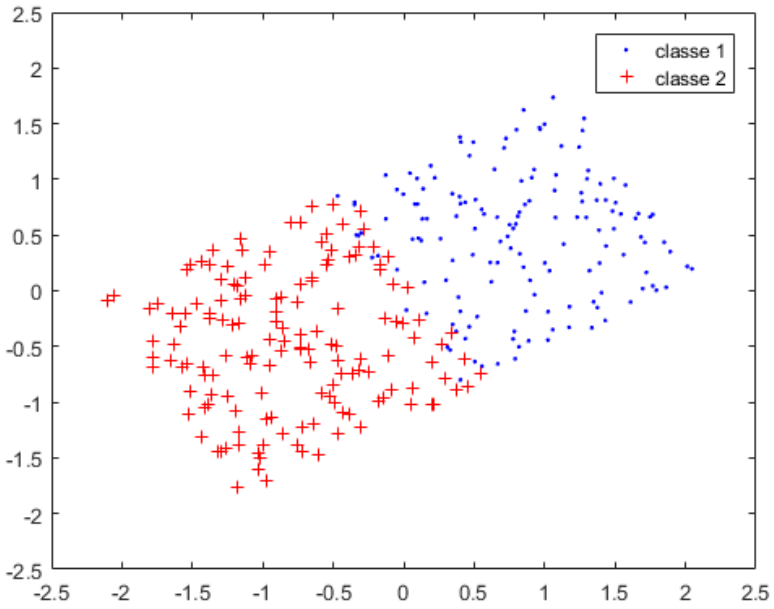
2.2 Máquina de vetores-suporte com margem flexível

Um problema da abordagem descrita na Seção 2.1 é que conjuntos de dados, principalmente os oriundos de problemas reais, geralmente não são linearmente separáveis. É comum existir padrões mal classificados ou sobreposição de elementos de classes diferentes, conforme ilustrado na Figura 5.

Para contornar a não separabilidade dos dados, Cortes e Vapnik (1995) sugeriram a adição de variáveis de folga ao Problema (2.5), levando à seguinte formulação:

$$\begin{aligned} \min_{(w,b,\xi)} \quad & \frac{1}{2} w^\intercal w + CF\left(\sum_{i=1}^{\ell} \xi_i\right) \\ \text{s.a.} \quad & y_i(w^\intercal x_i + b) \geq 1 - \xi_i, \quad i = 1,2,\ldots,\ell, \\ & \xi_i \geq 0, \qquad\qquad\quad i = 1,2,\ldots,\ell, \end{aligned} \tag{2.21}$$

Figura 5 – Exemplo de um conjunto de dados com sobreposição dos elementos de classes diferentes.



Fonte: elaborada pelo autor.

sendo que C é uma constante positiva, chamada constante de regularização, $F(u)$ é uma função monotônica convexa e ξ_i, $i = 1, \ldots, \ell$, são as variáveis de folga. Percebe-se que são justamente essas variáveis de folga que permitem que nem todos os padrões de treino estejam do lado certo do hiperplano separador ótimo. Os padrões para os quais $\xi_i = 0$ estão do lado correto do hiperplano e estão sobre a margem de separação, ou além dela. Quando $0 < \xi_i < 1$ os padrões correspondentes ainda estão do lado correto do hiperplano, mas dentro da margem. Padrões para os quais $\xi_i > 1$ estão do lado oposto do hiperplano, em relação aos elementos de sua classe.

Resolver o Problema (2.21) descreve (para C suficientemente grande) o problema de construir o hiperplano que minimiza a soma dos desvios dos erros e maximiza a margem de separação para os vetores corretamente classificados (CORTES; VAPNIK, 1995). Se o conjunto de dados de treino for linearmente separável, o hiperplano construído coincide com o hiperplano encontrado pelo SVM-HM. Esse método é conhecido como Máquina de Vetores-Suporte com Margem Flexível, tradução livre de *Soft Margin Support-Vector Machine* (SVM-SM).

Similarmente ao que foi feito para o SVM-HM, a solução do Problema (2.21) pode ser encontrada pelo método dos multiplicadores de Lagrange. Aqui considera-se apenas o caso $F(u) = u$. A solução do caso geral, quando $F(u)$ é qualquer função monotônica convexa, pode ser encontrada em Cortes e Vapnik (1995).

Primeiramente, constrói-se a função Lagrangeana

$$L(w,b,\xi,\alpha,\beta) = \frac{1}{2}w^{\intercal}w + C\sum_{i=1}^{\ell}\xi_i - \sum_{i=1}^{\ell}\alpha_i[y_i(w^{\intercal}x_i+b)-1+\xi_i] - \sum_{i=1}^{\ell}\beta_i\xi_i, \tag{2.22}$$

sendo que $\alpha^{\intercal} = [\alpha_1,\ldots,\alpha_\ell]$ e $\beta^{\intercal} = [\beta_1,\ldots,\beta_\ell]$ são vetores de multiplicadores de Lagrange não negativos e $\xi^{\intercal} = [\xi_1,\ldots,\xi_\ell]$ é o vetor das variáveis de folga não negativas.

Em seguida, deve-se minimizar a função Lagrangeana em relação às variáveis da formulação primal (w, b e ξ), e maximizá-la em relação às variáveis da formulação dual (α e β). Assim, tem-se

$$\left.\frac{\partial L}{\partial w}\right|_{w=w_0} = w_0 - \sum_{i=1}^{\ell}\alpha_i y_i x_i = 0, \tag{2.23}$$

$$\left.\frac{\partial L}{\partial b}\right|_{b=b_0} = -\sum_{i=1}^{\ell}\alpha_i y_i = 0, \tag{2.24}$$

$$\left.\frac{\partial L}{\partial \xi_i}\right|_{\xi_i=\xi_i^0} = C - \alpha_i - \beta_i = 0. \tag{2.25}$$

Das Equações (2.23)$-$(2.25), encontram-se as relações:

$$w_0 = \sum_{i=1}^{\ell}\alpha_i y_i x_i, \tag{2.26}$$

$$\sum_{i=1}^{\ell}\alpha_i y_i = 0, \tag{2.27}$$

$$C = \alpha_i + \beta_i, \quad i = 1,\ldots,\ell. \tag{2.28}$$

As Equações (2.26)-(2.28) podem ser utilizadas para simplificar a função Lagrangeana:

$$L(\alpha,\beta) = \frac{1}{2}w_0^{\intercal}w_0 + \sum_{i=1}^{\ell}(\alpha_i+\beta_i)\xi_i^0 - \sum_{i=1}^{\ell}\alpha_i y_i w_0^{\intercal}x_i + b_0\sum_{i=1}^{\ell}\alpha_i y_i + \sum_{i=1}^{\ell}\alpha_i - \sum_{i=1}^{\ell}\alpha_i\xi_i^0 - \sum_{i=1}^{\ell}\beta_i\xi_i^0. \tag{2.29}$$

O segundo termo da equação acima se anula com o sexto e sétimo termos. Além disso, de acordo com a Equação (2.27), o quarto termo é zero. Assim, obtém-se:

$$L(\alpha) = \frac{1}{2}w_0^{\intercal}w_0 - \sum_{i=1}^{\ell}\alpha_i y_i w_0^{\intercal}x_i + \sum_{i=1}^{\ell}\alpha_i. \tag{2.30}$$

Por fim, utilizando a Equação (2.26), chega-se à equação

$$L(\alpha) = -\frac{1}{2}w_0^{\intercal}w_0 + \sum_{i=1}^{\ell}\alpha_i, \tag{2.31}$$

a qual pode ser reescrita como

$$L(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} \alpha_i \alpha_j y_i y_j x_i^{\intercal} x_j. \tag{2.32}$$

Com isso, resta maximizar $L(\alpha)$ em relação às variáveis da formulação dual. Portanto, a solução que se procura é encontrada resolvendo o seguinte PPQ:

$$\begin{aligned} \max_{\alpha} \quad & L(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} \alpha_i \alpha_j y_i y_j x_i^{\intercal} x_j \\ \text{s.a.} \quad & \sum_{i=1}^{\ell} \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C, \qquad\qquad i = 1, 2, \ldots, \ell. \end{aligned} \tag{2.33}$$

O fato de que $\alpha_i \leq C$, $i = 1, 2, \ldots, \ell$, decorre diretamente da Equação (2.28). Como o menor valor que β_i pode assumir é 0, tem-se que o maior valor que α_i pode assumir é C.

As condições de complementaridade de Karush-Kuhn-Tucker desse problema são

$$\alpha_i^0 [y_i (w_0^{\intercal} x_i + b_0) - 1 + \xi_i^0] = 0 \tag{2.34}$$

e

$$(C - \alpha_i^0) \xi_i^0 = 0. \tag{2.35}$$

Note que no SVM-SM existirão dois tipos de vetores-suporte. Para $0 \leq \alpha_i^0 < C$, o vetor suporte x_i correspondente satisfaz as equações $y_i(w^{\intercal} x_i + b) = 1$ e $\xi_i^0 = 0$. Para $\alpha_i^0 = C$, ξ_i não necessariamente é zero e x_i não necessariamente satisfaz a equação $y_i(w^{\intercal} x_i + b) = 1$.

Assim como no SVM-HM, w_0 pode ser escrito como uma combinação linear dos vetores-suporte

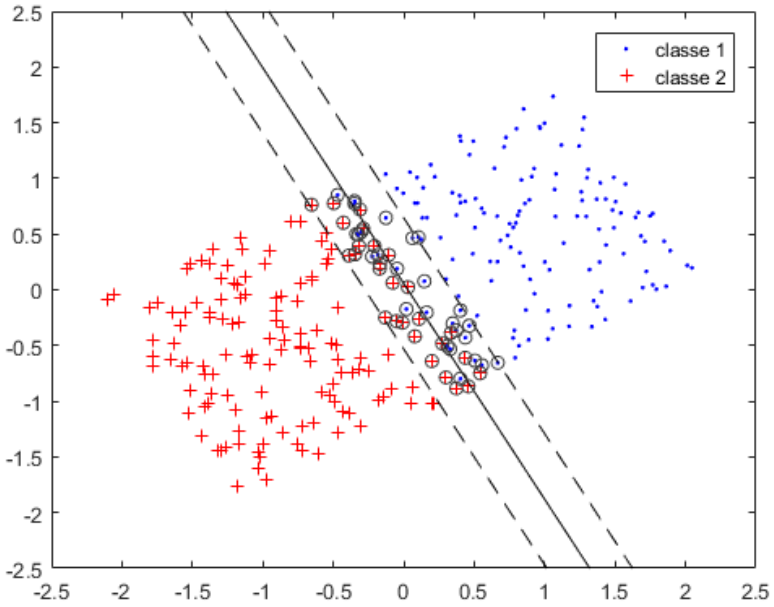
$$w_0 = \sum_{i \in SV} \alpha_i^0 y_i x_i, \tag{2.36}$$

sendo que SV representa o conjunto com os índices dos vetores-suporte. A função de decisão para o SVM-SM é idêntica à do SVM-HM, ou seja

$$f(x) = sign \left(\sum_{i \in SV} \alpha_i^0 y_i x_i^{\intercal} x + b_0 \right). \tag{2.37}$$

Na Figura 6 pode-se observar o hiperplano separador ótimo, encontrado pelo SVM-SM, para o conjunto de dados apresentado na Figura 5.

Figura 6 – Hiperplano separador ótimo (linha sólida) encontrado pelo SVM-SM para o conjunto de dados apresentado na Figura 5. Os vetores-suporte estão destacados com círculos cinzas. As linhas tracejadas representam as margens do HSO.



Fonte: elaborada pelo autor.

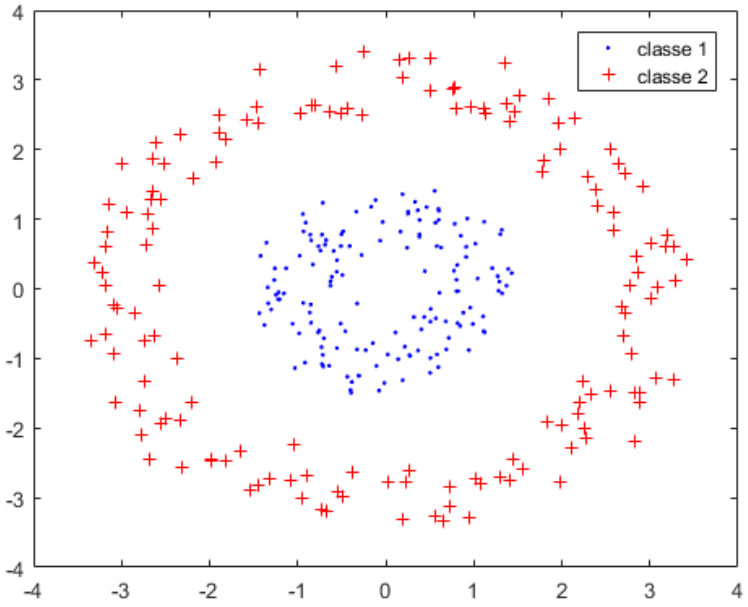
2.3 Máquinas de vetores-suporte não-lineares

Apesar do SVM-SM poder ser utilizado em qualquer conjunto de dados, há casos em que o hiperplano separador ótimo construído por esse método não consegue separar os dados de forma satisfatória (e.g. Figura 7).

Para separar adequadamente as classes do conjunto de dados exposto na Figura 7, faz-se necessário uma superfície de decisão não-linear. A construção desse tipo de superfície de decisão no espaço de entrada dos dados pode ser uma tarefa árdua. Contudo, ao invés de se construir uma superfície de decisão não-linear no espaço de entrada, o conjunto de dados pode ser mapeado para um espaço com dimensão maior que o espaço de entrada dos dados, chamado de espaço de características, onde espera-se que um hiperplano possa separar satisfatoriamente as classes do conjunto de dados. Por exemplo, na Figura 8 é apresentado um mapeamento para o $\mathbb{R}^3$ do conjunto de dados apresentado na Figura 7. Note que o conjunto de dados mapeado é linearmente separável em $\mathbb{R}^3$.

A princípio, um hiperplano separador ótimo pode ser construído pelo SVM-HM ou

Figura 7 – Exemplo de um conjunto de dados onde um hiperplano não consegue separar satisfatoriamente os elementos de classes opostas.



Fonte: elaborada pelo autor.

SVM-SM para o conjunto de dados transformados (e.g., Figura 9)

$$\phi(x_i) = (\phi_1(x_i), \phi_2(x_i), \ldots, \phi_M(x_i)), \quad i = 1, \ldots, \ell. \tag{2.38}$$

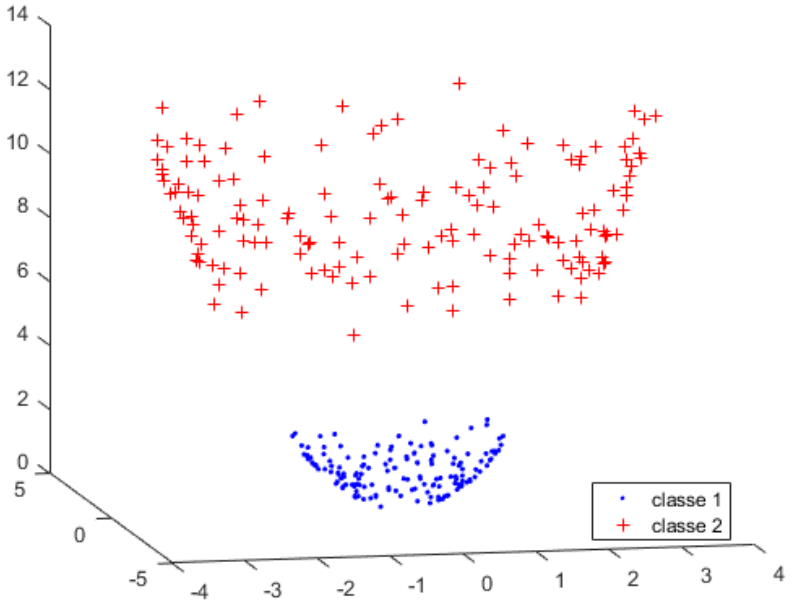
Nesse caso, a classificação de um novo elemento, x, seria feita pela seguinte função de decisão:

$$f(x) = sign(w_0^{\intercal}\phi(x) + b_0). \tag{2.39}$$

Porém, além de se ter que conhecer explicitamente o mapeamento ϕ, essa abordagem pode ser inviável computacionalmente. Em uma transformação polinomial de grau d, por exemplo, a quantidade de atributos no espaço de características é $\binom{p+d-1}{d}$, sendo p a quantidade de atributos no espaço de entrada dos dados. Ou seja, em um problema de reconhecimento de caracteres, cujos valores usuais para d e p são 7 e $28 \times 28 = 784$, respectivamente, o número de atributos em cada vetor transformado seria de aproximadamente $3,7 \times 10^{16}$ (SMOLA; SCHÖLKOPF, 2004). Para armazenar cada um dos vetores transformados seriam necessários aproximadamente $1,48 \times 10^{17}$ bytes $= 131,45$ petabytes (PB), o que inviabiliza o armazenamento do conjunto de dados na memória do computador.

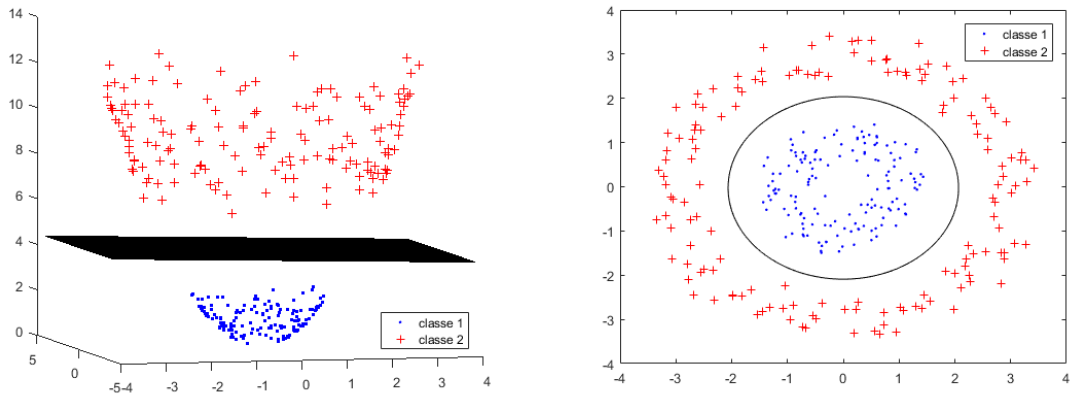
Em 1992, Boser, Guyon e Vapnik encontraram uma maneira eficaz de construir, para um conjunto linearmente separável, o hiperplano separador ótimo em um espaço de Hilbert, sem

Figura 8 – Mapeamento para o $\mathbb{R}^3$ do conjunto de dados apresentado na Figura 7. O mapeamento utilizado foi $(x_1, x_2) \Longrightarrow (x_1, x_2, x_1^2 + x_2^2)$.



Fonte: elaborada pelo autor.

Figura 9 – (a) hiperplano separador ótimo encontrado para o conjunto de dados mapeado para o $\mathbb{R}^3$. (b) projeção desse hiperplano para o espaço original dos dados.



(a) (b)

a necessidade de mapear explicitamente o conjunto de dados para esse espaço (BOSER *et al.*, 1992). Essa ideia foi generalizada em Cortes e Vapnik (1995) para conjuntos não linearmente separáveis.

Em ambos os casos, isso pôde ser feito considerando formas gerais do produto

interno em espaços de Hilbert

$$K(u,v) \equiv \langle \phi(u), \phi(v) \rangle, \tag{2.40}$$

sendo que $K(\cdot,\cdot)$ é chamada de função *kernel*.

Qualquer função *kernel* $K(u,v)$, contínua, simétrica e que satisfaça o Teorema de Mercer (MERCER; FORSYTH, 1909) define um produto interno no espaço de características. O Teorema de Mercer exige que

$$\int \int K(u,v)g(u)g(v)\ dudv > 0 \tag{2.41}$$

seja satisfeito para toda função g tal que

$$\int g^2(u)\ du < \infty. \tag{2.42}$$

A Equação (2.40) é conhecida como **truque de *kernel***, pois permite trabalhar implicitamente no espaço de características, sem a necessidade de realizar mapeamentos tais como $\phi(u)$ e $\phi(v)$. Para isso, é necessário apenas conhecer uma função que descreve o produto interno $\langle \phi(u), \phi(v) \rangle$ no espaço em que se deseja trabalhar. Alguns exemplos de funções *kernel* são apresentados na Tabela 1.

Tabela 1 – Lista de alguns *kernels* usuais em SVM, sendo d, σ, α e c constantes que devem ser definidas pelo usuário.

Kernel	**Descrição**
Linear	$K(x_1,x_2) = x_1^\intercal x_2$
Polinomial	$K(x_1,x_2) = (x_1^\intercal x_2 + 1)^d$
Gaussiano (RBF)	$K(x_1,x_2) = \exp\left(-\frac{\|x_1-x_2\|^2}{\sigma^2}\right)$
Sigmoidal	$K(x_1,x_2) = \tanh(\alpha x_1^\intercal x_2 + c)$
Log	$K(x_1,x_2) = -\log(\|x-y\|^d + 1)$
Cauchy	$K(x_1,x_2) = \frac{1}{1+\frac{\|x_1+x_2\|^2}{\sigma^2}}$

Com o uso de *kernels*, pode-se generalizar as Equações (2.15) e (2.32) referentes, respectivamente, aos classificadores SVM-HM e SVM-SM. Para isso, convém substituir cada

uma das funções objetivo desses problemas por

$$\sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} \alpha_i \alpha_j y_i y_j K(x_i, x_j), \tag{2.43}$$

tal que o produto interno $x_i^\intercal x_j$ foi substítuido pela função *kernel* $K(x_i x_j)$. Uma substituição similar também deve ser feita nas funções de decisão desses classificadores, de forma que as funções apresentadas nas Equações (2.20) e (2.37) devem ser substituídas pela equação

$$f(x) = sign\left(\sum_{i \in SV} \alpha_i^0 y_i K(x_i, x) + b_0\right), \tag{2.44}$$

sendo que SV representa o conjunto com os índices dos vetores-suporte.

2.4 Máquinas de vetores-suporte com opção de rejeição

Em problemas de classificação binária, o objetivo é aprender, a partir de um conjunto de dados de treino, uma regra de classificação que associa um padrão observado a uma das duas possíveis classes. Contudo, em muitas aplicações, como, por exemplo, em decisões médicas, o custo de se cometer um erro pode ser elevado. Nestes casos, quando a confiança na classificação de um padrão não é satisfatória, pode ser preferível não tomar uma decisão do que arriscar uma classificação errada. A opção de não tomar uma decisão é comumente chamada de *opção de rejeição*.

Cuidados devem ser tomados em classificadores com opção de rejeição, uma vez que quando a opção de rejeição é exercida, alguns padrões que seriam classificados corretamente podem ser convertidos em rejeições. De acordo com Chow (1970), um classificador com opção de rejeição é ótimo se, para uma determinada taxa de erro, a taxa de rejeição é minimizada. A taxa de erro é definida como a quantidade de erros de classificação dividida pela quantidade de elementos do conjunto de testes. Similarmente, a taxa de rejeição é definida como a quantidade de elementos rejeitados dividida pela quantidade de elementos do conjunto de testes.

Na classificação binária, existem dois tipos de erros: (1) falso-positivo, quando um padrão da classe negativa é classificado como positivo e (2) falso-negativo, quando um padrão da classe positiva é classificado como negativo.

Sejam $c_- > 0$ o custo de um erro falso-positivo, $c_+ > 0$ o custo de um erro falso-negativo, $r_- > 0$ o custo de rejeitar um elemento da classe negativa e $r_+ > 0$ o custo de rejeitar um elemento da classe positiva. Além disso, seja 0 a classe dos elementos rejeitados. Na

classificação com opção de rejeição, o risco, de acordo com (GRANDVALET *et al.*, 2009), é dado por:

$$R(d) = c_+ E_{XY}[Y = 1,\ d(X) = -1] + c_- E_{XY}[Y = -1,\ d(X) = 1] + \\ + r_+ E_{XY}[Y = 1,\ d(X) = 0] + r_- E_{XY}[Y = -1,\ d(X) = 0] \quad (2.45)$$

sendo que X e Y representam as variáveis aleatórias referentes aos padrões de treino e suas classes, respectivamente, e $d(X) \in \{-1,\ 0,\ 1\}$ representa a função de decisão. Esta função além de poder atribuir o elemento X às classes 1 ($d(X) = 1$) ou -1 ($d(X) = -1$), pode optar por rejeitá-lo ($d(X) = 0$).

Denotando-se por d_1, d_{-1} e d_0 as decisões de atribuir um elemento, respectivamente, às classes 1, -1 e 0, o risco condicional dessas decisões é dado por:

$$R(d_1 \mid x) = c_- P(Y = -1 \mid x),$$
$$R(d_{-1} \mid x) = c_+ P(Y = 1 \mid x),$$
$$R(d_0 \mid x) = r_+ P(Y = 1 \mid x) + r_- P(Y = -1 \mid x).$$

De acordo com a Teoria da Decisão Bayesiana (BERGER, 2013), a função de decisão ótima, $d^\star$, é aquela que minimiza o risco $R(d)$. Ou seja, um elemento, x, será atribuído à classe 1 se

$$R(d_1 \mid x) < R(d_{-1} \mid x) \quad \text{e} \quad R(d_1 \mid x) < R(d_0 \mid x). \quad (2.46)$$

Similarmente, x será atribuído à classe -1 se

$$R(d_{-1} \mid x) < R(d_1 \mid x) \quad \text{e} \quad R(d_{-1} \mid x) < R(d_0 \mid x). \quad (2.47)$$

O elemento será rejeitado se não satisfizer as Inequações (2.46) e (2.47). Trabalhando essas inequações, encontra-se que a regra de decisão ótima, também conhecida como Regra de Chow (CHOW, 1970), pode ser definida como:

$$d^\star(x) = \begin{cases} +1, & \text{se } P(Y = 1 \mid x) > p_+ \\ -1, & \text{se } P(Y = 1 \mid x) < p_- \\ 0, & \text{caso contrário} \end{cases}, \quad (2.48)$$

sendo

$$p_+ = \frac{c_- - r_-}{c_- - r_- + r_+} \quad \text{e} \quad p_- = \frac{r_-}{c_+ - r_+ + r_-}.$$

É razoável assumir que $p_+ > 0,5$, uma vez que não faz sentido atribuir um elemento à classe 1 se a probabilidade desse elemento pertencer a essa classe for menor do que $0,5$. Similarmente, assume-se que $p_- < 0,5$. Disso, encontram-se as seguintes relações entre os custos de erros e de rejeição:

$$r_- + r_+ < c_- \quad \text{e} \quad r_- + r_+ < c_+. \tag{2.49}$$

Diversos métodos foram propostos para introduzir a opção de rejeição em classificadores SVM (PLATT *et al.*, 1999; MUKHERJEE *et al.*, 1999; KWOK, 1999; FUMERA; ROLI, 2002; BARTLETT; WEGKAMP, 2008; GRANDVALET *et al.*, 2009). A seguir discute-se a formulação de alguns desses métodos.

No Capítulo 2, viu-se que a saída de um classificador SVM é uma função de decisão do tipo

$$f(x) = w^\intercal x + b \quad \text{ou} \quad f(x) = \sum_{i \in SV} \alpha_i y_i K(x_i, x) \tag{2.50}$$

tal que $sign(f(x))$ pode ser utilizado para classificar os padrões de teste. Como $f(x)$ é proporcional à distância do vetor x ao hiperplano separador ótimo, esse valor pode ser interpretado como uma medida de confiança para a classificação. Elementos mais distantes do hiperplano são classificados com uma maior confiança do que aqueles mais próximos ao hiperplano. Neste contexto, uma forma simples de implementar uma regra de rejeição é introduzindo um limiar, h, de forma que um elemento, x, é rejeitado se $|f(x)| \leq h$ (MUKHERJEE *et al.*, 1999). Assim, a classificação dos padrões de teste é feita de acordo com a seguinte regra

$$d^\star(x) = \begin{cases} 0, & \text{se } |f(x)| \leq h \\ sign(f(x)), & \text{caso contrário} \end{cases}. \tag{2.51}$$

Em Platt *et al.* (1999) foi proposto um modelo paramétrico para estimar a probabilidade a posteriori de um padrão pertencer à classe positiva, $P(Y = 1 \mid x)$. Isso é feito aproximando a probabilidade a posteriori por uma função sigmoidal:

$$P(Y = 1 \mid x) \approx P_{A,B}(f(x)) \equiv \frac{1}{1 + \exp(Af(x) + B)}. \tag{2.52}$$

Os parâmetros A e B são encontrados minimizando a função log-verossimilhança negativa dos dados de treino:

$$\min_p \; -\sum_i t_i \; \log(p_i) + (1 - t_i) \; \log(1 - p_i), \tag{2.53}$$

sendo

$$t_i = \frac{y_i + 1}{2} \quad \text{e} \quad p_i = \frac{1}{1 + \exp(Af(x_i) + B)}, \tag{2.54}$$

de forma que y_i representa a classe do elemento x_i. Neste caso, definidos os custos de erros e rejeições, a Regra de Chow (apresentada na Equação (2.48)) pode ser aplicada diretamente para a classificação dos padrões de teste.

É possível notar que a região de rejeição, nos dois métodos apresentados, é obtida após o treinamento do classificador e consiste de um par de hiperplanos paralelos e equidistantes ao hiperplano separador ótimo.

Um método em que a região de rejeição é obtida durante o treinamento do classificador foi proposto em Grandvalet *et al.* (2009). Nesse método, o modelo SVM-SM é modificado de forma que a Regra de Chow possa ser aplicada diretamente na saída do classificador. O modelo resultante é o seguinte:

$$\begin{aligned}
\min_{w,\, b,\, \xi,\, \eta} \quad & \frac{1}{2} w^\intercal w + \sum_{i=1}^{\ell} C_i \xi_i + D \sum_{i=1}^{\ell} \eta_i \\
\text{s.a.} \quad & y_i (w^\intercal x_i + b) \geq t_i - \xi_i, \quad i = 1, 2, \ldots, \ell \\
& y_i (w^\intercal x_i + b) \geq \tau_i - \eta_i, \quad i = 1, 2, \ldots, \ell \\
& \xi_i \geq 0, \eta_i \geq 0,
\end{aligned} \tag{2.55}$$

sendo que C é uma constante positiva, ξ_i e η_i, $i = 1, \ldots, \ell$, são variáveis de folga e $D = C(p_+ - p_-)$. Além disso, sendo $H(p) = -p\log(p) - (1-p)\log(1-p)$, define-se, para padrões da classe positiva:

- $C_i = C(1 - p_+)$,
- $t_i = H(p_+)/(1 - p_+)$,
- $\tau_i = -(H(p_-) - H(p_+))/(p_- - p_+)$,

e, para padrões da classe negativa:

- $C_i = Cp_-$,
- $t_i = H(p_-)/p_-$,
- $\tau_i = (H(p_-) - H(p_+))/(p_- - p_+)$.

Por fim, para este método, a classificação dos padrões de teste é feita de acordo com a seguinte

regra:

$$d^{\star}(x) = \begin{cases} +1, & \text{se } f(x) > f_{+} \\ -1, & \text{se } f(x) < f_{-} \\ 0, & \text{caso contrário} \end{cases}, \tag{2.56}$$

sendo $f_{+} = \log(p_{+}/(1-p_{+}))$, $f_{-} = \log(p_{-}/(1-p_{-}))$ e $f(x) = \sum_{i=1}^{\ell} \alpha_i y_i K(x, x_i)$.

Outros métodos SVMs que buscam obter a região de rejeição diretamente na fase de treino foram propostos em Fumera e Roli (2002) e Bartlett e Wegkamp (2008).

2.5 Máquinas de vetores-suporte nebulosas

Em problemas reais de classificação, alguns padrões de treino podem ser mais importantes para uma classe do que outros. Por exemplo, padrões corrompidos por ruído são menos importantes e poderiam ser descartados. No SVM tradicional, cada padrão de treino pertence única e exclusivamente à uma das duas classes possíveis. Desta forma, cada padrão tem a mesma importância na busca pelo hiperplano separador ótimo.

Em Lin e Wang (2002) foi proposto o primeiro SVM Nebuloso (FSVM), o qual introduziu uma função de pertinência nebulosa para cada padrão de treino. Com isso, padrões diferentes podem ter contribuições diferentes no processo de construção da superfície de decisão, o que pode melhorar a capacidade do SVM em lidar com *outliers* e ruídos no conjunto de dados. Sua formulação matemática é apresentada a seguir.

Considere um conjunto de dados

$$(x_i, y_i, s_i), \quad i = 1, \ldots, \ell \tag{2.57}$$

sendo que $x_i \in \mathbb{R}^p$ são os padrões de treino, y_i suas classes e $s_i \in (\sigma, 1]$ a pertinência de x_i à classe y_i, com $\sigma > 0$ suficientemente pequeno. O FSVM busca o vetor w e viés b que otimizam o problema

$$\begin{aligned} \min_{(w,b,\xi)} \quad & \frac{1}{2} w^{\mathsf{T}} w + C \sum_{i=1}^{\ell} s_i \xi_i \\ \text{s.a.} \quad & y_i(w^{\mathsf{T}} x_i + b) \geq 1 - \xi_i, \qquad i = 1, 2, \ldots, \ell, \\ & \xi_i \geq 0, \qquad\qquad\qquad\quad i = 1, 2, \ldots, \ell, \end{aligned} \tag{2.58}$$

sendo que C é uma constante positiva e ξ_i, $i = 1, \ldots, \ell$, são variáveis de folga. O termo $s_i \xi_i$ pode ser visto como uma medida dos desvios com pesos diferentes. A influência do termo ξ_i no

processo de minimização dependerá do valor de pertinência s_i; quanto maior s_i mais influente será ξ_i. Em outras palavras, é mais importante separar bem os elementos com valores altos de s_i, uma vez que o custo de desvios nesses elementos é maior. No caso em que a pertinência de todos os elementos do conjunto de dados é igual à 1, o FSVM reduz-se ao classificador SVM-SM convencional.

O método dos multiplicadores de Lagrange pode ser utilizado para encontrar a formulação dual do Problema (2.58):

$$\begin{aligned} \max_{\alpha} \quad & L(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} \alpha_i \alpha_j y_i y_j x_i^{\mathsf{T}} x_j \\ \text{s.a.} \quad & \sum_{i=1}^{\ell} \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq s_i C, \; i = 1, 2, \ldots, \ell, \end{aligned} \tag{2.59}$$

sendo que $\alpha^{\mathsf{T}} = [\alpha_1, \ldots, \alpha_\ell]$ é o vetor dos multiplicadores de Lagrange não-negativos. Note que na formulação dual do FSVM os multiplicadores de Lagrange devem ser menores ou iguais a $s_i C$, enquanto que na formulação dual do SVM-SM os multiplicadores de Lagrange devem ser menores ou iguais a C.

Em Hajiloo *et al.* (2013) o FSVM foi aplicado em conjunto de dados de leucemia, câncer de próstata e câncer de cólon. Os autores constataram que o método em questão é robusto e tem uma boa capacidade de generalização em problemas de classificação de microarranjos de expressão genética, que permitem analisar toda a atividade transcricional em uma amostra biológica.

Em Wang *et al.* (2005) foi apresentado um FSVM ponderado bilateralmente. Nesse modelo, considera-se que cada ponto do conjunto de dados pertence tanto à classe positiva quanto à negativa, porém com diferentes pertinências. Para isso, o primeiro passo é criar um novo conjunto de dados, tal que, para cada ponto do conjunto de dados original

$$(x_i, \; y_i), \quad i = 1, \ldots, \ell, \tag{2.60}$$

criam-se dois pontos no novo conjunto

$$(x_i, \; 1, \; s_i), \; (x_i, \; -1, \; 1 - s_i), \quad i = 1, \ldots, \ell. \tag{2.61}$$

Em seguida, o problema de classificação é modelado pelo seguinte problema de programação

quadrática

$$\begin{aligned} \min_{(w,b,\xi,\eta)} \quad & \frac{1}{2}w^{\intercal}w + C\sum_{i=1}^{\ell}[s_i\xi_i + (1-s_i)\eta_i] \\ \text{s.a} \quad & w^{\intercal}x_i + b \geq 1 - \xi_i,\ i = 1,2,\ldots,\ell, \\ & w^{\intercal}x_i + b \leq -1 + \eta_i,\ i = 1,2,\ldots,\ell, \\ & \xi_i \geq 0,\ \eta_i \geq 0,\ i = 1,2,\ldots,\ell. \end{aligned} \tag{2.62}$$

O modelo proposto por Wang *et al.* (2005) foi estendido em Hao *et al.* (2007), com o uso de conjuntos vagos. Conjunto vagos são generalizações de conjuntos nebulosos, onde as pertinências deixam de ser pontuais e passam a ser baseadas em intervalos.

Definição 2.5 *Sejam U o conjunto universo e x um elemento genérico de U. Um conjunto vago, V, em U é caracterizado por uma função de pertinência verdadeira, $t_V(x)$, e uma função de pertinência falsa, $f_V(x)$, com $0 \leq t_V(x) + f_V(x) \leq 1$. $t_V(x)$ é um limite inferior à pertinência de x, derivado da evidência em favor de x. Similarmente, $f_V(x)$ é um limite inferior à negação de x, derivado da evidência contrária a x. Desta forma, a pertinência de x pertence ao subintervalo $[t_V(x),\ 1 - f_V(x)]$ de $[0,\ 1]$.*

No modelo proposto em Hao *et al.* (2007), também considera-se que todos os pontos do conjunto de dados pertencem tanto à classe positiva quanto à negativa. Para isso, cria-se um novo conjunto de dados, tal que, para cada ponto do conjunto de dados original

$$(x_i,\ y_i), \quad i = 1,\ldots,\ell \tag{2.63}$$

criam-se dois pontos no novo conjunto

$$(x_i,\ 1,\ t_i),\ (x_i,\ -1,\ f_i), \quad i = 1,\ldots,\ell, \tag{2.64}$$

sendo que t_i e f_i representam as pertinências verdadeira e falsa, respectivamente, de cada ponto do conjunto de dados. É importante ressaltar que neste modelo não existe dependência entre t_i e f_i, diferentemente do modelo proposto por Wang *et al.* (2005) - onde a pertinência a uma das classes é o complementar da pertinência a outra classe. O problema de classificação é então

formulado da seguinte forma:

$$\begin{aligned} \min_{(w,b,\xi,\eta)} \quad & \frac{1}{2}w^{\mathsf{T}}w + C\sum_{i=1}^{\ell}[t_i\xi_i + f_i\eta_i] \\ \text{s.a} \quad & w^{\mathsf{T}}x_i + b \geq 1 - \xi_i,\ i = 1,2,\ldots,\ell, \\ & w^{\mathsf{T}}x_i + b \leq -1 + \eta_i,\ i = 1,2,\ldots,\ell, \\ & \xi_i \geq 0,\ \eta_i \geq 0,\ i = 1,2,\ldots,\ell. \end{aligned} \tag{2.65}$$

Tanto o modelo proposto em Wang *et al.* (2005) quanto o proposto em Hao *et al.* (2007) obtiveram bons resultados em problemas de avaliação de crédito.

Em An e Liang (2013) foi proposto um novo classificador SVM Nebuloso, o qual além de considerar funções de pertinência nebulosas para cada padrão de treino, também procura minimizar a dispersão intra-classe. Para isso, um termo que trata a dispersão intra-classe é incorporado ao modelo proposto por Lin e Wang (2002). O modelo resultante é o seguinte:

$$\begin{aligned} \min_{(w,b,\xi)} \quad & \frac{1}{2}w^{\mathsf{T}}w + \frac{1}{2}\beta w^{\mathsf{T}}S_w w + C\sum_{i=1}^{\ell}s_i\xi_i \\ \text{s.a.} \quad & w^{\mathsf{T}}y_i(w^{\mathsf{T}}x_i + b) \geq 1 - \xi_i,\ i = 1,2,\ldots,\ell, \\ & \xi_i \geq 0,\ i = 1,2,\ldots,\ell. \end{aligned} \tag{2.66}$$

sendo que $\beta \geq 0$ e $C > 0$ são parâmetros de regularização. A matriz S_w representa o espalhamento dentro da classe e é definida como:

$$S_w = \sum_{i=1,2} P(C_i) \sum_{x \in C_i} (x - m_i)(x - m_i)^{\mathsf{T}}, \tag{2.67}$$

tal que $m_i = \frac{1}{\ell_i}\sum_{x \in C_i} x$ e $P(C_i) = \ell_i/\ell$, $i \in 1,\ 2$, representam o vetor médio e a probabilidade à priori de cada classe, respectivamente.

Os autores mostraram que esse método, chamado de *FSVM with minimum within-class scatter* (WCS-FSVM), conseguiu melhorar a acurácia na classificação e a habilidade de generalização, e também apresentou uma melhor acurácia ao lidar com problemas de classificação com ruído ou *outliers*.

Em problemas com n-classes, para se utilizar o SVM, primeiramente o problema é convertido em n problemas binários. Em cada problema binário é determinada uma função de decisão que separa a i-ésima ($i = 1,\ldots,n$) classe das demais. Uma forma de fazer a classificação de um padrão é atribuí-lo à classe i somente quando o valor da i-ésima função de decisão (e somente ela) for positiva. Caso mais de uma função de decisão apresente valores positivos, ou se

todas apresentarem valores negativos, o padrão em questão não será atribuído a nenhuma classe. Métodos assim descritos são conhecidos como *one-against-all*.

Em Inoue e Abe (2001) e Abe (2015) foram propostos métodos SVM nebulosos para lidar com problemas de classificação com n-classes. A partir das i-ésimas funções de decisões geradas pelo SVM, funções de pertinência nebulosas são definidas para cada classe. Desta forma, os métodos propostos conseguem lidar com regiões não classificáveis e apresentaram melhor acurácia e capacidade de generalização em relação a métodos *one-against-all* SVM.

Como métodos *one-against-all* fogem do escopo deste trabalho, os métodos propostos por Inoue e Abe (2001) e Abe (2015) não serão considerados aqui.

2.6 Conclusão

Este capítulo apresentou conceitos fundamentais para SVMs, tais como os conceitos de margem de separação e hiperplano separador ótimo. Além disso, foram apresentadas as formulações clássicas de classificadores SVM, conhecidas como SVM-HM e SVM-SM. A Seção 2.4 introduziu o conceito de opção de rejeição, além de apresentar formulações de SVMs com opção de rejeição. Ademais, classificadores que introduzem conceitos da lógica nebulosa às SVMs foram vistos na Seção 2.5.

O próximo capítulo apresenta uma proposta de um novo classificador SVM nebuloso. O classificador proposto é uma generalização do SVM-HM, o qual incorpora restrições nebulosas a esse classificador para lidar com incertezas presentes nos conjuntos de dados. Além disso, o classificador resultante atribui valores de pertinência aos elementos do conjunto de dados, permitindo que uma classe de rejeição possa ser introduzida sem maiores dificuldades.

3 MÁQUINA DE VETORES-SUPORTE NEBULOSA: UMA NOVA PROPOSTA

Este capítulo apresenta uma proposta de um novo classificador SVM, que utiliza a lógica nebulosa para lidar com incertezas nos conjuntos de dados. Assim como o SVM-SM, o classificador em questão é uma generalização do SVM-HM. Contudo, essa generalização é feita relaxando as restrições do SVM-HM, sem que seja necessário a introdução de variáveis de folga. Além disso, o classificador proposto atribui valores de pertinência aos elementos do conjunto de dados, permitindo que uma classe de rejeição possa ser introduzida ao modelo.

Na Seção 3.1 é feita uma motivação para o desenvolvimento desse classificador. Na Seção 3.2 são apresentadas a formulação matemática e algumas propriedades desse classificador. A variante com opção de rejeição é discutida na Seção 3.3.

3.1 Motivação

Como visto no Capítulo 2, em um primeiro momento, os classificadores SVM foram propostos apenas para conjuntos de dados linearmente separáveis. Nesse tipo de conjunto, elementos de classes opostas podem ser facilmente diferenciados. Para lidar com conjuntos de dados mais complexos, quando a diferença entre os elementos de classes opostas não é tão clara, Cortes e Vapnik (1995) sugeriram a introdução de variáveis de folga ao problema original, assim como a de um parâmetro regularizador. Um problema dessa abordagem é que, além de ser necessário calibrar um parâmetro extra, é comum ocorrer um aumento na quantidade de vetores-suporte, aumentando o custo computacional de se classificar padrões de teste.

O classificador FSVM proposto por Lin e Wang (2002), o qual é uma generalização do SVM-SM, introduz a pertinência de cada padrão de treino à sua classe como uma forma de ponderar a importância desses padrões. Por ser baseado no SVM-SM, o FSVM apresenta os mesmos problemas que esse método.

Em Cruz *et al.* (2011), os autores propuseram uma abordagem paramétrica para resolver problemas de programação quadrática com relação de ordem nebulosa no conjunto de restrições. Nesse tipo de problema as restrições são relaxadas, podendo ser violadas até um limite pré-definido pelo usuário. Além disso, variáveis de folga não são necessárias para o relaxamento das restrições.

Neste capítulo, esta abordagem é adaptada para o desenvolvimento de um novo classificador SVM nebuloso. Ao trabalhar com o SVM-HM em um ambiente com restrições

com a relação de ordem nebulosa, permite-se que esse classificador seja utilizado em conjuntos de dados não linearmente separáveis, ao mesmo tempo em que não são criados novos tipos de vetores-suporte. Esse novo classificador será doravante chamado de Máquina de Vetores-Suporte com Abordagem Paramétrica, tradução livre de *Support-Vector Machine with Parametric Approach* (SVM-PA).

3.2 Desenvolvimento do método

Considere um conjunto de dados de treino $T = \{(x_i, y_i) \mid i = 1, \ldots, \ell\}$, sendo que $x_i \in \mathbb{R}^p$ representa o i-ésimo padrão de treino e $y_i \in \{-1, 1\}$ representa a classe à qual x_i pertence. O hiperplano separador ótimo para T, de acordo com o SVM-HM, é aquele que maximiza a margem de separação entre as classes 1 e -1. Esse hiperplano é encontrado através da solução do seguinte problema de programação quadrática:

$$\begin{aligned} \min_{w,\, b} \quad & \frac{1}{2} w^{\mathsf{T}} w \\ \text{s.a.} \quad & y_i(w^{\mathsf{T}} x_i + b) \geq 1, \qquad i = 1, 2, \ldots, \ell. \end{aligned} \tag{3.1}$$

Devido as restrições desse problema, uma solução será encontrada se, e somente se, T for linearmente separável. Visando atenuar a condição de separabilidade linear para T, pode-se generalizar o Problema (3.1) substituindo as restrições $y_i(w^{\mathsf{T}} x_i + b) \geq 1, i = 1, 2, \ldots, \ell$ por restrições com a relação de ordem nebulosa. Desta forma, obtém-se um problema de programação quadrática com relação de ordem nebulosa no conjunto de restrições

$$\begin{aligned} \min_{w,\, b} \quad & \frac{1}{2} w^{\mathsf{T}} w \\ \text{s.a.} \quad & y_i(w^{\mathsf{T}} x_i + b) \geq^{f} 1, \qquad i = 1, 2, \ldots, \ell \end{aligned} \tag{3.2}$$

cuja solução, de acordo com Cruz *et al.* (2011), é encontrada resolvendo um problema de programação quadrática paramétrico, equivalente ao Problema (3.2), descrito abaixo

$$\begin{aligned} \min_{w,\, b} \quad & \frac{1}{2} w^{\mathsf{T}} w \\ \text{s.a.} \quad & y_i(w^{\mathsf{T}} x_i + b) \geq 1 - d_i(1 - \lambda), \qquad i = 1, 2, \ldots, \ell, \\ & \lambda \in [0,\ 1]. \end{aligned} \tag{3.3}$$

No Problema (3.3), λ indica a pertinência da solução e $d_i \geq 0$ representa a violação máxima da i-ésima restrição, ou seja, quão incerto é o i-ésimo padrão de treino (x_i, y_i). O desenvolvimento algébrico para obter o Problema (3.3) a partir do Problema (3.2) está detalhado na Seção A.2.

Vale ressaltar que a pertinência λ controla o quanto cada restrição poderá ser violada. Para $\lambda = 0$ é permitido o máximo de violação, enquanto que para $\lambda = 1$ nenhuma violação é permitida. Ou seja, para $\lambda = 1$ o SVM-PA é idêntico ao SVM-HM.

Além disso, o conjunto de soluções viáveis é determinado pelos valores de d_i, $i = 1,\ldots,\ell$ e de λ, de tal maneira que o hiperplano separador ótimo será aquele com a menor norma dentro do conjunto de soluções viáveis. A seguir é apresentada uma propriedade interessante sobre os conjuntos de soluções viáveis e as pertinências das soluções.

Proposição 3.1 *Seja S_λ o conjunto de soluções viáveis com pertinência λ e considere $0 \leq \lambda_2 \leq \lambda_1 \leq 1$. Assim, tem-se que*

$$S_{\lambda_1} \subseteq S_{\lambda_2}. \tag{3.4}$$

Essa propriedade decorre diretamente das restrições do Problema (3.3).

Assim como foi feito na Seção 2.1, pode-se utilizar o método dos multiplicadores de Lagrange para resolver esse problema. Para isso, constrói-se a função Lagrangeana

$$L(w,b,\alpha) = \frac{1}{2}w^\intercal w - \sum_{i=1}^{\ell} \alpha_i(y_i(w^\intercal x_i + b) - 1 + d_i(1-\lambda)), \tag{3.5}$$

sendo $\alpha = [\alpha_1,\ldots,\alpha_\ell] \geq 0$ são os multiplicadores de Lagrange correspondentes as restrições $y_i(w^\intercal x_i + b) \geq 1 - d_i(1-\lambda),\ i = 1,2,\ldots,\ell$. A solução do Problema (3.5) é determinada pelo ponto que satisfaz o conjunto de restrições, minimiza a função L em relação às variáveis do problema primal (w e b) e maximiza L em relação aos multiplicadores de Lagrange.

Para encontrar o mínimo de L em relação a w e b, calcula-se as derivadas parciais de L em relação a essas variáveis e as equações resultantes são igualadas a zero, ou seja

$$\left.\frac{\partial L}{\partial w}\right|_{w=w_0} = w_0 - \sum_{i=1}^{\ell} \alpha_i y_i x_i = 0, \tag{3.6}$$

$$\left.\frac{\partial L}{\partial b}\right|_{b=b_0} = -\sum_{i=1}^{\ell} \alpha_i y_i = 0. \tag{3.7}$$

Das Equações (3.6) e (3.7) encontram-se as seguintes relações

$$w_0 = \sum_{i=1}^{\ell} \alpha_i y_i x_i, \tag{3.8}$$

$$\sum_{i=1}^{\ell} y_i \alpha_i = 0, \tag{3.9}$$

as quais podem ser utilizadas para simplificar L:

$$L(w,b,\alpha) = \frac{1}{2}w_0^\intercal w_0 - w_0^\intercal \sum_{i=1}^{\ell} y_i\alpha_i x_i - b\sum_{i=1}^{\ell}\alpha_i y_i - (1-\lambda)\sum_{i=1}^{\ell}\alpha_i d_i + \sum_{i=1}^{\ell}\alpha_i \tag{3.10}$$

$$= \frac{1}{2}w_0^\intercal w_0 - w_0^\intercal w_0 - (1-\lambda)\sum_{i=1}^{\ell}\alpha_i d_i + \sum_{i=1}^{\ell}\alpha_i \tag{3.11}$$

$$= -\frac{1}{2}w_0^\intercal w_0 - (1-\lambda)\sum_{i=1}^{\ell}\alpha_i d_i + \sum_{i=1}^{\ell}\alpha_i \tag{3.12}$$

Reordenando a equação acima e utilizando novamente a relação (3.8), obtém-se

$$L(\alpha) = \sum_{i=1}^{\ell}\alpha_i - (1-\lambda)\sum_{i=1}^{\ell}\alpha_i d_i - \frac{1}{2}\sum_{i=1}^{\ell}\sum_{j=1}^{\ell}\alpha_i\alpha_j y_i y_j x_i^\intercal x_j. \tag{3.13}$$

Portanto, a formulação dual do Problema (3.3) pode ser modelada da seguinte forma

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{\ell}\alpha_i(1-(1-\lambda)d_i) - \frac{1}{2}\sum_{i=1}^{\ell}\sum_{j=1}^{\ell}\alpha_i\alpha_j y_i y_j x_i^\intercal x_j \\ \text{s.a.} \quad & \sum_{i=1}^{\ell}\alpha_i y_i = 0 \\ & \alpha_i \geq 0, \qquad i = 1,2,\ldots,\ell. \end{aligned} \tag{3.14}$$

A solução do Problema (3.14) serão os multiplicadores de Lagrange associados ao ponto que minimiza o Problema (3.3).

De acordo com as condições de complementaridade de KKT, no ponto de sela (w_0, b_0, α^0), a equação

$$\alpha_i^0\left[y_i(w_0^\intercal x_i + b_0) - 1 + d_i(1-\lambda)\right] = 0$$

é válida para todo $i = 1,2,\ldots,\ell$. Assim, os vetores-suporte da SVM-PA são os elementos que satisfazem a equação

$$y_i(w_0^\intercal x_i + b) = 1 - d_i(1-\lambda).$$

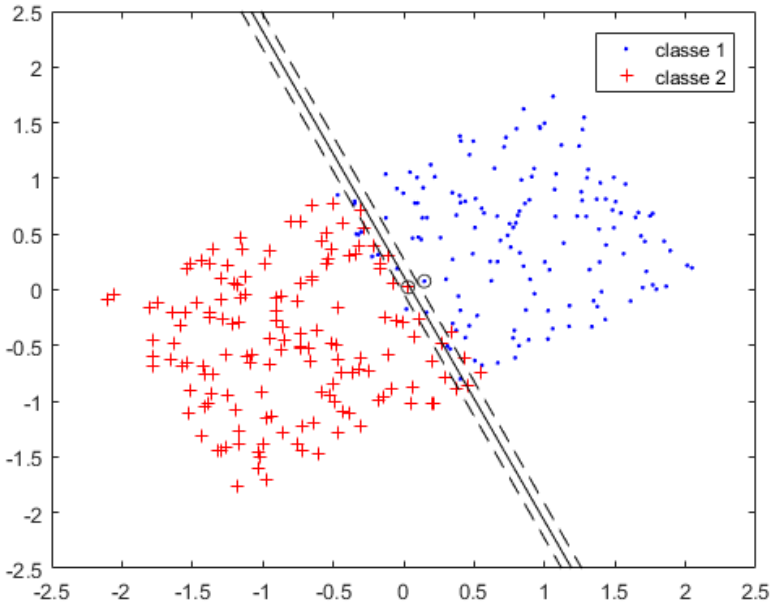
A classificação de novos padrões é feita verificando o sinal da seguinte função:

$$f(x) = \sum_{i\in SV}\alpha_i^0 y_i x_i^\intercal x + b_0, \tag{3.15}$$

sendo que SV representa o conjunto com os índices dos vetores-suporte.

A Figura 10 descreve o hiperplano separador ótimo para o conjunto de dados apresentado na Figura 5, encontrado pelo SVM-PA com pertinência $\lambda = 0,1$.

Figura 10 – Hiperplano separador ótimo encontrado pelo SVM-PA com pertinência $\lambda = 0,1$ para o conjunto de dados apresentado na Figura 5. Os vetores-suporte estão destacados com círculos cinzas.



Fonte: elaborada pelo autor.

É importante ressaltar que o SVM-HM, diferentemente do SVM-PA e do SVM-SM, não encontraria uma solução para o conjunto de dados apresentado na Figura 5, uma vez que tal conjunto não é linearmente separável. Ademais, a solução encontrada pelo SVM-PA (Figura 10) difere daquela encontrada pelo SVM-SM (Figura 6). Além de os hiperplanos separadores ótimos serem diferentes, a quantidade de vetores-suporte obtida pelo SVM-PA foi consideravelmente menor do que a quantidade de vetores-suporte obtida pelo SVM-SM.

3.3 Variante com opção de rejeição

O modelo apresentado na seção anterior busca pelo hiperplano que maximiza a margem de separação das classes dentro do conjunto de possíveis soluções, definido pelas restrições do modelo. Uma interpretação que pode ser dada a saída desse modelo é a de que os elementos que estão entre as margens do hiperplano separador ótimo possuem pertinência positiva a ambas as classes.

Considerando que os elementos dentro das margens do HSO possuem pertinência λ a classe oposta à sua, pode-se definir o espaço entre as margens do HSO como uma região de

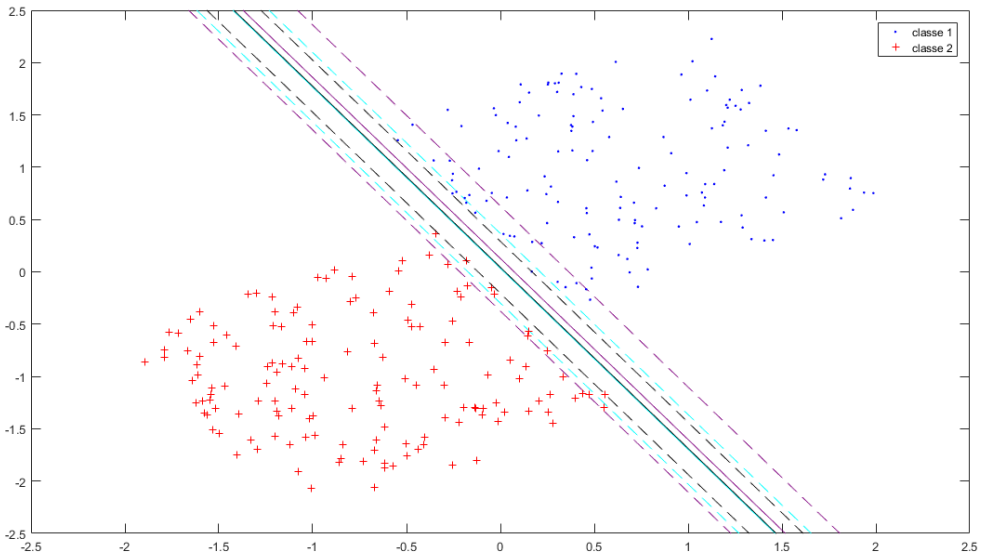
rejeição. Neste caso, a classificação de novos padrões é feita de acordo com a seguinte regra:

$$d^{\star}(x) = \begin{cases} +1, & \text{se } f(x) \geq \frac{1}{\|w_0\|} \\ -1, & \text{se } f(x) \leq -\frac{1}{\|w_0\|} \\ 0, & \text{caso contrário} \end{cases}, \tag{3.16}$$

sendo que w_0 é o vetor normal ao HSO. Esta formulação será doravante chamada de Máquina de Vetores-Suporte com Abordagem Paramétrica e Opção de Rejeição, tradução livre de *Support-Vector Machine with Parametric Approach and Rejection Option* (SVM-PA-RO).

De acordo com a relação (3.4), percebe-se que quanto menor a pertinência da solução, menor será a norma de w_0. Como a margem de separação é inversamente proporcional a norma de w_0, pode-se concluir que quanto menor a pertinência, maior será a margem de separação. Desta forma, quanto menor a pertinência da solução, espera-se que uma maior quantidade de elementos serão rejeitados. Esta propriedade é exemplificada na Figura 11.

Figura 11 – As linhas contínuas são os hiperplanos separadores ótimos encontrados pelo SVM-PA com pertinências $\lambda_1 = 1$ (em preto), $\lambda_2 = 0,75$ (em azul claro) e $\lambda_3 = 0,5$ (em roxo). As linhas tracejadas são as margens desses hiperplanos. As linhas contínuas em preto e em azul claro estão sobrepostas.



Fonte: elaborada pelo autor.

Note que a função de decisão apresentada na Equação (3.16) é similar as funções de decisões apresentadas na Seção 2.4. Assim como os modelos propostos em Mukherjee *et al.* (1999) e Platt *et al.* (1999), a região de rejeição é definida como o espaço entre dois

hiperplanos equidistante do HSO. A região de rejeição do método proposto em Grandvalet *et al.* (2009) também é definida como o espaço entre dois hiperplanos, porém não necessariamente equidistantes do HSO.

3.4 Conclusão

Neste capítulo foi apresentado um novo classificador SVM nebuloso, baseado em uma abordagem paramétrica desenvolvida para resolver problemas de programação quadrática com relação de ordem nebulosa no conjunto de restrições. Além disso, foi apresentado uma variante desse classificador com opção de rejeição.

No próximo capítulo é feita uma análise da eficiência das duas versões do classificador proposto, com e sem opção de rejeição. Esses classificadores são testados em conjuntos de dados reais e comparadas a outros classificadores SVMs.

4 RESULTADOS E DISCUSSÃO

Para verificar a eficácia dos modelos SVM-PA e SVM-PA-RO, foram feitos testes em diversos conjuntos de dados oriundos da área médica. A Seção 4.1 detalha a metodologia utilizada para o treinamento e teste dos classificadores. Os conjuntos de dados utilizados nos testes são descritos na Seção 4.2. Os resultados são apresentados na Seção 4.3. Na Subseção 4.3.1 é feita uma comparação entre os métodos SVM-PA e SVM-SM. Já na Subseção 4.3.2, o SVM-PA-RO é comparado com os métodos SVM com opção de rejeição apresentados na Seção 2.4.

4.1 Metodologia de treinamento e teste

Para cada conjunto de dados foram feitas 50 rodadas de treino e teste. Em cada rodada, o conjunto de dados é dividido em dois subconjuntos: o conjunto de treino, com 70% dos elementos, e o conjunto de teste, com os demais elementos. Essa divisão é feita de forma a manter a proporção de elementos de cada classe em ambos os subconjuntos. Além disso, o conjunto de treino é normalizado, de forma a equalizar as ordens de grandezas dos seus atributos. A normalização é feita mudando a escala original de cada variável para o intervalo $[-1,\ 1]$. O conjunto de teste é então normalizado de acordo com os parâmetros da normalização do conjunto de treino[1].

Para o SVM-PA também é feito um passo adicional antes do treinamento: a remoção de outliers[2] no conjunto de treino. No caso, um elemento é considerado um *outlier* se sua distância[3] média até 5%[4] elementos mais próximos de sua classe é maior que a distância média até 5% elementos mais próximos da classe oposta. As violações máximas para o SVM-PA são estimadas de acordo com o método descrito no Apêndice B.

Além disso, foi utilizado *kernel* gaussiano (Tabela 1) em todos os classificadores e para todos os conjuntos de dados.

Os códigos necessários para gerar os resultados apresentados neste capítulo foram desenvolvidos em Matlab®, utilizando as funções *fitcsvm* e *quadprog*. A primeira delas utiliza

[1] Se um padrão do conjunto de treino foi normalizado de $(-10, 10)$ para $(-1, 1)$ e um elemento do conjunto de treino é $(-12, 7)$, então esse padrão será normalizado para $(-1.2, 0.7)$.

[2] A remoção de outliers faz com as estimativas feitas pelo método utilizado para estimar a violações máximas, apresentado no Apêndice B, sejam mais precisas.

[3] A distância utilizada é a euclidiana.

[4] Valor definido de forma empírica.

o algoritmo *Sequential Minimal Optimization* (PLATT, 1999) para encontrar a solução ótima do problema (2.33). Já a segunda é um *solver* para problemas de programação quadrática com restrições lineares.

4.2 Conjuntos de dados utilizados

A seguir, é apresentada uma breve descrição dos conjuntos de dados utilizados, obtidos no Repositório UCI (DUA; GRAFF, 2017):

- **Câncer de mama:** Conjunto de dados obtido e disponibilizado pelo Dr. William H. Wolberg da Universidade de Wisconsin (MANGASARIAN, 1990). O conjunto possui dados de 699 pacientes com tumores na mama, dos quais 458 eram tumores benignos. Cada paciente é representado nos dados por 10 atributos biomédicos. O objetivo é identificar se um tumor na mama do paciente é benigno ou maligno.
- **Haberman:** Conjunto formado por casos de um estudo, realizado entre 1958 e 1970 no Hospital Billings da Universidade de Chicago, sobre a sobrevivência de pacientes submetidos a cirurgia para câncer de mama (HABERMAN, 1976). Esse estudo acompanhou 306 pacientes, sendo que 225 deles sobreviveram por pelo menos 5 anos após a cirurgia. Os pacientes são representados por três atributos: idade, ano da operação e quantidade de gânglios axilares positivos detectados. O objetivo é prever se determinado paciente sobreviveu (sobriverá) mais de 5 anos após a cirurgia.
- **Parkinsons:** Conjunto de dados criado por Max Little, da Universidade de Oxford, em colaboração com o *National Centre for Voice and Speech*, em Denver, Estados Unidos (LITTLE *et al.*, 2009). O conjunto é composto de uma série de medições biomédicas da voz de 31 pessoas, 23 delas com doença de Parkinson. Ao todo foram realizadas 195 gravações, 147 destas da voz de pessoas com Parkinson e 48 de pessoas sem esta doença. Neste conjunto, o objetivo é identificar, através da gravação da voz do paciente, se o mesmo possui ou não a doença de Parkinson.
- **Coluna vertebral:** Conjunto de dados biomédicos construído pelo Dr. Henrique da Mota durante um período de residência médica no *Group of Applied Research in Orthopaedics (GARO)* do *Centre médico-chirurgical de réadaptation des Massues*, Lyon, França. O objetivo é classificar os pacientes como pertencentes a uma das duas categorias: Normal (100 pacientes) ou Anormal (210 pacientes) (ROCHA NETO; BARRETO, 2009). Cada paciente é representado nos dados por seis atributos biomecânicos derivados da forma e

orientação da pelve e da coluna lombar.

4.3 Resultados

4.3.1 Comparação entre classificadores SVM sem opção de rejeição

Esta seção apresenta uma comparação entre os classificadores SVM-SM e SVM-PA. A eficácia de ambos os métodos é verificada nos quatro conjuntos de dados descritos previamente. Aqui, a eficácia é obtida verificando-se a quantidade de elementos do conjunto de teste que é classificada corretamente, a partir da superfície de decisão obtida no conjunto de treino. A quantidade média de vetores-suporte obtida por cada método também é considerada.

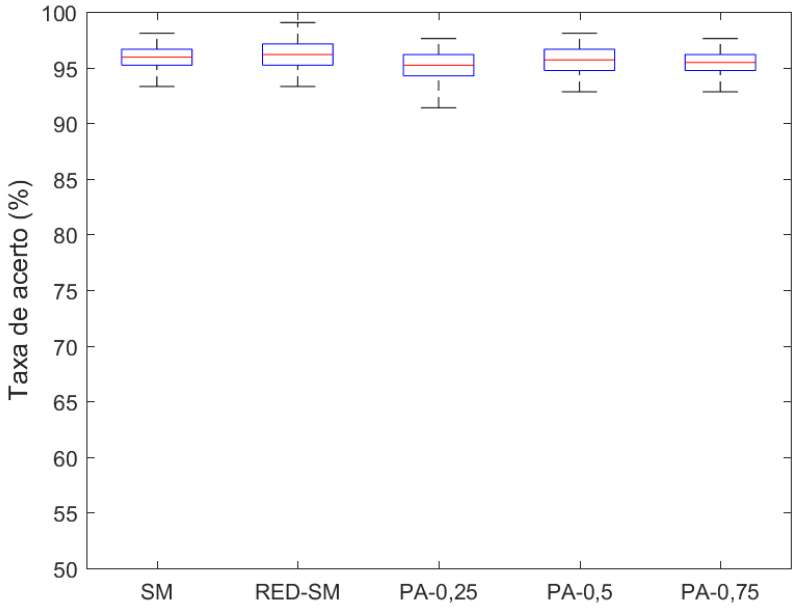
Como para os testes realizados foi utilizado *kernel* gaussiano, dois parâmetros devem ser escolhidos para o SVM-SM: a abertura do *kernel* (σ) e a constante de regularização (C). Já para o SVM-PA é necessário escolher apenar um parâmetro: a abertura do *kernel*.

Os melhores parâmetros para cada modelo são encontrados através de uma busca em grade, com base em uma estratégia de validação cruzada de 5-dobras, realizada no conjunto de treinamento. Para σ a busca é feita no conjunto $\{2^{-3}, 2^{-2}, 2^{-1}, 2^{-0.5}, 2^{0}, 2^{0.5}, 2^{1}, 2^{2}, 2^{3}\}$, enquanto que para C a busca é feita no conjunto $\{2^{-3}, 2^{-2}, 2^{-1}, 2^{0}, 2^{1}, 2^{2}, 2^{3}, 2^{5}, 2^{7}\}$. Como o SVM-PA possui apenas um parâmetro, a busca é feita em apenas uma dimensão.

Na Tabela 2 são apresentados os resultados obtidos pelos classificadores SVM-SM e SVM-PA, com pertinências $0,75$, $0,5$ e $0,25$, nos conjuntos de dados câncer de mama, haberman, parkinsons e coluna vertebral. Além disso, como para a aplicação do SVM-PA alguns elementos são removidos do conjunto de dados, o classificador chamado de RED-SM aplica o SVM-SM nesse conjunto de dados reduzido. O desempenho dos classificadores é avaliado de acordo com a taxa de acerto média, desvio padrão e quantidade média de vetores-suporte.

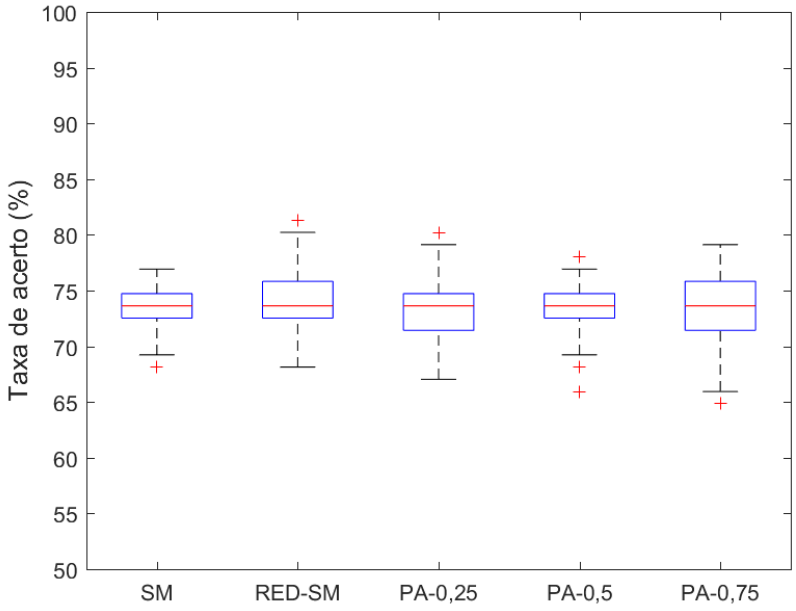
Outras informações do desempenho dos métodos são apresentadas em diagramas de caixa (*boxplots*), nas Figuras 12, 13, 14 e 15. Em cada caixa, a marca central indica a mediana, e as bordas inferior e superior indicam o primeiro e o terceiro quartis, respectivamente. Os bigodes estendem-se até os pontos de dados mais extremos que não são considerados outliers, e os outliers são plotados individualmente usando o símbolo $+$.

Figura 12 – Diagramas de caixa das taxas de acerto dos classificadores SVM-SM, RED-SM e SVM-PA com pertinências $0,75$, $0,5$ e $0,25$, em 50 rodadas de treino e teste no conjunto de dados câncer de mama.



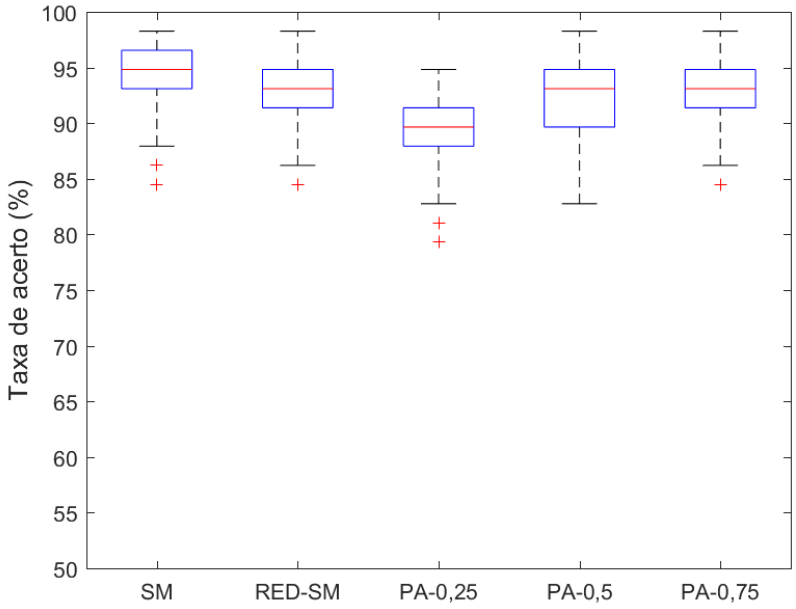
Fonte: elaborada pelo autor.

Figura 13 – Diagramas de caixa das taxas de acerto dos classificadores SVM-SM, RED-SM e SVM-PA com pertinências $0,75$, $0,5$ e $0,25$, em 50 rodadas de treino e teste no conjunto de dados haberman.



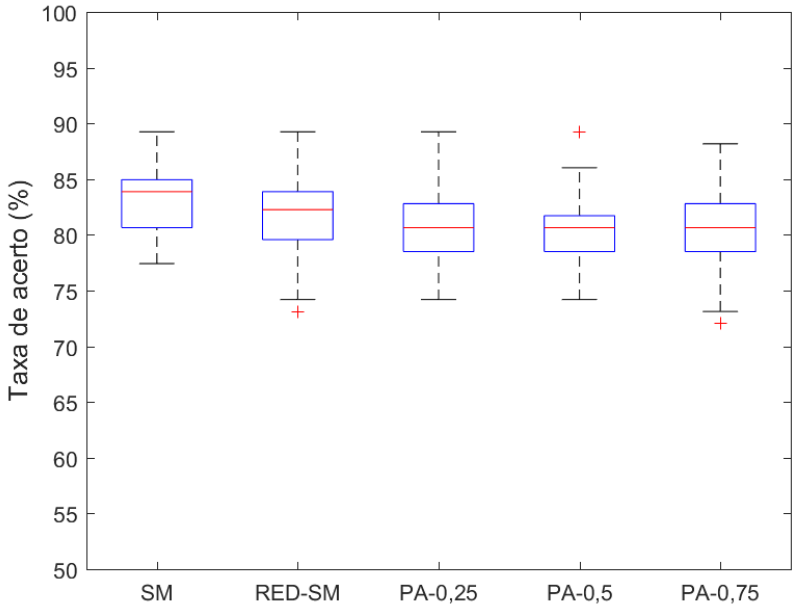
Fonte: elaborada pelo autor.

Figura 14 – Diagramas de caixa das taxas de acerto dos classificadores SVM-SM, RED-SM e SVM-PA com pertinências $0,75$, $0,5$ e $0,25$, em 50 rodadas de treino e teste no conjunto de dados parkinsons.



Fonte: elaborada pelo autor.

Figura 15 – Diagramas de caixa das taxas de acerto dos classificadores SVM-SM, RED-SM e SVM-PA com pertinências $0,75$, $0,5$ e $0,25$, em 50 rodadas de treino e teste no conjunto de dados coluna vertebral.



Fonte: elaborada pelo autor.

Tabela 2 – Tabela com a taxa de acerto média, desvio padrão e quantidade média de vetores-suporte (VS) obtidos pelos métodos SVM-SM, RED-SM e SVM-PA com pertinências $0,75$, $0,5$ e $0,25$ nos conjuntos de dados câncer de mama (BC), haberman (HB), parkinsons (PK) e coluna vertebral (CV). Na coluna "Redução VS" é apresentada a porcentagem de redução na quantidade média de vetores-suporte em relação ao SVM-SM.

	Classificador	**Taxa de Acerto (%)**	**Desvio Padrão (%)**	**VS**	**Redução VS (%)**
BC	SVM-SM	95,90	1,16	78,02	0,00
	RED-SM	96,25	1,34	42,96	44,94
	SVM-PA-0,75	95,43	1,12	40,16	48,53
	SVM-PA-0,5	95,62	1,29	27,58	64,65
	SVM-PA-0,25	95,35	1,30	18,16	76,72
HB	SVM-SM	73,58	1,83	120,94	0,00
	RED-SM	73,69	2,72	33,34	72,43
	SVM-PA-0,75	73,47	2,83	21,94	81,86
	SVM-PA-0,5	73,23	2,57	39,14	67,67
	SVM-PA-0,25	73,18	2,70	16,42	86,42
PK	SVM-SM	94,03	3,10	80,36	0,00
	RED-SM	92,41	3,44	77,09	4,07
	SVM-PA-0,75	93,10	3,03	67,42	16,10
	SVM-PA-0,5	92,10	3,41	49,14	38,85
	SVM-PA-0,25	89,03	3,54	34,38	57,22
CV	SVM-SM	83,35	2,99	92,26	0,00
	RED-SM	81,61	3,45	55,18	40,19
	SVM-PA-0,75	80,49	3,60	34,64	62,45
	SVM-PA-0,5	80,70	3,34	29,34	68,20
	SVM-PA-0,25	80,81	3,10	25,18	72,71

Fonte: elaborada pelo autor.

Nos conjuntos câncer de mama e haberman, a taxa de acerto dos cinco classificadores foi muito próxima. Em relação a quantidade de vetores-suporte, o SVM-PA-0,25 foi superior aos demais classificadores. Em ambos os conjuntos, essa quantidade foi mais do que duas vezes menor do que a obtida pelo classificador RED-SM. Em comparação ao SVM-SM, a quantidade de vetores-suporte foi cerca de quatro vezes menor no conjunto câncer de mana e cerca de sete vezes menor no conjunto haberman.

No conjunto parkinsons, as melhores taxas de acerto foram obtidas pelos classificadores SVM-SM e SVM-PA-0,75, seguidos pelos classificadores RED-SM e SVM-PA-0,5. O SVM-PA-0,25 foi o método que obteve a menor quantidade de vetores-suporte, sendo essa quantidade cerca de duas vezes menor do que a dos classificadores SVM-SM e SVM-RED. Porém, a acurácia do SVM-PA-0,25 foi consideravelmente inferior à acurácia dos demais classificadores.

No conjunto coluna vertebral, a melhor acurácia foi obtida pelo SVM-SM, seguido do RED-SM. Novamente, o método que apresentou a menor quantidade de vetores-suporte foi o

SVM-PA-0,25, sendo esta cerca de quatro vezes menor do que a do SVM-SM e cerca de duas vezes menor do que a do RED-SM.

Em geral, pode-se observar que o SVM-PA obteve taxas de acerto similares as taxas de acerto do SVM-SM, mas com uma menor quantidade de vetores-suporte. Uma quantidade reduzida de vetores-suporte é desejável, principalmente em problemas de grande porte, pois faz com que o processo de classificação de novos padrões seja mais rápido, além de ser necessário armazenar uma quantidade menor de informação para ser aplicada no conjunto de teste.

Essa redução na quantidade de vetores-suporte deve-se, em parte, ao fato de que para o treinamento do SVM-PA alguns elementos são removidos do conjunto de dados. Contudo, a quantidade de vetores-suporte obtida pelo SVM-PA-0,25 foi, em todos os conjuntos de dados, pelo menos duas vezes menor do que a obtida pelo RED-SM.

4.3.2 Comparação entre classficadores SVM com opção de rejeição

A análise de classificadores com opção de rejeição é comumente feita através de uma curva, denominada curva A-R, que leva em consideração a taxa de acerto em relação à taxa de rejeição. Cada valor dessa curva corresponde a uma taxa de acerto e sua taxa de rejeição correspondente.

Observando a Regra de Chow, apresentada na Equação (2.48), é possível perceber que a taxa de rejeição depende dos custos dos possíveis erros (falso-positivo, c_-, e falso-negativo, c_+) e dos seus custos de rejeição (r_- e r_+). Visando satisfazer a relação (2.49) e obter diferentes valores de taxas de acerto e rejeição, nos testes realizados foram utilizados os valores $c_- = c_+ = 1$ e $r_- = r_+ = w_R \in \{0,07; 0,14; \ldots; 0,41; 0,49\}$.

A seguinte nomenclatura será utilizada no decorrer desta seção:

- SVM-PA-RO: modelo proposto neste trabalho, na Seção 3.3;
- SVM-NV: primeiro modelo apresentado na Seção 2.4 (Equações (2.50) e (2.51));
- SVM-1C: segundo modelo apresentado na Seção 2.4 (Equações (2.52), (2.53) e (2.54));
- SVM-GV: terceiro modelo apresentado na Seção 2.4 (Equações (2.55) e (2.56)).

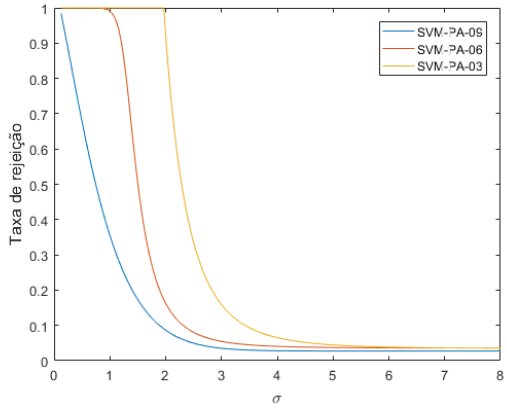
Os métodos SVM-1C e SVM-NV utilizam as saídas do SVM-SM - vide Equação (2.50) - para encontrar a região de rejeição. Assim, como foi utilizado *kernel* gaussiano para os testes desta seção, é necessário escolher os parâmetros do SVM-SM. Esses parâmetros são encontrados conforme a estratégia apresentada na terceiro parágrafo da Seção 4.3.1. Uma estratégia similar é utilizada para encontrar os parâmetros do SVM-GV (C_{GV} e σ_{GV}), porém

buscando minimizar a função $g(w_R) = |\mathbb{E}| + w_R|\mathbb{R}|$, sendo que $|\mathbb{E}|$ representa a quantidade de erros, $|\mathbb{R}|$ a quantidade de rejeições e w_R é o custo de rejeição.

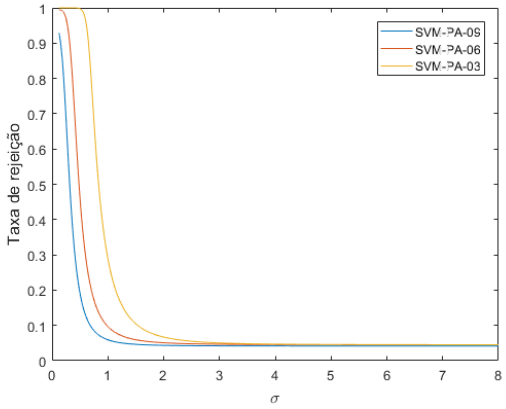
Diferentemente desses modelos, a taxa de rejeição do SVM-PA-RO depende da pertinência da solução e da dispersão dos dados no espaço de características, o que, por sua vez, depende do *kernel* utilizado. Isso deve-se ao fato de que a região de pertinência do SVM-PA-RO é definida como o espaço entre as margens do HSO. Além disso, como visto no Capítulo 3, quanto menor a pertinência da solução maior é o espaço entre as margens. Ou seja, a medida que a pertinência da solução diminui, espera-se que uma maior quantidade de elementos sejam rejeitados.

Na Figura 16 encontram-se exemplos da relação entre a abertura do *kernel* gaussiano e a taxa de rejeição do SVM-PA-RO, com pertinências $0,9$, $0,6$ e $0,3$, nos quatro conjuntos de dados avaliados.

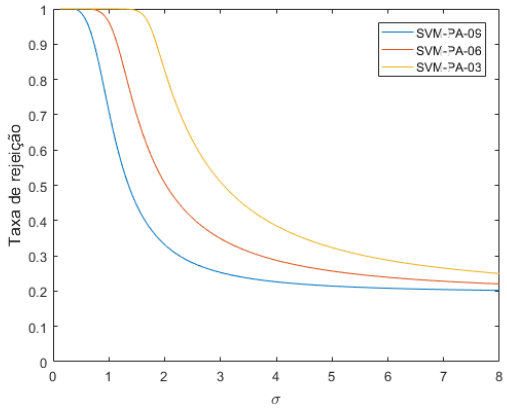
Figura 16 – Relação entre a abertura do *kernel* gaussiano (σ) e a taxa de rejeição para o modelo SVM-PA-RO, com pertinências $0,9$, $0,6$ e $0,3$ nos conjuntos de dados câncer de mama, haberman, parkinsons e coluna vertebral.



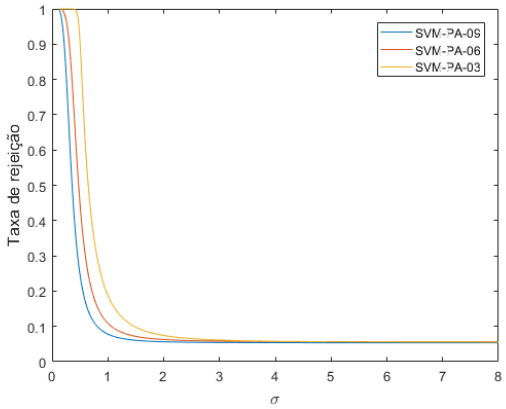
(a) Câncer de mama.



(b) Haberman.



(c) Parkinsons.



(d) Coluna vertebral.

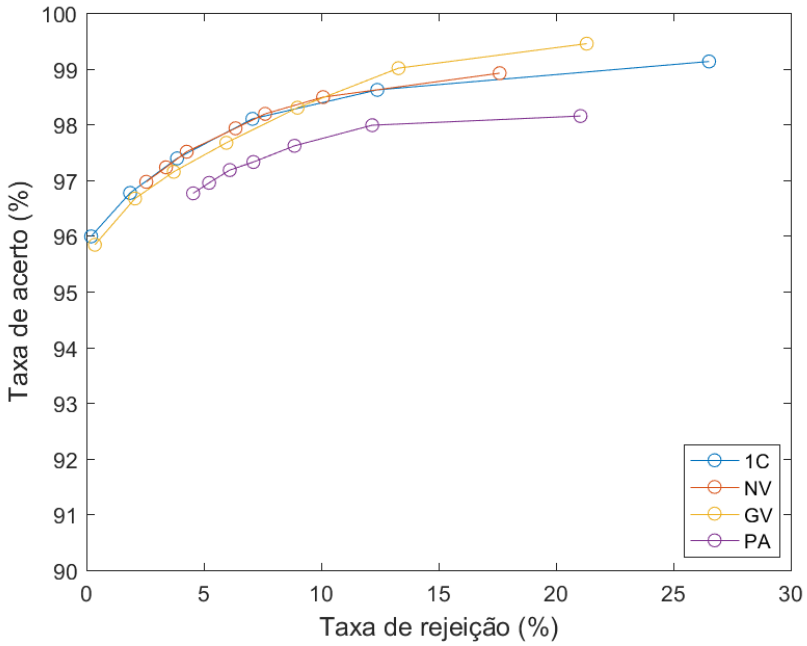
Fonte: elaborada pelo autor.

As Figuras 17, 18, 19 e 20 apresentam as curvas A-R para os classificadores SVM-1C, SVM-NV, SVM-GV e SVM-PA-RO nos conjuntos de dados câncer de mama, haberman, parkinsons e coluna vertebral, respectivamente. Cada círculo nessas figuras representa a média da taxa de acerto e de rejeição obtidas nos conjuntos de testes, dentre as 50 rodadas de treino e teste, com custos de rejeição fixos. Foram considerados apenas valores de rejeição menores do que 30%, uma vez que geralmente estes são os valores de interesse em aplicações práticas (FUMERA; ROLI, 2002). Por isso, nessas figuras, nem todos os classificadores possuem a mesma quantidade de pontos. Para o SVM-PA-RO foram utilizadas pertinências $\lambda = \{0,3; 0,4; \ldots; 0,8; 0,9\}$.

A Tabela 3 apresenta a quantidade média de vetores-suporte obtidas pelos classificadores SVM-1C, SVM-NV, SVM-GV e SVM-PA-RO. Para os classificadores SVM-1C e

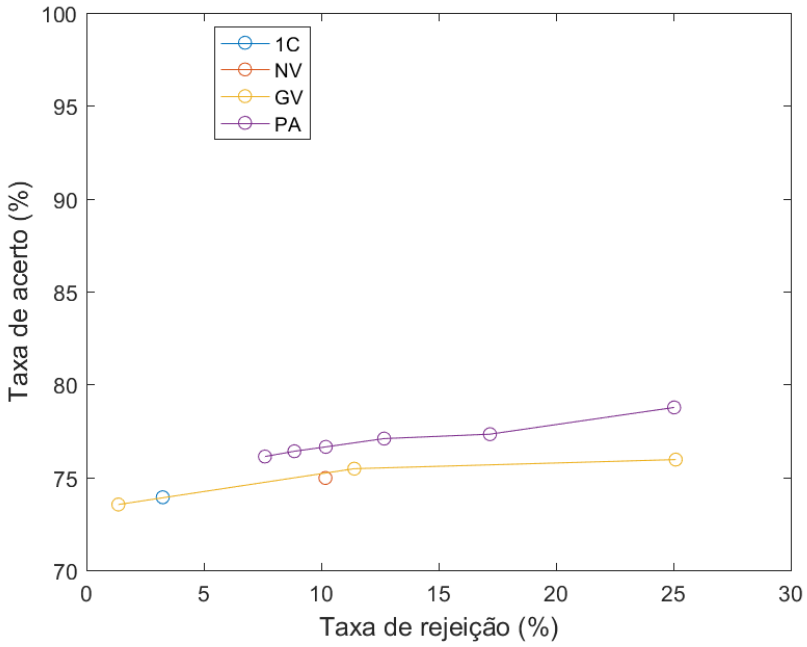
SVM-NV essa quantidade não depende do custo de rejeição, uma vez que tais modelos são baseados na saída do SVM-SM. Já para o SVM-GV, a quantidade de vetores-suporte tende a aumentar a medida que o custo de rejeição diminui. Para o SVM-PA-RO há uma variação na quantidade de vetores-suporte para soluções com pertinência diferente, mas essa variação é mínima.

Figura 17 – Curvas A-R dos métodos SVM-1C, SVM-NV, SVM-GV e SVM-PA-RO no conjunto de dados câncer de mama.



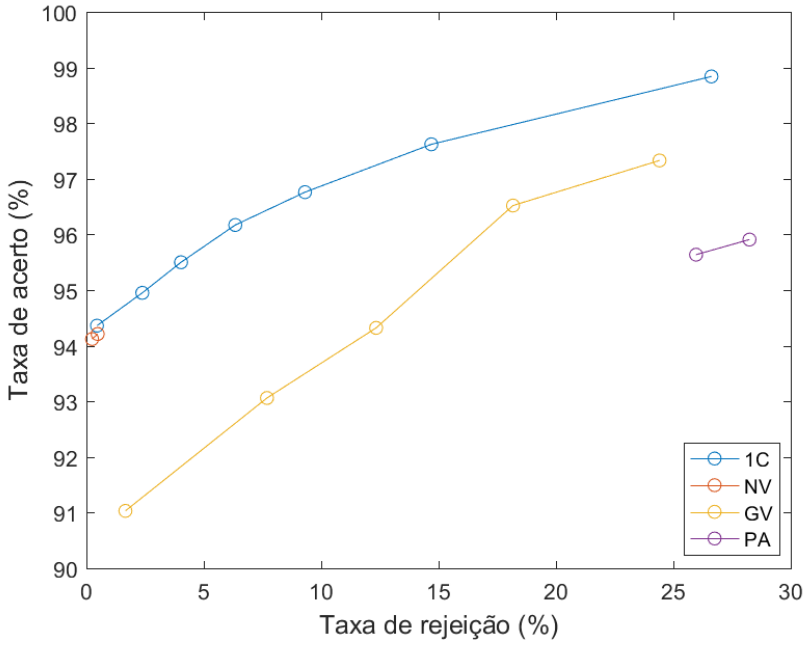
Fonte: elaborada pelo autor.

Figura 18 – Curvas A-R dos métodos SVM-1C, SVM-NV, SVM-GV e SVM-PA-RO no conjunto de dados haberman.



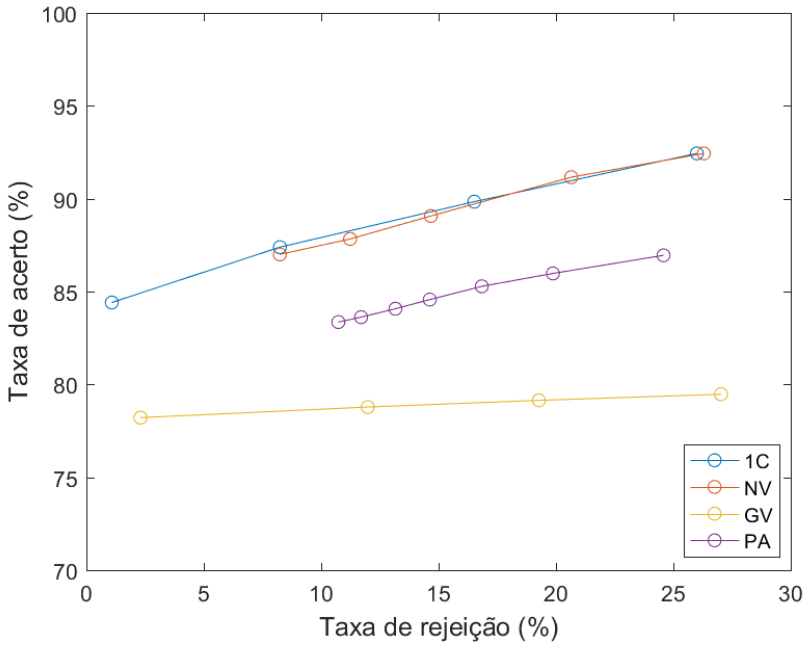
Fonte: elaborada pelo autor.

Figura 19 – Curvas A-R dos métodos SVM-1C, SVM-NV, SVM-GV e SVM-PA-RO no conjunto de dados parkinsons.



Fonte: elaborada pelo autor.

Figura 20 – Curvas A-R dos métodos SVM-1C, SVM-NV, SVM-GV e SVM-PA-RO no conjunto de dados coluna vertebral.



Fonte: elaborada pelo autor.

Tabela 3 – Quantidade média de vetores-suporte dos classificadores SVM-1C, SVM-NV, SVM-GV e SVM-PA-RO nos conjuntos de dados câncer de mama, haberman, parkinsons e coluna vertebral. Em parênteses, para o SVM-GV, encontra-se o desvio padrão da quantidade de vetores-suporte. Essa informação não é apresentada para os demais classificadores já que a variação para o SVM-PA-RO, em todos os conjuntos, foi menor do que um e essa variação é inexistente (i.e., zero) para os classificadores SVM-1C e SVM-NV.

Classificador	**Câncer de mama**	**Haberman**	**Parkinsons**	**Coluna vertebral**
SVM-1C e SVM-NV	79,58	120,28	76,92	99,18
SVM-GV	139,66 ($\pm 77,61$)	154,89 ($\pm 7,91$)	58,34 ($\pm 3,13$)	106,03 ($\pm 8,84$)
SVM-PA-RO	21,10	18,09	34,48	27,98

Fonte: elaborada pelo autor.

Analisando as Figuras 17, 18, 19 e 20, pode-se observar que o SVM-1C teve o melhor desempenho, dentre os métodos analisados, nos conjuntos parkinsons e coluna vertebral, e também apresentou bons resultados no conjunto câncer de mama. À parte do conjunto parkinsons, o desempenho do SVM-NV foi similar ao do SVM-1C. No conjunto haberman, tanto o SVM-1C quanto o SVM-NV apresentaram apenas um resultado com taxa de rejeição menor do que 30%.

Em relação aos métodos com opção de rejeição embutida, percebe-se que o SVM-

GV apresentou bons resultados nos conjuntos câncer de mama e parkinsons, enquanto que seu desempenho não foi bom no conjunto coluna vertebral. Já o SVM-PA-RO apresentou a melhor relação acurácia $\times$ taxa de rejeição no conjunto haberman. Além disso, para valores baixos de rejeição, seu desempenho foi similar ao dos demais métodos no conjunto câncer de mama. No conjunto coluna vertebral, seu desempenho foi melhor que o desempenho do SVM-GV, mas inferior ao desempenho dos métodos SVM-1C e SVM-NV. No conjunto parkinsons, o SVM-PA-RO apresentou apenas dois resultados com taxa de rejeição menor do que 30%.

Na Tabela 3 pode-se verificar que a quantidade de vetores-suporte obtida pelo SVM-PA-RO é bem menor do que àquela obtida pelos outros classificadores, principalmente em relação ao SVM-GV. Como mencionado anteriormente, uma quantidade reduzida de vetores-suporte torna o processo de classificação de novos elementos mais eficiente.

5 CONSIDERAÇÕES FINAIS

Neste trabalho foram propostos dois modelos, baseados na primeira formulação de uma SVM, que incorporam um método de otimização quadrática nebulosa para tratar dados incertos e/ou vagos. O segundo modelo proposto difere do primeiro por incorporar uma opção de rejeição, muito usada em problemas onde o custo de se cometer erros é elevado, como por exemplo na classificação de dados médicos. Ademais, os modelos propostos foram comparados com outros classificadores SVM.

5.1 Conclusões sobre as variantes propostas

A versão proposta sem opção de rejeição, SVM-PA, foi comparada com o classificador SVM-SM. Em relação a taxa de acerto, o desempenho dos métodos foi similar, salvo em um dos conjuntos testados onde a acurácia do SVM-SM foi razoavelmente melhor. Porém, a quantidade de vetores-suporte obtida pelo SVM-PA nos quatro conjuntos de dados testados foi menor do que a obtida pelo SVM-SM. Uma quantidade reduzida de vetores-suporte é desejado pois acelera o processo de classificação de novos padrões e necessita de menos espaço para armazenamento, o que é de fundamental importância em conjuntos com grande quantidade de dados.

A versão proposta com opção de rejeição, SVM-PA-RO, foi comparada com três classificadores SVM com opção de rejeição. Apesar do SVM-PA-RO apresentar melhores taxas de acerto que o SVM-PA, o que é esperado para um método com opção de rejeição, em três dos quatro conjuntos de dados usados neste trabalho seu desempenho foi um pouco inferior ao outros classificadores com essa característica. Contudo, a quantidade de vetores-suporte obtida pelo SVM-PA-RO foi consideravelmente menor do que a dos outros classificadores com opção de rejeição.

5.2 Trabalhos futuros

Como trabalhos futuros pode-se citar o desenvolvimento de algoritmos eficientes para encontrar a solução ótima do SVM-PA, de forma a viabilizar a utilização dos modelos propostos em problemas de grande porte.

Além disso, como as violações máximas tem um papel muito importante nos modelos propostos, a implementação de métodos que façam boas estimativas para esses valores pode

melhorar consideravelmente a acurácia do SVM-PA.

Ademais, pretende-se generalizar o modelo proposto para conjuntos de dados com *n*-classes.

REFERÊNCIAS

ABE, S. Fuzzy support vector machines for multilabel classification. **Pattern Recognition**, Elsevier, v. 48, n. 6, p. 2110–2117, 2015.

AN, W.; LIANG, M. Fuzzy support vector machine based on within-class scatter for classification problems with outliers or noises. **Neurocomputing**, Elsevier, v. 110, p. 101–110, 2013.

BARROS, L. C. de; BASSANEZI, R. C. **Tópicos de lógica fuzzy e biomatemática**. [S.l.]: Grupo de Biomatemática, Instituto de Matemática, Estatística e Computação Científica (IMECC), Universidade Estadual de Campinas (UNICAMP), 2010.

BARTLETT, P. L.; WEGKAMP, M. H. Classification with a reject option using a hinge loss. **Journal of Machine Learning Research**, v. 9, n. Aug, p. 1823–1840, 2008.

BERGER, J. O. **Statistical decision theory and Bayesian analysis**. New York, NY, USA: Springer Science & Business Media, 2013.

BISHOP, C. M. **Pattern Recognition and Machine Learning (Information Science and Statistics)**. Berlin, Heidelberg: Springer, 2011.

BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. A training algorithm for optimal margin classifiers. In: **Proceedings of the Fifth Annual Workshop on Computational Learning Theory**. New York, NY, USA: ACM, 1992. (COLT '92), p. 144–152.

BURGES, C. J. A tutorial on support vector machines for pattern recognition. **Data Mining and Knowledge Discovery**, v. 2, n. 2, p. 121–167, Jun 1998.

BYVATOV, E.; SCHNEIDER, G. Support vector machine applications in bioinformatics. **Applied bioinformatics**, v. 2, n. 2, p. 67–77, 2003.

CHAPELLE, O.; HAFFNER, P.; VAPNIK, V. N. Support vector machines for histogram-based image classification. **IEEE transactions on Neural Networks**, IEEE, v. 10, n. 5, p. 1055–1064, 1999.

CHOW, C. On optimum recognition error and reject tradeoff. **IEEE Transactions on information theory**, IEEE, v. 16, n. 1, p. 41–46, 1970.

CHOW, C.-K. An optimum character recognition system using decision functions. **IRE Transactions on Electronic Computers**, IEEE, n. 4, p. 247–254, 1957.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, v. 20, n. 3, p. 273–297, Sep 1995.

CRUZ, C.; SILVA, R. C.; VERDEGAY, J. L. Extending and relating different approaches for solving fuzzy quadratic problems. **Fuzzy Optimization and Decision Making**, Springer, v. 10, n. 3, p. 193–210, 2011.

DÉNIZ, O.; CASTRILLON, M.; HERNÁNDEZ, M. Face recognition using independent component analysis and support vector machines. **Pattern recognition letters**, Elsevier, v. 24, n. 13, p. 2153–2157, 2003.

DUA, D.; GRAFF, C. **UCI Machine Learning Repository**. 2017. Disponível em: <http://archive.ics.uci.edu/ml>.

DUMAIS, S.; PLATT, J.; HECKERMAN, D.; SAHAMI, M. Inductive learning algorithms and representations for text categorization. In: **Proceedings of the Seventh International Conference on Information and Knowledge Management**. New York, NY, USA: ACM, 1998. (CIKM '98), p. 148–155.

FOODY, G. M.; MATHUR, A. The use of small training sets containing mixed pixels for accurate hard image classification: Training on mixed spectral responses for classification by a svm. **Remote Sensing of Environment**, Elsevier, v. 103, n. 2, p. 179–189, 2006.

FUMERA, G.; ROLI, F. Support vector machines with embedded reject option. In: SPRINGER. **International Workshop on Support Vector Machines**. Berlin, Heidelberg, 2002. p. 68–82.

GRANDVALET, Y.; RAKOTOMAMONJY, A.; KESHET, J.; CANU, S. Support vector machines with a reject option. In: **Advances in Neural Information Processing Systems 21**. [S.l.]: Curran Associates, Inc., 2009. p. 537–544.

GROTHER, P. J. **NIST Handprinted Forms and Characters, NIST Special Database 19.** [S.l.]: National Institute of Standards and Technology, 1995.

HABERMAN, S. J. Generalized residuals for log-linear models. In: **Proceedings of the 9th international biometrics conference**. Boston, MA, USA: Soc., 1976. p. 104–122.

HAJILOO, M.; RABIEE, H. R.; ANOOSHAHPOUR, M. Fuzzy support vector machine: an efficient rule-based classification technique for microarrays. **BMC bioinformatics**, BioMed Central, v. 14, n. 13, p. S4, 2013.

HAO, Y.-Y.; CHI, Z.-X.; YAN, D.-Q. Fuzzy support vector machine based on vague sets for credit assessment. In: IEEE. **Fourth International Conference on Fuzzy Systems and Knowledge Discovery, 2007 (FSKD 2007)**. Haikou, Hainan, China, 2007. v. 1, p. 603–607.

HAYKIN, S. **Neural networks: a comprehensive foundation**. [S.l.]: Prentice Hall PTR, 1994.

HEISELE, B.; HO, P.; POGGIO, T. Face recognition with support vector machines: Global versus component-based approach. In: IEEE. **Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on**. [S.l.], 2001. v. 2, p. 688–694.

INOUE, T.; ABE, S. Fuzzy support vector machines for pattern classification. In: IEEE. **Neural Networks, 2001. Proceedings. IJCNN'01. International Joint Conference on**. [S.l.], 2001. v. 2, p. 1449–1454.

JOACHIMS, T. Text categorization with support vector machines: Learning with many relevant features. **Machine learning: ECML-98**, Springer, p. 137–142, 1998.

KARUSH, W. Minima of functions of several variables with inequalities as side constraints. **Master's thesis**, 01 1939.

KUHN, H. W.; TUCKER, A. W. Nonlinear programming. In: **Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability**. Berkeley, Calif.: University of California Press, 1951. p. 481–492. Disponível em: <https://projecteuclid.org/euclid.bsmsp/1200500249>.

KWOK, J.-Y. Moderating the outputs of support vector machine classifiers. In: IEEE. **IJCNN'99. International Joint Conference on Neural Networks. Proceedings (Cat. No. 99CH36339)**. [S.l.], 1999. v. 2, p. 943–948.

LIN, C.-F.; WANG, S.-D. Fuzzy support vector machines. **IEEE Transactions on Neural Networks**, v. 13, n. 2, p. 464–471, Mar 2002. ISSN 1045-9227.

LITTLE, M. A.; MCSHARRY, P. E.; HUNTER, E. J.; SPIELMAN, J.; RAMIG, L. O. *et al.* Suitability of dysphonia measurements for telemonitoring of parkinson's disease. **IEEE transactions on biomedical engineering**, IEEE, v. 56, n. 4, p. 1015–1022, 2009.

MANGASARIAN, O. L. Cancer diagnosis via linear programming. **SIAM news**, v. 23, n. 5, p. 1–18, 1990.

MERCER, J.; FORSYTH, A. R. Xvi. functions of positive and negative type, and their connection the theory of integral equations. **Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences**, The Royal Society, v. 209, n. 441-458, p. 415–446, 1909. ISSN 0264-3952.

MIKA, S.; RATSCH, G.; WESTON, J.; SCHOLKOPF, B.; MULLERS, K.-R. Fisher discriminant analysis with kernels. In: IEEE. **Neural networks for signal processing IX: Proceedings of the 1999 IEEE signal processing society workshop (cat. no. 98th8468)**. [S.l.], 1999. p. 41–48.

MUKHERJEE, S.; TAMAYO, P.; SLONIM, D.; VERRI, A.; GOLUB, T.; MESIROV, J.; POGGIO, T. Support vector machine classification of microarray data. AI Memo 1677, Massachusetts Institute of Technology, 1999.

OSUNA, E.; FREUND, R.; GIROSIT, F. Training support vector machines: an application to face detection. In: **Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 1997. v. 97, n. 130-136, p. 99.

PLATT, J. *et al.* Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. **Advances in large margin classifiers**, Cambridge, MA, v. 10, n. 3, p. 61–74, 1999.

PLATT, J. C. Advances in kernel methods. In: SCHöLKOPF, B.; BURGES, C. J. C.; SMOLA, A. J. (Ed.). Cambridge, MA, USA: MIT Press, 1999. cap. Fast Training of Support Vector Machines Using Sequential Minimal Optimization, p. 185–208.

QUINLAN, J. R. Induction of decision trees. **Machine learning**, Springer, v. 1, n. 1, p. 81–106, 1986.

ROCHA NETO, A.; BARRETO, G. Wci 04 on the application of ensembles of classifiers to the diagnosis of pathologies of the vertebral column: A comparative analysis. **IEEE Latin America Transactions**, v. 7, p. 487–496, 09 2009.

SINGH, H.; GUPTA, M. M.; MEITZLER, T.; HOU, Z.-G.; GARG, K. K.; SOLO, A. M.; ZADEH, L. A. Real-life applications of fuzzy logic. **Advances in Fuzzy Systems**, Hindawi Publishing Corporation, v. 2013, 2013.

SMOLA, A. J.; SCHÖLKOPF, B. A tutorial on support vector regression. **Statistics and Computing**, v. 14, n. 3, p. 199–222, Aug 2004.

VAPNIK, V. **Estimation of dependences based on empirical data**. [S.l.]: Springer Science & Business Media, 2006.

WANG, Y.; WANG, S.; LAI, K. K. A new fuzzy support vector machine to evaluate credit risk. **IEEE Transactions on Fuzzy Systems**, IEEE, v. 13, n. 6, p. 820–831, 2005.

WEBB, A. R. **Statistical pattern recognition**. West Sussex, England; New Jersey: John Wiley & Sons, 2003.

WHITNEY, C. R. Jeanne calment, world's elder, dies at 122. 1997. Disponível em: <https://www.nytimes.com/1997/08/05/world/jeanne-calment-world-s-elder-dies-at-122.html>.

ZADEH, L. Fuzzy sets. **Information and Control**, v. 8, n. 3, p. 338 – 353, 1965. ISSN 0019-9958.

APÊNDICE A – LÓGICA NEBULOSA

A Teoria dos Conjuntos Nebulosos foi introduzida em 1965 pelo matemático Lotfi Asker Zadeh, com a publicação do artigo "*Fuzzy Sets*" (ZADEH, 1965), sendo sua principal intenção dar um tratamento matemático a certos termos linguísticos subjetivos (BARROS; BASSANEZI, 2010).

Na Seção A.1 são apresentados alguns conceitos e definições básicos da Teoria de Conjuntos Nebulosos. Além disso, na Seção A.2 é apresentado um método paramétrico próprio para resolver problemas de programação quadrática de natureza nebulosa.

A.1 Definições e conceitos básicos

Um conjunto nada mais é do que um coleção de elementos. Um conjunto clássico e seus elementos possuem uma relação binária de pertinência, i.e., um elemento qualquer pertence ou não ao conjunto em questão. Essa relação pode ser caracterizada por uma função, chamada de função característica, cuja definição é:

Definição A.1 *Seja $\tilde{A}$ um subconjunto do conjunto universo U. A função característica de $\tilde{A}$, $\chi_{\tilde{A}}(x) : U \to \{0,\ 1\}$, é dada por*

$$\chi_{\tilde{A}}(x) = \begin{cases} 1, & \text{se } x \in \tilde{A} \\ 0, & \text{se } x \notin \tilde{A} \end{cases}.$$

A definição de um subconjunto nebuloso é obtida ampliando-se o contra-domínio da função característica para o intervalo $[0,1]$:

Definição A.2 *Seja U um conjunto clássico. Um subconjunto nebuloso $\tilde{A}$ de U é um conjunto de pares ordenados:*

$A = \{(x, \chi_A(x)) \text{ tal que } x \in U\}$

onde $\chi_{\tilde{A}}$ é conhecida como função de pertinência de x em A e $\chi_{\tilde{A}} : U \to [0,1]$.

O valor da função $\chi_{\tilde{A}}(x)$ indica o grau com que o elemento $x \in U$ pertence ao conjunto A. Esse valor é conhecido como grau de pertinência. A exclusão completa de um elemento x ao conjunto nebuloso A é descrito por $\chi_{\tilde{A}}(x) = 0$, enquanto que a pertinência completa é descrita por $\chi_{\tilde{A}}(x) = 1$.

Diversas funções podem ser utilizadas para expressar um conjunto nebuloso, tais como função triangular, função trapezoidal, função gaussiana, dentre outras (Pedrycz e Gomide 98). Por exemplo, a função triangular pode ser definida como:

$$\chi_{\tilde{A}}(x) = \begin{cases} \frac{x-\underline{a}}{a-\underline{a}}, & \text{se } x \in [\underline{a}, a] \\ \frac{\bar{a}-x}{\bar{a}-a}, & \text{se } x \in [a, \bar{a}] \\ 0, & \text{caso contrário,} \end{cases}$$

sendo que a é o valor modal e $\underline{a}$ e $\bar{a}$ são os limitantes inferior e superior, respectivamente.

A seguir são apresentadas algumas definições que também podem ser utilizadas para caracterizar subconjunto nebulosos.

Definição A.3 *Define-se como suporte de $\tilde{A}$ o subconjunto clássico definido por*

$$S_{\tilde{A}} = \{x \in U : \chi_{\tilde{A}}(x) > 0\}.$$

Definição A.4 *Um subconjunto nebuloso $\tilde{A}$ é convexo se*

$$\chi_{\tilde{A}}(\lambda x_1 + (1-\lambda)x_2) \geq \lambda \chi_{\tilde{A}}(x_1) + (1-\lambda)\chi_{\tilde{A}}(x_2), \tag{A.1}$$

para todo $x_1,\ x_2 \in S_{\tilde{A}}$ e $\lambda \in [0,1]$.

Definição A.5 *Seja $\tilde{A}$ um subconjunto nebuloso de U e $\alpha \in [0,1]$. O α-corte de $\tilde{A}$ é o subconjunto clássico de U definido por*

$$\tilde{A}_\alpha = \{x \in U : \chi_{\tilde{A}}(x) \geq \alpha\},\ \alpha \in [0,1]. \tag{A.2}$$

A.2 Programação quadrática nebulosa

A programação quadrática representa uma classe especial da programação não-linear, a qual pode ser vista como uma generalização da programação linear. Problemas dessa classe possuem função objetivo quadrática e restrições lineares, sendo formalizado como

$$\begin{aligned} \min \quad & c^\intercal x + \frac{1}{2} x^\intercal Q x \\ \text{s.a.} \quad & Ax \leq b \\ & x \geq 0, \end{aligned} \tag{A.3}$$

sendo $c \in \mathbb{R}^p$, $b \in \mathbb{R}^\ell$, $A \in \mathbb{R}^{\ell \times p}$ e $Q \in \mathbb{R}^{p \times p}$ uma matriz simétrica.

Problemas de programação quadrática nebulosa são uma generalização dos problemas de programação quadrática usuais, onde incertezas nebulosas podem aparecer na função objetivo, nas restrições de igualdade e/ou nas restrições de desigualdade. Aqui é considerado apenas o caso onde as incertezas nebulosas aparecem no conjunto de restrições.

Um problema de programação quadrática com relação de ordem nebulosa no conjunto de restrições pode ser formalizado como

$$\begin{aligned} \min \quad & c^\intercal x + \frac{1}{2} x^\intercal Q x \\ \text{s.a.} \quad & Ax \leq^f b \\ & x \geq 0, \end{aligned} \tag{A.4}$$

sendo que o símbolo $\leq^f$ quer dizer que essas restrições podem ser satisfeitas com algumas violações. As funções de pertinência $\chi_i : \mathbb{R}^\ell \to [0,1]$, $i = 1, \ldots, \ell$, definem o quanto cada restrição pode ser violada.

Denotando-se cada restrição $\sum_{j=1}^{p} a_{ij} x_j$ por $(Ax)_i$, $i = 1, \ldots, \ell$, no Problema (A.4) pode ser reescrito como

$$\begin{aligned} \min \quad & c^\intercal x + \frac{1}{2} x^\intercal Q x \\ \text{s.a.} \quad & (Ax)_i \leq^f b_i, \qquad i = 1, \ldots, \ell \\ & x \geq 0, \end{aligned} \tag{A.5}$$

sendo as funções de pertinência

$$\chi_i : \mathbb{R}^p \to [0,1], \quad i = 1, \ldots, \ell \tag{A.6}$$

nas restrições nebulosas devem ser definidas por um especialista. Cada função de pertinência indica o grau de pertinência com o qual qualquer $x \in \mathbb{R}^p$ satisfaz a restrição nebulosa correspondente. Esse grau é 1 quando a restrição é satisfeita sem violações e, a medida que tende a zero, violações cada vez maiores são permitidas. No caso linear, essas funções de pertinência podem ser definidas como

$$\chi_i(x) = \begin{cases} 1, & \text{se } (Ax)_i \leq b_i \\ 1 - \frac{(Ax)_i - b_i}{d_i}, & \text{se } b_i \leq (Ax)_i \leq b_i + d_i \\ 0, & \text{se } (Ax)_i > b_i + d_i \end{cases},$$

sendo d_i a violação máxima permitida para a i-ésima restrição, $i = 1, \ldots, \ell$.

Em Cruz *et al.* (2011), os autores desenvolveram um método de duas etapas para resolver o Problema (A.4). A primeira etapa consiste em parametrizar o Problema (A.4). Para isso, primeiro define-se, para cada restrição nebulosa, um conjunto com os vetores $0 \leq x \in \mathbb{R}^p$ que satisfazem essa restrição

$$X_i = \left\{ x \in \mathbb{R}^p \mid (Ax)_i \leq^f b_i,\ x \geq 0 \right\}. \tag{A.7}$$

Denotando por X a intersecção dos conjuntos X_i, $i = 1, \ldots, \ell$, o Problema (A.4) pode ser reescrito de forma compacta como

$$\min \left\{ c^\intercal x + \frac{1}{2} x^\intercal Q x \mid x \in X \right\}. \tag{A.8}$$

Para cada $\alpha \in (0,1]$, um α-corte do conjunto de restrições nebulosa será o conjunto clássico

$$X(\alpha) = \{ x \in \mathbb{R}^p \mid \chi_X(x) \geq \alpha \} \tag{A.9}$$

sendo que para todo $x \in \mathbb{R}^p$

$$\chi_X(x) = \min_{i=1,\ldots,\ell} \chi_i(x). \tag{A.10}$$

Assim, se para todo $\alpha \in (0,1]$

$$S(\alpha) = \left\{ x \in \mathbb{R}^P \mid c^\intercal x + \frac{1}{2} x^\intercal Q x = \min c^\intercal z + \frac{1}{2} z^\intercal Q z,\ z \in X(\alpha) \right\} \tag{A.11}$$

a solução do problema será o conjunto nebuloso definido pela seguinte função de pertinência

$$\chi_S(x) = \begin{cases} \sup \{ \alpha \ :\ x \in S(\alpha) \}, & \text{se } x \in \bigcup_\alpha S(\alpha), \\ 0, & \text{caso contrário.} \end{cases} \tag{A.12}$$

Note que para todo $\alpha \in (0,1]$,

$$X(\alpha) = \bigcap_{i=1}^{\ell} \{ x \in \mathbb{R}^p \mid (Ax)_i \leq b_i + d_i(1-\alpha),\ 0 \leq x \in \mathbb{R}^p \}. \tag{A.13}$$

Assim, a solução operacional do Problema (A.8) pode ser encontrada, α-corte por α-corte, através da solução do seguinte problema de programação quadrática paramétrico auxiliar

$$\begin{aligned} \min \quad & c^\intercal x + \frac{1}{2} x^\intercal Q x \\ \text{s.a.} \quad & (Ax)_i \leq b_i + d_i(1-\alpha), \qquad i = 1, \ldots, \ell \\ & x \geq 0 \\ & \alpha \in [0,1] \end{aligned} \tag{A.14}$$

A segunda etapa do método é resolver, com técnicas convencionais de programação quadrática, o Problema (A.14) para cada valor de α.

APÊNDICE B – ESTIMAÇÃO DAS VIOLAÇÕES MÁXIMAS DO SVM-PA

Seja $T = \{(x_i, y_i) \mid i = 1, \ldots, \ell\}$ um conjunto de dados tal que $x_i \in \mathbb{R}^p$ é o i-ésimo vetor de atributos e $y_i \in \{-1, 1\}$ representa a classe de x_i.

Sejam $m(x,\ C)$ a média das distâncias, no espaço de características, usando como base a distância euclidiana, de x até os 5% elementos mais próximos da sua classe e $m(x,\ O)$ a média das distâncias até os 5% elementos mais próximos da classe oposta. Além disso, seja dm_1 (dm_{-1}) a distância média dos elementos da classe positiva (negativa) até o elemento central dessa classe.

O conjunto de dados, T, é dividido em três subconjuntos disjuntos:

1. $\mathfrak{I}$: conjunto composto pelos elementos de T tais que $m(x,\ C) > m(x, O)$;
2. $\mathfrak{F}$: um elemento da classe positiva (negativa) entra para a fronteira sua distância para, pelo menos, 25 % dos elementos da classe oposta for menor do que dm_1 (dm_{-1}).
3. $\mathfrak{R}$: demais elementos.

Definidos esses subconjuntos, as violações máximas são obtidas da seguinte forma:

- Para elementos dos conjuntos $\mathfrak{I}$ e $\mathfrak{F}$, a violação máxima é definida como sendo a média das distâncias de x até os 25 % elementos mais próximos de sua classe que pertencem à $\mathfrak{F}$.
- Para um elemento do conjunto $\mathfrak{R}$, a violação máxima é igual a menor violação entre os elementos de sua classe pertencentes aos conjuntos $\mathfrak{I}$ ou $\mathfrak{F}$.