UNIVERSIDADE ABERTA

Aplicação de Máquinas de Vector Suporte para classificação de ratos transgénicos através de imagem da retina

Érick Braga Valentim

Mestrado em Estatística, Matemática e Computação

2020

UNIVERSIDADE ABERTA

Aplicação de Máquinas de Vector Suporte para classificação de ratos transgénicos através de imagem da retina

Érick Braga Valentim

Mestrado em Estatística, Matemática e Computação

Dissertação orientada pelo Professor Doutor Pedro Miguel Picado de Carvalho Serranho, Universidade Aberta

Dissertação coorientada pelo Professor Doutor Rui Manuel Dias Cortesão dos Santos Bernardes, Universidade de Coimbra

2020

Resumo

O objetivo deste trabalho consistiu na criação de modelos de aprendizagem supervisionada baseados nas técnicas de *Support Vector Machine* (SVM) e *Support Vector Machine* com informação privilegiada (SVM+) capazes de distinguir entre ratos saudáveis (C) e transgénicos (D) por meio de análise de textura da imagem de tomografia de coerência óptica (OCT) de retinas do olho direito.

A amostra é composta por 74 ratos, sendo 40 saudáveis e 34 transgénicos.

A tomografia de coerência óptica foi utilizada para obtenção da imagem da retina dos ratos que, por sua vez, foi dividida em 4 quadrantes. A partir destes, obteve-se uma imagem de fundo 2D e foram aplicados 20 indicadores de análise de textura de imagem de fundo, usados como *features* para o modelo SVM.

As *features* com maior capacidade de separação entre grupos e que possuem coeficiente de correlação inferior a 0,7 entre elas foram *Inertia* (primeiro, segundo e quarto quadrantes), INN (*Inverse difference normalized*; terceiro quadrante), IMC2 (*Information measure of correlation*; terceiro quadrante) e *ClusterShade* (terceiro quadrante).

Considerando as 6 *features* mais relevantes foram criados os modelos SVM e SVM+ cujos parâmetros foram afinados de maneira a obter os modelos com a melhor precisão na classificação dos ratos nas categorias saudável e transgénico. A técnica de validação cruzada em 5 grupos foi utilizada para validar os resultados dos modelos criados.

Tanto para o conjunto de teste como para o conjunto de dados total o modelo SVM obteve 100% precisão, enquanto que a precisão obtida pelo modelo SVM+ foi de 93,33% (erro de apenas 1 caso em 15 – conjunto de teste) na classificação dos dados do conjunto de teste e 98,65% (erro de apenas 1 caso em 74 – conjunto de dados total) no conjunto de dados total.

Palavras-chave: modelos de aprendizagem supervisionada, SVM, SVM+, validação cruzada, *features*

Abstract

The aim of this work was to create supervised learning models based on the Support Vector Machine (SVM) and Support Vector Machine with privileged information (SVM +) capable of distinguishing between healthy (C) and transgenic (D) mice through texture analysis of the optical coherence tomography (OCT) image of the retinas of the right eye.

The sample consists of 74 mice, 40 healthy and 34 transgenic.

Optical coherence tomography was used to obtain the image of the mice's retina, which in turn was divided into 4 quadrants. From these, a 2D background image was obtained and 20 background image texture analysis indicators were applied, used as features for the SVM model.

The features with greater separation capacity between groups and which have a less than 0.7 correlation coefficient between each other were Inertia (first, second and fourth quadrants), INN (*Inverse difference normalized*; third quadrant), IMC2 (*Information measure of correlation*; third quadrant) and ClusterShade (third quadrant).

Regarding the 6 most relevant features, the SVM and SVM + models were created, whose parameters were adjusted in order to obtain the models with the best precision in the classification of mice in the healthy and transgenic categories. The 5 fold cross-validation technique was used to validate the results of the models created.

For both, the test set and the total data set, the SVM model obtained 100% accuracy, while the precision obtained by the SVM + model was 93.33% (error of only 1 case in 15 - test set) in the classification of the test set data and 98.65% (error of only 1 case in 74 - total data set) in the total data set.

Keywords: supervised learning models, SVM, SVM+, cross-validation, features

Dedicatória

Dedico este trabalho à minha amada mãe, Irismar Braga. Seus valores, esforço e dedicação, desde o meu nascimento, trouxeram-me até aqui e são fonte de inspiração para que eu possa atingir todas as minhas metas.

Agradecimentos

A Deus, criador do mundo e tudo o que nele há. Sem a ajuda Dele este trabalho não seria realizado.

À minha família, fonte de motivação e que sempre proveu o apoio necessário para atingir meus objetivos.

Ao excelente Professor Doutor Pedro Serranho pela orientação, conselhos, paciência e sabedoria na condução deste trabalho.

Ao Professor Doutor Rui Bernardes, coorientador desta dissertação, pela disponibilização da base de dados utilizada e discussão do trabalho.

Aos professores da UAb pelos conhecimentos transmitidos, em especial aos professores Doutores Pedro Serranho, Catarina Nunes e Maria do Rosário, pela didática simples, completa e clara.

Ao amigo Tomas Ferreira, sempre disposto a ajudar e que muito me auxiliou neste Mestrado.

Aos colegas do MEMC, que enriqueciam as discussões nos fóruns de dúvidas.

A presente dissertação é feita no âmbito do projeto de investigação PTDC/EMD-EMD/28039/2017, financiado pela Fundação para a Ciência e Tecnologia (FCT), Fundo Europeu de Desenvolvimento Regional (FEDER) e Orçamento de Estado de Portugal (OE).

Índice

Lista de tabelas

Lista de figuras

Lista de abreviaturas

2D – Duas dimensões

3D – Três dimensões

AA – Aprendizagem automática

FCT – Fundação para a Ciência e Tecnologia

FEDER – Fundo Europeu de Desenvolvimento Regional

GCL - *Ganglion cell laye*r

IDN - *Inverse difference moment normalized*

IMC – *Information measure of correlation*

INN – *Inverse difference normalized*

IPL - *Inner plexiform layer*

KKT - Karush-Kuhn-Tucker

LUPI – *Learning Using Privileged Information*

MCP - McCulloch-Pitts

MEMC – Mestrado em Estatística, Matemática e Computação

OCT - *Optical Coherence Tomography*

OE – Orçamento de Estado

RBF - *Radial Basis Function*

RNA – Rede neural artificial

RNFL - *Retinal nerve fiber layer*

SVM – *Support Vector Machine*

SVM+ - *Support Vector Machine* com informação privilegiada

UAb – Universidade Aberta

1. Introdução

Devido ao avanço da tecnologia, tem-se a cada dia uma quantidade maior de dados produzida. Entretanto, não basta apenas ter os dados. É necessário transformá-los em informação útil. Em muitos casos, tal transformação passa pela classificação de dados em classes, de preferência automaticamente.

Para os casos em que as classes dos dados são conhecidas, algoritmos de Aprendizagem Automática (AA) surgem como boa alternativa para a obtenção de um classificador que, com base nos dados conhecidos, classifique corretamente novos dados em suas respetivas classes (Mitchel, 1997).

As técnicas de AA utilizadas neste trabalho foram as de *Support Vector Machine* (SVM) e SVM com informação privilegiada (SVM+). Conforme exposto por Cristianini e Shawe-Taylor (2000) e por Phangtriastu et al. (2017), o estudo e aplicação da técnica de SVM aumentaram devido aos seus bons resultados em problemas de classificação.

O presente estudo versa sobre a classificação de ratos nas classes saudável e transgénico. Assim, trata-se de um problema de classificação binária. De maneira semelhante, também há problemas relacionados à classificação de dados em mais de duas classes, ou seja, multiclasses, os quais não são abordados nesta dissertação.

Por meio dos dados relativos às características da retina dos ratos saudáveis e transgénicos, desejou-se saber quais os atributos calculados que mais influenciavam na correta classificação dos ratos nas classes saudável e transgénico através do uso da técnica de classificação SVM.

Tal técnica cria um modelo de classificação utilizando um conjunto de treino, retirado da população, e o aplica na classificação dos dados num conjunto de teste (que não se interseta com o conjunto de treino). Desta forma, almeja-se que o modelo obtido tenha a capacidade de predizer, com grande precisão, a classificação de novos dados cujos rótulos são desconhecidos, porém que pertencem ao domínio dos rótulos dos dados originais.

O objetivo deste trabalho consistiu na criação de modelos de aprendizagem automática supervisionada baseados nas técnicas de SVM e SVM+ capazes de distinguir

entre ratos saudáveis (C, de controlo) e transgénicos (D, de doente) por meio da análise de textura das imagens obtidas a partir de tomografia de coerência óptica (OCT).

De maneira a separar os temas tratados, o presente trabalho foi dividido em cinco secções: introdução, revisão bibliográfica sobre as técnicas de SVM e SVM+, metodologia, resultados e conclusão.

A segunda secção é responsável por introduzir os conceitos de SVM, SVM linearmente separável, escolha do melhor hiperplano separador, definição de SVM com margens suaves, SVM não linearmente separáveis, Kernel Trick, tipos de Kernel e uma explanação sobre o SVM com informação privilegiada.

A terceira secção dispõe sobre a metodologia utilizada para o atingimento do objetivo desta dissertação. Nela, constam a caracterização da amostra, a maneira pela qual as imagens foram obtidas por meio da técnica de imagiologia por tomografia de coerência óptica, quais os vinte indicadores de análise de textura de imagem de fundo aplicados e que foram usados como *features* para o modelo SVM e, terminando a secção de metodologia, explica-se o procedimento de aplicação dos testes estatísticos que suportaram a escolha das *features* mais relevantes para o SVM.

Na quarta secção são apresentados os resultados do presente estudo. Como ponto de partida, apresenta-se a listagem das *features* escolhidas como mais relevantes para o modelo SVM, seguida da descrição dos dados, da técnica de validação cruzada, dos resultados advindos da aplicação da técnica de SVM e, finalmente, dos resultados obtidos por meio da aplicação do SVM com informação privilegiada.

Por fim, a quinta secção apresenta as conclusões alcançadas nesta dissertação bem como inclui uma sugestão para trabalhos futuros referentes à matéria.

2. Revisão Bibliográfica

A Aprendizagem Automática (AA) pode ser explicada como a capacidade de um algoritmo aprender sem ser, de facto, programado para tal. Seguindo este raciocínio, podem ser construídos algoritmos que, por meio da inserção de novos dados, aprendam e produzam melhores resultados.

Considerando-se o grande volume de dados atualmente produzido, técnicas de AA podem ser utilizadas para identificar e classificar tais informações. Nesse sentido, podem ser usadas diferentes tipos de técnicas de AA para realizar a classificação dos dados, como por exemplo SVM, cujo estudo e aplicação seguem em alta devido aos bons resultados em problemas de classificação (Phangtriastu et al., 2017).

No que tange às redes neurais, atribui-se a McCulloch & Pitts (1943) o pioneirismo das pesquisas relacionadas a esta área. Em 1943, os autores apresentaram um cálculo em termos de lógica matemática para representar a atividade referente aos neurónios. Este modelo ficou conhecido como neurónio MCP (McCulloch-Pitts) e é definido como um conjunto de *n* entradas em que cada entrada é multiplicada por um determinado peso e, em seguida, os resultados são somados de forma ponderada (considerando os valores nos neurónios de entrada e os pesos de cada ligação) e determinados os valores num novo nível. Nesse nível é repetido o processo, e assim sucessivamente nos vários níveis considerados, até ser obtido um valor de saída. Haykin (2001) definiu rede neural como um processador maciço, paralelamente distribuído, cuja constituição envolve unidades de processamento simples com propensão natural para armazenar conhecimento experimental e que pode ser disponibilizado ao uso. Concluindo a definição, Haykin comparou a rede neural ao cérebro humano pelo fato da rede adquirir conhecimento proveniente de um ambiente e, valendo-se de um processo de aprendizagem, produzir resultado. Assim, uma rede neural artificial (RNA) seria uma rede neural simulada que aprenderia e erraria cada vez menos a partir da inclusão de novos dados.

Outra importante contribuição para a área de redes neurais foi dada pelo psicólogo Donald Hebb (1949), que demonstrou como as redes neurais poderiam aprender por meio da variação dos pesos de entrada neuronais.

Em 1958, Rosenblatt (1958) publicou um estudo sobre uma RNA simples, chamada Perceptron, cujo objetivo era demonstrar uma nova maneira de lidar com os problemas relacionados ao reconhecimento de padrões. Rosenblatt utilizou o nome Perceptron para descrever um sistema nervoso hipotético, com o intuito de fazer uma analogia entre os sistemas biológicos naturais e o Perceptron. Assim, pode-se dizer que o Perceptron foi desenvolvido para simular, de maneira geral, algumas propriedades de sistemas inteligentes. Rosenblatt notou que a utilização de linguagem de símbolos lógicos e álgebra booleana, empregada em estudos anteriores por McCulloch & Pitts (1943) e Minsky (1956), por exemplo, não era a mais adequada para lidar com eventos relativos a sistemas em que a estrutura exata não se conhece, ou seja, possuem apenas uma organização em linhas gerais. Além disso, os modelos anteriormente propostos falhavam em aspectos como ausência de equipotencialidade e excessiva especificidade de conexões. Assim, Rosenblatt optou por formular o Perceptron sob o paradigma da Teoria da probabilidade.

2.1. Support Vector Machine (SVM)

De acordo com Vapnik (1995), as Máquinas de Vectores de Suporte são técnicas de Aprendizagem Automática cuja base é proveniente da Teoria de Aprendizagem Estatística (*Statistical learning*). Segundo Mitchell (1997), o campo da AA possui como objetivo a construção de programas de computador capazes de aprender e errar cada vez menos, valendo-se da inclusão de novos dados, de forma que um computador aprendia por meio da experimentação.

Ainda de acordo com o autor, tais modelos de Aprendizagem Automáticas são construídos com o intuito de resolver os mais diversos tipos de problemas. Dentre seus usos, podem ser citados: classificação de dados em categorias, em que pode ser incluída a classificação por patologia de imagens de retina obtidas por tomografia de coerência óptica como será feito neste trabalho, reconhecimento óptico de caracteres, classificação de texto (se um e-mail é spam ou não), reconhecimento de género por meio de sinais de voz, estado de saúde de determinado paciente (com doença ou não), determinação de diferentes tipos de doenças (diferente tipos de cancro, por exemplo), detecção de expressões faciais, entre outros. Ainda de acordo com Mitchell (1997), a partir do conhecimento das classes de dados

das observações obtidas, os algoritmos de aprendizagem de máquina podem ser usados com o intuito de realizar a classificação de novas observações referentes à população alvo. Braga et al. (2000) afirmam que os resultados alcançados devido à utilização de algoritmos de AA superam, em alguns casos, os resultados obtidos por outras formas de aprendizado, tais como os obtidos por meio de Redes Neurais Artificiais (RNA). Hearst et al. (1998) exemplificam alguns tipos de utilização da técnica, como reconhecimento de padrões faciais e categorização de texto.

Haykin (1999) apresenta dois paradigmas fundamentais de aprendizagem: aprendizagem com professor (aprendizagem supervisionada) e aprendizagem sem professor (aprendizagem não supervisionada).

Ainda de acordo com Haykin (1999), na aprendizagem supervisionada o professor demonstra o seu conhecimento do ambiente por meio de conjuntos de exemplos de entrada/saída. O propósito da utilização de algoritmos de AA é classificar tão bem quanto possível os dados fornecidos pelo, fazendo com que as respostas dadas para novos elementos sejam corretas.

Por sua vez, na aprendizagem não supervisionada não há o professor para orientar o processo de aprendizagem e, desta forma, não há exemplos a serem aprendidos. Souto et al. (2003) diz que no aprendizado sem professor estão disponíveis apenas os atributos de entrada e que a referida técnica de aprendizado é utilizada quando a meta a ser alcançada for encontrar dados padrões ou algum tipo de tendência que seja útil no entendimento dos dados.

Para esta dissertação será utilizada a abordagem de aprendizagem supervisionada, nomeadamente através do uso de SVM. Pretende-se o desenvolvimento de um algoritmo para a classificação por patologia de imagens de retina obtidas por OCT de ratos saudáveis e transgénicos por meio de SVM com informação privilegiada. O objetivo será implementar um algoritmo em software R que se adapte às necessidades da aplicação específica, de forma a obter uma classificação automática das imagens por patologia a partir das características essenciais de imagens de retina. Antes da apresentação dos conceitos e metodologias inerentes ao SVM com informação privilegiada, inicia-se pela introdução do caso mais simples de SVM na secção seguinte e, então, prossegue-se a partir daí em termos de aumento de complexidade do método.

2.2. SVM linearmente separável

Seja S um conjunto de treino, composto por pares (x_i,y_i) em que x_i é um vetor no espaço de *features* e y_i é o rótulo para cada elemento x_i. De acordo com Russel e Norvig (1995), S será linearmente separável e seus classificadores serão chamados de lineares, quando existir um hiperplano capaz de separar os padrões das classes distintas contidas em *S*. Tais hiperplanos são definidos da seguinte maneira

$$f(x) = w \cdot x + b$$

em que $w \cdot x$ representa o produto escalar entre os vetores w e x, $w \in X$ é o vetor normal ao hiperplano, $b \in \mathbb{R}$, $\frac{b}{\|w\|}$ é a distância perpendicular do plano à origem e X é o espaço dos n dados $x_i \in X$ com rótulos dados por $y_i \in Y$, para $Y = \{-1,1\}$ no caso binário.

O espaço X dos dados pertencentes é dividido pelo hiperplano nas regiões $w \cdot x + b > 0$ e $w \cdot x + b < 0$, de forma que as classificações das classes de S podem ser obtidas por meio da aplicação de uma função sinal $g(x) = sgn(f(x))$, resultando em (Smola et al., 1999)

$$g(x) = sgn(f(x)) = \begin{cases} 1 \; se \; w \cdot x + b > 0 \\ -1 \; se \; w \cdot x + b < 0 \end{cases}$$

Representando graficamente a equação acima, tem-se o esquema apresentado na figura 1:

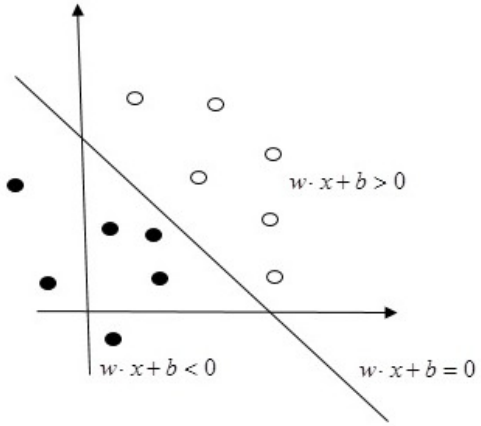

Figura 1: Esquema de hiperplano de separação para dados linearmente separáveis em duas classes.

Como conclusão, S será linearmente separável sempre que for possível obter pelo menos um $par\ (w, b)$ de forma que a função sinal acima exposta classifique, corretamente, as classes do conjunto de dados.

Para mais fácil visualização, expõe-se um caso bidimensional. Sob duas dimensões, um classificador linear é uma linha reta, conforme ilustrado na figura 1. A linha representada possui a forma $w_1 x_1 + w_2 x_2 = b$, para $(x_1, x_2) \in \mathbb{R}^2$, e as categorias são classificadas de acordo com o sinal da função $g(x) = sgn\big(f(x)\big)$ (Manning et al., 2009). Ainda de acordo com os autores, existe um número infinito de separadores lineares, uma vez que, em geral, o número de possíveis hiperplanos separadores é infinito, no caso de dados linearmente separáveis. A figura 2 retrata esta situação, apresentando várias possibilidades para hiperplanos de separação.

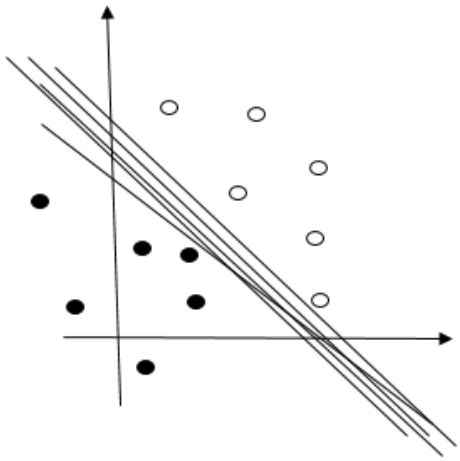

Figura 2: Exemplos de hiperplanos de separação para dados separáveis.

Estando diante de um problema linearmente separável e, usando o facto da existência de infinitos hiperplanos que separam as classes do conjunto de treino, a seguinte pergunta vem à tona: como escolher o melhor hiperplano? Isso será explorado na secção seguinte.

2.2.1. Escolha do melhor hiperplano

De acordo com Manning et al. (2009), o método de SVM possui como critério a busca por uma superfície de decisão com maior margem possível para os pontos mais próximos de cada uma das classes. Campbell (2000) define a margem do classificador como a distância perpendicular entre o hiperplano separador e o hiperplano paralelo sobre os pontos de dados mais próximos de cada classe. Desta forma, o método SVM determina a posição do separador, com base nos dados do conjunto de treino. Estes pontos mais próximos ao hiperplano são conhecidos como vectores de suporte (Campbell, 2000). O melhor hiperplano é aquele que possui a maior margem de separação entre as duas classes (Hearst et al., 1998).

Partindo da hipótese de que o conjunto de treino S é linearmente separável, $w\ e\ b$ podem ser escolhidos tais que os pontos mais próximos ao hiperplano separador definido por $w \cdot x + b = 0$ atendam a condição $|w \cdot x_i + b| = 1$ (Müller et al., 2001). Em outras palavras, trata-se de obter a representação canônica do hiperplano.

As seguintes inequações derivam da condição acima exposta e representam as duas classes do conjunto de treino

$$\begin{cases} w \cdot x_i + b \geq +1, se\ y_i = \ \ 1 \\ w \cdot x_i + b \leq -1, se\ y_i = -1 \end{cases}$$

Conforme descrito por Campbell (2000), se o conjunto de treino é separável, então os dados serão corretamente classificados se $y_i(w \cdot x_i + b) > 0, \forall i$. Ainda, a relação anterior é invariante sob uma mudança de escala positiva do argumento dentro da função sinal, o que permite definir o hiperplano canônico de forma que $w \cdot x + b = 1$ para os pontos mais próximos de um lado do hiperplano e $w \cdot x + b = -1$ para os pontos mais próximos do outro lado. Os pontos mais próximos ao hiperplano separador e que satisfazem as igualdades anteriores são conhecidos como Vectores de Suporte (*Support Vectors*). O menor caminho entre um ponto e um hiperplano é perpendicular ao plano, logo, paralelo ao vetor w. Assim, para o hiperplano separador, tem-se que um vetor unitário nesta direção é dado por $\frac{w}{\|w\|}$. Calculando-se a distância entre as retas paralelas $w \cdot x + b = 1$ e $w \cdot x + b = 0$, chega-se ao valor para a distância que separa o ponto de S mais próximo ao hiperplano

separador, conhecida como margem (Cristianini e Schölkopf, 2002), cujo valor é $\frac{1}{\|w\|}$. Esta distância também é conhecida como margem geométrica do classificador linear. A figura 3 ilustra os vectores suporte de ambas as classes:

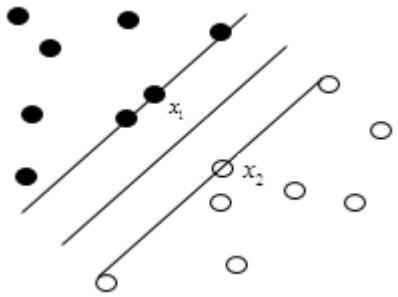

Figura 3: Exemplos de vectores de suporte de diferentes classes. (Campbell, 2000)

Como o melhor hiperplano é aquele que garante maior margem e a margem é dada por $\frac{1}{\|w\|}$, maximizar a margem significa minimizar $\|w\|$ (ou, equivalentemente, $w \cdot w$). Este problema pode ser formulado como um problema de optimização quadrática (Müller et al., 2001)

$$\underset{w,b}{\text{Minimizar}} \frac{1}{2}\|w\|^2$$

$$Sujeito\ a\ y_i(w \cdot x_i + b) - 1 \geq 0, para\ i = 1, \ldots, n$$

Lorena (2006) explica que o classificador SVM obtido pode ser chamado de SVM com margens rígidas, pois as restrições do problema apresentado são impostas de forma a garantir que não haja elementos do conjunto de treino entre as margens de separação das classes.

Burges (1998) esclarece que há duas razões para o uso da formulação Lagrangeana do problema. A primeira, deve-se ao facto de as restrições apresentadas poderem ser trocadas por restrições sobre os próprios multiplicadores de Lagrange, o que é mais fácil de tratar. A segunda, refere-se à reformulação do problema, que faz com que os dados do conjunto de

treino apareçam somente na forma de produtos internos entre vectores. Tal facto permitirá a generalização do procedimento para o caso não linear. Assim, são introduzidos multiplicadores de Lagrange não negativos $\alpha_i, i = 1, \dots, n,$ para cada uma das restrições em $y_i(w \cdot x_i + b) - 1 \geq 0, \forall i.$

O método dos multiplicadores de Lagrange foi desenvolvido por Joseph-Louis Lagrange com o objetivo de encontrar os pontos óptimos de uma função sujeita a restrições.

Resumidamente, para levar a cabo o método dos multiplicadores de Lagrange são necessários os seguintes passos:

- Construir a função lagrangeana $\mathcal{L}$ introduzindo um multiplicador para cada restrição.

- Obter o gradiente $\nabla\mathcal{L}$ da função lagrangeana.

- Resolver $\nabla\mathcal{L}(x, \alpha) = 0$

Assim, seguindo os passos anteriores, pode-se introduzir a função lagrangeana. A função objetivo do problema é dada por $f(w) = \frac{1}{2}\|w\|^2$ e as n funções de restrição são $g_i(w, b) = y_i(w \cdot x_i + b) - 1, i = 1, \dots, n.$ A função lagrangeana é dada por

$$\mathcal{L}(w, b, \alpha) = f(w) - \sum_{i=1}^{n} \alpha_i g_i(w, b)$$

que, ao efetuar as substituições, converte-se em

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i [y_i(w \cdot x_i + b) - 1].$$

O problema lagrangeano, em sua forma primal, passa pela minimização em $w\ e\ b$ e pela maximização em α. Conforme explica Kowalczyk (2017), o problema lagrangeano geralmente é resolvido usando sua forma dual. Logo, o problema que anteriormente era de minimização passa a ser de maximização. Desta forma, o problema em causa é de programação quadrática convexa, pois a função objetivo, que é uma norma, é convexa. Para demonstrar isso, tem-se o seguinte teorema.

Teorema 1. Seja $x_1, x_2 \in \mathbb{R}^n$ e $t \in [0,1]$. Seja também $f: R^n \to R^n$ uma norma. Então, f é convexa. Isto é

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2).$$

Demonstração: Como f é uma norma, pela desigualdade triangular vem

$$f(tx_1 + (1-t)x_2) \leq f(tx_1) + f\big((1-t)x_2\big).$$

Pela homogeneidade positiva de f, o lado direito da inequação acima é dado por

$$f(tx_1) + f\big((1-t)x_2\big) = tf(x_1) + (1-t)f(x_2)$$

$$0 \leq t \leq 1,$$

ou seja, f é convexa, como se queria demonstrar. □

Finalmente, de acordo com Boyd (2004), f é convexa e $\alpha \geq 0$ uma constante, então a função αf é convexa. Mais ainda, a função lagrangeana relativa ao problema de SVM envolve a soma da função objetivo com o produto da restrição do problema de SVM pelo multiplicador de Lagrange. Burges (1998), afirma que os pontos que satisfazem as restrições também formam um conjunto convexo, pois qualquer restrição linear define um conjunto convexo e um conjunto de n restrições lineares simultâneas define a intersecção dos n conjuntos convexos, o qual também é um conjunto convexo. Desta forma, a função lagrangeana envolve a soma de duas funções convexas e, por isso, o lagrangeano $\mathcal{L}(w, b, \alpha)$ é convexo.

Mais ainda, o problema é quadrático, pois a norma usada está ao quadrado. Logo, a função objetivo além de ser convexa é quadrática. Cabe destacar que os problemas de optimização não convexos apresentam grandes desafios na resolução, grande parte devido às limitações em métodos de programação matemática e à capacidade de processamento computacional. Tamanha importância em obter um problema convexo pode ser explicada pelo facto de estes apresentarem um mínimo global e que em problemas de programação convexa todo o óptimo local é um óptimo global (Minoux, 1986). Além disso, Fletcher (1987) expôs que, para problemas convexos, as condições de otimalidade de Karush-Kuhn-Tucker (KKT) são necessárias e suficientes para que w, b e α possam ser a solução do problema de SVM e, assim, resolver o problema de SVM seria equivalente a encontrar uma solução para as condições de KKT, as quais serão vistas adiante. Por fim, como o máximo do problema dual é igual ao mínimo do problema primal, resolver o dual equivaleria a resolver o primal, porém de forma menos trabalhosa. Esta formulação particular do problema

dual é conhecida na literatura como Dual de Wolfe (Fletcher, 1987). Assim, nas próximas linhas serão comparadas as resoluções destes problemas de otimização na forma primal e dual.

Recorde-se que o próximo passo consiste em resolver $\nabla\mathcal{L}(x,\alpha)=0$. Assim, calculando as derivadas parciais da função lagrangeana em relação a $w\ e\ b$ e igualando o resultado a zero, tem-se

$$\frac{\partial\mathcal{L}(w,b,\alpha)}{\partial w}=w-\sum_{i=1}^{n}\alpha_i y_i x_i=0\Leftrightarrow w=\sum_{i=1}^{n}\alpha_i y_i x_i$$

$$\frac{\partial\mathcal{L}(w,b,\alpha)}{\partial b}=-\sum_{i=1}^{n}\alpha_i y_i=0$$

Substituindo o valor de w acima em $\mathcal{L}(w,b,\alpha)$, obtém-se uma nova função W (que depende apenas de a e b) dada por

$$W(\alpha,b)=\frac{1}{2}\left(\sum_{i=1}^{n}\alpha_i y_i x_i\right)\cdot\left(\sum_{j=1}^{n}\alpha_j y_j x_j\right)-\sum_{i=1}^{n}\alpha_i\left[y_i\left(\left(\sum_{j=1}^{n}\alpha_j y_j x_j\right)\cdot x_j+b\right)-1\right]$$

Simplificando a expressão anterior, chega-se a

$$W(\alpha,b)=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i\cdot x_j-\sum_{i=1}^{n}\alpha_i y_i\left(\left(\sum_{j=1}^{n}\alpha_j y_j x_j\right)\cdot x_i+b\right)+\sum_{i=1}^{n}\alpha_i$$

$$=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i\cdot x_j-\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i\cdot x_j-b\sum_{i=1}^{n}\alpha_i y_i+\sum_{i=1}^{n}\alpha_i$$

$$=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i\cdot x_j-b\sum_{i=1}^{n}\alpha_i y_i$$

Como $\sum_{i=1}^{n}\alpha_i y_i=0$, da expressão anterior pode ser obtida a função lagrangeana dual de Wolfe na forma

$$W(\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i\cdot x_j$$

O problema de otimização fica escrito de forma que a função objetivo W depende apenas dos multiplicadores de lagrange. Assim, apresenta-se o problema de maximização

$$\underset{\alpha}{Maximizar} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i \cdot x_j$$

$$Sujeito\ a\ \alpha_i \geq 0, \forall i = 1, \dots, n$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

Importante notar que tanto o problema primal quanto o dual têm origem na mesma função objetivo, porém com diferentes restrições. Além disso, enquanto o problema primal é resolvido por meio de minimização, o problema dual é solucionado com uso de maximização, embora a solução seja equivalente. Conforme pode ser verificado em $w = \sum_{i=1}^{n} \alpha_i y_i x_i$, existe um multiplicador de Lagrange α_i para cada elemento do conjunto de treino. Além disso, a expressão para w demonstra que a solução óptima para o hiperplano pode ser escrita como uma combinação linear dos pontos do conjunto de treino. Os pontos para os quais o multiplicador de Lagrange é positivo são chamados de vectores de suporte e todos os elementos remanescentes do conjunto de treino que possuem multiplicador de Lagrange nulo são irrelevantes (Schölkopf, 1997). De maneira semelhante, Cortes e Vapnik (1995), baseados no teorema de Karush-Kuhn-Tucker, explicam que, no ponto de sela em $w, b\ e\ \alpha$, qualquer multiplicador de Lagrange e sua restrição correspondente estão conectados pela igualdade $\alpha_i[y_i(w \cdot x_i + b) - 1] = 0, i = 1, \dots, n$ e que os α_i diferentes de zero somente são atingidos nos casos em que $y_i(w \cdot x_i + b) - 1 = 0$. Em outras palavras, $\alpha_i \neq 0$ ocorre somente para os casos em que a desigualdade é satisfeita como uma igualdade e, assim, os vectores de suporte são os vectores x_i para os quais os multiplicadores de Lagrange são diferentes de zero.

O método dos multiplicadores de Lagrange foi desenvolvido para problemas que possuem restrições de igualdade. Entretanto, nos passos anteriores, o método foi utilizado em um problema que continha restrições de desigualdade. Isto só é possível quando a solução satisfaz as condições de KKT, que são as condições de primeira ordem necessárias para a solução de um problema de optimização ser óptima (Kowalczyk, 2017).

As condições de KKT para o problema primal são as seguintes (Fletcher, 1983)

$$\frac{\partial \mathcal{L}(w, b, \alpha)}{\partial w} = w - \sum_{i=1}^{n} \alpha_i y_i x_i = 0$$

$$\frac{\partial \mathcal{L}(w, b, \alpha)}{\partial b} = -\sum_{i=1}^{n} \alpha_i y_i = 0$$

$$y_i(w \cdot x_i + b) - 1 \geq 0, \forall i = 1, \dots, n$$

$$\alpha_i \geq 0, \forall i = 1, \dots, n$$

$$\alpha_i[y_i(w \cdot x_i + b) - 1] = 0, \forall i = 1, \dots, n$$

De acordo com Fletcher (1987) e McCormick (1983), as condições de KKT são satisfeitas na solução de qualquer problema de optimização restrito (convexo ou não), independentemente do tipo de restrição, somente se a intersecção do conjunto de direções viáveis com o conjunto de direções de descida coincida com a intersecção do conjunto de possíveis direções para restrições linearizadas com o conjunto de direções de descida. Como as restrições para os problemas envolvendo SVM são lineares, a suposição é válida. Além disso, Fletcher (1987) expôs que, como o problema relativo a SVM é convexo e, para problemas convexos, as condições de KKT são necessárias e suficientes para $w, b\ e\ \alpha$ ser solução, resolver o problema do SVM seria equivalente a encontrar uma solução para as condições de KKT.

De posse dos multiplicadores de Lagrange, encontrados ao resolver o problema dual de Wolfe, resta obter o valor para b (w já pode ser obtido por meio da expressão $w = \sum_{i=1}^{n} \alpha_i y_i x_i$).

Conforme já exposto neste trabalho, multiplicadores de Lagrange diferentes de zero somente ocorrem para os casos em que $y_i(w \cdot x_i + b) = 1$. Escolhendo qualquer i para o qual o multiplicador de Lagrange é diferente, pode-se encontrar o valor de b. Logo, usando $y_i(w \cdot x_i + b) = 1$ e multiplicando ambos os lados por y_i, chega-se a ${y_i}^2(w \cdot x_i + b) = y_i$. Considerando o facto de $y_i \in Y$, para $Y = \{-1,1\}$, tem-se que ${y_i}^2 = 1$ e então $b = y_i - w \cdot x_i$. Burges (1998) afirma que é numericamente seguro considerar para o valor de b a média dos seus valores encontrados para cada i, de maneira que seu cálculo seria dado por $b =$

$\frac{1}{m}\sum_{i=1}^{m}(y_i - w \cdot x_i)$, em que m representa o número de vectores de suporte. Para o cálculo de b, Ng (2018) e Vapnik (1998) utilizam a média entre o vector de suporte positivo mais próximo e o vector de suporte negativo mais próximo, dada por

$$b = -\frac{max_{y_i=-1}(w \cdot x_i) + min_{y_i=1}(w \cdot x_i)}{2}$$

2.3. SVM com margens suaves

Na secção anterior foi desenvolvido o suporte teórico para o caso linearmente separável, ou seja, quando há um hiperplano óptimo que separa perfeitamente (sem erros) os dados do conjunto de treino em classes.

Segundo Schölkopf (1997), na prática são raros os casos de existência de hiperplano separador. Por outras palavras, raros são os casos de dados linearmente separáveis sem erros. Dentre os factores que causam essa ausência de linearidade, podem ser destacados: presença de *outliers*, instrumento de medição não calibrado adequadamente, erro de digitação ao transpor dados de medição para formato digital e a própria estrutura dos dados entre outros.

Cortes e Vapnik (1995) definem a SVM com margem suave como o caso em que se deseja separar o conjunto de treino com o menor número possível de erros. Trata-se de uma modificação na SVM com margem rígida para permitir que alguns dados do conjunto de treino violem a restrição $y_i(w \cdot x_i + b) - 1 \geq 0$. Para tal, são introduzidas variáveis de folga $\xi_i \geq 0, i = 1, \dots, n$. A restrição anteriormente apresentada para o caso separável pode ser dividida nas duas desigualdades

$$\begin{cases} w \cdot x_i + b \geq +1, se\ y_i = \ \ 1 \\ w \cdot x_i + b \leq -1, se\ y_i = -1 \end{cases}$$

Desse modo, a introdução das variáveis de folga faz com que tais restrições passem a ser

$$\begin{cases} w \cdot x_i + b \geq +1 - \xi_i, se\ y_i = \ \ 1 \\ w \cdot x_i + b \leq -1 + \xi_i, se\ y_i = -1 \\ \xi_i \geq 0\ \forall i \end{cases}$$

As restrições com as variáveis de folga podem ser combinadas numa única desigualdade, transformando-se em $y_i(w \cdot x_i + b) \geq 1 - \xi_i$. Ao analisar as desigualdades acima expostas, Burges (1998) explica que, para um erro ocorrer, a correspondente ξ_i deve ser maior do que 1, de maneira que $\sum_i \xi_i$ é um limite superior para o número de erros de treino.

Vapnik (1998) introduziu o conceito de hiperplano óptimo generalizado para o caso de SVM com margens suaves. Segundo ele, o hiperplano óptimo generalizado é encontrado por meio da minimização da função objetivo

$$\underset{w,\xi}{\text{Minimizar}} \ \frac{1}{2}\|w\|^2 + C\left(\sum_{i=1}^{n} \xi_i\right)$$

$$Sujeito\ a\ y_i(w \cdot x_i + b) \geq 1 - \xi_i, i = 1, \ldots, n$$

Cabe notar que o somatório de ξ_i foi incluído na função objetivo. Isso quer dizer que a solução será o hiperplano que maximiza a margem e possui o menor erro possível. Além disso, a constante C representa um termo de regularização, cuja função é definir um equilíbrio, determinado pelo utilizador, entre a margem de separação e o número de casos incorrectamente classificados ou pelo menos dentro da margem de separação, que são os casos correctamente classificados. Passerini (2004) diz que o parâmetro de regularização C estabelece um *trade-off* entre erro empírico pequeno e a suavidade da solução, de maneira que a correta escolha de C previne o sobreajuste. Desta forma, grandes valores para a constante C significam maior penalização para os erros.

Uma vez mais, tem-se um problema de optimização quadrático, com restrições lineares e sua resolução segue os mesmos passos demonstrados no caso linearmente separável. Em outras palavras, trata-se de introduzir uma função lagrangeana e encontrar os pontos para os quais suas derivadas parciais são nulas.

A função objetivo do problema é dada por

$$f(w, \xi) = \frac{1}{2}\|w\|^2 + C\left(\sum_{i=1}^{n} \xi_i\right)$$

e as funções de restrição são

$$g_i(w, b, \xi) = y_i(w \cdot x_i + b) - 1 + \xi_i \; e \; h_i(\mu, \xi) = \mu_i \xi_i, \; i = 1, \ldots, n.$$

A função lagrangeana é dada por

$$\mathcal{L}(w, b, \alpha, \xi, \mu) = f(w, \xi) - \sum_{i=1}^{n} \alpha_i g_i(w, b, \xi) - \sum_{i=1}^{n} \mu_i \xi_i.$$

Ao efetuar as substituições, a função lagrangeana converte-se em

$$\mathcal{L}(w, b, \alpha, \xi, \mu) = \frac{1}{2}\|w\|^2 + C\left(\sum_{i=1}^{n} \xi_i\right) - \sum_{i=1}^{n} \alpha_i [y_i(w \cdot x_i + b) - 1 + \xi_i] - \sum_{i=1}^{n} \mu_i \xi_i$$

e as condições de KKT para o problema primal são (Burges, 1998)

$$\frac{\partial \mathcal{L}(w, b, \alpha, \xi, \mu)}{\partial w} = w - \sum_{i=1}^{n} \alpha_i y_i x_i = 0 \Rightarrow w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

$$\frac{\partial \mathcal{L}(w, b, \alpha, \xi, \mu)}{\partial b} = -\sum_{i=1}^{n} \alpha_i y_i = 0$$

$$\frac{\partial \mathcal{L}(w, b, \alpha, \xi, \mu)}{\partial \xi_i} = C - \alpha_i - \mu_i = 0$$

$$y_i(w \cdot x_i + b) - 1 + \xi_i \geq 0$$

$$\xi_i \geq 0$$

$$\alpha_i \geq 0$$

$$\mu_i \geq 0$$

$$\alpha_i [y_i(w \cdot x_i + b) - 1 + \xi_i] = 0$$

$$\mu_i \xi_i = 0$$

de forma que o problema dual é expresso como (Vapnik, 1998)

$$\underset{\alpha}{Maximizar} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i \cdot x_j$$

$$Sujeito \; a \; 0 \leq \alpha_i \leq C, \forall i = 1, \ldots, n$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

Nota-se que, para o presente caso, os α_i estão limitados superiormente pela constante C.

De maneira análoga ao que foi exposto para o caso das margens rígidas, Burges (1998) afirma que para o cálculo do valor de b, pode ser tomada a média dos seus valores encontrados para cada ponto do conjunto de treino.

2.4. SVM não linearmente separável

A teoria exposta até o presente momento utiliza um hiperplano separador linear como superfície de decisão. Tal hiperplano separa linearmente os dados do conjunto de treino em duas regiões. Todavia, nem sempre será possível separar linearmente os dados do conjunto de treino, conforme apresentado na figura 4, logo abaixo.

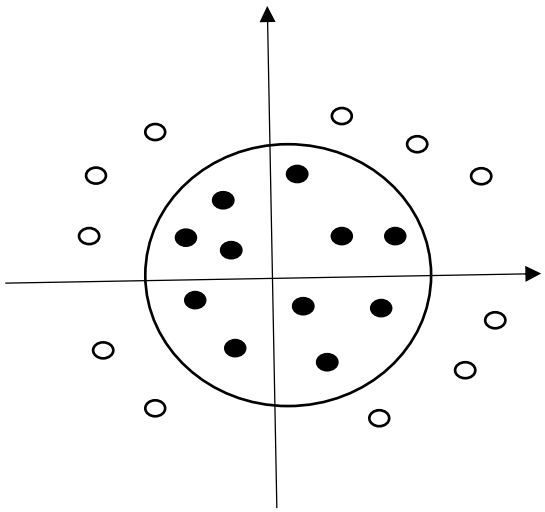

Figura 4: Esquema de dados não linearmente separáveis

Desta forma, caso os dados não sejam separáveis, pode-se mapeá-los de forma não linear para um espaço de dimensão superior, determinando nesse espaço de dimensão superior uma superfície de separação, da mesma forma que no caso linearmente separável com margens rígidas ou suaves. Ao voltar ao espaço de características inicial - *feature space*

– (Schölkopf, 1997) invertendo a transformação não linear, permite-se superfícies de decisão mais gerais, no sentido em que não são necessariamente lineares devido ao inverso do mapeamento utilizado não ser obrigatoriamente linear. Este *modus operandi* está baseado no Teorema de Cover (Haykin, 1999) que, resumidamente, afirma que se a transformação a ser utilizada for não linear e o espaço de dimensão superior possuir dimensão suficientemente elevada, então existe elevada probabilidade de que um espaço de entradas não linearmente separável χ (espaço dos dados originais) seja transformado em um espaço com dados linearmente separáveis. Além disso, o SVM possui uma característica interessante, que é a de não precisar conhecer explicitamente o mapa não linear e, ainda assim, conseguir aplicar um algoritmo linear ao problema de classificação no espaço de dimensão mais alta. Esta característica é conhecida como *Kernel Trick*.

De maneira mais formal, Burges (1998) define o mapeamento dos dados do conjunto de treino para um espaço euclidiano de dimensão superior, $\mathcal{H}$, usando um mapeamento chamado de Φ, da forma

$$\Phi: \mathbb{R}^d \rightarrow \mathcal{H}$$

Conforme explicado em Cortes e Vapnik (1995), a classificação desconhecida de um vector x é feita transformando primeiro o vector no espaço de separação $\big(x \mapsto \Phi(x)\big)$ e então tomando o sinal da função

$$f(x) = w \cdot \Phi(x) + b$$

De forma a melhor tratar inconsistências porventura presentes nos dados, utiliza-se o SVM linear com margens suaves e, por meio de suas propriedades, pode-se escrever o vector w na forma de combinação linear dos vectores de suporte no espaço de características, ou seja

$$w = \sum_{i=1}^{n} y_i \alpha_i \Phi(x_i)$$

Conforme determinado pela aplicação de SVM, os dados do conjunto de treino aparecem na formulação do problema do caso com separador linear como produtos internos $x_i \cdot x_j$. Utilizando a linearidade do produto interno na função de classificação f, tem-se

$$f(x) = w \cdot \Phi(x) + b = \sum_{i=1}^{n} y_i \alpha_i \Phi(x) \cdot \Phi(x_i) + b$$

E o classificador fica da forma

$$g(x) = sgn\big(f(x)\big) = sgn\left(\sum_{i=1}^{n} y_i \alpha_i \Phi(x) \cdot \Phi(x_i) + b\right)$$

De maneira análoga, o Dual encontrado para a SVM de margem suave, após a aplicação do mapeamento Φ, fica escrito como

$$\underset{\alpha}{Maximizar} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \big(\Phi(x_i) \cdot \Phi(x_j)\big)$$

$$Sujeito\ a\ 0 \leq \alpha_i \leq C, \forall i = 1, \dots, n$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

2.4.1. Kernel Trick

A secção anterior mostrou que, para o caso não linearmente separável, é necessário realizar uma transformação do conjunto de treino num espaço de dimensão superior. Este mapeamento faz que com que os dados sejam levados do espaço de entradas para um espaço de dimensão superior, exigindo o cálculo dos produtos internos $\Phi(x) \cdot \Phi(x_i)$ num espaço Euclidiano $\mathcal{H}$ de maior dimensão (Schölkopf, 1997), podendo inclusive ser de dimensão infinita (Burges, 1998), tornando o cálculo de tais produtos uma tarefa hercúlea e, por vezes, inexequível. Assim, surge a pergunta: existe alguma forma de conseguir os valores relativos ao produto interno $\langle \Phi(x_i), \Phi\big(x_j\big) \rangle$ no espaço de dimensão superior sem recorrer à transformação dos vetores? Sim, usando o denominado Kernel Trick, que será explanado de seguida.

De acordo com a definição de Herbrich (2002), Kernel é uma função que calcula o produto interno, no espaço de dimensão superior, para todos os x_i e x_j pertencentes ao espaço de entrada (espaço original dos dados). Matematicamente, pode ser escrito como

$$K\big(x_i, x_j\big) = \langle \Phi(x_i), \Phi\big(x_j\big) \rangle$$

Desta forma, a ideia por trás da utilização do Kernel é escolher uma função Kernel K, ao invés de um mapeamento, antes de aplicar um algoritmo de aprendizagem (Herbrich, 2002). Por outras palavras, o truque está numa escolha adequada de uma função kernel K, que possa ser utilizada por forma a satisfazer as propriedades do produto interno e sem necessidade de explicitar o mapeamento Φ.

Ao efetuar a substituição do produto interno $x_i \cdot x_j$ realizado no espaço de origem dos dados pelo produto interno $\langle \Phi(x_i), \Phi(x_j) \rangle$ realizado no espaço $\mathcal{H}$, sem a necessidade do mapeamento Φ, tem-se o truque conhecido como *"Kernel Trick"*, que aplicado à função de decisão f, ao classificador e ao Dual obtido para a SVM de margem suave conduz a

$$f(x) = \sum_{i=1}^{n} y_i \alpha_i K(x, x_i) + b$$

$$g(x) = sgn\big(f(x)\big) = sgn\left(\sum_{i=1}^{n} y_i \alpha_i K(x, x_i) + b\right)$$

$$\underset{\alpha}{Maximizar} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K\big(x_i, x_j\big)$$

$$Sujeito\ a\ 0 \leq \alpha_i \leq C, \forall i = 1, \dots, n$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

Assim, a teoria do caso linear também se aplica ao caso não linear obtido por meio do Kernel K adequado ao invés de utilizar o produto interno Euclidiano (Schölkopf, 1997).

É importante esclarecer que não pode ser usada como Kernel qualquer função k, uma vez que deve satisfazer as condições do Teorema de Mercer, nomeadamente uma função simétrica $k: \chi \times \chi \to \mathbb{R}$ é um Kernel de Mercer, se e só se, para cada $n \in \mathbb{N}$ e $x = (x_1, \dots, x_n) \in \chi^n$, a matriz $M = \left(K\big(x_i, x_j\big)\right)_{i,j=1,\dots,n}$, quadrada de ordem n, é semi-definida positiva.

2.4.2. Tipos de Kernel

Dentre os Kernels mais utilizados, tem-se (Hsu et al.,2010)

Linear: $k(x_i, x_j) = x_i \cdot x_j$

Polinomial de grau d: $k(x_i, x_j) = (\gamma x_i \cdot x_j + r)^d, \gamma > 0$

Gaussiano: $k(x_i, x_j) = exp\left(-\gamma \|x_i - x_j\|^2\right), \gamma > 0$

Sigmoidal: $k(x_i, x_j) = tanh(\gamma x_i \cdot x_j + r)$

em que $r, \gamma \; e \; d$ são parâmetros dos Kernels. Conforme destacam Amami et al. (2013), o parâmetro r serve como parâmetro de deslocamento que controla o limite do mapeamento. Já γ atua como um parâmetro de escala. Por fim, d é o grau do polinômio.

Adicionalmente aos Kernels acima mencionandos, Vapnik e Izmailov (2015) destacam os seguintes Kernels

INK-spline Kernel

Este Kernel é para spline de ordem zero com número infinito de nós.

$$K_{INK_0}(x, y) = \prod_{k=1}^{d} (min(x^k, y^k) + \delta)$$

Já o Kernel do spline de ordem 1 é definido por

$$K_{INK_1}(x, y) = \prod_{k=1}^{d} \left(\delta + x^k y^k + \frac{|x^k - y^k| min\{x_k, y^k\}}{2} + \left(\frac{min\{x^k, y^k\}}{3} \right)^3 \right)$$

em que σ é um parâmetro livre, $x^k \geq 0$ e $y^k \geq 0$ são as k coordenadas do vector d-dimensional x.

Kernel quadrático homogêneo

$K_{Pol_2} = (x, y)^2$, em que (x, y) representa o produto interno entre os vectores x e y.

2.5. SVM com informação privilegiada (SVM+)

Vapnik e Vashist (2009) introduziram o conceito de aprendizagem com informações privilegiadas (*LUPI – Learning Using Privileged Information*). Esta aprendizagem possui como objetivo não só melhorar a predição dos algoritmos de aprendizagem, mas também reduzir o número de dados de treino a serem utilizados. O modelo proposto pelos autores adiciona informação privilegiada na fase de treino, informação essa que se supõe não estar disponível para os elementos de teste posteriormente classificados. Ao utilizar o LUPI é necessário organizar as representações do conjunto de treino em informação fundamental e informação privilegiada, as quais são descritas de seguida.

Informação fundamental: Informação obtida por meio de processo técnico de baixo custo, facilmente acessível, seja em relação a recursos humanos ou computacionais utilizados na geração dessa informação.

Informação privilegiada: Diferentemente do caso anteriormente exposto, esta informação, de característica adicional, é obtida por meio de um processo de alto custo (um perito ou especialista da área). Por isso, é definida como privilegiada e, geralmente, representa apenas um pequeno conjunto dos dados de treino (Marcacini, 2014). Tal informação também está sujeita a inconsistências.

Para ilustrar a diferença entre os dois tipos de informação, Vapnik e Vashist (2009) descreveram um problema de classificação de imagens de biópsia que possuía duas classes: cancro e não cancro. De forma a possibilitar a aprendizagem por um modelo preditivo, apenas um dos grupos de imagens possui um relatório manual produzido por um perito, enquanto o outro grupo não possui tal informação. O espaço de características χ é dado pelas informações visuais que foram automaticamente extraídas dos pixels das imagens (pré-processamento de imagem). Ainda, possuem uma segunda descrição (diagnóstico textual,

análise, comentários), χ^*, somente os elementos do conjunto de treino que possuem um relatório do especialista. Assim, χ é o espaço de características com informação fundamental, χ^* representa a informação privilegiada e o modelo deve classificar os elementos do conjunto de treino descritos pela informação fundamental usando a informação privilegiada para melhorar o classificador. Reforça-se que a informação privilegiada está disponível apenas durante o momento de treino, não no momento do teste.

Os rótulos utilizados para a informação fundamental são os mesmos utilizados para a informação privilegiada e, mesmo sendo possível aperfeiçoar o modelo apenas com a informação fundamental, os comentários, análises e relatórios, concernentes às informações privilegiadas, podem melhorar a aprendizagem.

Segundo Pechyony e Vapnik (2010), o algoritmo do SVM+ implementa o paradigma LUPI, isto é, generaliza o SVM padrão. Para tanto, as variáveis de folga ξ_i são parametrizadas como função das informações privilegiadas, ou seja, na forma

$$\xi_i(w^*, b^*) = w^* \cdot x_i^* + b^*$$

Acrescentando a informação privilegiada e utilizando raciocínio análogo ao caso do SVM padrão, quer-se minimizar a seguinte função objetivo

$$\underset{w,b,w^*,b^*}{\text{minimizar}} \frac{1}{2}(\|w\|^2 + \gamma\|w^*\|^2) + C\left(\sum_{i=1}^{n} w^* \cdot x_i^* + b^*\right)$$

$$Sujeito\ a\ y_i(w \cdot x_i + b) \geq 1 - (w^* \cdot x_i^* + b^*)$$

$$w^* \cdot x_i^* + b^* \geq 0$$

em que γ controla o peso da informação privilegiada.

Já o problema da forma primal referente ao SVM+ é escrito como

$$\mathcal{L}(w, b, w^*, b^*, \alpha, \beta) = \frac{1}{2}(\|w\|^2 + \gamma\|w^*\|^2) + C\left(\sum_{i=1}^{n} w^* \cdot x_i^* + b^*\right) -$$

$$\sum_{i=1}^{n} \alpha_i[y_i(w \cdot x_i + b) - 1 + (w^* \cdot x_i^* + b^*)] - \sum_{i=1}^{n} \beta_i(w^* \cdot x_i^* + b^*)$$

em que $\alpha_i, \beta_i \geq 0$ são os multiplicadores de Lagrange. Da mesma maneira que no SVM padrão é usado o produto interno $x_i \cdot x_j$ no espaço de origem, sendo que este é substituído na forma do Kernel de Mercer $K(x_i, x_j)$ na formulação quando se pretende obter separadores não lineares. Nesse caso, a formulação dual para o problema é definida como (Gavrilov, 2012)

$$\underset{\alpha,\beta}{\text{maximizar}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i x_j) - \frac{1}{2\gamma} \sum_{i,j=1}^{n} (\alpha_i + \beta_i - C)(\alpha_j + \beta_j - C)\, K^*(x_i^*, x_j^*)$$

$$Sujeito\ a$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0\,; \qquad \sum_{i=1}^{n} (\alpha_i + \beta_i - C) = 0; \qquad \alpha_i \geq 0, \beta_i \geq 0$$

em que $\alpha_i \geq 0, \beta_i \geq 0$ são multiplicadores de Lagrange.

Nos próximos capítulos será apresentada a aplicação desta metodologia ao caso em estudo, nomeadamente à classificação de imagem da retina de ratos nas classes saudável ou transgénico.

3. Metodologia

Neste capítulo será apresentada a metodologia utilizada para atingir o objetivo do estudo, iniciando a apresentação pela caracterização da amostra.

3.1. Objetivo de Aplicação e Caracterização da Amostra

O objetivo deste estudo consiste na criação de modelos de aprendizagem de máquina baseados nas técnicas de SVM e SVM+ capazes de distinguir entre ratos saudáveis (C) e transgénicos (D) por meio da análise de textura da imagem de tomografia de coerência óptica (OCT). O modelo aplicado no presente estudo é um modelo animal transgénico 3xTg-AD, aplicado para a doença de Alzheimer. Desta forma, por meio das conclusões desse estudo, pretende-se que este trabalho sirva como o primeiro passo para o rastreio de Alzheimer em fases iniciais e por métodos não invasivos. A utilização da imagem da retina é justificada por seu tecido ser semelhante ao do cérebro, pelo que se espera conseguir demonstrar que as alterações no tecido do cérebro por doenças neurodegenerativas provocam alterações nos tecidos da retina em fases iniciais da doença.

Os dados tratados nesta dissertação foram obtidos no âmbito do projeto PTDC/EMD-EMD/28039/2017, financiado pela Fundação para a Ciência e Tecnologia (FCT), Fundo Europeu de Desenvolvimento Regional (FEDER) e Orçamento de Estado de Portugal (OE), cumprindo todas as disposições legais quanto ao tratamento dos animais e aquisição da imagem.

A técnica de imagiologia utilizada para a obtenção de imagem foi a tomografia de coerência óptica (*optical coherence tomography - OCT)*, que é uma técnica de diagnóstico por imagem não invasiva. Conforme explicam Palazzi et. al (2015), a OCT tem sido usada para análise da microestrutura de tecidos biológicos. Para tanto, o sistema da OCT se baseia na interferometria de baixa coerência óptica para produzir mapas em profundidade da dispersão da luz pela retina, gerando imagens que definem as diferentes camadas que compõem os tecidos. Para mais detalhes sobre a técnica utilizada, pode-se consultar Serranho, Morgado e Bernardes (2012).

No mesmo sentido, Serranho, Morgado e Bernardes (2012) comentam que a técnica de OCT permite obter uma representação 3D do fundo do olho com base nas diferentes refletividades das camadas da retina. Como vantagem da alta resolução de imagem proporcionada pela técnica de OCT, os autores citam a possibilidade de obtenção da informação estrutural da retina e de suas camadas ao nível histológico e o facto de a OCT ser uma técnica confortável para os pacientes, uma vez que não há contato durante o exame e, assim, não há necessidade de anestesia. Desta forma, a OCT se apresenta como uma boa solução de rastreio, no caso em estudo, potencialmente para o caso da Doença de Alzheimer.

As recolhas foram efetuadas apenas no olho direito dos ratos, por forma a evitar qualquer tipo de viés nos resultados que tenha relação ao olho de recolha ou à inclusão de dois olhos do mesmo rato na amostra, causando potencial dependência nos dados.

3.2. Aquisição e Pré-processamento de imagem

Para a aquisição da imagem foram recolhidos os volumes de dados OCT para o olho direito de cada um dos 74 ratos. De posse dos volumes OCT obtidos, realizou-se a segmentação automática das 3 camadas anteriores da retina, nomeadamente RNFL, GCL e IPL. De notar que as fibras nervosas são encontradas na camada anterior da retina. Assim, tais camadas são as potencialmente mais afetadas por doenças neurodegenerativas, como o Alzheimer. Por outro lado, os dados OCT obtidos por meio da segmentação das referidas camadas nem sempre é muito evidente, ou seja, há um problema em relação à qualidade da imagem, fazendo com que sua aquisição se torne, por vezes, muito difícil. Com o intuito de resolver este problema, considera-se a camada agregada (constituída pelas 3 camadas) como a fonte para a informação fundamental do SVM e os dados de cada uma das 3 camadas como os dados para a informação privilegiada. O racional por trás desta decisão é que é relativamente fácil segmentar automaticamente a camada agregando as 3 camadas superiores da retina (RNFL, GCL e IPL) – pelo que esta é considerada informação *standard* -, porém se torna mais difícil segmentar cada uma das 3 camadas – considerando-se esta, portanto, a informação privilegiada.

Cada volume de dados foi dividido em 4 quadrantes e, para cada quadrante e cada camada, foi obtida uma imagem de fundo por meio da média em profundidade dos valores de intensidade dos dados OCT na camada, transformando a informação em três dimensões (3D) numa imagem em duas dimensões (2D). Finalmente, a cada imagem de fundo obtida em cada quadrante foram aplicados 20 indicadores de análise de textura de imagem de fundo, que serão usados com *features* para o modelo SVM. O procedimento para o processamento da imagem e posterior recolha de dados de textura pode ser consultado em detalhe em Ferreira et al. (2020). Também em Ferreira et al. (2020) podem ser encontradas as definições para os indicadores de análise de textura utilizados nesta dissertação, quais sejam: *Homogeneity, Uniformity, Entropy, Dissimilarity, Inertia, Correlation, Autocorrelation, ClusterShade, ClusterProminence, MaximumProbability, SumOfSquares, SumAverage, SumVariance, SumEntropy, DifferenceVariance, DifferenceEntropy,* IMC1, IMC2, INN e IDN (*Inverse difference moment normalized*).

Seguindo o raciocínio acima, nota-se que, para cada sujeito, existem 20x4=80 *features* de informação fundamental e 3x20x4=240 *features* de informação privilegiada, pelo que se tornou necessário reduzir o número de *features* dado o pequeno número de ratos constantes na amostra. Liu e Zheng (2006) explicam a importância da redução do número de *features* na aprendizagem de máquina e citam duas razões para que isso seja feito. São elas: melhoria da capacidade de generalização do classificador, já que diminui o risco de sobreajuste, e diminuição da complexidade computacional, pois menos *features* implicam menos tempo de execução para treinar e aplicar o classificador.

3.3. Seleção de *features* mais relevantes

É importante, conforme exposto na secção anterior, reduzir o número de *features* a serem utilizadas. Assim, objetivou-se escolher as que melhor separam entre as classes saudável e transgénico.

Primeiramente, realizou-se um teste de normalidade por grupo, de cada *feature*, com significância de 10% para saber que grupos têm ou não distribuição Normal. Devido ao tamanho amostral superior a 30 em cada grupo, procedeu-se ao teste de Kolmogorov-

Smirnov. Em relação à escolha do nível de significância de 10%, esta deve-se ao facto de se querer fazer uma análise conservadora e só "aceitar" como normais as distribuições com *p-value* superior a 0,1.

A importância da avaliação exposta no parágrafo anterior baseia-se na escolha do teste estatístico a ser utilizado na comparação das *features* entre grupos, nomeadamente para perceber quais teriam mais poder discriminatório entre grupos. Assim, o objetivo foi descobrir se o teste estatístico a usar seria paramétrico ou não. Para os casos em que foi verificada distribuição Normal, o teste utilizado para a comparação das *features* entre os grupos foi o teste T, paramétrico. Nos casos em que a hipótese de distribuição Normal foi rejeitada para pelo menos um dos grupos de *features f*oi utilizado o teste não paramétrico de Mann-Whitney. Destaca-se que ambos os testes foram realizados para cada *feature*, nos dois grupos, para as camadas RNFL, GCL, IPL e para o agregado das 3 camadas.

Em outras palavras, procurou-se saber o nível de distinção dos grupos. Assim, uma grande separação entre grupos de ratos saudáveis e ratos transgénicos indica que os dados sob análise não possuem a mesma distribuição, não sendo possível aceitar a hipótese nula, mostrando que a *feature* é importante na distinção entre os grupos.

As hipóteses objeto de teste foram

H_0: A média/mediana de ratos saudáveis e ratos transgénicos é igual.

H_1: A média/mediana de ratos saudáveis e ratos transgénicos é diferente.

Após a aplicação dos testes acima citados, obteve-se para cada *feature* 4 *p-values* (1 para cada umas das 3 camadas e outro para a camada agregada). A regra de decisão para considerar a *feature* relevante utilizou estes 4 *p-values*, que por definição representa o nível de significância mínimo para o qual a hipótese nula é aceite. Caso os 4 *p-values* encontrados para as *features* fossem, conjuntamente, inferiores ao nível de significância adotado (10%), a *feature* era considerada significativa.

Finalmente, procedeu-se à análise de correlação para excluir *features* altamente correlacionadas com as anteriormente escolhidas para o modelo. Para tal, elas foram ordenadas por *p-value* para a camada agregada, uma vez que é esta a informação fundamental. Como critério de exclusão, definiu-se o limiar de 0,7 para o valor absoluto do

coeficiente de correlação de uma *feature* com qualquer das outras previamente escolhidas, obtendo-se a listagem ordenada de *features* para o modelo SVM.

3.4.Técnica de Validação Cruzada

A validação de resultados é uma etapa necessária, pois ao criar um modelo de aprendizagem de máquina, almeja-se que ele obtenha boas predições não só na base de dados utilizada, mas também diante de novas observações, evitando-se o sobreajuste. Em outras palavras, o modelo criado deve ser capaz de fazer boas predições para dados desconhecidos, que não foram anteriormente empregados.

Com o intuito de validar o desempenho dos modelos de aprendizagem de máquina criados nesta dissertação, fez-se uso da técnica de validação cruzada com *k* grupos.

A mencionada técnica consiste em dividir, aleatoriamente, o conjunto de dados total em *k* subconjuntos, mutuamente exclusivos. Assim, de posse dos *k* subconjuntos, pode-se separar um deles para ser o conjunto de teste e, os *k-1* restantes são unidos para comporem o conjunto de treino. Este procedimento é feito *k* vezes, alterando-se o conjunto de teste de forma a utilizar *k* diferentes conjuntos de teste. O conjunto de treino é utilizado para treinar o modelo, enquanto que o conjunto de teste serve para verificar a precisão do modelo na presença de dados que, como não foram empregados no processo de treino, podem ser considerados dados novos.

4. Resultados

Este capítulo é dedicado à demonstração dos resultados alcançados por meio da aplicação da metodologia apresentada no secção 03. A ferramenta computacional utilizada para a criação de gráficos, realização de testes estatísticos e aplicação das técnicas de classificação SVM e SVM+ foi o software R, versão 3.6.1, nomeadamente com recursos aos pacotes e1071 (Meyer et al., 2019), caret (Kuhn, 2019) e svmplus (Gauraha e Spjuth, 2018).

Como ponto de partida, apresenta-se a descrição dos dados objeto de estudo. Após, a listagem das *features* mais relevantes, a forma de validação cruzada utilizada e, por fim, os resultados obtidos por meio da técnica de classificação.

4.1. Descrição dos dados

A figura 5 ilustra a distribuição dos 74 elementos da base de dados em dois grupos: saudável (C) e transgénico (D). Desta forma, a dimensão da amostra em cada grupo é semelhante, não comprometendo a aplicação das técnicas de aprendizagem automática.

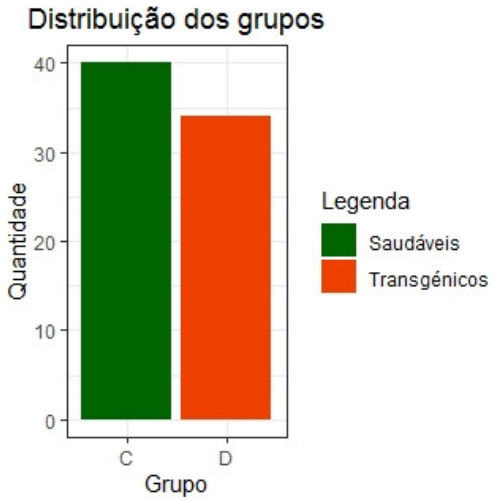

Figura 5: Distribuição dos dados entre grupos

4.2. Listagem ordenada das *features* mais relevantes

Das 80 *features* iniciais, apenas 35 apresentaram *p-value* significativo (a 10%) para as 4 camadas da retina que, após ordenadas por *p-value* na distinção na camada agregada, obtém-se a seguinte lista: *Inertia* (segundo quadrante), *Dissimilarity* (segundo quadrante), IDN (segundo quadrante), *Correlation* (segundo quadrante), *Correlation* (terceiro quadrante), IDN (terceiro quadrante), INN (terceiro quadrante), *Dissimilarity* (primeiro quadrante), *Inertia* (primeiro quadrante), IMC1 (primeiro quadrante), IMC1 (segundo quadrante), *Correlation* (primeiro quadrante), IDN (primeiro quadrante), IMC1 (terceiro quadrante), *Dissimilarity* (terceiro quadrante), INN (primeiro quadrante). *Inertia* (terceiro quadrante), *ClusterProminence* (primeiro quadrante), *ClusterProminence* (segundo quadrante), INN (segundo quadrante), *ClusterProminence* (terceiro quadrante), *Correlation* (quarto quadrante), IMC2 (primeiro quadrante), IDN (quarto quadrante), IMC2 (segundo quadrante), IMC2 (terceiro quadrante), *Inertia* (quarto quadrante), IMC2 (quarto quadrante), *Dissimilarity* (quarto quadrante), INN (quarto quadrante), *ClusterShade* (terceiro quadrante), IMC1 (quarto quadrante), *SumAverage* (terceiro quadrante), *MaximumProbability* (segundo quadrante) e *SumVariance* (quarto quadrante).

Finalmente, foram introduzidas as *features* cuja correlação com as anteriormente escolhidas era inferior a 0,7, conforme indicado na figura 6.

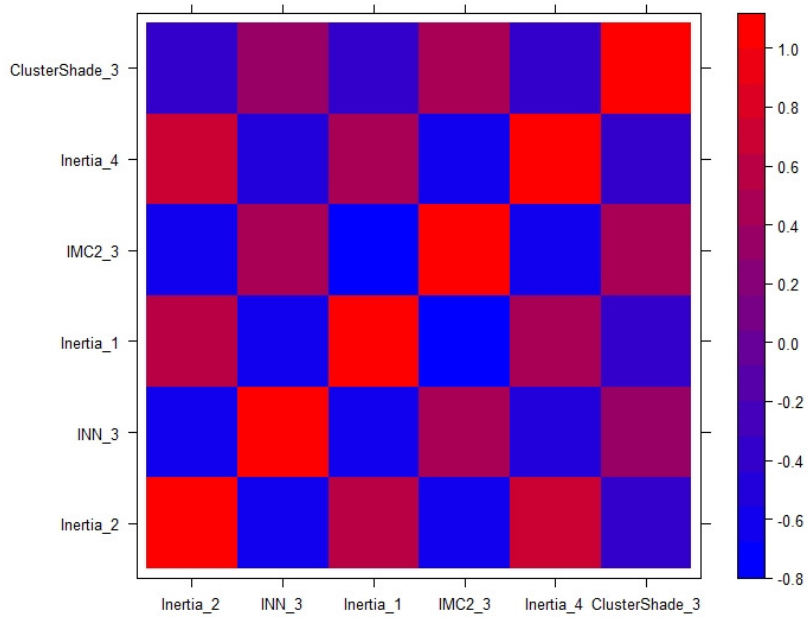

Figura 6: Correlação entre as *features* escolhidas como mais relevantes

Assim, por ordem, as *features* são: *Inertia* (segundo quadrante), INN (terceiro quadrante), *Inertia* (primeiro quadrante), IMC2 (terceiro quadrante), *Inertia* (quarto quadrante) e *ClusterShade* (terceiro quadrante), cuja correlação é indicada na figura 6.

4.3. Aplicação da técnica de Validação Cruzada

Neste trabalho foi utilizado $k = 5$, ou seja, procedeu-se à validação cruzada com 5 grupos. Assim, dividiu-se o conjunto de dados total em 5 partes e, para cada execução do algoritmo, um dos cinco grupos foi utilizado como conjunto de teste, enquanto o treino foi realizado nos outros quatro grupos. Tal procedimento se repetiu por 5 vezes, de maneira que houvesse 5 conjuntos de treino diferentes. No final, calculou-se a média das precisões obtidas nos 5 distintos conjuntos de teste. O procedimento exposto nesta secção está ilustrado na figura 7, logo abaixo.

Execução *k*				
1	**2**	**3**	**4**	**5**
Treino	Treino	Treino	Treino	Teste
Treino	Treino	Treino	Teste	Treino
Treino	Treino	Teste	Treino	Treino
Treino	Teste	Treino	Treino	Treino
Teste	Treino	Treino	Treino	Treino

Figura 7: Procedimento de Validação Cruzada em 5 grupos

4.4. Resultados por meio da técnica de SVM

Após separados os dados em conjunto de treino e conjunto de teste em cada passo da validação cruzada, criou-se o modelo cuja variável dependente é o grupo ao qual os ratos pertencem (saudável ou transgénico). As variáveis independentes são as *features* que demonstraram retratar melhor as diferenças entre os grupos e a base de dados empregada foi o conjunto de treino (80% do conjunto de dados). Assim, foram treinados cinco modelos cujos parâmetros foram afinados e utilizados na classificação dos elementos do conjunto de

teste. Este procedimento foi realizado $k = 5$ vezes, uma vez que em cada treino um dos cinco conjuntos era escolhido como conjunto de teste.

Criou-se a matriz de confusão de cada modelo. A matriz de confusão é uma tabela que compara as classificações previstas pelo modelo com as classificações reais. A soma dos elementos de sua diagonal principal representa a quantidade de valores classificados corretamente.

A tabela 1 ilustra a matriz de confusão de um dos cinco modelos SVM criados para a classificação dos 15 elementos do conjunto de teste, obtida por meio de um dos passos da validação cruzada e sem ajuste de parâmetros, utilizando os por defeito da função implementada em R.

Tabela 1: Matriz de confusão para o modelo SVM sem ajuste de parâmetros para os dados do conjunto de teste

		Estimado	
		Saudável	**Transgénico**
Real	**Saudável**	7	1
	Transgénico	1	6

Por meio da tabela 1, nota-se que o modelo apresentou precisão de 86,67% na classificação dos dados do conjunto de teste.

Com o intuito de melhorar a performance do modelo, realizou-se o ajuste dos seus parâmetros por meio da utilização de uma grelha de valores para as constantes C (parâmetro de penalização) e γ.

Permaneceu-se com o mesmo Kernel (Radial) no processo de ajuste dos parâmetros dos modelos criados. A grelha de valores testados para C foi: 1, 10^1, 10^2, 10^3, 10^4 e 10^5. Já os valores testados para γ foram: 10^{-3}, 10^{-2}, 10^{-1}, 1, 10^1, 10^2, 10^3. Os resultados para o conjunto de treino são apresentados na tabela 2, a seguir.

Tabela 2: Resultados do ajuste dos parâmetros do modelo do conjunto de treino

Melhores parâmetros			
Cost (C)		γ	
10000		0.001	
Melhor performance: 0.219697			
Cost (C)	γ	**error**	**dispersion**
1.00E+00	1.00E-03	0.457576	0.072189
1.00E+01	1.00E-03	0.221212	0.077702
1.00E+02	1.00E-03	0.254546	0.059796
1.00E+03	1.00E-03	0.287879	0.125287
1.00E+04	**1.00E-03**	**0.2197**	**0.07227**
1.00E+05	1.00E-03	0.236364	0.066347
1.00E+00	1.00E-02	0.221212	0.077702
1.00E+01	1.00E-02	0.271212	0.069102
1.00E+02	1.00E-02	0.254546	0.059796
1.00E+03	1.00E-02	0.25303	0.079744

Conforme exposto na tabela 2, acima, os parâmetros C e γ que apresentaram o melhor ajuste do modelo foram iguais a 10.000 e 0,001, respectivamente.

De posse dos parâmetros que apresentaram o melhor ajuste no modelo criado, realizou-se novamente a matriz de confusão, cujo resultado está na tabela 3, a seguir.

Tabela 3: Matriz de confusão para o modelo SVM para os dados do conjunto de teste com utilização de validação cruzada e ajuste de parâmetros

		Estimado	
		Saudável	**Transgénico**
Real	**Saudável**	8	0
	Transgénico	0	7

Assim, verifica-se a precisão de 100% na classificação dos dados do conjunto de teste para este modelo.

A tabela 4, abaixo, ilustra a soma das matrizes de confusão nos respetivos conjuntos de teste dos 5 modelos SVM criados. Desta forma, pode-se verificar a precisão agregada de 87,84% na classificação dos dados do conjunto de teste.

Tabela 4: Soma das matrizes de confusão para os cinco modelos SVM para os dados do conjunto de teste com utilização de validação cruzada e ajuste de parâmetros

		Estimado	
		Saudável	Transgénico
Real	Saudável	34	6
	Transgénico	3	31

No que se refere à precisão de cada modelo, individualmente, elas foram de

Tabela 5: Precisão apurada em cada um dos cinco modelos para o SVM

Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 5
86,67%	100%	100%	80%	73,33%

Por fim, a tabela 6 apresenta a Matriz de Confusão obtida por meio da aplicação no conjunto de dados total (74 ratos) do melhor modelo SVM obtido.

Tabela 6: Matriz de Confusão do conjunto total para o SVM

		Estimado	
		Saudável	Transgénico
Real	Saudável	40	0
	Transgénico	0	34

Após a aplicação no conjunto total do melhor modelo obtido para a classificação do elementos do conjunto de teste, obteve-se 100% de precisão.

4.5. Resultados obtidos por meio da técnica de SVM+

Os dados foram separados em conjunto de treino e conjunto de teste, similarmente ao realizado para o SVM sem utilização de informação privilegiada. Cinco modelos SVM+ foram criados, cujos parâmetros, após afinados manualmente, compuseram a classificação dos elementos do conjunto de teste. Ressalta-se que, tanto para o SVM quanto para o SVM+, o kernel utilizado foi o Radial.

A tabela 7 ilustra a matriz de confusão de um dos cinco modelos SVM+ criados para a classificação dos elementos do conjunto de teste, obtida por meio de um dos passos da validação cruzada e sem ajuste de parâmetros.

Tabela 7: Matriz de confusão para o modelo SVM+ sem ajuste de parâmetros para os dados do conjunto de teste

		Estimado	
		Saudável	**Transgénico**
Real	**Saudável**	8	0
	Transgénico	2	5

Por meio da tabela 7, nota-se que o modelo apresentou precisão de 86,67% na classificação dos dados do conjunto de teste.

Com o intuito de melhorar a performance, realizou-se o ajuste dos parâmetros do modelo SVM+. Diferentemente do modelo SVM, em que o pacote caret, do software R, possui uma função específica para fazer o ajuste dos parâmetros C e γ, para o SVM+ esta função não foi encontrado. Assim, tal ajuste foi realizado manualmente, alterando-se os valores de C e γ de forma a encontrar os modelos com maior precisão.

A tabela 8 ilustra a matriz de confusão obtida para o modelo SVM+, com a utilização das técnicas de validação cruzada e de ajuste de parâmetros, que apresentou a melhor precisão na classificação dos elementos do conjunto de teste.

Tabela 8: Matriz de confusão para o modelo SVM+ com validação cruzada para os dados do conjunto de teste

		Estimado	
		Saudável	**Transgénico**
Real	**Saudável**	8	0
	Transgénico	1	6

Pela análise da tabela 7, verifica-se que a precisão do melhor modelo SVM+ na classificação dos dados do conjunto de teste foi de 93,33%.

A tabela 9, abaixo, ilustra a soma das matrizes de confusão dos cinco modelos SVM+ criados. Desta forma, nota-se a precisão agregada de 83,78% na classificação dos dados do conjunto de teste, inferior ao percentual de 87,84% obtido pela aplicação do SVM sem utilização de informação privilegiada.

Tabela 9: Soma das matrizes de confusão para os cinco modelos SVM+ para os dados do conjunto de teste com utilização de validação cruzada e ajuste de parâmetros

		Estimado	
		Saudável	**Transgénico**
Real	**Saudável**	37	3
	Transgénico	9	25

No que diz respeito a cada um dos 5 modelos SVM+ criados, as precisões são as apresentadas na tabela 10.

Tabela 10: Precisão apurada em cada um dos modelos SVM+

Modelo 1	**Modelo 2**	**Modelo 3**	**Modelo 4**	**Modelo 5**
86,67%	93,33%	78,57%	86,67%	73,33%

Na tabela 11 é mostrada a Matriz de Confusão obtida por meio da aplicação no conjunto de dados total do melhor modelo SVM+ criado.

Tabela 11: Matriz de Confusão do conjunto total para o SVM+

		Estimado	
		Saudável	**Transgénico**
Real	**Saudável**	40	0
	Transgénico	1	33

Após a aplicação no conjunto total, ou seja, nos 74 ratos, do melhor modelo gerado para a classificação do elementos do conjunto de teste, obteve-se 98,65% de precisão, resultado ligeiramente inferior ao percentual de 100% alcançado por meio da aplicação do

SVM sem utilização de informação privilegiada. No entanto, este erro se refere ao erro da classificação de apenas um rato.

5. Conclusão

O objetivo deste trabalho consistiu na construção de um modelo fundamentado no algoritmo SVM+, capaz de classificar ratos transgénicos por meio da textura de imagem OCT extraída do olho direito. Para tal, foram considerados dados de ratos saudáveis e transgénicos.

A tomografia de coerência óptica foi utilizada para obtenção da imagem da retina dos ratos. Esta, por sua vez, foi dividida em 4 quadrantes. Após, obteve-se uma imagem de fundo por camada da retina considerada que transformou a informação em 3D em 2D. Por fim, foram aplicados 20 indicadores de análise de textura em cada um dos 4 quadrantes da imagem de fundo que foram usados como *features* para o modelo SVM.

As *features* que melhor separaram os dados foram *Inertia* (segundo quadrante), INN (terceiro quadrante), *Inertia* (primeiro quadrante), IMC2 (terceiro quadrante), *Inertia* (quarto quadrante) e *ClusterShade* (terceiro quadrante), encontradas por meio do teste-t ou de Mann-Whitney, conforme os dados tivessem ou não tivessem distribuição Normal, respetivamente, no grupo dos ratos saudáveis e no grupo dos ratos transgénicos. O teste de Kolmogorov-Smirnov foi utilizado para verificar se a distribuição das *features* entre os grupos era Normal.

Quanto aos resultados do trabalho, pôde-se verificar que, comparativamente ao SVM+, a técnica SVM sem utilização de informação privilegiada alcançou resultados melhores na classificação de ratos em saudáveis ou transgénicos. Assim, a informação relativa ao SVM+, para o caso em análise, não foi útil, mas sim prejudicou o poder de predição do modelo.

Com o intuito de explicar o resultado inferior do SVM+, alguns factores podem ser destacados: não existência de uma implementação de função de *tuning* para o SVM+; o facto de as *features* consideradas terem elevado poder discriminatório; e não se ter testado os resultados considerando menos *features* no modelo, caso em que o SVM+ parece produzir melhores resultados.

A não existência na biblioteca de R de uma implementação de função para realizar o *tuning* no SVM+ prejudicou a comparação entre os resultados das técnicas de SVM e

SVM+. Para o SVM, o pacote caret, do software R, possui uma função específica para fazer o ajuste dos parâmetros C e γ. Assim, por meio da utilização de listas de valores para os citados parâmetros, a função cria diferentes modelos SVM cujos parâmetros são as distintas combinações de C e γ e, no final, exibe C e γ que forneceram o melhor ajuste ao modelo. Para o SVM+, este ajuste foi realizado manualmente. Desta forma, o total de modelos criados por meio da combinação entre C, γ e os parâmetros referentes à informação privilegiada foi inferior ao empregado para o SVM.

No que diz respeito ao alto poder discriminatório das *features*, patente no facto de existirem várias *features* significativas, as diferenças entre grupos foram tão evidentes que o SVM obteve 100% precisão. Como consequência, não foi possível avaliar correctamente o desempenho relativo entre o SVM e o SVM+.

Em relação ao terceiro factor acima citado, destaca-se que, devido ao exíguo tempo para elaboração deste trabalho, não foi realizada a análise da capacidade de predição do SVM com menos *features*. Nesta dissertação foram utilizadas 6 *features*. Talvez o uso de menos *features* diminuísse a precisão do modelo SVM e, portanto, evidenciasse a diferença entre a utilização das técnicas de SVM com e sem informação privilegiada. A comprovação desta hipótese fica como sugestão de trabalhos futuros.

6. Bibliografia

Amami, R., Ben Ayed, D., Ellouze, N.: *Practical Selection of SVM Supervised Parameters with Different Feature Representations for Vowel Recognition*. International Journal of Digital Content Technology and its Applications 7, 418–424, 2013.

Boyd, S. P., Vandenberghe, L. *Convex Optimization.* Cambridge University Press, 2004.

Braga, A., de Carvalho, A. C. P. L. F., Ludermir, T. B. Redes Neurais Artificiais: Teoria e aplicações. Editora LTC, 2000.

Burges, J.C., "*A tutorial on Support Vector Machines for Pattern Recognition*", Data Mining and knowledge Discovery, Vol 2, pp. 121-167, 1998.

Campbell, C. An Introduction to Kernel Methods. In RJ. Howlett, & LC. Jain (Eds.), *Radial Basis Function Networks: Design and Applications* (pp. 155 - 192). Springer-Verlag Berlin, 2000.

Corinna, C., Vladimir, V. *Support vector networks*. Machine Learning, 20:273–297. Kluwer Academic Publishers, Boston, 1995

Cristianini, N., Schölkopf, B. *Support Vector Machines and other Kernel methods*. *The New Generation of Learning Machines*. American Association for Artificial Intelligence, Volume 23, Number 23, 2002.

Ferreira, H. et al. *Characterization of the Retinal Changes of the 3xTg-AD Mouse Model of Alzheimer's Disease*. In: Henriques, J., Neves, N., de Carvalho, P. (eds) XV Mediterranean Conference on Medical and Biological Engineering and Computing – MEDICON 2019. MEDICON 2019. IFMBE Proceedings, vol 76. Springer, Cham, 2019.

Fletcher, R. *Practical Methods of Optimization*. John Willey and Sons, Inc., 2nd edition, 1987.

Gauraha, N., Spjuth, O. *svmplus: Implementation of Support Vector Machines Plus* (SVM+). R package version 1.0.1, 2018. <https://CRAN.R-project.org/package=svmplus>

Gavrilov, Z., *Learning using privileged information - Learning with teacher*, MIT online. Acesso em 21/07/2019, 2012. <http://web.mit.edu/zoya/www/SVM+.pdf>

Haykin, S. *Neural Networks - A Compreensive Foundation*. Prentice-Hall, New Jersey, 2nd edition, 1999.

Haykin, S. Redes Neurais: Princípios e prática. Trad. P.M. Engel. 2. Ed. Porto Alegre: Bookman, 2001.

Hearst, M. A., Schölkopf, B., Dumais, S., Osuna, E., Platt, J. *Trends and controversies – support vector machines. IEEE Intelligent Systems*, 13(4):18-28, 1998.

Hebb, D. O. *The Organization of Behaviour: A Neuropsychological Theory*, Wiley, New York, 1948.

Herbrich, R. *Learning Kernel Classifiers: Theory and Algorithms*. MIT Press, 2002.

Hsu, C., Chang, C., Lin, C. *A practical guide to support vector classification*, 1(1), 1-16, 2010.

Kowalczyk, A. *Support Vector Machines Succinctly*. Syncfusion, Morrisville, 2017.

Kuhn, M. Contribuições de Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan and Tyler Hunt, 2019. *caret: Classification and Regression Training. R package version* 6.0-84. <https://CRAN.R-project.org/package=caret>

Liu, Y., Zheng, Y. F. *FS_SFS: A novel feature selection method for support vector machines*. Pattern Recognition, vol. 39, pp. 1333–1345, 2006.

Lorena, A.C. Investigação de estratégias para a geração de máquinas de vetores de suporte multiclasses. Tese de doutorado. Universidade de São Paulo, São Carlos, 2006.

Manning, C. D., Raghavan P., Schütze H. *An Introduction to Information Retrieval*. Cambridge University Press. Cambridge, 2009.

Marcacini, R. M. Aprendizado de máquina com informação privilegiada: abordagens para agrupamento hierárquico de textos. Tese de doutorado. Universidade de São Paulo, São Carlos, 2014.

McCormick, G.P. *Non Linear Programming: Theory, Algorithms and Applications*. John Wiley and Sons, Inc., 1983.

McCulloch, W.S.; Pitts, W. *A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics. Vol. 5, p.115-133,* 1943.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version* 1.7-3, 2019. <https://CRAN.R-project.org/package=e1071>

Minicursos de Inteligência Artificial, Jornada de Atualização Científica em Inteligência Artificial, XXIII Congresso da Sociedade Brasileira de Computação, 2003.

Minoux, M. *Mathematical Programming*. John Wiley & Sons Ltd, 1986.

MINSKY, M. L. *Some universal elements for finite automata*. In C. E. Shannon & J. McCarthy (Eds.), Automata studies. Princeton: Princeton Univer. Press. p. 117-128, 1956.

Mitchell, T. *Machine Learning*. McGraw Hill, 1997.

Müller, K. R., Mika, S., Rätsch, G., Tsuda, K., Schölkopf, B. *An Introduction to kernel-based learning algorithms. IEEE Transactions on Neural Networks,* 12(2):181-201, 2001.

Ng, A. *CS229 Lecture notes - Part V Support Vector Machines*, 2018. Acesso em 21/07/2019 <http://cs229.stanford.edu/notes/cs229-notes3.pdf>

Nied, A. Treinamento de redes neurais artificiais baseado em sistemas de estrutura variável com taxa de aprendizado adaptativa, 2007.

Palazzi, M. A., Abreu, H. F. H., Freitas, A. C. L. H., Quagliato, L. B., Freitas, J. A. H. Tomografia de coerência óptica na avaliação do retinoblastoma macular. Rev Bras Oftalmol; 74 (5): 275-8, 2015

Passerini, A., Pontil, M., Frasconi, P. *New results on error correctiong output codes of kernel machines*. IEEE Transactions on Neural Networks, 15:45-54, 2004.

Phangtriastu, M., Harefa, J., & Tanoto, D. *Comparison Between Neural Network and Support Vector Machine in Optical Character Recognition*, 2017.

Rosenblatt, F. *The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. Psychological Review, 65:386-408,* 1958.

Russell, S.J., Norvig, P. *Artificial Intelligence – a Modern Approach*. Prentice Hall, 1995.

Sarafianos, N., Vrigkas, M., Kakadiaris, I.A. *Adaptive SVM+: Learning with Privileged Information for Domain Adaptation*. Computational Biomedicine Lab, University of Houston, 2017.

Schölkopf, B. *Support Vector Learning*. R. Oldenbourg Verlag, Munich, 1997

Serranho, P., Morgado, A. M., Bernardes, R. *Optical Coherence Tomography: A Concept review*. Berlin. Springer Berlin Heidelberg, 2012.

Smola, A.J., Barlett, P., Schölkopf, B., Schuurmans, D. *Introduction to large margin classifiers*. In Smola, A. J., Barlett, P., Schölkopf, B., Schuurmans, D. Editors, *Advances in Large Margin Classifiers*, pages 1-28. MTI Press, Cambridge, 1999.

Souto, M. C. P., Lorena, A. C., Delbem, A. C. B., Carvalho, A. C. P. L. F. Técnicas de Aprendizado de Máquina para problemas de Biologia Molecular, pages 103–152, 2008.

Vapnik, V. *Estimation of dependencies based on empirical data*. Springer–Verlag, 2nd edition, 2006.

Vapnik, V. *Statistical Learning Theory*. John Wiley and Sons, 1998

Vapnik, V., Izmailov, R. *A new learning paradigm: Learning using privileged information*. Journal of Machine Learning Research 16. 2023-2049, 2015.

Vapnik, V., Vashist, A. *"A new learning paradigm: Learning using privileged information"*.Neural Networks, vol. 22, no. 5, pp. 544–557, 2009.