



**INSTITUTO
FEDERAL**

Tocantins

**INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO
TOCANTINS**

CAMPUS PALMAS

CURSO DE LICENCIATURA EM MATEMÁTICA

JODÃ GOMES BRITO

**ESTUDO DO ALGORITMO OTIMIZAÇÃO SEQUENCIAL MÍNIMA APLICADO NO
TREINAMENTO DE MÁQUINAS DE VETORES DE SUPORTE**

PALMAS - TO

2021

JODÃ GOMES BRITO

ESTUDO DO ALGORITMO OTIMIZAÇÃO SEQUENCIAL MÍNIMA APLICADO NO
TREINAMENTO DE MÁQUINAS DE VETORES DE SUPORTE

Trabalho de Conclusão de Curso apresentado
ao Curso de Licenciatura em Matemática
do *campus* Palmas do Instituto Federal de
Educação, Ciência e Tecnologia do Tocantins,
como requisito parcial à obtenção do grau de
licenciado em Matemática.

Orientador: Prof. Dr. Carlos Eduardo da
Silva Santos

PALMAS - TO

2021

Dados Internacionais de Catalogação na Publicação (CIP)
Bibliotecas do Instituto Federal do Tocantins

B862e Brito, Jodã Gomes
Estudo do Algoritmo Otimização Sequencial Mínima Aplicado no
Treinamento de Máquinas de Vetores de Suporte / Jodã Gomes Brito.
– Palmas, TO, 2021.
69 p.

Trabalho de Conclusão de Curso (Licenciatura em Matemática) –
Instituto Federal de Educação, Ciência e Tecnologia do Tocantins,
Campus Palmas, Palmas, TO, 2021.

Orientador: Dr. Carlos Eduardo da Silva Santos

1. Aprendizagem Supervisionada de Máquinas. 2. Máquinas de
Vetores de Suporte. 3. Otimização Matemática. I. Santos, Carlos
Eduardo da Silva. II. Título.

CDD 510

A reprodução total ou parcial, de qualquer forma ou por qualquer meio, deste documento é autorizada para fins
de estudo e pesquisa, desde que citada a fonte.

**Elaborado pelo sistema de geração automática de ficha catalográfica do IFTO com os dados fornecidos
pelo(a) autor(a).**



Ministério da Educação
Secretaria de Educação Profissional e Tecnológica
Instituto Federal de Educação, Ciência e Tecnologia do Tocantins
CAMPUS PALMAS
COORDENAÇÃO DO CURSO DE LICENCIATURA EM MATEMÁTICA

FOLHA DE APROVAÇÃO DE TRABALHO DE CONCLUSÃO DE CURSO DE GRADUAÇÃO

1. **TÍTULO:** ESTUDO DO ALGORITMO OTIMIZAÇÃO SEQUENCIAL MÍNIMA APLICADO NO TREINAMENTO DE MÁQUINAS DE VETORES DE SUPORTE.
2. **AUTORA:** JODÃ GOMES BRITO.
3. **ORIENTADOR:** Prof. Dr. Carlos Eduardo da Silva Santos

Trabalho de Conclusão de Curso apresentado ao Instituto Federal de Educação, Ciência e Tecnologia do Tocantins, *campus* Palmas, como parte das exigências para a conclusão do Curso de Licenciatura em Matemática.

Aprovado em 03/12/2021.

Prof. Dr. Carlos Eduardo da Silva Santos
IFTO – Campus Palmas
Presidente/Orientador

Prof. Me. Edson Luiz Kraemer
IFTO – Campus Palmas
Avaliador 1

Prof. Dr. Paulo Alexandre Oliveira
UFT – Campus Palmas
Avaliador Externo



Documento assinado eletronicamente por **Carlos Eduardo da Silva Santos, Servidor**, em 03/12/2021, às 11:32, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).

Documento assinado eletronicamente por **Edson Luiz Kraemer, Servidor**, em 03/12/2021, às 11:33, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **Paulo Alexandre Oliveira, Usuário Externo**, em 09/12/2021, às 11:24, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



A autenticidade deste documento pode ser conferida no site http://sei.iftto.edu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1462576** e o código CRC **AD43BEF9**.

Quadra 310 Sul, Lo 5, s/n, esquina com a Avenida NS 10 - Plano Diretor Sul — CEP 77.021-090
Palmas/TO — (63) 3236-4000
portal.iftto.edu.br — reitoria@iftto.edu.br

Referência: Processo nº 23236.020201/2021-06

SEI nº 1462576

Dedico este trabalho aos meus pais, Seu Josué e Dona Valmirene e às minhas super irmãs Joelza, Joildy e Jozeildes.

AGRADECIMENTOS

Agradeço primeiramente a Deus, por tudo o que já me proporcionou até este momento. Por se revelar nas pessoas que tive e continuarei a ter a oportunidade de conhecer me permitindo ser uma pessoa melhor a cada dia.

Um agradecimento muito especial à minha Mãe e meu Pai, Dona Valmirene e Seu Josué. Por me incentivarem desde cedo a estudar e me mostrarem através dos seus exemplos de superação que tenho sempre que lutar pelo meus sonhos e encarar com seriedade a vida que me foi dada.

Às minhas irmãs, Joelza, Joildy e Jozeildes. Por me fazerem uma pessoa melhor. Pela parceria que temos e pelos muitos momentos incríveis que já vivemos e que carregarei comigo sempre.

À Marina, por todos os valiosos ensinamentos que nossa relação nos proporcionou. Pelo apoio ao longo do curso e na contribuição em minha evolução pessoal.

Ao Prof. Dr. Carlos, pela incrível parceria ao longo deste ano em que trabalhamos juntos. Por sempre se mostrar disposto a me ajudar nas diversas dúvidas que tive. O Senhor foi mais que um orientador, nossos encontros me renderam aprendizagens que vão muito além do tema deste TCC.

À Lisandra, por ter tido o grande prazer de conhecer um pouco mais de sua pessoa. Pela parceria no Residência e ao longo de todo o curso.

À Mirene, pela pessoa determinada e inspiradora que é, cuja simples observação já nos permite pensar em como podemos ser melhores.

À Dyana, por ser sempre tão simpática e se mostrar disposta a me ajudar. É sempre bom ter alguém a quem contar.

Ao Mateus, por ter me fornecido os códigos do LaTeX do seu trabalho a fim de que eu pudesse verificar alguns pontos relativos à formatação do meu. E também à parceria que firmamos no PIBID.

À cada um dos professores do curso de Licenciatura em Matemática, pelos ensinamentos passados e por me ajudarem a ter a certeza de que escolhi o curso certo para mim.

Por fim, agradeço a todas as pessoas que de forma direta ou indireta fizeram a conclusão deste trabalho possível.

“Um especialista em resolver problemas deve ser dotado de duas capacidade incompatíveis - uma imaginação inquieta e uma paciente obstinação.”

(Howard W. Eves)

RESUMO

Os modelos de aprendizagem de máquina fazem parte da sociedade moderna, seja em redes sociais ou em equipamentos de uso cotidiano como, por exemplo, celulares, computadores câmeras de segurança e outros. Entretanto, os conceitos matemáticos envolvidos são negligenciados e mesmo aqueles que possuem conhecimento nas áreas envolvidas na Aprendizagem de Máquina têm dificuldades em compreender como obter ou mesmo manipular esses modelos matemáticos. Diante do exposto, este trabalho visa reunir em um único documento os conceitos básicos para obtenção de modelos baseado em Máquinas de Vetores de Suporte (*Support Vectors Machines* (SVM)) e, assim, tornar o domínio desta técnica acessível ao público interessado. O trabalho foi desenvolvido com análise de publicações clássicas, desenvolvendo exemplos didáticos passo a passo, com gráficos e tabelas dos conceitos e algoritmos envolvidos no processo. Este documento apresenta como resultado um material sucinto porém completo em relação a conceitos necessários à obtenção de modelos de SVM, desde os dados de treinamento até o modelo final propriamente dito.

Palavras-chave: Aprendizagem Supervisionada de Máquinas, Máquinas de Vetores de Suporte, *Sequential Minimal Optimization*, Otimização Matemática.

ABSTRACT

Machine learning models are part of modern society, whether in social networks or in devices such as smart phones, computers, surveillance camera and others. However, the mathematical concepts involved are neglected and even those who have knowledge in the mathematics areas involved in Machine Learning have difficulties in understanding how to obtain or even manipulate these mathematical models. Therefore, this work aims to bring together in a single document the basic concepts for obtaining models based on Support Vectors Machines (*Support Vectors Machines* (SVM)) and, thus, make this technique accessible to the interested public. This work was developed with analysis of classic publications, developing didactic examples step by step, with graphs and tables of the concepts and algorithms involved in the process. This document shows a objective and complete text about concepts necessary to get SVM models, from training data to the final model itself.

Keywords: Supervised Machine Learning, Support Vector Machines, Sequential Minimal Optimization, Mathematical Optimization.

LISTA DE FIGURAS

Figura 1 – Projeção de um vetor $\mathbf{v}$ sobre um vetor $\mathbf{u}$	18
Figura 2 – Equação do plano	18
Figura 3 – Ortogonalidade do vetor gradiente	20
Figura 4 – Gradiente: taxa de crescimento	20
Figura 5 – Pontos críticos	21
Figura 6 – Paralelismo dos vetores gradiente	22
Figura 7 – Otimização: fábrica de sapatos e bolsas (espaço de busca)	29
Figura 8 – Formação do preditor em aprendizagem supervisionada	34
Figura 9 – SVMs de Margens rígidas	35
Figura 10 – Hiperplano canônico	37
Figura 11 – SVMs de Margens Suaves	42
Figura 12 – Mapeamento do conjunto de treinamento	46
Figura 13 – Representação gráfica das restrições impostas a α_1 e α_2	50
Figura 14 – Significado gráfico da determinação dos limites L e H	51
Figura 15 – Disposição dos exemplos no $\mathbb{R}^2$	56
Figura 16 – Função (4.21) após primeira atualização	60
Figura 17 – Hiperplano após 1 iteração	61
Figura 18 – Hiperplano após 2 iterações	64
Figura 19 – Hiperplano após 2 iterações (com as margens)	66

LISTA DE TABELAS

Tabela 1 – Condições de Otimalidade	30
Tabela 2 – Funções <i>Kernels</i> mais utilizadas	47
Tabela 3 – Conjunto de treinamento	55
Tabela 4 – <i>Gram matrix kernel</i> , $K(x,y) = x \cdot y$	57
Tabela 5 – Produto dos rótulos	57
Tabela 6 – Produto dos rótulos com o produto escalar	57

SUMÁRIO

1	INTRODUÇÃO	13
2	CONCEITOS PRELIMINARES	17
2.1	Geometria Analítica	17
2.2	Cálculo Diferencial	19
2.3	Álgebra Linear	23
2.4	Otimização	27
3	ASPECTOS TEÓRICOS: MÁQUINAS DE VETORES DE SUPORTE .	33
3.1	Introdução às SVMs	33
3.2	SVM Lineares: dados linearmente separáveis	35
3.3	As SVMs aplicadas à conjuntos de dados não linearmente separáveis . .	45
3.4	Algumas Aplicações das SVMs	47
4	TREINANDO UMA SVM: PROBLEMA DE OTIMIZAÇÃO QUADRÁ- TICO	49
4.1	O Algoritmo <i>Sequential Minimal Optimization</i> (SMO)	49
4.2	Exemplo: Obtenção de uma SVM com emprego do algoritmo SMO . . .	55
5	CONSIDERAÇÕES FINAIS	67
	REFERÊNCIAS	69

1 INTRODUÇÃO

Problemas que envolvem classificação são comuns ao dia a dia do ser humano. A própria distinção visual dos objetos configura um problema de classificação. Outros exemplos são em exames clínicos na detecção de doenças, ranqueamento de candidatos em processos seletivos e classificação dos países como desenvolvidos ou subdesenvolvidos. O fato é que neste requisito o cérebro humano tem uma enorme capacidade de realizar o procedimento, visto que um grande número de variáveis é analisado quase de modo instantâneo e por conseguinte uma resposta é obtida rapidamente. Esta capacidade humana sempre impressionou a comunidade científica que estuda o assunto.

Por outro lado, quando se deseja automatizar estas tarefas percebe-se que o problema gerado tem certo grau de complexidade matemática devido principalmente à grande quantidade de variáveis envolvidas e também à necessidade de um elevado número de elementos para que se possa obter um modelo matemático eficiente quando apresentados novos dados. Neste trabalho, modelo se refere a função preditora final.

Ao longo dos anos surgiram diferentes técnicas de reconhecimento de padrões para utilização em máquinas. De acordo com Ales (2008), data de 1936, o primeiro algoritmo com esta finalidade, desenvolvido pelo geneticista, estatístico britânico Ronald Aylmer Fisher é denominado de Discriminante de Fisher e se baseia na utilização de conceitos estatísticos.

Posteriormente, nas décadas de 50 e 60, o psicólogo americano Frank Rosenblatt desenvolveu um algoritmo utilizando a ideia de aprendizagem máquina tendo como base o funcionamento do próprio cérebro humano. O algoritmo foi chamado de *Perceptron de Rosenblatt* e marcou o início dos estudos às Redes Neurais Artificiais (RNA).

No decorrer das décadas seguintes outras técnicas foram criadas para a obtenção de resultados melhores. Dentre elas se encontram as Máquinas de Vetores de Suporte (SVMs) desenvolvida por Vladimir Vapnik a partir de seus estudos da *Teoria da Aprendizagem Estatística* (TAE) ¹ (VAPNIK, 1995). É baseada na metodologia de aprendizagem supervisionada, onde a partir de um conjunto de dados de treinamento previamente definidos tem-se a geração de um modelo matemático que visa a separação desses dados por meio de hiperplanos.

O problema central consiste em definir um hiperplano separador que gere uma margem ótima de modo a minimizar erros cometidos pelo preditor. A técnica vem sendo objeto de extenso

¹ Pode ser entendida como um ramo da Aprendizagem de Máquinas que explora conceitos estáticos na investigação dos modelos de predição estudados.

estudo ao longo das últimas décadas, obtendo-se com isso diversos avanços (ALES, 2008).

O treinamento de uma SVM consiste em resolver um problema de otimização quadrático convexo e que tem como característica, uma solução ótima única (LORENA; CARVALHO, 2007). Porém devido à grande quantidade de variáveis envolvidas no processo é preciso escolher uma metodologia que apresente resultados satisfatórios quanto a sua resolução. Normalmente busca-se decompor o problema geral em subproblemas menores. É neste momento que o algoritmo SMO pode ser introduzido no processo, sua função será encontrar a solução do problema de otimização gerado.

O algoritmo *Sequential Minimal Optimization* (SMO) reduz o problema de treinamento a duas variáveis (o mínimo possível), assim o método do *gradiente descendente*² é empregado com sucesso em cada par de variáveis por vez (sequencial), tornando o algoritmo computacionalmente eficiente. Este algoritmo foi desenvolvido pelo cientista da computação John Platt (PLATT, 1998). Neste trabalho é detalhado o SMO, tendo em vista seu amplo emprego pela comunidade científica para o treinamento de SVMs. É apresentado também, um exemplo simples de sua aplicação a partir de um conjunto artificial de dados contendo quatro elementos.

No processo de reconhecimento de padrões são utilizados diversos conceitos de Cálculo Diferencial e Integral, Álgebra Linear, Geometria Analítica e Estatística. Portanto, a reunião destas áreas em uma única aplicação torna o processo de compreensão das técnicas de reconhecimento de padrões inacessível a maioria dos interessados, dessa forma, o desenvolvimento de um material de caráter didático visa tornar acessível este conjunto de conceitos para pessoas da área de Matemática, Engenharia e/ou Computação, de modo a contribuir para divulgação de técnicas de Aprendizagem de Máquinas. Configurando esta como uma das justificativas para a elaboração deste trabalho.

Ademais, outra justificativa reside no interesse acadêmico em aprofundar os conhecimentos relacionados às aplicações das áreas da Matemática citadas. Também é buscado compreender a respeito de assuntos relacionados à área de Ciências da Computação como uma forma de inclusão na sociedade cada vez mais tecnológica.

Este trabalho tem sua importância também no fato de que detalha alguns conceitos como por exemplo, Espaço Vetorial, Produto Interno e Produto Escalar que normalmente são apresentados aos acadêmicos, durante a graduação, sem ou com pouca associação às aplicações. Desse modo, configura um documento em que os estudantes podem visualizar uma aplicação

² Trata-se de um método de otimização onde, a cada iteração é tomada a direção do vetor gradiente (direção de maior taxa de variação possível da função).

desses conceitos em uma área de estudo relativamente recente.

O objetivo geral do estudo consiste na compreensão do algoritmo SMO aplicado às Máquinas de Vetores de Suporte. Para tanto, o texto foi dividido nos seguintes objetivos específicos:

1. Associar propriedades matemáticas a solução de problemas de otimização.
2. Compreender como a otimização matemática se relaciona com o problema de treinamento das SVMs e com o algoritmo SMO.
3. Descrever o algoritmo SMO e o treinamento das SVMs de forma didática e detalhada.
4. Descrever de forma sucinta os conceitos de Álgebra Linear, Cálculo Diferencial e Integral e Otimização empregados na construção do problema de otimização, cuja resolução consiste no treinamento de uma SVM.

Para alcançar cada um dos objetivos específicos, a metodologia empregada no trabalho foi uma pesquisa bibliográfica tendo em vista a realização de estudos de artigos publicados em periódicos sobre o assunto, assim como em livros de Matemática de nível universitário. A pesquisa em artigos foi necessária para o entendimento do funcionamento das SVMs e do Algoritmo SMO tendo em vista que a maioria dos trabalhos publicados nesta área estão no idioma inglês. Já o estudo dos livros didáticos promoveu uma melhor compreensão de conceitos matemáticos relativos à otimização. Desse modo, o estudo teve um caráter exploratório, o que segundo Gil (2002) se trata de um modelo de pesquisa usado quando a finalidade é familiarizar-se com o objeto de estudo de forma a torná-lo mais claro.

Ainda sobre a metodologia, os cálculos apresentados foram refeitos de forma detalhada e, dentro da proposta do trabalho, apresentados neste documento. Isto porque nota-se uma pouca preocupação na literatura em explicar os caminhos percorridos até se chegar aos resultados. Utilizou-se o *Software* Geogebra para gerar os gráficos e acompanhar visualmente o comportamento do algoritmo SMO e ainda foi buscado esclarecer termos técnicos por meio de notas de rodapé.

A raciocínio que fundamentou a elaboração deste trabalho baseou-se primeiramente em compreender a ideia inicial quanto ao funcionamento das SVMs, o conceito para margens rígidas, suaves e dados não linearmente separáveis. Na sequência, como ocorre o desenvolvimento do problema de treinamento das SVMs e quais suas características. Em seguida, compreender como acontece a resolução do problema usando o algoritmo SMO. Por fim, para o fechamento do caráter didático do trabalho, apresentar um exemplo simples, resolvido manualmente para melhorar a compreensão acerca do assunto.

O presente trabalho foi dividido em quatro capítulos. O primeiro é composto por esta introdução onde o tema é apresentado assim como os objetivos, metodologia e justificativas. O Capítulo 2 é dedicado a apresentar conceitos matemáticos que estão diretamente relacionados às SVMs. No Capítulo 3 é apresentado a técnica de aprendizado de máquinas, SVMs, e as adaptações do problema de otimização para cada caso, além de exemplos de aplicações dessa técnica no cotidiano. O Capítulo 4 explica o algoritmo SMO, como o mesmo resolve o problema de otimização gerado pelo treinamento de uma SVMs e apresenta um exemplo de resolução para um exemplo simples composto de apenas quatro vetores.

2 CONCEITOS PRELIMINARES

Este capítulo é dedicado à apresentação de conceitos gerais necessários para um melhor entendimento do problema otimização das SVMs e resolução deste por meio do algoritmo SMO. Desse modo, será feito um resumo dos principais conceitos relacionados às disciplinas de Geometria Analítica, Cálculo Diferencial, Álgebra Linear e Otimização, intrinsecamente relacionados a temática abordada. Acredita-se que esta postura contribuirá para o caráter didático deste relatório.

2.1 Geometria Analítica

Nesta seção será definido o produto escalar entre vetores, apresentado a equação que determina o vetor projeção ortogonal, e por fim o raciocínio para dedução da equação geral de um plano tendo como fonte Winterle (2014).

O produto escalar pode ser entendido como um operador binário que associa dois vetores a um número real. Estes vetores são tomados em um *Espaço Vetorial* cuja definição será feita na Seção 2.3.

Definição 2.1.1 *Dados dois vetores $\mathbf{u} = (x_1, x_2, \dots, x_n)$, $\mathbf{v} = (y_1, y_2, \dots, y_n)$ o produto escalar de $\mathbf{u}$ com $\mathbf{v}$ ($\mathbf{u} \cdot \mathbf{v}$) é o número real $\mathbf{u} \cdot \mathbf{v} = x_1 y_1 + x_2 y_2 + \dots + x_n y_n$.*

Para quaisquer três vetores $\mathbf{u}, \mathbf{v}, \mathbf{w}$ e um número real α , são válidas as seguintes propriedades:

P.1 $\mathbf{u} \cdot \mathbf{v} = \mathbf{v} \cdot \mathbf{u}$.

P.2 $\mathbf{u} \cdot (\mathbf{v} + \mathbf{w}) = \mathbf{u} \cdot \mathbf{v} + \mathbf{u} \cdot \mathbf{w}$ e $(\mathbf{u} + \mathbf{v}) \cdot \mathbf{w} = \mathbf{u} \cdot \mathbf{w} + \mathbf{v} \cdot \mathbf{w}$.

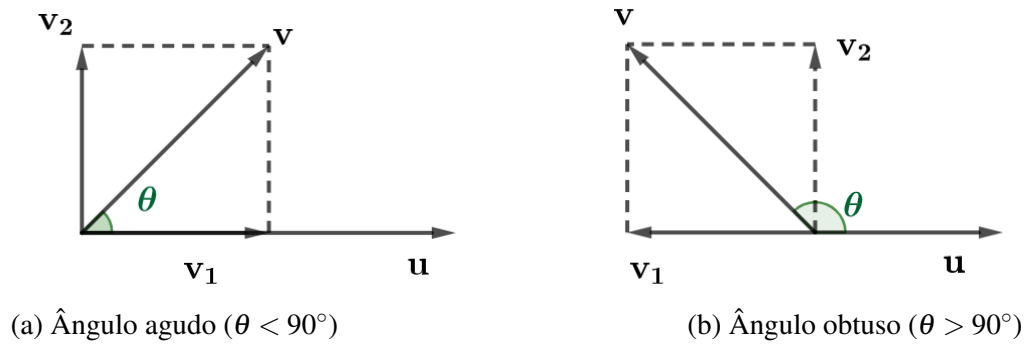
P.3 $\alpha(\mathbf{u} \cdot \mathbf{v}) = (\alpha\mathbf{u}) \cdot \mathbf{v} = \mathbf{u} \cdot (\alpha\mathbf{v})$.

P.4 $\mathbf{u} \cdot \mathbf{u} > 0$ se $\mathbf{u} \neq \mathbf{0}$ e $\mathbf{u} \cdot \mathbf{u} = 0$ se $\mathbf{u} = \mathbf{0}$, onde $\mathbf{0}$ é o vetor nulo.

P.5 $\mathbf{u} \cdot \mathbf{u} = \|\mathbf{u}\|^2$.

Sejam os vetores não nulos $\mathbf{v}$ e $\mathbf{u}$ e θ o ângulo entre eles (Figura 1). Tem-se o interesse em decompor os vetores de tal forma que $\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2$. Chama-se o vetor $\mathbf{v}_1$ de projeção ortogonal de $\mathbf{v}$ sobre $\mathbf{u}$, indicado por ($\mathbf{v}_1 = \text{proj}_{\mathbf{u}} \mathbf{v}$), o mesmo é definido pela Equação (2.1).

$$\text{proj}_{\mathbf{u}} \mathbf{v} = \left(\frac{\mathbf{v} \cdot \mathbf{u}}{\mathbf{u} \cdot \mathbf{u}} \right) \mathbf{u} \quad (2.1)$$

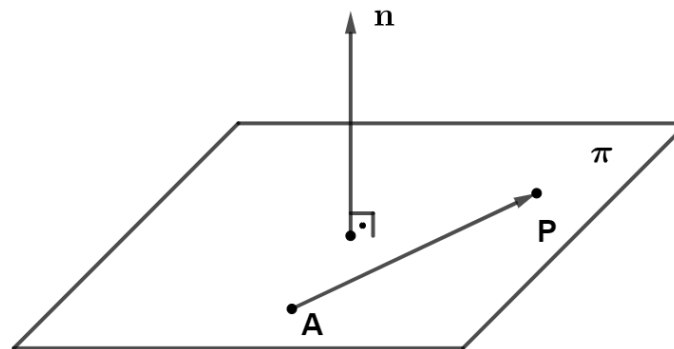
Figura 1 – Projeção de um vetor $\mathbf{v}$ sobre um vetor $\mathbf{u}$ 

Fonte: Próprio autor.

Um último conceito apresentado nesta seção diz respeito a expressão que define equação de um plano. Para efeitos didáticos considerou-se o $\mathbb{R}^3$, contudo o mesmo raciocínio é empregado para qualquer dimensão $k \in \mathbb{N}$ perdendo-se, no entanto, a capacidade visualização da geometria dos elementos.

Como pode observado na Figura 2, dado um plano π é preciso definir um vetor normal a ele ($\mathbf{n} \neq \mathbf{0}$) e um ponto A pertencente a esse plano. Pelo fato de $\mathbf{n} \perp \pi$, então $\mathbf{n}$ será ortogonal a todo vetor paralelo a π . Logo um ponto P genérico, pertencerá a π se, e somente se, o vetor $\mathbf{P} - \mathbf{A}$ for ortogonal a $\mathbf{n}$. Desse modo, qualquer ponto pertencente ao plano π satisfará a Equação (2.2), isto é, o produto escalar de $\mathbf{n}$ com $\mathbf{P} - \mathbf{A}$ será nulo.

Figura 2 – Equação do plano



Fonte: Próprio autor.

$$\mathbf{n} \cdot (\mathbf{P} - \mathbf{A}) = 0 \quad (2.2)$$

Os três conceitos abordados nesta seção são especialmente importantes para o desenvolvimento do problema de otimização gerado pelas SVMs, como poderá ser visto no Capítulo 3. Em particular, as propriedades de produto escalar são úteis para as manipulações algébricas necessárias a realização da transição do problema de otimização primal para sua forma dual. Por

sua vez, para encontrar a distância entre as margens do hiperplano separador foi utilizado a ideia de projeção de um vetor sobre o outro. O último tópico constitui a premissa central ligada às SVMs, isto é, a definição da equação de um hiperplano para qual se tenham pontos localizados a uma distância máxima do mesmo.

2.2 Cálculo Diferencial

Neste momento serão destacados três conceitos importantes para o entendimento da temática. A definição de vetor gradiente, de pontos críticos e do método dos multiplicadores de Lagrange, este último, uma ferramenta poderosa para a determinação dos pontos críticos e consequentemente os máximos e mínimos de uma função, quando condicionada a uma ou mais restrições. Ao longo deste capítulo consideramos como função, as funções contínuas definidas em um intervalo aberto $I = (a, b)$ em que $a, b \in \mathbb{R}$, $f : I \rightarrow \mathbb{R}$. A bibliografia consultada para a construção desta seção foi o livro de Thomas *et al.* (2012).

Definição 2.2.1 *O vetor gradiente (ou simplesmente gradiente) de uma função $f(x, y)$ em um ponto $P_0 = (x_0, y_0)$ é o vetor obtido por meio do cálculo das derivadas parciais de f em P_0 , sua notação é ∇f , vide Equação (2.3).*

$$\nabla f = \frac{\partial f}{\partial x} \mathbf{i} + \frac{\partial f}{\partial y} \mathbf{j} \quad (2.3)$$

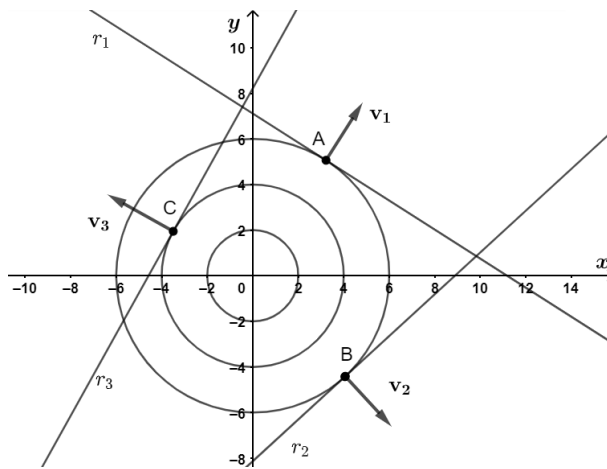
A Equação (2.3) pode ser estendida para uma função de n variáveis, isto é, o gradiente de uma função é o vetor formado, em cada um de seus componentes, por suas derivadas parciais de primeira ordem, com relação às variáveis do domínio. Além disso, o vetor gradiente, em um ponto, é sempre normal à curva de nível que passa por este ponto, informação explorada na técnica dos multiplicadores de Lagrange.

A Figura 3, destaca a ortogonalidade de ∇f para o parabolóide $f(x, y) = \frac{x^2}{4} + \frac{y^2}{4}$, cujas curvas de nível são circunferências. Os vetores $\mathbf{v}_1, \mathbf{v}_2$, e $\mathbf{v}_3$ são gradientes de f escolhidos aleatoriamente.

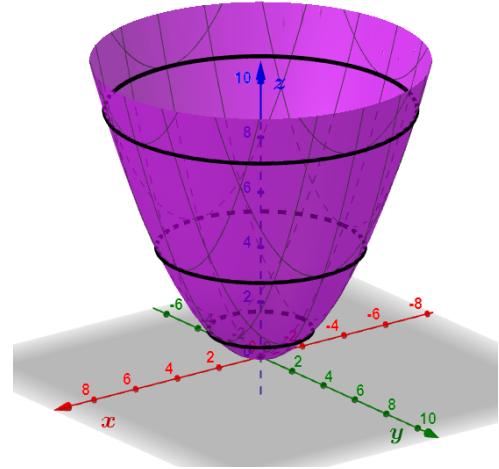
Uma função cresce mais rapidamente, em cada ponto de seu domínio, na direção do vetor gradiente e, consequentemente, decresce mais rapidamente na direção oposta ao deste vetor (THOMAS *et al.*, 2012). Isto indica que o sentido do vetor gradiente representa o menor caminho pelo qual se encontra um ponto de máximo ou mínimo local de uma função.

Na Figura 4 foi avaliada a taxa de variação da função $z = f(x, y) = \frac{x^2}{2} + \frac{y^2}{2}$ no ponto de coordenada $(1, 1, 1)$, detectando que a direção de maior variação positiva é $\nabla f = \mathbf{i} + \mathbf{j}$ enquanto

Figura 3 – Ortogonalidade do vetor gradiente



(a) Curvas de nível de $f(x,y)$. As retas r_1, r_2 e r_3 são ortogonais a v_1, v_2 e v_3 nos pontos A, B e C, respectivamente.

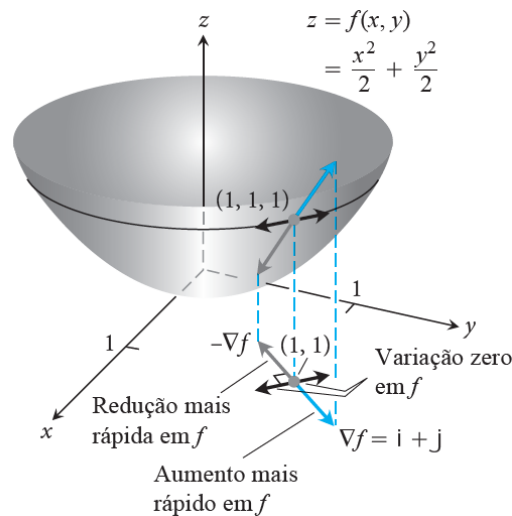


(b) Gráfico de $f(x,y)$ com destaque para as curvas de contorno em $z = 1, z = 4$ e $z = 9$.

Fonte: Próprio autor.

que a de redução mais rápida $-\nabla f$. As curvas de nível da função também são circunferências.

Figura 4 – Gradiente: taxa de crescimento



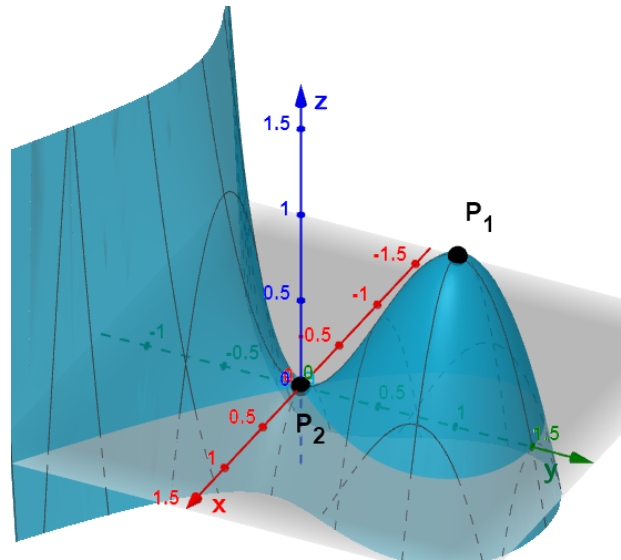
Fonte: (THOMAS *et al.*, 2012).

Definição 2.2.2 Um ponto no interior do domínio de uma função $f(x,y)$, onde $\frac{\partial f}{\partial x} = \frac{\partial f}{\partial y} = 0$ ou $\frac{\partial f}{\partial x}$ e $\frac{\partial f}{\partial y}$ não existam é chamado de ponto crítico de f .

A identificação dos pontos críticos de uma função é importante, pois indicam locais onde se encontram os valores extremos (mínimos ou máximos) ou pontos de sela de $f(x,y)$. Na Figura

5 está o gráfico da função $f(x,y) = 3y^2 - 2y^3 - 3x^2$, onde é possível observar dois pontos críticos $P_1(0, 1, 1)$ e $P_2(0, 0, 0)$, o primeiro um ponto de máximo local, o segundo um ponto de sela.

Figura 5 – Pontos críticos



Fonte: Próprio autor.

Conforme destaca Thomas *et al.* (2012) os valores de máximo e mínimo de uma função $f(x,y)$ sujeitos a uma determinada restrição $g(x,y,z) = 0$ correspondem aos pontos onde às linhas de contorno de f são tangentes às linhas de contorno da função restrição g . Este fato conduz a observação de que nestes pontos de tangência (caso existam), os vetores gradientes de f e g são paralelos e portanto múltiplos escalares. A Figura 6 destaca esta propriedade ¹ para o plano $f(x,y) = 3x + 4y$ sujeita a restrição do cilindro $g(x,y) = x^2 + y^2 = 1$.

Desse modo, é anunciado o método dos multiplicadores de Lagrange baseado na Observação 2.2.1.

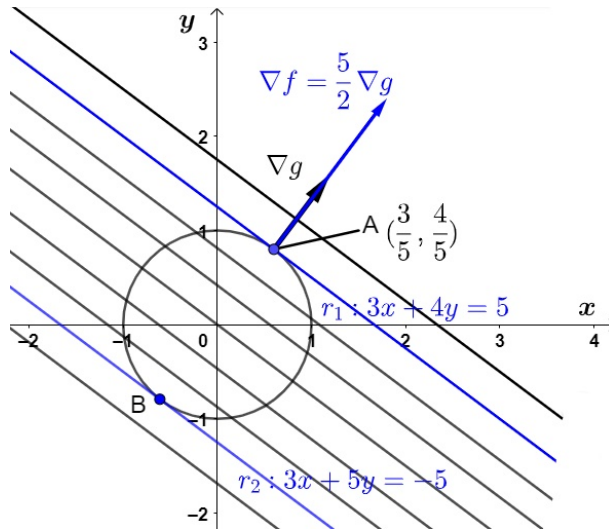
Observação 2.2.1 *Supondo que $f(x,y,z)$ e $g(x,y,z)$ sejam diferenciáveis e $\nabla g \neq \mathbf{0}$ quando $g(x,y,z) = 0$. Para determinar os valores de máximo e mínimos de f sujeito à restrição de $g(x,y,z) = 0$ (se existir), deve-se encontrar os valores x,y,z e λ que satisfaçam simultaneamente as equações (2.4) e (2.5).*

$$\nabla f = \lambda \nabla g \quad (2.4)$$

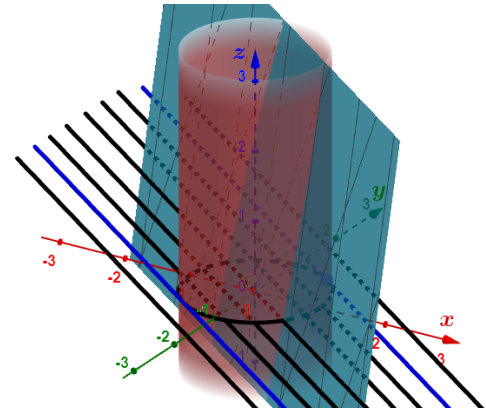
$$g(x,y,z) = 0 \quad (2.5)$$

¹ Foi enfatizado na figura somente o ponto A que maximiza $f(x,y)$, no entanto, o mesmo ocorre para o ponto B que minimiza a função.

Figura 6 – Paralelismo dos vetores gradiente



(a) Paralelismo do vetor gradiente de $f(x,y)$ (função objetivo) e de $g(x,y)$ (função restrição) no ponto A, onde as curvas se tangenciam.



(b) Representação gráfica da função objetivo (plano descrito por $f(x,y)$) e da função restrição (cilindro descrito por $g(x,y)$).

Fonte: Próprio autor.

O matemático italiano Joseph-Louis Lagrange (1736 - 1813) resumiu as condicionantes descritas nas equações (2.4) e (2.5) na forma de uma função, atualmente conhecida como *função lagrangiana* (Equação (2.6)). As condicionantes são obtidas anulando as derivadas parciais da referida função, ou seja, encontrando seus pontos críticos. Na Equação (2.6), λ é um escalar denominado *multiplicador de Lagrange*. Thomas *et al.* (2012) alerta para o fato de que é necessário cautela no uso deste método, uma vez que, um valor extremo pode não existir.

$$\mathcal{L}(x, y, z, \lambda) = f(x, y, z) - \lambda g(x, y, z) \quad (2.6)$$

A título de exemplo será aplicado o método dos multiplicadores de Lagrange para o caso da Figura 6. O problema de otimização² neste caso é representado pela Equação (2.7).

$$\begin{aligned} \text{Max } f(x, y) &= 3x + 4y, \\ \text{s.a} \\ x^2 + y^2 &= 1 \end{aligned} \quad (2.7)$$

Define-se a função lagrangiana conforme a Equação (2.8), agora com 3 variáveis no domínio.

$$\mathcal{L}(x, y, \lambda) = 3x + 4y - \lambda(x^2 + y^2 - 1) \quad (2.8)$$

² Em problemas de otimização, a notação *Max* indica que a função objetivo está sendo maximizada, assim como a notação *Min* indica a minimização da função objetivo. Por sua vez, a notação *s.a* deve ser lida como “sujeita à”.

As derivadas parciais anuladas resultam, respectivamente, nas equações (2.9), (2.10) e (2.11). Encontrá-las é necessário para determinar os pontos críticos da função lagrangeana.

$$\frac{\partial \mathcal{L}}{\partial x} = 3 - \lambda 2x = 0 \Rightarrow x = \frac{3}{2\lambda} \quad (2.9)$$

$$\frac{\partial \mathcal{L}}{\partial y} = 4 - \lambda 2y = 0 \Rightarrow y = \frac{2}{\lambda} \quad (2.10)$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = x^2 + y^2 = 1 \quad (2.11)$$

Com o objetivo de determinar o valor de λ é feita a substituição das equações (2.9) e (2.10) na Equação (2.11) obtendo-se a Equação (2.12).

$$\begin{aligned} \frac{9}{4\lambda^2} + \frac{4}{\lambda^2} &= 1 \\ 25 &= 4\lambda^2 \\ \lambda &= \pm \frac{5}{2} \end{aligned} \quad (2.12)$$

Com o resultado obtido é feita a substituição dos valores de λ nas equações (2.9) e (2.10) encontrando-se $x = \pm \frac{3}{5}$ e $y = \pm \frac{4}{5}$. Investigando estes valores em $f(x, y) = 3x + 4y$ é possível concluir que a solução para o problema de otimização é o par $\left(\frac{3}{5}, \frac{4}{5}\right)$. Ou seja, este é o ponto comum as duas funções e que maximiza $f(x, y) = 3x + 4y$.

Os três conceitos explorados nesta seção são essenciais para a obtenção do problema de otimização na forma dual, gerado pelo processo de treinamento das SVMs. Como poderá ser analisado no Capítulo 3, após a obtenção do problema primal é introduzido uma *função lagrangeana*, onde seu *vetor gradiente* é anulado a fim de se obter os *pontos críticos* da mesma. Em resumo, a aplicação desses conceitos associado a uma série de manipulações algébricas é fundamental para a obtenção da função preditora, haja vista que permite a transição do problema de otimização na sua forma primal para a sua forma dual, viabilizando com isso a aplicação do algoritmo SMO.

2.3 Álgebra Linear

Nesta seção serão descritos as definições de espaço vetorial, produto interno, matrizes simétricas, matrizes definidas positivas, matrizes semidefinidas positivas e matrizes indefinidas. Os dois primeiros conceitos foram consultados no livro de Coelho e Lourenço (2005), apenas

restringindo as definições para o conjunto dos números reais. Para a definição de matrizes simétricas usou-se o livro de Steinbruch e Winterle (1987) enquanto que para as demais foi consultada a tese de Alves (2016).

Definição 2.3.1 *Um conjunto não vazio V será denominado espaço vetorial sobre $\mathbb{R}$, se em seus elementos (vetores), forem definidas as seguintes operações:*

(A) *A soma entre vetores, isto é, para cada par $\mathbf{u}, \mathbf{v} \in V$, $\mathbf{u} + \mathbf{v} \in V$, de modo que vale as seguintes propriedades:*

A.1 $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}, \forall \mathbf{u}, \mathbf{v} \in V$ (comutatividade).

A.2 $(\mathbf{u} + \mathbf{v}) + \mathbf{w} = \mathbf{u} + (\mathbf{v} + \mathbf{w}), \forall \mathbf{u}, \mathbf{v}, \mathbf{w} \in V$ (Associatividade).

A.3 *Existe um vetor nulo em V , denotado por $\mathbf{0}$, de modo que $\mathbf{0} + \mathbf{v} = \mathbf{v}, \forall \mathbf{v} \in V$ (elemento neutro).*

A.4 *Para cada $\mathbf{v} \in V$ existe um vetor em V denotado por $-\mathbf{v}$, de modo que $\mathbf{v} + (-\mathbf{v}) = \mathbf{0}$ (simétrico).*

(M) *Para cada par $\alpha \in \mathbb{R}$ e $\mathbf{v} \in V$, corresponde um vetor $\alpha\mathbf{v} \in V$, denominado produto por escalar de α por $\mathbf{v}$, de modo que vale as seguintes propriedades:*

M.1 $(\alpha\beta)\mathbf{v} = \alpha(\beta\mathbf{v}), \forall \alpha, \beta \in \mathbb{R} \text{ e } \forall \mathbf{v} \in V$ (Associatividade).

M.2 $1\mathbf{v} = \mathbf{v}, \forall \mathbf{v} \in V$ (onde 1 é o elemento identidade de $\mathbb{R}$).

(D) *As operações dadas em (A) e (M) se distribuem, isto é, vale as seguintes propriedades:*

D.1 $\alpha(\mathbf{u} + \mathbf{v}) = \alpha\mathbf{u} + \alpha\mathbf{v}, \forall \alpha \in \mathbb{R} \text{ e } \forall \mathbf{u}, \mathbf{v} \in V$.

D.2 $(\alpha + \beta)\mathbf{v} = \alpha\mathbf{v} + \beta\mathbf{v}, \forall \alpha, \beta \in \mathbb{R} \text{ e } \forall \mathbf{v} \in V$.

Definição 2.3.2 *Sendo V um espaço vetorial definido sobre $\mathbb{R}$. Um produto interno sobre V é uma função $\langle \cdot, \cdot \rangle : V \times V \rightarrow \mathbb{R}$ em que as quatro propriedades seguintes são satisfeitas:*

P.1 $\langle \mathbf{u} + \mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{u}, \mathbf{w} \rangle + \langle \mathbf{v}, \mathbf{w} \rangle, \forall \mathbf{u}, \mathbf{v}, \mathbf{w} \in V$ (Distributividade).

P.2 $\langle \lambda\mathbf{u}, \mathbf{v} \rangle = \lambda \langle \mathbf{u}, \mathbf{v} \rangle, \forall \lambda \in \mathbb{R}, \forall \mathbf{u}, \mathbf{v} \in V$ (Associatividade).

P.3 $\langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle, \forall \mathbf{u}, \mathbf{v} \in V$ (Comutatividade).

P.4 $\langle \mathbf{u}, \mathbf{u} \rangle > 0$ se $\mathbf{u} \neq \mathbf{0}$ (Definido positivo).

Os conceitos de *Espaço Vetorial* e *Produto Interno*, são explorados ao aplicar a técnica de aprendizagem das SVMs em dados não linearmente separáveis (mais detalhe no Seção 3.3). Nestas situações faz-se necessário realizar um mapeamento do conjunto de treinamento do espaço vetorial original para um outro espaço vetorial de dimensão superior (espaço de características),

denominado *Espaço de Hilbert*³. Este mapeamento somente pode ser feito considerando que se está trabalhando em um espaço vetorial em que é definido um produto interno.

Definição 2.3.3 Dada uma matriz quadrada $A_{n \times n}$, diz-se que é uma matriz simétrica se $A^T = A$, onde A^T é a matriz transposta de $A_{n \times n}$, obtida desta através da permutação das linhas pelas colunas de mesmo índice. Em outras palavras, a matriz $A_{n \times n}$ será simétrica se os elementos dispostos simetricamente em relação à diagonal principal têm mesmo valor, isto é, $a_{ij} = a_{ji}$, com $i, j = 1, 2, 3, \dots, n$. $\forall (i \neq j)$.

A Definição 2.3.3 no contexto do treinamento de SVMs é associada a propriedade comutatividade do produto interno, vide P.3.

Definição 2.3.4 Tomando uma matriz simétrica $A_{n \times n}$ e um vetor $\mathbf{x} = (x_1, \dots, x_n)^T$ o produto $\mathbf{x}^T A \mathbf{x}$ é denominado uma forma quadrática e pode ser representado como segue na Equação (2.13).

$$\mathbf{x}^T A \mathbf{x} = \begin{pmatrix} x_1 & \dots & x_n \end{pmatrix} \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \sum_{i=1}^n \sum_{j=1}^n a_{ij} x_i x_j \quad (2.13)$$

Do conceito de formas quadráticas segue as seguintes definições:

Definição 2.3.5 Se a forma quadrática for tal que $\mathbf{x}^T A \mathbf{x} > 0$ para todo vetor $\mathbf{x}$ não nulo, então esta forma quadrática é dita definida positiva e A é uma matriz definida positiva⁴.

A matriz identidade é um exemplo clássico de matriz definida positiva, isto porque resultará sempre na soma de quadrados. O Exemplo 2.3.1 destaca o fato para o caso de I_2 .

Exemplo 2.3.1 Matriz definida positiva

$$\mathbf{x}^T I_2 \mathbf{x} = \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_1^2 + x_2^2$$

Definição 2.3.6 Se a forma quadrática for tal que $\mathbf{x}^T A \mathbf{x} \geq 0$ para todo vetor $\mathbf{x}$ não nulo, então esta forma quadrática é dita semidefinida positiva e A é uma matriz semidefinida positiva⁵.

³ O conceito de *Espaço de Hilbert* é comumente explorado em Análise Funcional e pode ser objeto de estudo posterior.

⁴ $A_{n \times n}$ é definida negativa se $-A_{n \times n}$ é definida positiva.

⁵ $A_{n \times n}$ é semidefinida negativa se $-A_{n \times n}$ é semidefinida positiva.

A matriz que resulta em uma forma quadrática representada pelo quadrado da soma finita de termos será semidefinida positiva. A matriz cujos termos são formados por valores unitários tem essa característica. No Exemplo 2.3.2 foi evidenciado este fato para o caso de uma matriz quadrada de ordem 2.

Exemplo 2.3.2 *Matriz semidefinida positiva*

$$\mathbf{x}_1^T A \mathbf{x}_1 = \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_1^2 + 2x_1x_2 + x_2^2 = (x_1 + x_2)^2$$

Definição 2.3.7 *Se a forma quadrática for tal que $\mathbf{x}^T A \mathbf{x} > 0$ para um ou mais vetores não nulos $\mathbf{x}$ e para outros $\mathbf{x}^T A \mathbf{x} < 0$ então esta forma quadrática é dita indefinida e A é uma matriz indefinida.*

No exemplo 2.3.3 tem-se que para $\mathbf{x}_2 = (1, 1)$, $\mathbf{x}_2^T A \mathbf{x}_2 = 6$. Já para $\mathbf{x}_3 = (-1, 1)$ obtém-se $\mathbf{x}_3^T A \mathbf{x}_3 = -2$.

Exemplo 2.3.3 *Matriz Indefinida*

$$\mathbf{x}_3^T A \mathbf{x}_3 = \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_1^2 + 4x_1x_2 + x_2^2$$

O conhecimento acerca das definições 2.3.5, 2.3.6 e 2.3.7 é importante para a compreensão de como é obtido o problema de otimização a ser resolvido pelo algoritmo SMO, isto é, da Equação (3.26). O somatório duplo $\sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle$ é obtido na forma do produto das matrizes $(\alpha^T)(A)(\alpha)$. Onde (α^T) é a matriz linha formada pelos valores de α_i , com $i = 1, \dots, N$, e (A) é uma matriz semidefinida positiva, também conhecida como *Gram matrix kernel* (SANTOS, 2019). Em particular, (A) é formada por todas as combinações de $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$, ou seja, $A = a_{ij}, \forall i, j = 1, \dots, N$. Enfatizamos também que devido a propriedade comutativa do produto interno a matriz A é simétrica. Por fim, (α) é o vetor coluna cujos valores são compostos por todos os α_i com $i = 1, \dots, N$.

Na Seção 2.4 são apresentados alguns conceitos importantes a respeito de otimização matemática. Sendo parte dos conceitos envolvidos na construção do problema de treinamento das SVMs, assim sendo, de leitura fundamental para leitores não familiarizados com os termos desta área do conhecimento.

2.4 Otimização

Nesta seção serão destacados alguns conceitos básicos relacionados à otimização e o que determina o Teorema de KKT (Karush-Kuhn-Tucker). O primeiro tópico faz uma breve generalização, explica-se o que é um problema de otimização com a finalidade de promover uma contextualização ao objeto de estudo deste trabalho. O segundo, diz respeito a uma consideração importante, pois permite determinar a forma dual do problema de otimização gerado pelo modelo de treinamento das SVMs.

Problemas de otimização matemática são problemas onde é requerido a maximização ou minimização de uma função definida em um domínio. No geral são estabelecidas certas restrições para as variáveis desse domínio. A título de exemplo destaca-se, na Equação (2.14), uma generalização feita por Haeser (2016).

$$\begin{aligned}
 &\text{Min } f(\mathbf{x}), \\
 &\text{s.a} \\
 &h_i(\mathbf{x}) = 0, \quad i = 1, 2, \dots, m. \\
 &g_j(\mathbf{x}) \leq 0, \quad j = 1, 2, \dots, p.
 \end{aligned} \tag{2.14}$$

Na Equação (2.14), $f(\mathbf{x})$ é denominada de função objetivo enquanto $h_i(\mathbf{x})$ são restrições de igualdade e $g_j(\mathbf{x})$ são restrições de desigualdade, todas são definidas em um mesmo domínio, $h_i(\mathbf{x})$ e $g_j(\mathbf{x})$ determinam o chamado espaço de soluções.

A Equação (2.14) representa um problema em que se busca minimizar função $f(\mathbf{x})$ dentro do espaço solução $S = \mathbb{R}^n$ definido por $h_i(\mathbf{x})$ e $g_j(\mathbf{x})$. Denominamos de solução ótima $\mathbf{x}^* \in \mathbb{R}^n$: $f(\mathbf{x}^*) \leq f(\mathbf{x}), \forall \mathbf{x} \in \mathbb{R}^n$. Conforme destaca Haeser (2016), problemas com esta formatação são comuns em física, engenharia, química, finanças, pesquisa operacional, tecnologia espacial, entre outros.

No Exemplo (2.4.2) apresentamos um caso particular da Equação (2.14) adaptado de Hamdy (2008).

Exemplo 2.4.1 *Fábrica de produção de bolsas e sapatos*

Considere uma indústria que confecciona bolsas e sapatos de couro. Esta indústria deseja maximizar o lucro L da venda de bolsas e sapatos utilizando da melhor forma possível seu estoque de linha e couro. O modelo do problema proposto é dado pela Equação (2.15), em que

as restrições se refere aos estoques de couro e de linha e $L(x,y)$ é o lucro líquido resultante da venda, de bolsas (x) e sapatos (y).

$$\begin{aligned} \text{Max } L(x,y) &= 4y + 5x \\ \text{s.a} \\ 4y + 6x &\leq 24 \quad (a) \\ 2y + x &\leq 6 \quad (b) \\ x \geq 0, y &\geq 0 \quad (c) \end{aligned} \tag{2.15}$$

Para a Equação (2.15), $L(x,y)$ representa a função objetivo, (a) se refere a restrição de estoque de couro, (b) a restrição de estoque de linha e (c) são restrições que advém do fato de não existir valor negativo para a quantidade de bolsas (x) ou sapatos (y).

Em qualquer problema de otimização é necessário distinguir três pontos importantes para iniciar o processo de busca de uma solução. Estes pontos são: compreender quais são as variáveis de decisão, qual o objetivo do problema e quais as restrições. Na Equação (2.15), as variáveis de decisão são (x) e (y) , isto é, são as variáveis que se busca determinar. Já o objetivo se refere a maximizar a função objetivo $L(x,y)$ e por fim as restrições, como já destacado, estão expressas em (a) , (b) e (c) e podemos entender como fruto de um estudo das condições do estoque da fábrica.

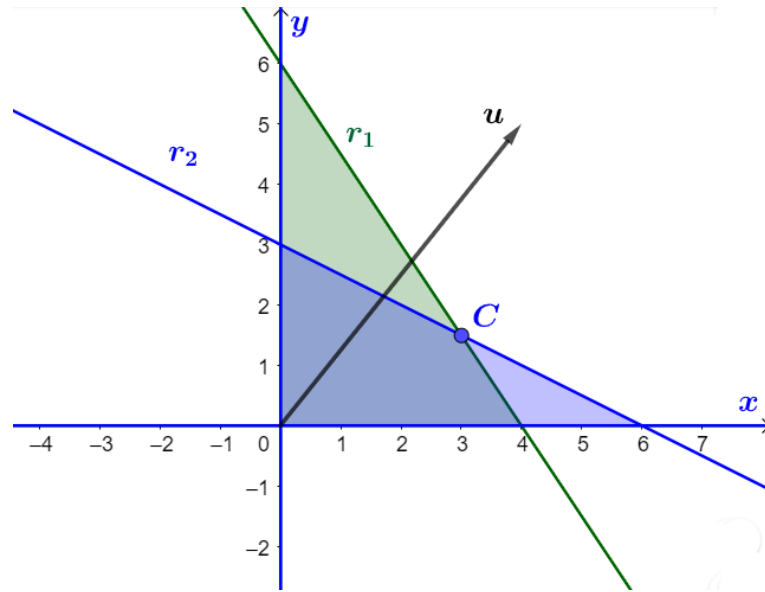
A Figura 7 evidencia as restrições da Equação (2.15). A reta r_1 refere-se a restrição (a) enquanto que a reta r_2 refere-se a restrição (b) . As restrições de não negatividade expressas em (c) reduzem o espaço de busca ao primeiro quadrante do plano xOy . Portanto, o espaço de busca final para este problema é a intersecção entre a a região azul e verde.

Uma metodologia comumente usada para a resolução deste tipo de problema é, como já mencionado neste trabalho, explorando a técnica do gradiente descendente. Na Figura 7 o gradiente está representado pelo vetor $\mathbf{u} = (4,5)$. Em específico a resolução para este problema de otimização⁶ é destacada por Hamdy (2008) como sendo o ponto $C(3.01, 1.49)$.

No geral, a resolução de problemas de otimização envolve a utilização de algoritmos computacionais devido a grande quantidade de variáveis, restrições e dificuldade em avaliar a função objetivo. Neste sentido, vários métodos vêm sendo estudados e aplicados ao longo dos anos, sendo que para resolver estes problemas é requerido a possível identificação de uma solução ótima, ou seja, que se tenha condições de avaliar a otimalidade de uma solução. Desse

⁶ Não é o objetivo aqui explicar a resolução do problema e sim como interpretá-lo.

Figura 7 – Otimização: fábrica de sapatos e bolsas (espaço de busca)



Fonte: Próprio autor.

modo, o emprego de métodos adequados, cuja escolha resulta de uma análise do espaço de busca $\mathbb{R}^n$ e da função $f(\mathbf{x})$, assim como, da capacidade de convergência desse algoritmo, torna-se fundamental.

Na Seção 2.2 foi destacado a técnica dos multiplicadores de Lagrange como uma ferramenta para resolução de problemas de otimização aplicado quando as restrições são equações. É possível estender o método de Lagrange para restrições na forma de inequações. Respeitando algumas condições (faz - se as restrições serem do tipo $\leq$), as desigualdades podem ser convertidas em equações com a utilização de variáveis de folga não negativas e assim, a solução ótima será a mesma do problema original.

O Teorema de Karush-Kuhn-Tucker (KKT), conforme destaca Hamdy (2008), desenvolve condições necessárias para a definição dos pontos críticos de uma função. Essas condições também são suficientes sob certas regras que trataremos adiante.

Seja um problema de otimização inicial como descrito na Equação (2.16), a introdução de uma variável de folga $S \geq 0$ permite definir a função lagrangiana na forma da Equação (2.17), com $\lambda_i \geq 0$. Em resumo, com a utilização do método dos multiplicadores de Lagrange a Equação (2.16) foi convertida na Equação (2.17), de modo que a resolução da segunda implica na resolução da primeira.

$$\begin{aligned} &\text{Max } f(\mathbf{x}), \\ &\text{s.a} \\ &g_i(\mathbf{x}) \leq 0, \quad i = 1, 2, \dots, m. \end{aligned} \tag{2.16}$$

$$\text{Max } \mathcal{L}(\mathbf{x}, S, \lambda) = f(\mathbf{x}) - \lambda_i [g_i(\mathbf{x}) + S_i] \quad (2.17)$$

Segundo o Teorema de KKT, uma condição necessária para a otimalidade é que λ seja não negativa para problemas de maximização e não positiva para problemas de minimização (HAMDY, 2008). As demais condições necessárias para a determinação do ponto ótimo são descritas nas equações (2.18), (2.19) e (2.20).

$$\nabla f(\mathbf{x}) - \lambda_i \nabla g_i(\mathbf{x}) = \mathbf{0} \quad (2.18)$$

$$g_i(\mathbf{x}) \leq 0, \quad (2.19)$$

$$\lambda_i g_i(\mathbf{x}) = 0, \quad (2.20)$$

A Equação (2.18) refere-se ao cálculo de $\nabla \mathcal{L}(\mathbf{x}, S, \lambda) = 0$, ou seja, o ponto ótimo, solução do problema de otimização, será um ponto crítico. Por sua vez, as inequações (2.19) refere-se às restrições do problema que devem ser atendidas. Por fim, as equações (2.20) são chamadas, como destaca Santos (2019) de *folga complementar*.

De acordo com Hamdy (2008), as condições necessárias de KKT são também suficientes quando a função objetivo e o espaço de soluções satisfazem um dos dois conjuntos de condições resumidos na Tabela 1. Para o caso do problema primal gerado pelo treinamento das SVMs, este é de minimização, a função objetivo é convexa e o espaço de soluções, um conjunto convexo, logo, as condições de KKT são suficientes e necessárias, o que implica que a solução será um ótimo global.

Tabela 1 – Condições de Otimalidade

Sentido de Otimização	Condições	
	Função objetivo	Espaço de soluções
Maximização	Côncava	Conjunto convexo
Minimização	Convexa	Conjunto convexo

Fonte: Adaptado de Hamdy (2008).

Na Equação (2.21) temos um exemplo numérico de aplicação do Teorema de KKT para um problema de minimização (HAMDY, 2008), em que $\mathbf{x} \in \mathbb{R}^3$.

Exemplo 2.4.2 *Aplicação do Teorema de KKT*

$$\begin{aligned}
& \text{Min } f(\mathbf{x}) = x_1^2 + x_2^2 + x_3^2, \\
& \text{s.a} \\
& g_1(\mathbf{x}) = 2x_1 + x_2 - 5 \leq 0 \\
& g_2(\mathbf{x}) = x_1 + x_3 - 2 \leq 0 \\
& g_3(\mathbf{x}) = 1 - x_1 \leq 0 \\
& g_4(\mathbf{x}) = 2 - x_2 \leq 0 \\
& g_5(\mathbf{x}) = -x_3 \leq 0
\end{aligned} \tag{2.21}$$

$$(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5) \leq 0 \tag{2.22}$$

A restrição destacada na Equação (2.18) pode ser trabalhada com auxílio do produto de matrizes, em que a matriz linha, a se realizar a multiplicação, é composta pelos multiplicadores de Lagrange λ_i (com $i \in \mathbb{N}$ variando de 1 a 5) e a matriz $A_{5 \times 3}$ é gerada, em cada linha, pelos coeficientes do vetor gradiente das restrições correspondentes. Dessa forma, a Equação (2.18) passa a ter a forma da Equação (2.23), o que nos permite trabalhar com a restrição da Equação (2.18) somente a partir de três expressões (equações (2.28), (2.29) e (2.30)). Cada uma relativa a uma das coordenadas x_1, x_2 ou x_3 .

$$(2x_1, 2x_2, 2x_3) - (\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5) \begin{pmatrix} 2 & 1 & 0 \\ 1 & 0 & 1 \\ -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix} = \mathbf{0} \tag{2.23}$$

Hamdy (2008) opta por trabalhar por meio do produto de matrizes, no entanto, o que a Equação (2.23) está mostrando é basicamente o anulamento das derivadas parciais da função lagrangiana (2.24) definida a partir da Equação (2.21). Ou seja, estabelecendo as condições para a determinação dos pontos críticos da mesma.

$$\begin{aligned}
\text{Min } \mathcal{L}(x_1, x_2, x_3, \lambda_i) &= x_1^2 + x_2^2 + x_3^2 - (\lambda_1(2x_1 + x_2 - 5) + \lambda_2(x_1 + x_3 - 2) + \lambda_3(1 - x_1) \\
&+ \lambda_4(2 - x_2) + \lambda_5(-x_3))
\end{aligned} \tag{2.24}$$

Prosseguindo com a análise das demais condições de KKT tem-se que de (2.19) e (2.20) segue respectivamente as equações (2.25) e as inequações (2.26).

$$\lambda_1 g_1(\mathbf{x}) = \lambda_2 g_2(\mathbf{x}) = \lambda_3 g_3(\mathbf{x}) = \lambda_4 g_4(\mathbf{x}) = \lambda_5 g_5(\mathbf{x}) = 0 \quad (2.25)$$

$$g_1(\mathbf{x}) \leq 0; g_2(\mathbf{x}) \leq 0; g_3(\mathbf{x}) \leq 0; g_4(\mathbf{x}) \leq 0; g_5(\mathbf{x}) \leq 0 \quad (2.26)$$

Por fim as condições de KKT se reduzem a:

$$\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5 \leq 0 \quad (2.27)$$

$$2x_1 - 2\lambda_1 - \lambda_2 + \lambda_3 = 0 \quad (2.28)$$

$$2x_2 - \lambda_1 + \lambda_4 = 0 \quad (2.29)$$

$$2x_3 - \lambda_2 + \lambda_5 = 0 \quad (2.30)$$

$$\lambda_1(2x_1 + x_2 - 5) = 0 \quad (2.31)$$

$$\lambda_2(x_1 + x_3 - 2) = 0 \quad (2.32)$$

$$\lambda_3(1 - x_1) = 0 \quad (2.33)$$

$$\lambda_4(2 - x_2) = 0 \quad (2.34)$$

$$\lambda_5(-x_3) = 0 \quad (2.35)$$

$$2x_1 + x_2 \leq 5 \quad (2.36)$$

$$x_1 + x_3 \leq 2 \quad (2.37)$$

$$x_1 \geq 1 \quad (2.38)$$

$$x_2 \geq 2 \quad (2.39)$$

$$x_3 \geq 0 \quad (2.40)$$

As condições expressas em (2.27) advém de (2.22), as condições (2.28), (2.29) e (2.30) resultam de (2.23). Por sua vez, (2.25) gera as restrições de (2.31) a (2.35). Por fim, as inequações presentes em (2.36) a (2.40) existem devido (2.26). Hamdy (2008) destaca que a solução⁷ para este problema é $x_1 = 1, x_2 = 2, x_3 = 0, \lambda_1 = \lambda_2 = \lambda_5 = 0, \lambda_3 = -2, \lambda_4 = -4$. Sendo que esta é a solução que satisfaz as condições de KKT. O valor mínimo é obtido substituindo x_1, x_2, x_3 na função objetivo do problema (2.21).

⁷ Não é o objetivo aqui detalhar como o resultado foi obtido, mas sim como as condições de KKT se materializam no problema de otimização.

3 ASPECTOS TEÓRICOS: MÁQUINAS DE VETORES DE SUPORTE

Este capítulo é dedicado a descrever como é obtido o modelo matemático de treinamento das SVMs, desde a aplicação mais simples a partir de dados linearmente separáveis até situações em que os dados são não linearmente separáveis, onde é necessário a introdução de funções denominadas *Kernels*. Porém, para um melhor entendimento, é feita uma introdução acerca dos conhecimentos iniciais inerentes ao contexto de Aprendizagem de Máquinas (AM).

3.1 Introdução às SVMs

As técnicas de Aprendizagem de Máquinas são baseadas em um processo indutivo onde a partir de um conjunto inicial de dados consegue-se encontrar um algoritmo com capacidade de classificar novas informações as quais a máquina não teve contato. Segundo Lorena e Carvalho (2007), o aprendizado pode acontecer de dois modos distintos: de forma *não supervisionada* ou de forma *supervisionada*.

Para o primeiro caso, como o próprio nome sugere, não existe a presença de um professor. A aprendizagem acontece com a máquina agrupando dados, o que é feito por meio da verificação de características compartilhadas pelos mesmos. Trata-se de um método de grande uso quando o objetivo é reconhecer padrões ou predisposições no objeto de estudo que não possui rótulos *a priori*.

No caso da aprendizagem supervisionada tem-se a figura de um agente externo, um professor, que tem acesso a um conjunto de dados já devidamente classificados, os quais são denominados de conjunto de treinamento. Esta técnica de aprendizagem de máquina utiliza dados de treinamento cujo rótulos são conhecidos gerando um modelo matemático capaz de classificar dados novos, ou seja, que não pertencem ao conjunto de treinamento. A esta capacidade denominamos de capacidade de generalização.

O papel do professor no processo é avaliar a função classificadora e realizar os ajustes necessários. Os modelos matemáticos, durante o treinamento, são avaliados sistematicamente por um critério, com base nessa avaliação são modificados com o objetivo de melhorar a capacidade de generalização. As SVMs fazem parte do grupo em que o treinamento ocorre de modo supervisionado.

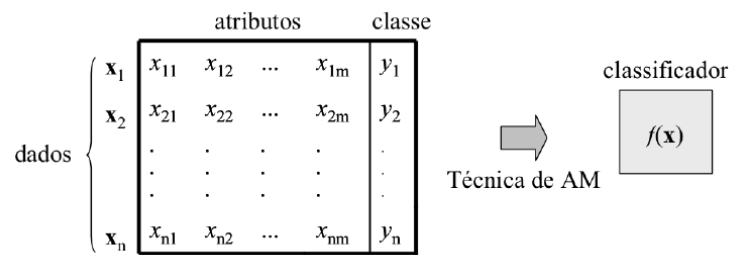
Para a aprendizagem supervisionada o conjunto de treinamento é composto por uma quantidade N de vetores denominados de *exemplos*, cada um dos vetores possuem m componentes

também chamadas de *atributos*. Cada atributo determina uma característica do exemplo.

Por fim, cada exemplo é rotulado com uma única classe, ou seja, tem sua classe devidamente identificada. Cabe destacar que o número de *atributos* dos exemplos pode variar a depender da natureza do problema, sendo uma etapa de estudo anterior ao treinamento e, de grande importância para a obtenção de uma função classificadora com boa capacidade de generalização.

Em linguagem matemática podemos indicar o conjunto de treinamento na forma $X = (\mathbf{x}_i, y_i)$ em que $\mathbf{x}_i$ é o exemplo e y_i sua respectiva classe/rótulo, com $i = 1, 2, 3, \dots, N$. O vetor $\mathbf{x}_i$ possui m atributos, ou seja, tem a forma de m coordenadas $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{i(m-1)}, x_{im})$. Ao final do processo de treinamento tem-se a máquina treinada na forma de uma função preditora $\hat{f}(\mathbf{x}_i) = \hat{y}_i$ com capacidade de classificar novos dados. Para o caso das SVMs $y_i \in \{-1, 1\}$, ou seja, o modelo de classificação é binário. A Figura 8, extraída de Lorena e Carvalho (2007), simplifica graficamente o processo que culmina no preditor para o caso da aprendizagem supervisionada.

Figura 8 – Formação do preditor em aprendizagem supervisionada



Fonte: (LORENA; CARVALHO, 2007).

Um aspecto importante que precisa ser evidenciado é a relevância que o conjunto de treinamento exerce no processo de formação da função classificadora. Escolher um conjunto de dados que seja representativo torna-se primordial para construção de um bom modelo de classificação. Neste sentido, duas situações devem ser evitadas, a presença de exemplos muito específicos e um conjunto de treinamento com poucos exemplos.

No primeiro cenário a função preditora obtida apresentará uma baixa taxa de acerto quando apresentada à novos dados, uma vez que se especializou excessivamente nos exemplos da etapa de treinamento. Quando isto ocorre diz-se que a função $\hat{f}(\mathbf{x})$ encontra-se em uma condição de super-ajustamento (*overfitting*). Por outro lado, em situações onde o preditor não consegue nem mesmo classificar corretamente os dados do conjunto de treinamento é dito que a função se encontra em uma condição de sub-ajustamento (*underfitting*), o que pode ser provocado também

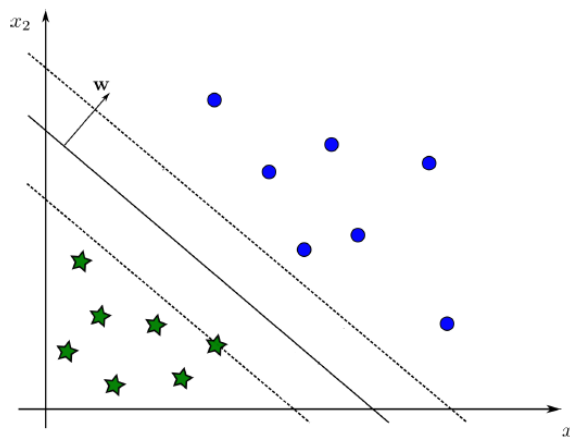
por um modelo de predição excessivamente simplório.

Em uma comparação didática Burges (1998) diz que, em condição de *overfitting*, uma função classificadora é como um botânico com memória fotográfica. Quando lhe é apresentado uma nova árvore, este conclui logo que não se trata de uma árvore pois possui um número de folhas diferente de qualquer outra da qual ele tem conhecimento. Por sua vez, uma máquina treinada em condição de *underfitting* seria como o irmão preguiçoso deste botânico. Ao ver uma nova árvore declara que realmente se trata de uma pelo simples fato de ser verde, ou seja, usando esse critério ele facilmente pode cometer erros. Em resumo, nenhuma das duas situações são desejadas pois conduzem a uma hipótese de predição com baixa capacidade de generalização tendo, portanto, a necessidade de ser encontrado um meio termo a essas duas condições.

3.2 SVM Lineares: dados linearmente separáveis

Um conjunto de dados é dito linearmente separável quando é possível determinar duas regiões em que cada exemplo fique perfeitamente localizado em uma dessas regiões. Este se configura o modelo mais simples das SVMs onde é possível definir uma margem de separação máxima chamada de hiperplano tal como representado na Figura 9. Neste modelo não existirá vetores no interior das margens, motivo pelo qual são chamados de SVMs de margens rígidas.

Figura 9 – SVMs de Margens rígidas



Fonte: Adaptado de Santos (2019).

A Equação (3.1) descreve um hiperplano, $\mathbf{w} \cdot \mathbf{x}$ representa o produto escalar entre esses dois vetores, sendo que $\mathbf{w}$ é o vetor perpendicular (normal) ao hiperplano e $b \in \mathbb{R}$.

$$\hat{f}(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b = 0 \quad (3.1)$$

A Equação (3.1) divide o espaço onde se localizam os dados de treinamento em duas regiões a saber: $\mathbf{w} \cdot \mathbf{x} + b > 0$ e $\mathbf{w} \cdot \mathbf{x} + b < 0$. Ao final do treinamento a máquina treinada pode ser entendida como uma função sinal conforme Equação (3.2).

$$g(x) = \text{sgn}(\hat{f}(\mathbf{x})) = \begin{cases} +1 & \text{se } \mathbf{w} \cdot \mathbf{x} + b > 0 \\ -1 & \text{se } \mathbf{w} \cdot \mathbf{x} + b < 0 \end{cases} \quad (3.2)$$

A seguir serão detalhados os raciocínios empregados para a obtenção do problema de otimização primal das SVMs, assim como as manipulações usadas para se alcançar sua forma dual.

A partir da Equação (3.1) é possível conseguir uma infinidade de hiperplanos equivalentes pela simples multiplicação da expressão por uma constante. Tem-se o interesse em trabalhar especificamente com o canônico, definido como aquele onde os exemplos situados mais próximos do hiperplano satisfaçam a Equação (3.3). Isto implica no fato de que qualquer ponto do conjunto de treinamento irá atender as inequações (3.4), cujo significado é o mesmo da desigualdade (3.5).

$$|\mathbf{w} \cdot \mathbf{x}_i + b| = 1 \quad (3.3)$$

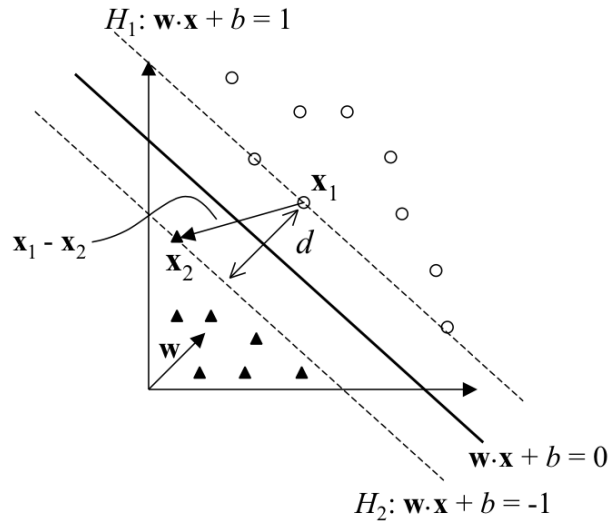
$$\begin{cases} \mathbf{w} \cdot \mathbf{x}_i + b \geq +1 & \text{se } y_i = +1 \\ \mathbf{w} \cdot \mathbf{x}_i + b \leq -1 & \text{se } y_i = -1 \end{cases} \quad (3.4)$$

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1 \geq 0 \quad (3.5)$$

Para um melhor entendimento da situação é apresentado na Figura 10 a imagem do hiperplano referido, H_1 e H_2 são suas margens. O objetivo é determinar uma expressão para a distância d , entre essas margens, o que pode ser feito projetando o vetor $\mathbf{x}_1 - \mathbf{x}_2$ sobre o vetor $\mathbf{w}$, perpendicular ao hiperplano canônico.

Como já visto na Seção 2.1, dados dois vetores quaisquer $\mathbf{u}$ e $\mathbf{v}$. A projeção do vetor $\mathbf{v}$ sobre o vetor $\mathbf{u}$ ($\text{proj}_{\mathbf{u}} \mathbf{v}$) é dado pela Equação (2.1). Alterando a notação para este caso e realizando as devidas manipulações tem-se:

Figura 10 – Hiperplano canônico



Fonte: (LORENA; CARVALHO, 2007).

$$proj_{\mathbf{w}}(\mathbf{x}_1 - \mathbf{x}_2) = \left(\frac{(\mathbf{x}_1 - \mathbf{x}_2) \cdot \mathbf{w}}{\mathbf{w} \cdot \mathbf{w}} \right) \mathbf{w}$$

$$proj_{\mathbf{w}}(\mathbf{x}_1 - \mathbf{x}_2) = \left(\frac{(\mathbf{x}_1 - \mathbf{x}_2) \cdot \mathbf{w}}{\|\mathbf{w}\|^2} \right) \mathbf{w}$$

Aplicando a distributiva no produto escalar do numerador (isto porque os pontos também podem ser considerados como vetores que partem da origem) consegue-se:

$$(\mathbf{x}_1 - \mathbf{x}_2) \cdot \mathbf{w} = \mathbf{x}_1 \cdot \mathbf{w} - \mathbf{x}_2 \cdot \mathbf{w}$$

Do hiperplano canônico segue que:

$$\mathbf{w} \cdot \mathbf{x}_1 + b - (\mathbf{w} \cdot \mathbf{x}_2 + b) = 1 - (-1)$$

$$\mathbf{w} \cdot \mathbf{x}_1 - \mathbf{w} \cdot \mathbf{x}_2 = 2$$

Assim podemos prosseguir com os cálculos:

$$proj_{\mathbf{w}}(\mathbf{x}_1 - \mathbf{x}_2) = \left(\frac{2}{\|\mathbf{w}\|^2} \right) \mathbf{w}$$

$$proj_{\mathbf{w}}(\mathbf{x}_1 - \mathbf{x}_2) = 2 \left(\frac{\mathbf{w}}{\|\mathbf{w}\| \|\mathbf{w}\|} \right)$$

$$proj_{\mathbf{w}}(\mathbf{x}_1 - \mathbf{x}_2) = 2 \left(\frac{1}{\|\mathbf{w}\|} \frac{\mathbf{w}}{\|\mathbf{w}\|} \right)$$

Temos interesse no módulo do vetor projeção, pois este representa a distância entre as margens, desse modo faz-se:

$$\begin{aligned} d &= \left\| 2 \left(\frac{1}{\|\mathbf{w}\|} \frac{\mathbf{w}}{\|\mathbf{w}\|} \right) \right\| \\ d &= \frac{2}{\|\mathbf{w}\|} \end{aligned} \quad (3.6)$$

Do desenvolvimento é possível concluir portanto que a menor distância do hiperplano separador aos dados de treinamento é $d = \frac{1}{\|\mathbf{w}\|}$. Santos (2019) adota uma metodologia diferente para determinar d , parte-se da diferença das expressões que computam a distância de cada margem em relação à origem. A maximização das margens é obtida minimizando $\|\mathbf{w}\|$. Dessa forma tem-se o problema de otimização (3.7), na sua forma conhecida como primal. É interessante observar que $\|\mathbf{w}\|$ passou a ser elevado ao quadrado e multiplicado por $\frac{1}{2}$, um artifício usado para garantir a existência do ótimo global e também para facilitar as manipulações que levam à obtenção da forma dual do mesmo. Note que não há, com esta adaptação, prejuízo quando se pensa no objetivo de minimização do módulo deste vetor.

$$\begin{aligned} \text{Min} \quad & \frac{1}{2} \|\mathbf{w}\|^2 \\ \text{s.a.} \quad & \\ & y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1 \geq 0 \quad i = 1, \dots, N, \end{aligned} \quad (3.7)$$

Conforme destaca Lorena e Carvalho (2007), o problema de otimização (3.7) é quadrático, além disso, como a função objetivo a ser minimizada é convexa e os vetores que atendem as restrições formam um conjunto convexo, existe a garantia da unicidade do mínimo global, conforme o Teorema de Karush-Kuhn-Tucker (KKT). Problemas como esse podem ser resolvidos usando a técnica dos Multiplicadores de Lagrange. Considerando isso, tem-se a *função lagrangiana* (3.8) que engloba as restrições à função objetivo e deve ser minimizada. Os pontos críticos da mesma são obtidos anulando as derivadas parciais.

$$\begin{aligned} \text{Min} \quad & \mathcal{L}(\mathbf{w}, b, \alpha_i) = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^N \alpha_i (y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1) \\ \text{s.a.} \quad & \\ & \alpha_i \geq 0 \quad i = 1, \dots, N. \end{aligned} \quad (3.8)$$

envolvidas no processo. A forma Dual, sendo positiva também justifica a necessidade dos multiplicadores de Lagrange serem não negativos como mencionado na Seção 2.4.

$$\begin{aligned}
 \text{Max} \quad & \mathcal{L}(\alpha_i) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) \\
 \text{s.a} \quad & \\
 & \sum_{i=1}^N \alpha_i y_i = 0 \\
 & \alpha_i \geq 0 \quad i = 1, \dots, N.
 \end{aligned} \tag{3.15}$$

A resolução do problema (3.15) permite encontrar os valores de α_i^* ótimos, por meio dos quais é possível calcular o parâmetro $\mathbf{w}^*$ através da Equação (3.16).

$$\mathbf{w}^* = \sum_{i=1}^N \alpha_i^* y_i \mathbf{x}_i \tag{3.16}$$

Por sua vez, o parâmetro b^* é calculado a partir da Equação $\alpha_i^* [y_i (\mathbf{w} \cdot \mathbf{x}_i + b) - 1] = 0$, resultante do Teorema de KKT, mais especificamente da Equação (2.20).

$$\begin{aligned}
 \alpha_i^* [y_i (\mathbf{w} \cdot \mathbf{x}_i + b) - 1] &= 0 \\
 \alpha_i^* y_i \mathbf{w}^* \cdot \mathbf{x}_i + \alpha_i^* y_i b^* - \alpha_i^* &= 0 \\
 \alpha_i^* y_i b^* &= \alpha_i^* - \alpha_i^* y_i \mathbf{w}^* \cdot \mathbf{x}_i \\
 b^* &= \frac{1}{y_i} - \mathbf{w}^* \cdot \mathbf{x}_i
 \end{aligned} \tag{3.17}$$

Conforme enfatiza Santos (2019, p.123). “Como as margens são equidistantes do hiperplano classificador e o termo b^* determina uma translação em relação a origem, toma-se sua média”. Dessa forma, é obtida a Equação (3.18).

$$b^* = \frac{1}{N_{SV}} \sum_{\mathbf{x}_i \in SV} \left(\frac{1}{y_i} - \mathbf{w}^* \cdot \mathbf{x}_i \right) \tag{3.18}$$

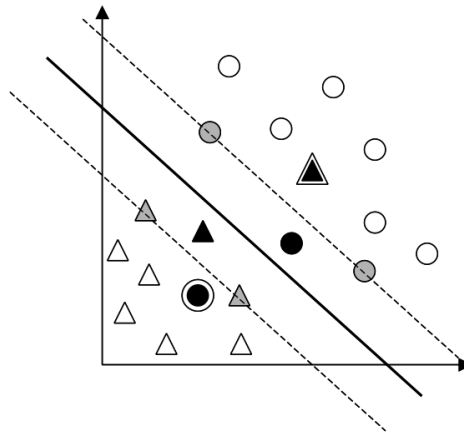
Conforme destaca Lorena e Carvalho (2007), na equação $\alpha_i^* [y_i (\mathbf{w} \cdot \mathbf{x}_i + b) - 1] = 0$, os valores de α_i somente serão diferentes de zero para exemplos posicionados sobre as margens H_1 e H_2 . Estes dados são chamados de Vetores de Suporte (SVs), apenas eles participam da determinação da equação do hiperplano separador (Note que para $\alpha_i^* = 0$, a parcela correspondente no somatório (3.16) é nula). Logo os SVs são todos os exemplos para os quais $\alpha_i^* > 0$.

Como consequência das equações (3.16) e (3.18) tem-se a função classificadora (3.19).

$$\hat{f}(\mathbf{x}) = \text{sgn} \left(\sum_{\mathbf{x}_i \in SV} \alpha_i^* y_i (\mathbf{x}_i \cdot \mathbf{x}) + b^* \right) \tag{3.19}$$

Em situações reais é raro os casos onde os exemplos podem ser perfeitamente separados pelo hiperplano. Para contornar esta limitação da aplicação das SVMs é introduzido uma variável de folga. Com a inclusão dessa variável a máquina passará a aceitar dados que estejam entre as margens do hiperplano separador assim como alguns erros de classificação. A Figura 11 destaca este caso, onde pode ser visto que há dois exemplos no interior das margens classificados corretamente e dois erros.

Figura 11 – SVMs de Margens Suaves



Fonte:(LORENA; CARVALHO, 2007).

Nestas situações as SVMs são denominadas de *margens suaves*. O problema de otimização (3.7) é adaptado para a expressão (3.20). Onde ξ_i são variáveis de folga e C um parâmetro de regularização do erro empírico¹ definido pelo projetista (SANTOS, 2019). A medida que C cresce a SVM obtida irá tender ao caso das margens rígidas, por outro lado, se o valor do parâmetro de regularização tender a zero haverá uma flexibilização excessiva das margens ocasionando perda de generalização do preditor.

$$\begin{aligned}
 \text{Min} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^N \xi_i \\
 \text{s.a.} \quad & y_i(\mathbf{w} \cdot \mathbf{x}_i + b) + \xi_i - 1 \geq 0 \\
 & \xi_i \geq 0.
 \end{aligned} \tag{3.20}$$

As manipulações matemáticas para o desenvolvimento do problema dual no caso das margens suaves são similares ao apresentado para as margens rígidas. Por esse motivo, será

¹ O erro empírico, conforme destaca Lorena e Carvalho (2007), é responsável por avaliar o desempenho da função em relação ao conjunto de treinamento.

tratado de forma mais direta.

Por meio da técnica dos multiplicadores de Lagrange é obtido a Equação (3.21) cujos pontos críticos são definidos anulando as derivadas parciais em relação a $\mathbf{w}$ (Equação (3.22)), em relação a b (Equação (3.23)) e em relação a ξ (Equação (3.24)).

$$\begin{aligned} \text{Min } \mathcal{L}(\mathbf{w}, b, \xi, \alpha_i, \mu_i) &= \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^N \xi_i - \sum_{i=1}^N \alpha_i (y_i (\mathbf{w} \cdot \mathbf{x}_i + b) + \xi_i - 1) - \sum_{i=1}^N \mu_i \xi_i \\ \text{s.a.} \\ \alpha_i, \mu_i &\geq 0 \quad i = 1, \dots, N. \end{aligned} \tag{3.21}$$

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{w}} &= \mathbf{w} - \sum_{i=1}^N \alpha_i y_i \mathbf{x}_i = 0 \\ \mathbf{w} &= \sum_{i=1}^N \alpha_i y_i \mathbf{x}_i \end{aligned} \tag{3.22}$$

$$\frac{\partial \mathcal{L}}{\partial b} = - \sum_{i=1}^N \alpha_i y_i = 0 \tag{3.23}$$

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \xi_i} &= -\alpha_i + C - \mu_i = 0 \\ \mu_i &= C - \alpha_i \\ 0 &\leq \alpha_i \leq C. \end{aligned} \tag{3.24}$$

Na Equação (3.21) realizando as substituições pelas equações (3.22) e (3.23) tem-se :

$$\begin{aligned}
\mathcal{L}(\alpha_i) &= \frac{1}{2} \mathbf{w} \cdot \mathbf{w} + C \sum_{i=1}^N \xi_i - \sum_{i=1}^N \alpha_i y_i \mathbf{w} \cdot \mathbf{x}_i - b \sum_{i=1}^N \alpha_i y_i - \sum_{i=1}^N \alpha_i \xi_i \\
&- \sum_{i=1}^N \mu_i \xi_i + \sum_{i=1}^N \alpha_i \\
\mathcal{L}(\alpha_i) &= \frac{1}{2} \left(\sum_{i=1}^N \alpha_i y_i \mathbf{x}_i \right) \cdot \left(\sum_{i=1}^N \alpha_i y_i \mathbf{x}_i \right) + C \sum_{i=1}^N \xi_i - \sum_{i=1}^N \alpha_i y_i \left(\sum_{i=1}^N \alpha_i y_i \mathbf{x}_i \right) \cdot \mathbf{x}_i \\
&- b \sum_{i=1}^N \alpha_i y_i - \sum_{i=1}^N \alpha_i \xi_i - \sum_{i=1}^N \mu_i \xi_i + \sum_{i=1}^N \alpha_i \\
\mathcal{L}(\alpha_i) &= \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) - \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) + C \sum_{i=1}^N \xi_i \\
&+ \sum_{i=1}^N \alpha_i - \sum_{i=1}^N \alpha_i \xi_i - \sum_{i=1}^N (C - \alpha_i) \xi_i \\
\mathcal{L}(\alpha_i) &= -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) - \sum_{i=1}^N (C - \alpha_i) \xi_i + \sum_{i=1}^N (C - \alpha_i) \xi_i + \sum_{i=1}^N \alpha_i \\
\mathcal{L}(\alpha_i) &= \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) \tag{3.25}
\end{aligned}$$

Desse modo, a partir das equações 3.25 e 3.24 consegue-se o problema de otimização quadrático descrito pela Equação (3.26) em sua forma dual. Os valores dos parâmetros $\mathbf{w}$ e b são obtidos de modo idêntico ao caso das SVMs de margens rígidas.

$$\begin{aligned}
\text{Max} \quad & \mathcal{L}(\alpha_i) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) \\
\text{s.a} \quad & \\
& \sum_{i=1}^N \alpha_i y_i = 0 \\
& \xi_i \geq 0 \quad i = 1, \dots, N \\
& 0 \leq \alpha_i \leq C \quad i = 1, \dots, N. \tag{3.26}
\end{aligned}$$

Como visto na Seção (2.4) o Teorema de KKT estabelece as condições necessárias e suficientes para que um problema de otimização convexo tenha solução ótima global. No caso particular da Equação (3.21) as condições de KKT são:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^N \alpha_i y_i \mathbf{x}_i = 0, \quad (3.27)$$

$$\frac{\partial \mathcal{L}}{\partial b} = - \sum_{i=1}^N \alpha_i y_i = 0, \quad (3.28)$$

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = C - \alpha_i - \mu_i = 0, \quad (3.29)$$

$$y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 + \xi_i \geq 0, \quad (3.30)$$

$$\xi_i \geq 0, \quad (3.31)$$

$$\alpha_i \geq 0, \quad (3.32)$$

$$\mu_i \geq 0, \quad (3.33)$$

$$\alpha_i [y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 + \xi_i] = 0, \quad (3.34)$$

$$\mu_i \xi_i = 0. \quad (3.35)$$

As equações de (3.27) a (3.29) são resultados da condição expressa na Equação (2.18), enquanto que as equações de (3.30) a (3.33) são resultado da Equação (2.19). Por fim, as equações (3.34) e (3.35) decorrem da Equação (2.20).

3.3 As SVMs aplicadas à conjuntos de dados não linearmente separáveis

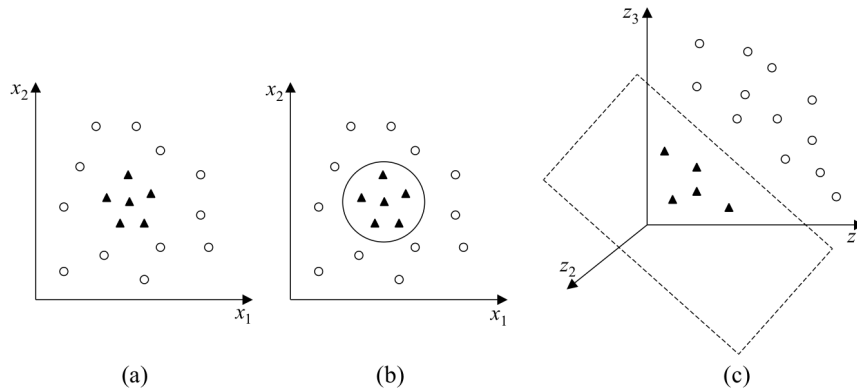
Há situações em que não é possível separar os dados por meio de um hiperplano, mesmo assumindo a ocorrência de erros por meio da inclusão de variáveis de folga. Conjuntos de dados com esta característica são chamados de não linearmente separáveis e, segundo Santos (2019), são comuns em problemas reais. O modelo das SVMs para margens suaves nestas situações se mostra insuficiente.

Quando isto acontece Lorena e Carvalho (2007) destaca que é feito um mapeamento dos dados do seu espaço original, a qual é referenciado como espaço de entrada, para um novo espaço de maior dimensão, chamado de *espaço de características* (*feature space*). Este procedimento é amparado pelo teorema de Cover que afirma:

Dado um conjunto de dados não linear no espaço de entradas X , esse teorema afirma que X pode ser transformado em um espaço de características $\mathfrak{S}$ no qual, com alta probabilidade, os dados são linearmente separáveis. Para isso duas condições devem ser satisfeitas. A primeira é que a transformação seja não linear, enquanto a segunda é que a dimensão do espaço de características seja suficientemente alta (LORENA; CARVALHO, 2007).

Um mapeamento adequado torna os dados linearmente separáveis no espaço de características como pode ser visto na Figura 12, onde houve um mapeamento do $\mathbb{R}^2$ para o $\mathbb{R}^3$. Em resumo, aplica-se uma transformação não linear $\Phi : X \rightarrow \mathfrak{S}$ aos dados de treinamento de modo a se obter um preditor que atenda satisfatoriamente a classificação de dados não lineares no espaço de entrada.

Figura 12 – Mapeamento do conjunto de treinamento



Fonte: (LORENA; CARVALHO, 2007).

Conforme destaca Santos (2019) o espaço de características Ω é um espaço de Hilbert, isto é, um espaço vetorial onde é definido o produto interno. Considerando que o algoritmo de treinamento para as SVMs de margens suaves depende do produto interno $\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$ (o produto escalar é um tipo especial de produto interno), o modelo para dados não lineares é adaptado resultando no problema de otimização (3.36).

$$\begin{aligned}
 \text{Max} \quad & \mathcal{L}(\alpha_i, \xi_i) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j (\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)) \\
 \text{s.a} \quad & \\
 & \sum_{i=1}^N \alpha_i y_i = 0 \\
 & \xi_i \geq 0 \quad i = 1, \dots, N \\
 & 0 \leq \alpha_i \leq C \quad i = 1, \dots, N.
 \end{aligned} \tag{3.36}$$

Calcular $\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$ pode não ser viável computacionalmente pelo fato da dimensão do espaço de característica, geralmente, ser elevada (SANTOS, 2019). Para contornar isto é introduzido uma função K , conhecida como núcleo *Kernel*, responsável por calcular o produto interno de $\mathbf{x}_i$ com $\mathbf{x}_j$ no espaço de características sem que seja necessário determinar $\Phi(\mathbf{x}_i)$ e $\Phi(\mathbf{x}_j)$. Assim o problema de otimização toma a forma da Equação (3.37) onde $K(\mathbf{x}_i, \mathbf{x}_j) =$

$$\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j).$$

$$\begin{aligned} \text{Max} \quad & \mathcal{L}(\alpha_i) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \\ \text{s.a} \quad & \\ & \sum_{i=1}^N \alpha_i y_i = 0 \\ & \xi_i \geq 0 \quad i = 1, \dots, N \\ & 0 \leq \alpha_i \leq C \quad i = 1, \dots, N. \end{aligned} \tag{3.37}$$

Importante notar que a introdução de uma função *kernel* não altera as características do problema de otimização, este continua quadrático e convexo e, portanto, possuindo um único ótimo global. O produto escalar também pode ser interpretado como um função *kernel*, no entanto não há mapeamento para um espaço de maior dimensão. Segundo Lorena e Carvalho (2007), as funções *kernels* mais comuns são as Polinomiais, os Gaussianos ou RBF (*Radial-Basis Function*) e os Sigmoidais (vide Tabela 2). Em cada uma delas, há parâmetros que devem ser determinados pelo projetista.

Tabela 2 – Funções *Kernels* mais utilizadas

Tipo do Kernel	Função $K(\mathbf{x}_i, \mathbf{x}_j)$	Parâmetros
Polinomial	$(\delta(\mathbf{x}_i \cdot \mathbf{x}_j) + \kappa)^d$	δ, κ, d
Gaussiano	$\exp(-\sigma \ \mathbf{x}_i - \mathbf{x}_j\ ^2)$	σ
Sigmoidal	$\tanh(\delta(\mathbf{x}_i \cdot \mathbf{x}_j) + \kappa)$	σ, κ

Fonte: Próprio autor.

A função preditora, após a substituição pela função núcleo, se apresenta na forma da Equação (3.38), (SANTOS, 2019):

$$\hat{f}(\mathbf{x}) = \text{sgn} \left(\sum_{\mathbf{x}_i \in SV} \alpha_i^* y_i K(\mathbf{x}_i, \mathbf{x}) + b \right) \tag{3.38}$$

3.4 Algumas Aplicações das SVMs

No trabalho desenvolvido por Santos (2002) há um resumo de algumas aplicações para as SVMs. Uma delas é na classificação de documentos de textos em um número pré-determinado de categorias. Neste caso, o problema maior é estudado como uma série de problemas de classificação binária, cada um referente a uma categoria de texto. Trata-se de algo útil para automação de gerenciamento de informações. A classificação é feita com base no conteúdo dos documentos.

O mesmo trabalho ainda cita o uso das Máquinas de Vetores de Suporte para reconhecimento de textos manuscritos. Destacando se tratar de uma das primeiras aplicações das SVMs. Desenvolvida inicialmente nos Estados Unidos a partir da necessidade do serviço postal americano automatizar a etapa classificação das correspondências cujos códigos de endereçamento postal eram escritos a mão.

Ainda uma terceira aplicação citada por Santos (2002) refere-se ao reconhecimento de imagens. Algo que pode se usado, por exemplo, para recuperação de informações, filtragem de dados na internet e detecção de objetos. Para esta aplicação a construção dos vetores é feita a partir da aquisição de fotos dos objetos retiradas sob diferentes ângulos e direções de iluminação.

Há também aplicações das SVMs na área da saúde. Collazo *et al.* (2009) descreve o desenvolvimento de um modelo de predição com capacidade de quantificar o risco de morte (alto ou baixo) de pacientes internados com síndrome coronariana aguda, uma doença causada pela obstrução de uma artéria coronária. As coordenadas dos vetores descrevem, neste caso, características que podem ou não contribuir para o aumento do risco de morte. Entre as estudadas estavam sexo, idade e Índice de Massa Corporal (IMC). O estudo teve sua importância no sentido de tornar mais eficaz a terapia para os pacientes classificados como sendo de alto risco. Estes são alguns exemplos de aplicações das SVMs mostrando assim a importância desta ferramenta matemática não apenas pelas suas propriedades, mas também, por sua versatilidade em diferentes aplicações.

No capítulo seguinte será detalhado o funcionamento algoritmo *Sequential Minimal Optimization* (SMO), dando ênfase ao aspecto matemático do código. Também será desenvolvido um exemplo para um conjunto artificial de 4 exemplos. O objetivo neste caso é puramente didático, uma vez que visa detalhar os processos matemáticos que ocorrem a cada iteração.

4 TREINANDO UMA SVM: PROBLEMA DE OTIMIZAÇÃO QUADRÁTICO

Treinar uma SVM consiste em solucionar um problema de programação quadrática. Neste contexto, existem variados algoritmos criados para resolver problemas com esta característica. De acordo Junior e Oliveira (2017), entre eles estão os métodos de Programação Quadrática Sequencial (SPQ, sigla em inglês). Mauri e Lorena (2009) destaca que outros algoritmos se baseiam na ideia de linearização do modelo quadrático original. Este trabalho tratará unicamente do algoritmo SMO, de modo que o mesmo será explicado em detalhes e na sequência será feito um exemplo fictício de sua aplicação para resolver um problema gerado pelo treinamento de uma SVM com base em um conjunto de treinamento contendo quatro vetores.

4.1 O Algoritmo *Sequential Minimal Optimization* (SMO)

Uma metodologia normalmente utilizada para a resolução do problema de otimização gerado a partir do treinamento das SVMs é baseado na subdivisão deste em problemas menores. Os subproblemas são então otimizados e, sob certas condições, há a garantia da convergência para a solução ótima. Seguindo este princípio, o cientista da computação americano John Platt (PLATT, 1998) desenvolveu o algoritmo o qual é conhecido como *Sequential Minimal Optimization* (SMO).

O algoritmo SMO conduz ao extremo, o princípio de divisão do problema maior em subproblemas. Conforme indica Platt (1998), é otimizado o menor problema de otimização possível. Para a expressão (3.37) o menor subproblema envolve tomar dois Multiplicadores de Lagrange. Em resumo o SMO escolhe dois multiplicadores para otimizar de forma agrupada, os outros são considerados como constantes. Encontra os valores ótimos desse par e então atualiza o sistema. Continua a fazer isso em um processo iterativo até que todos os multiplicadores sejam otimizados.

A utilização do algoritmo SMO apresenta algumas vantagens se comparado com outras metodologias de resolução. Computacionalmente se destaca, conforme enfatiza Noronha (2017), por permitir a resolução do problema na forma analítica evitando processos numéricos ou metaheurísticas¹. Outro aspecto positivo reside na pouca ocupação de memória computacional. Sobre isso Platt (1998) destaca que mesmo grandes problemas podem ser armazenados em um

¹ São técnicas que se baseiam em heurísticas capazes de resolver diferentes problemas. Enquanto que heurísticas, segundo Carvalho *et al.* (2020), são métodos de otimização que utilizam técnicas especializadas em um problema. Tem-se a intenção de escapar de pontos de máximo ou mínimo locais.

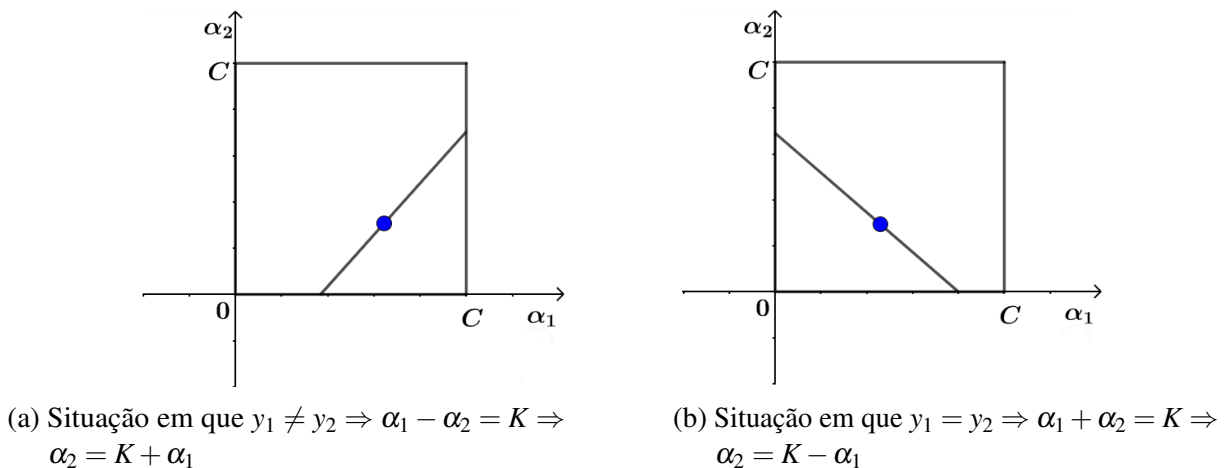
computador pessoal. Do ponto de vista matemático, em razão do problema ser quadrático e convexo, há a garantia de que existe solução ótima única, então, o teorema de KKT possibilita verificar, a cada iteração, se a solução é um ponto ótimo global.

O problema analítico resolvido pelo SMO representa apenas a segunda etapa do processo de otimização. A primeira consiste na escolha do par de Multiplicadores de Lagrange que serão atualizados. É nesta primeira etapa que reside a deficiência do algoritmo. Isto porque não existe uma maneira considerada “incorreta” ou “correta” para a escolha dos pares que serão otimizados. No entanto, conforme destaca Noronha (2017), a ordem com que são escolhidos pode acelerar ou retardar a convergência para a solução ótima. Em resumo, é escolhido a cada iteração dois multiplicadores de Lagrange dos n multiplicadores do problema, estes são as variáveis que serão otimizadas, todos os demais $(n - 2)$ alfas são considerados constantes.

Em vista do exposto, na sequência será detalhado como ocorre a otimização dos pares de multiplicadores de Lagrange pelo SMO, após estes terem sido escolhidos. Posteriormente será explicado uma das possíveis heurísticas a ser adotada para a escolha destes pares. As referências adotadas para a elaboração desta etapa do relatório são o artigo de Platt (1998) e a dissertação de Noronha (2017).

Em cada uma das etapas de otimização, os dois multiplicadores de Lagrange devem atender às duas restrições do problema geral descritas na Equação (3.37). A restrição de desigualdade os posicionam dentro de uma região quadrada de tamanho C , enquanto que a restrição de igualdade os limitam a um segmento de reta diagonal. Em outras palavras, ótimo de cada subproblema estará localizado em um segmento de linha diagonal, no interior de um quadrado de lado C (parâmetro de regularização) como destacado na Figura 13.

Figura 13 – Representação gráfica das restrições impostas a α_1 e α_2



Fonte: Próprio autor.

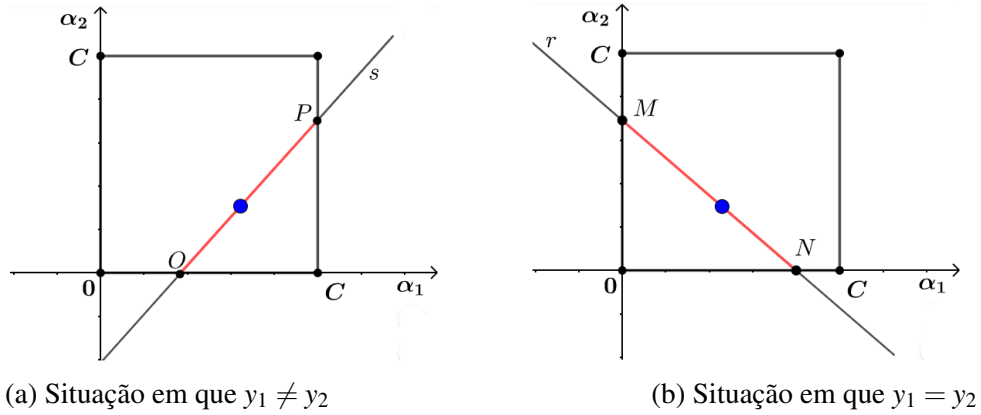
Uma vez escolhidos as duas variáveis a otimizar, o SMO primeiro calcula o segundo multiplicador de Lagrange (α_2) e em seguida determina α_1 em função do novo α_2 encontrado. O primeiro passo é definir os limites L e H de modo a buscar garantir o atendimento à restrição de desigualdade da Equação (3.37), ou seja, $0 \leq \alpha_2 \leq C$. Para tanto, caso $y_1 \neq y_2$, tem-se os limites estabelecidos pelas equações (4.1). Por sua vez, se $y_1 = y_2$ os limites tomam a forma das equações expressas em (4.2).

$$L = \max(0, \alpha_2 - \alpha_1), \quad H = \min(C, C + \alpha_2 - \alpha_1) \quad (4.1)$$

$$L = \max(0, \alpha_2 + \alpha_1 - C), \quad H = \min(C, \alpha_2 + \alpha_1) \quad (4.2)$$

A Figura 14 destaca o efeito das equações (4.1) e (4.2) no espaço de busca do par de multiplicadores de Lagrange a otimizar. Isto é, busca garantir que o valor de α_1 e α_2 permaneçam no segmento interior ao quadrado de lado C (em vermelho).

Figura 14 – Significado gráfico da determinação dos limites L e H



Fonte: Próprio autor.

A Equação (4.3) trata-se do desenvolvimento da função lagrangiana do problema (3.37) para os dois valores de α a serem otimizados. Suas derivadas parciais de segunda ordem estão destacadas nas equações (4.4), (4.5), (4.6) e (4.7).

$$\begin{aligned} \mathcal{L}(\alpha_i) &= \sum_{i=1}^2 -\frac{1}{2} \sum_{i=1}^2 \sum_{j=1}^2 \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \\ \mathcal{L}(\alpha_i) &= \alpha_1 + \alpha_2 - \frac{1}{2} (\alpha_1^2 y_1^2 K(\mathbf{x}_1, \mathbf{x}_1) + \alpha_2^2 y_2^2 K(\mathbf{x}_2, \mathbf{x}_2) + 2\alpha_1 \alpha_2 y_1 y_2 K(\mathbf{x}_1, \mathbf{x}_2)) \end{aligned} \quad (4.3)$$

$$\frac{\partial^2 \mathcal{L}}{\partial \alpha_1^2} = -k(\mathbf{x}_1, \mathbf{x}_1) \quad (4.4)$$

$$\frac{\partial^2 \mathcal{L}}{\partial \alpha_2^2} = -k(\mathbf{x}_2, \mathbf{x}_2) \quad (4.5)$$

$$\frac{\partial^2 \mathcal{L}}{\partial \alpha_1 \partial \alpha_2} = k(\mathbf{x}_1, \mathbf{x}_2) \quad (4.6)$$

$$\frac{\partial^2 \mathcal{L}}{\partial \alpha_2 \partial \alpha_1} = k(\mathbf{x}_1, \mathbf{x}_2) \quad (4.7)$$

Do somatório das derivadas parciais define a grandeza η descrita pela Equação (4.8).

$$\eta = 2k(\mathbf{x}_1, \mathbf{x}_2) - k(\mathbf{x}_1, \mathbf{x}_1) - k(\mathbf{x}_2, \mathbf{x}_2) \quad (4.8)$$

Por fim, o novo valor de α_2 que maximiza a Função Objetivo (3.37) é dado pela Equação (4.9). Onde E_1 representa o erro no exemplo $\mathbf{x}_1$, computado pela Equação (4.10) que, por sua vez, é advinda da Equação (3.38), porém antes de ocorrer a otimização dos multiplicadores de Lagrange.

$$\alpha_2^{novo} = \alpha_2 - \frac{y_2(E_1 - E_2)}{\eta} \quad (4.9)$$

$$E_1 = f(x_1) - y_1 \quad \Rightarrow \quad E_1 = \left(\sum_{\mathbf{x}_i \in SV} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}_1) + b \right) - y_1 \quad (4.10)$$

Por fim, caso ainda o valor de α_2^{novo} não esteja limitado por L e H é escolhido o limitante mais próximo, como explicitado na Equação (4.11).

$$\alpha_2^{novo} = \begin{cases} H & se \quad \alpha_2^{novo} > H \\ \alpha_2^{novo} & se \quad L \leq \alpha_2^{novo} \leq H \\ L & se \quad \alpha_2^{novo} < L \end{cases} \quad (4.11)$$

Uma vez determinado o valor de α_2^{novo} o cálculo de α_1^{novo} é feito conforme mostra a Equação (4.12).

$$\alpha_1^{novo} = \alpha_1 + y_1 y_2 (\alpha_2 - \alpha_2^{novo}) \quad (4.12)$$

Terminada a otimização de α_1 e α_2 prossegue-se para o cálculo do limiar b , este é feito em cada etapa visando o atendimento das condições de KKT. O limiar b_1 é dado pela Equação

(4.13) e forçará a saída da SVM a ser de fato y_1 quando a entrada for $\mathbf{x}_1$ (NORONHA, 2017).

$$b_1 = E_1 + y_1(\alpha_1^{novo} - \alpha_1)K(\mathbf{x}_1, \mathbf{x}_1) + y_2(\alpha_2^{novo} - \alpha_2)K(\mathbf{x}_1, \mathbf{x}_2) + b \quad (4.13)$$

Da mesma forma, o limiar b_2 é calculado pela Equação (4.14) e forçará a saída da SVM a ser y_2 quando a entrada for $\mathbf{x}_2$ (NORONHA, 2017).

$$b_2 = E_2 + y_1(\alpha_1^{novo} - \alpha_1)K(\mathbf{x}_1, \mathbf{x}_2) + y_2(\alpha_2^{novo} - \alpha_2)K(\mathbf{x}_2, \mathbf{x}_2) + b \quad (4.14)$$

É interessante observar que as equações (4.13) e (4.14) são oriundas, com algumas adaptações, da Equação (3.17). Para verificar, basta estender o somatório que determina parâmetro w^* na Equação (3.16) para os dois α s a serem otimizados. A Equação (4.15) apresenta o resultado deste processo para o caso de b_1 . A diferença reside no fato de que na Equação (4.13) foi introduzido o erro E_1 e o valor do limiar da iteração anterior b , também α_1 e α_2 são substituídos pela diferença em relação aos resultados dos multiplicadores obtidos do processo recursivo anterior. Acredita-se que estas adaptações são explicadas no estudo da teoria de convergência do algoritmo e portanto, excede o objeto de estudo deste trabalho.

$$\begin{aligned} b_1 &= y_1 - \mathbf{w} \cdot \mathbf{x}_1 \\ b_1 &= y_1 - (\alpha_1 y_1 \mathbf{x}_1 + \alpha_2 y_2 \mathbf{x}_2) \cdot \mathbf{x}_1 \\ b_1 &= y_1 - y_1 \alpha_1 (\mathbf{x}_1 \cdot \mathbf{x}_1) + \alpha_2 y_2 (\mathbf{x}_1 \cdot \mathbf{x}_2) \end{aligned} \quad (4.15)$$

Ainda, se ambos b_1 e b_2 são válidos, eles terão o mesmo valor. Por outro lado, se os novos α s estão na borda, ou seja $\alpha_1 = 0$ ou $\alpha_1 = C$ e $\alpha_2 = 0$ ou $\alpha_2 = C$, então todos os valores entre b_1 e b_2 são válidos, neste caso o SMO faz a média de b_1 e b_2 . Em resumo, o valor de b é escolhido conforme a Equação (4.16) (NORONHA, 2017).

$$b = \begin{cases} b_1 & \text{se } 0 < \alpha_1^{novo} < C \\ b_2 & \text{se } 0 < \alpha_2^{novo} < C \\ (b_1 + b_2)/2 & \text{caso contrário.} \end{cases} \quad (4.16)$$

Conforme destaca Noronha (2017), um ponto ótimo na equação (3.38) será global se, e somente se, as condições de Karush-Kuhn-Tucker (KKT) forem atendidas. Em particular as condições de KKT são determinadas por (4.17) e são uma extensão da Equação (3.34). Em resumo, $\alpha_i = 0$ para a função classificando corretamente os dados, $\alpha_i = C$ representa uma

punição para o erro da função e por fim, $0 < \alpha_i < C$ se os dados estiverem sobre as margens. O que o algoritmo SMO faz é iterar repetidas vezes até que todas as condições sejam satisfeitas.

$$\begin{aligned}\alpha_i = 0 &\Leftrightarrow y_i f(\mathbf{x}_i) \geq 1 \\ \alpha_i = C &\Leftrightarrow y_i f(\mathbf{x}_i) \leq -1 \\ 0 < \alpha_i < C &\Leftrightarrow y_i f(\mathbf{x}_i) = -1\end{aligned}\tag{4.17}$$

Foi explicado como o algoritmo SMO otimiza os pares de multiplicadores de lagrange, no entanto, anterior a este processo há a etapa de escolha dos pares a serem otimizados. Nesse sentido, não existe uma maneira ‘incorreta’ ou ‘correta’ de fazê-la, porém a ordem com que os multiplicadores são escolhidos pode alterar significativamente a velocidade de convergência para a solução ótima do problema de otimização. Na literatura são apresentados variados métodos, neste relatório será explicado o método original descrito por Platt (1998).

Existem duas heurísticas para a escolha dos pares, a primeira é composta de um loop externo sendo responsável pela escolha do primeiro multiplicador de lagrange (α_1). A segunda constitui-se de um loop interno e escolhe o segundo multiplicador de lagrange (α_2).

No loop externo, para a escolha do primeiro multiplicador de Lagrange, é feita uma iteração em todo o conjunto de treinamento a fim de encontrar algum elemento que não atenda as condições de KKT descritas na Equação (4.17). Quando um elemento é encontrado ele é então escolhido para ser otimizado. Com o intuito aumentar a velocidade da convergência, a cada loop externo, o SMO também itera nos elementos cujos multiplicadores não são limitados, isto é, possuem valores diferentes de 0 e C.

A heurística da segunda escolha está fundamentada em selecionar o α_2 que possibilite o maior passo em direção a solução do problema de otimização. Sendo o erro descrito pela Equação (4.10), o passo é entendido pelo módulo apresentado em (4.18). Assim o maior passo será dado pelos multiplicadores que possibilitarem obter o maior valor absoluto para a diferença dos erros.

$$|E_1 - E_2|\tag{4.18}$$

Noronha (2017) ainda destaca que caso a escolha do segundo multiplicador não resulte em melhora na convergência, é feito um loop em todos os α s não limitados até o momento em que o primeiro deles proporcione uma melhora na convergência. Ainda, se nenhum elemento melhorar a convergência itera-se sobre todo o conjunto de treinamento.

4.2 Exemplo: Obtenção de uma SVM com emprego do algoritmo SMO

Com a intenção de promover um maior entendimento dos aspectos matemáticos tratados até o momento, no decorrer das páginas seguintes será detalhado o treinamento de uma SVM utilizando o algoritmo SMO para um conjunto de treinamento fictício. Este, composto somente de quatro exemplos linearmente separáveis (caso mais simples das SVMs). Cabe destacar que o objetivo desta seção é puramente didático, acredita-se que irá contribuir para a fixação dos processos matemáticos generalizados ao longo das páginas anteriores. Ademais, em situações reais normalmente os dados não são linearmente separáveis e o conjunto de treinamento constitui-se de um número elevado de exemplos, o que torna o seu tratamento viável somente com o uso de recursos computacionais.

O conjunto de treinamento será composto por quatro exemplos de duas coordenadas e, portanto, podem ser visualizados no plano cartesiano. A Tabela 3 resume o conjunto de treinamento. A Figura 15 evidencia a disposição dos pontos no $\mathbb{R}^2$. Os exemplos em azul são os

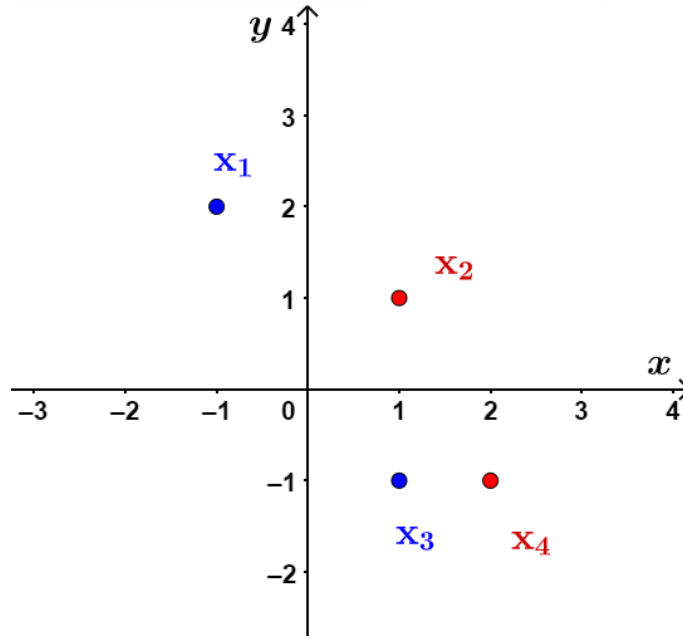
Tabela 3 – Conjunto de treinamento

EXEMPLOS ($\mathbf{x}_i$)	COORDENADAS	RÓTULOS (y_i)
$\mathbf{x}_1$	$(-1, 2)$	1
$\mathbf{x}_2$	$(1, 1)$	-1
$\mathbf{x}_3$	$(1, -1)$	1
$\mathbf{x}_4$	$(2, -1)$	-1

Fonte: Próprio autor.

que possuem rótulos iguais a 1 enquanto que os exemplos em vermelho têm rótulos iguais a -1.

É preciso definir o problema de otimização a ser resolvido pelo algoritmo SMO. Como trata-se de dados linearmente separáveis será dado pela função lagrangiana detalhada na Seção 3.2, de forma mais específica a Equação (3.15). Visando a aplicação do SMO definiu-se o problema como sendo de margens suaves admitindo $C = 1$, ou seja, $0 \leq \alpha_i \leq 1$. O primeiro passo será expandir a Equação (3.15) para o conjunto de treinamento explicitado na Tabela 3. O

Figura 15 – Disposição dos exemplos no $\mathbb{R}^2$ 

Fonte: Próprio autor.

que foi feito resultando na Equação (4.19).

$$\begin{aligned}
 \mathcal{L}(\alpha_i) &= \sum_{i=1}^4 \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^4 \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) \\
 \mathcal{L}(\alpha_i) &= \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 - \frac{1}{2} (\alpha_1 \alpha_1 y_1 y_1 (\mathbf{x}_1 \cdot \mathbf{x}_1) + \alpha_1 \alpha_2 y_1 y_2 (\mathbf{x}_1 \cdot \mathbf{x}_2) + \alpha_1 \alpha_3 y_1 y_3 (\mathbf{x}_1 \cdot \mathbf{x}_3) \\
 &+ \alpha_1 \alpha_4 y_1 y_4 (\mathbf{x}_1 \cdot \mathbf{x}_4) + \alpha_2 \alpha_1 y_2 y_1 (\mathbf{x}_2 \cdot \mathbf{x}_1) + \alpha_2 \alpha_2 y_2 y_2 (\mathbf{x}_2 \cdot \mathbf{x}_2) + \alpha_2 \alpha_3 y_2 y_3 (\mathbf{x}_2 \cdot \mathbf{x}_3) \\
 &+ \alpha_2 \alpha_4 y_2 y_4 (\mathbf{x}_2 \cdot \mathbf{x}_4) + \alpha_3 \alpha_1 y_3 y_1 (\mathbf{x}_3 \cdot \mathbf{x}_1) + \alpha_3 \alpha_2 y_3 y_2 (\mathbf{x}_3 \cdot \mathbf{x}_2) + \alpha_3 \alpha_3 y_3 y_3 (\mathbf{x}_3 \cdot \mathbf{x}_3) \\
 &+ \alpha_3 \alpha_4 y_3 y_4 (\mathbf{x}_3 \cdot \mathbf{x}_4) + \alpha_4 \alpha_1 y_4 y_1 (\mathbf{x}_4 \cdot \mathbf{x}_1) + \alpha_4 \alpha_2 y_4 y_2 (\mathbf{x}_4 \cdot \mathbf{x}_2) + \alpha_4 \alpha_3 y_4 y_3 (\mathbf{x}_4 \cdot \mathbf{x}_3) \\
 &+ \alpha_4 \alpha_4 y_4 y_4 (\mathbf{x}_4 \cdot \mathbf{x}_4))
 \end{aligned} \tag{4.19}$$

Para facilitar o desenvolvimento de $\mathcal{L}(\alpha_i)$ é conveniente estabelecer alguns cálculos auxiliares que também serão úteis durante a aplicação do algoritmo SMO. A intenção é dispor, em forma de tabelas, cada um dos valores $y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)$ a fim de que possam ser facilmente identificados. A Tabela 4 apresenta os resultados de $(\mathbf{x}_i \cdot \mathbf{x}_j)$, a Tabela 5 organiza os produtos $y_i y_j$ e, por fim, como resultado das duas anteriores, na Tabela 6 tem-se em cada linha os valores de $y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)$. É interessante observar que os valores obtidos na Tabela 4 são simétricos em relação à diagonal principal, tal fato é resultado da comutatividade, propriedade do produto escalar.

Realizada a organização em forma de tabelas, em especial com o auxílio da Tabela 6

Tabela 4 – Gram matrix kernel, $K(x,y) = x \cdot y$

$(\mathbf{x}_i \cdot \mathbf{x}_j)$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$
$\mathbf{x}_1$	5	1	-3	-4
$\mathbf{x}_2$	1	2	0	1
$\mathbf{x}_3$	-3	0	2	3
$\mathbf{x}_4$	-4	1	3	5

Fonte: Próprio autor.

Tabela 5 – Produto dos rótulos

$(y_i y_j)$	y_1	y_2	y_3	y_4
y_1	1	-1	1	-1
y_2	-1	1	-1	1
y_3	1	1	-1	-1
y_4	-1	1	-1	1

Fonte: Próprio autor.

Tabela 6 – Produto dos rótulos com o produto escalar

$y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)$	$y_i y_1 (\mathbf{x}_i \cdot \mathbf{x}_1)$	$y_i y_2 (\mathbf{x}_i \cdot \mathbf{x}_2)$	$y_i y_3 (\mathbf{x}_i \cdot \mathbf{x}_3)$	$y_i y_4 (\mathbf{x}_i \cdot \mathbf{x}_4)$
$y_1 y_j (\mathbf{x}_1 \cdot \mathbf{x}_j)$	5	-1	-3	4
$y_2 y_j (\mathbf{x}_2 \cdot \mathbf{x}_j)$	-1	2	0	1
$y_3 y_j (\mathbf{x}_3 \cdot \mathbf{x}_j)$	-3	0	2	-3
$y_4 y_j (\mathbf{x}_4 \cdot \mathbf{x}_j)$	4	1	-3	5

Fonte: Próprio autor.

substitui-se os valores da Equação (4.19) obtendo com este processo a Equação (4.20).

$$\begin{aligned} \mathcal{L}(\alpha_i) = & \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 - \frac{1}{2}(5\alpha_1\alpha_1 - \alpha_1\alpha_2 - 3\alpha_1\alpha_3 + 4\alpha_1\alpha_4 - \alpha_2\alpha_1 + 2\alpha_2\alpha_2 \\ & + 0\alpha_2\alpha_3 + \alpha_2\alpha_4 - 3\alpha_3\alpha_1 + 0\alpha_3\alpha_2 + 2\alpha_3\alpha_3 - 3\alpha_3\alpha_4 + 4\alpha_4\alpha_1 + \alpha_4\alpha_2 - 3\alpha_4\alpha_3 + 5\alpha_4\alpha_4) \end{aligned} \quad (4.20)$$

Por fim, o problema de otimização que será resolvido pelo SMO está expresso na Equação (4.21).

$$\begin{aligned} \text{Max } \mathcal{L}(\alpha_i) = & \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 - \frac{1}{2}(5\alpha_1^2 + 2\alpha_2^2 + 2\alpha_3^2 + 5\alpha_4^2 - 2\alpha_1\alpha_2 - 6\alpha_1\alpha_3 \\ & + 8\alpha_1\alpha_4 + 2\alpha_2\alpha_4 - 6\alpha_3\alpha_4) \\ \text{s.a} & \\ \alpha_1 - \alpha_2 + \alpha_3 - \alpha_4 = & 0 \\ 0 \leq \alpha_i \leq 1 & \quad i = 1, \dots, 4. \end{aligned} \quad (4.21)$$

As iterações iniciam considerando todos os multiplicadores de Lagrange nulos. Ao final de cada iteração a função classificadora é atualizada de modo que possa ser verificada as condições de KKT. A mesma, em fase de construção é dado pela Equação (4.22) e é resultado da Equação (3.38). Apenas não foi feita distinção a respeito dos Vetores de Suporte por se tratar

de uma quantidade pequena de vetores . Neste exemplo também não houve a utilização de uma função *kernel* não linear devido aos dados serem linearmente separáveis e, considerando que o enfoque será mostrar a rapidez da convergência do algoritmo SMO, não se preocupou com a determinação dos valores de ξ_i .

$$f(\mathbf{x}) = \left(\sum_{i=1}^4 \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{x}) + b \right) \quad (4.22)$$

• **Iteração 01**

– *Passo 01: Escolha do primeiro α_i*

Itera-se sobre o conjunto de treinamento a fim de encontrar algum $\mathbf{x}_i$ que viola as condições de KKT descritas em (4.17). Fazemos primeiro com $\mathbf{x}_1$. Nestas circunstâncias qualquer valor não atenderia às condições de KKT já que a função retornará valores nulos. Portanto $\mathbf{x}_1$ é escolhido.

$$\begin{aligned} y_1 f(\mathbf{x}_1) &= \sum_{i=1}^4 \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{x}_1) + b \\ y_1 f(\mathbf{x}_1) &= y_1 \alpha_1 (\mathbf{x}_1 \cdot \mathbf{x}_1) + y_2 \alpha_2 (\mathbf{x}_2 \cdot \mathbf{x}_1) + y_3 \alpha_3 (\mathbf{x}_3 \cdot \mathbf{x}_1) + y_4 \alpha_4 (\mathbf{x}_4 \cdot \mathbf{x}_1) + b \\ y_1 f(\mathbf{x}_1) &= 0 \end{aligned}$$

– *Passo 02: Escolha do segundo α_i*

Usando a Equação (4.10) busca-se o α_i que possibilite o maior valor para o módulo (4.18).

$$\begin{aligned} E_1 &= f(\mathbf{x}_1) - y_1 = 0 - 1 = -1 \\ E_2 &= f(\mathbf{x}_2) - y_2 = 0 - (-1) = 1 \\ |E_1 - E_2| &= |-1 - 1| = 2 \end{aligned}$$

Este é o maior valor que pode ser obtido (para $\mathbf{x}_3$ seria 0 e para $\mathbf{x}_4$ seria novamente 1), logo os multiplicadores escolhidos são α_1 e α_2 .

– *Passo 03: Otimizando α_1 e α_2*

Como os rótulos são diferentes analisa-se os limites com base na Equação (4.1).

$$L = \max(0, \alpha_2 - \alpha_1), \quad H = \min(C, C + \alpha_2 - \alpha_1)$$

Ambos os multiplicadores de Lagrange são nulos, logo: $0 \leq \alpha_2 \leq 1$. Na sequência é determinado a grandeza η conforme a Equação (4.8). Os valores dos produtos escalares podem ser obtidos facilmente na Tabela 4.

$$\begin{aligned} \eta &= 2(\mathbf{x}_1 \cdot \mathbf{x}_2) - (\mathbf{x}_1 \cdot \mathbf{x}_1) - (\mathbf{x}_2 \cdot \mathbf{x}_2) \\ \eta &= 2 \cdot 1 - 5 - 2 = -5 \end{aligned}$$

Após o cálculo de η , prossegue-se para a determinação de α_2^{novo} com base na Equação (4.9).

$$\begin{aligned}\alpha_2^{novo} &= \alpha_2 - \frac{y_2(E_1 - E_2)}{\eta} \\ \alpha_2^{novo} &= 0 - \frac{(-1)(-2)}{-5} \\ \alpha_2^{novo} &= \frac{2}{5}\end{aligned}$$

Após verificar, pela Equação (4.11), que o valor de α_2^{novo} não precisa ser alterado, o próximo passo é determinar o valor de α_1^{novo} , para tanto é usado a Equação (4.12).

$$\begin{aligned}\alpha_1^{novo} &= \alpha_1 + y_1 y_2 (\alpha_2 - \alpha_2^{novo}) \\ \alpha_1^{novo} &= 0 + 1 \cdot (-1) (0 - \frac{2}{5}) \\ \alpha_1^{novo} &= \frac{2}{5}\end{aligned}$$

– *Passo 04: Determinar o limiar b*

Com base nas equações (4.13) e (4.14) parte-se para a determinação do limiar b .

$$\begin{aligned}b_1 &= E_1 + y_1(\alpha_1^{novo} - \alpha_1)(\mathbf{x}_1 \cdot \mathbf{x}_1) + y_2(\alpha_2^{novo} - \alpha_2)(\mathbf{x}_1 \cdot \mathbf{x}_2) + b \\ b_1 &= -1 + \frac{2}{5} \cdot 5 - \frac{2}{5} + 0 \\ b_1 &= \frac{3}{5} \\ b_2 &= E_2 + y_1(\alpha_1^{novo} - \alpha_1)(\mathbf{x}_1 \cdot \mathbf{x}_2) + y_2(\alpha_2^{novo} - \alpha_2)(\mathbf{x}_2 \cdot \mathbf{x}_2) + b \\ b_2 &= 1 + \frac{2}{5} - \frac{2}{5} \cdot 2 + 0 \\ b_2 &= \frac{3}{5}\end{aligned}$$

Como α_1^{novo} e α_2^{novo} estão contidos no intervalo $[0, 1]$ os limiares b_1 e b_2 têm o mesmo valor.

– *Passo 05: Atualizar a função preditora (4.22).*

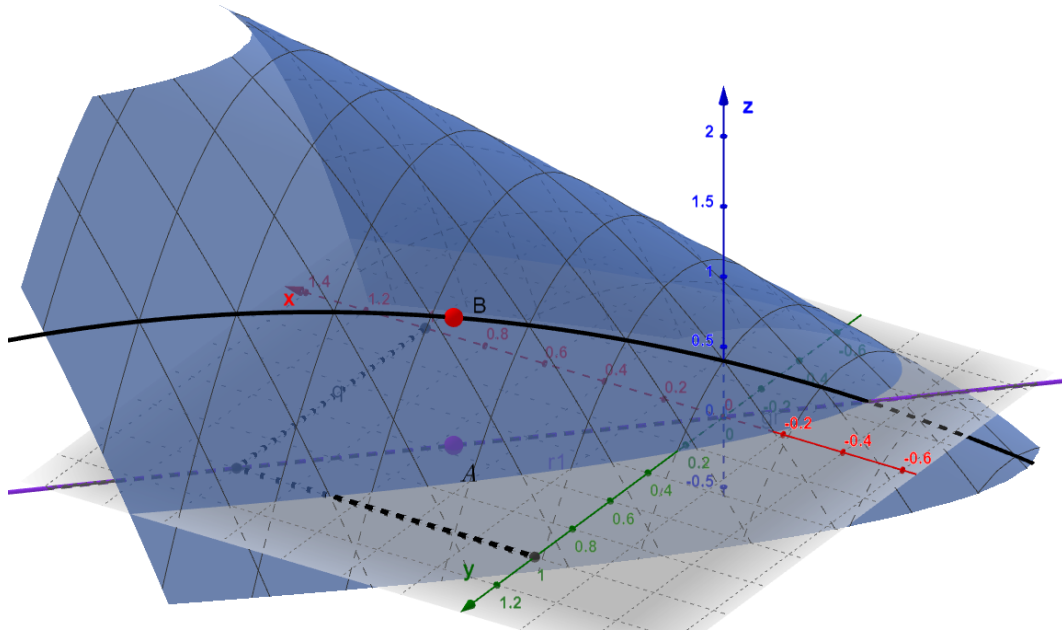
$$\begin{aligned}f(\mathbf{x}) &= y_1 \alpha_1 (\mathbf{x}_1 \cdot \mathbf{x}) + y_2 \alpha_2 (\mathbf{x}_2 \cdot \mathbf{x}) + y_3 \alpha_3 (\mathbf{x}_3 \cdot \mathbf{x}) + y_4 \alpha_4 (\mathbf{x}_4 \cdot \mathbf{x}) + b \\ f(\mathbf{x}) &= \frac{2}{5}(\mathbf{x}_1 \cdot \mathbf{x}) - \frac{2}{5}(\mathbf{x}_2 \cdot \mathbf{x}) + \frac{3}{5}\end{aligned}$$

Após a atualização de qualquer par de multiplicadores de Lagrange, uma boa maneira de compreender como de fato está acontecendo a otimização do problema, na forma dual descrito na Equação (4.21), é substituindo na função objetivo, os novos valores de α_1 e α_2 encontrados. Para este caso específico, em que se está trabalhando com 4 multiplicadores a função objetivo ficará definida em termos das variáveis α_3 e α_4 e pode ser visualizada no $\mathbb{R}^3$. Realizando as

devidas substituições na Equação (4.21), esta adquire a forma da Equação (4.23) a qual está representada na Figura 16 (considerou-se $\alpha_3 = x$ e $\alpha_4 = y$).

$$\mathcal{L}(\alpha_3, \alpha_4) = -\alpha_3^2 - \frac{5}{2}\alpha_4^2 + 3\alpha_3\alpha_4 + \frac{11}{5}\alpha_3 - \alpha_4 + \frac{2}{5} \quad (4.23)$$

Figura 16 – Função (4.21) após primeira atualização



Fonte: Próprio autor.

É interessante observar, na Figura 16, como se apresenta as restrições do problema de otimização. No domínio, a restrição de desigualdade limita o espaço de busca a uma região quadrada de lado unitário, uma vez que $C = 1$. Ainda mais, a restrição de igualdade adiciona a exigência de se trabalhar com valores que estão sobre o segmento de reta no interior deste quadrado. Este ponto é especialmente interessante, pois ao se pensar como próxima iteração a atualização dos valores de α_3 e α_4 (o que não necessariamente irá acontecer), em razão de $\alpha_1 = \alpha_2 = 0,4$, a restrição de igualdade conduz a expressão $\alpha_3 - \alpha_4 = 0$, o que gera uma reta sendo bissetriz dos quadrantes ímpares de $x0y$ como de fato pode ser observado na Figura 16. Outro ponto importante é a característica quadrática da função objetivo que garante sua maximização para a hipótese de qualquer mudança das variáveis α_3 e α_4 , respeitada as restrições do problema.

Com a finalidade de realizar uma análise de hiperplano obtido após esta primeira iteração calcula-se o vetor $\mathbf{w}$ conforme a Equação 3.16 e a variável independente b , conforme a Equação (3.18). Para o cálculo de b não é preciso considerar no somatório os rótulos y_i , uma vez que se

anularão.

$$\mathbf{w} = \sum_{i=1}^N \alpha_i y_i \mathbf{x}_i = \alpha_1 \mathbf{x}_1 - \alpha_2 \mathbf{x}_2 + \alpha_3 \mathbf{x}_3 - \alpha_4 \mathbf{x}_4$$

$$\mathbf{w} = \frac{2}{5}(-1, 2) - \frac{2}{5}(1, 1) + 0(1, -1) - 0(2, -1)$$

$$\mathbf{w} = \left(-\frac{4}{5}, \frac{2}{5}\right)$$

$$b = \frac{1}{N} \sum_{i=1}^4 \left(\frac{1}{y_i} - \mathbf{w} \cdot \mathbf{x}_i \right)$$

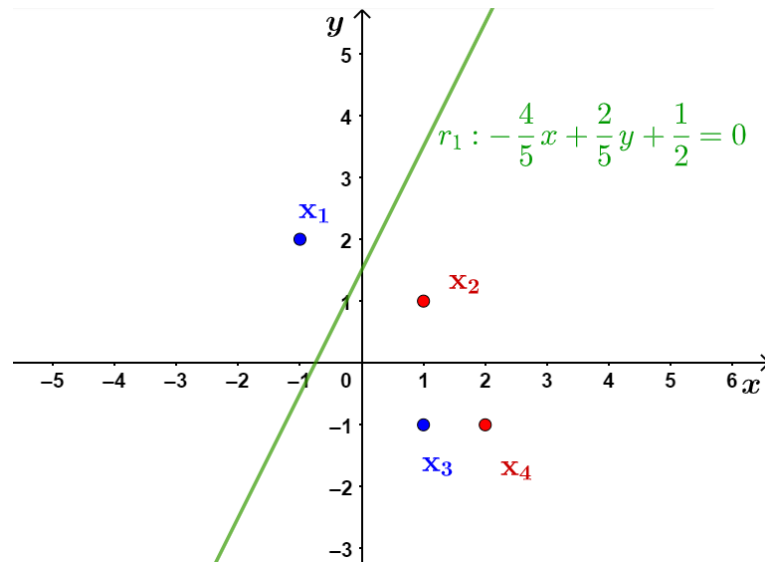
$$b = -\frac{\mathbf{w} \cdot (x_1 + x_2 + x_3 + x_4)}{4}$$

$$b = -\frac{1}{2}$$

Calculado os valores de $\mathbf{w}$ e b prossegue-se para a determinação da equação do hiperplano conforme Equação (3.1). O resultado obtido está descrito na Equação (4.24) e o seu gráfico será a reta r_1 destacada na Figura 17.

$$\begin{aligned} f(\mathbf{x}) &= \mathbf{w} \cdot \mathbf{x} + b = 0 \\ \left(-\frac{4}{5}, \frac{2}{5}\right) \cdot (x, y) - \frac{1}{2} &= 0 \\ -\frac{4}{5}x + \frac{2}{5}y + \frac{1}{2} &= 0 \end{aligned} \tag{4.24}$$

Figura 17 – Hiperplano após 1 iteração



Fonte: Próprio autor.

Ao observar a Figura 17 é possível perceber que o hiperplano obtido após a primeira iteração será deficiente quanto a sua capacidade de generalização. Isto fica evidente ao se imaginar novos dados com a mesma lógica distributiva dos quatro que compõem o conjunto de treinamento em questão.

Para as iterações subsequentes os passos serão apresentados de modo mais direto dispensando maiores detalhes (caso haja necessidade aconselha-se buscar as referências de cada cálculo consultando a descrição da **Iteração 01**). Nas frações optou-se por apresentá-las de modo a unificar os denominadores, ou seja, não estão na forma irredutível, esta metodologia facilita a revisão dos cálculos quando feitos manualmente.

• **Iteração 02**

– *Passo 01: Escolha do primeiro α_i*

Condições de KKT (α_1 é escolhido).

$$y_1 f(\mathbf{x}_1) = \frac{2}{5}(\mathbf{x}_1 \cdot \mathbf{x}_1) - \frac{2}{5}(\mathbf{x}_2 \cdot \mathbf{x}_1) + \frac{3}{5}$$

$$y_1 f(\mathbf{x}_1) = \frac{2}{5} \cdot 5 - \frac{2}{5} \cdot 1 + \frac{3}{5}$$

$$y_1 f(\mathbf{x}_1) = \frac{11}{5}$$

– *Passo 02: Escolha do segundo α_i*

O α_3 é o que possibilita o maior passo.

$$E_1 = f(\mathbf{x}_1) - y_1 = \frac{11}{5} - 1 = \frac{6}{5}$$

$$E_3 = f(\mathbf{x}_3) - y_3 = \frac{2}{5}(-3) - \frac{2}{5}(0) + \frac{3}{5} - 1$$

$$E_3 = -\frac{8}{5}$$

$$|E_1 - E_3| = \left| \frac{6}{5} + \frac{8}{5} \right| = \frac{14}{5}$$

– *Passo 03: Otimizando α_1 e α_3*

Analisando os limites para α_3 ($y_1 = y_3$).

$$L = \max(0, \alpha_1 + \alpha_3 - C), \quad H = \min(C, \alpha_1 + \alpha_3)$$

$L=0$ e $H=\frac{2}{5}$, assim $0 \leq \alpha_3 \leq \frac{2}{5}$.

$$\eta = 2(\mathbf{x}_1 \cdot \mathbf{x}_3) - (\mathbf{x}_1 \cdot \mathbf{x}_1) - (\mathbf{x}_3 \cdot \mathbf{x}_3)$$

$$\eta = 2 \cdot (-3) - 5 - 2 = -13$$

Determinando α_3^{novo} .

$$\alpha_3^{novo} = \alpha_3 - \frac{y_3(E_1 - E_3)}{\eta}$$

$$\alpha_3^{novo} = 0 - \frac{\frac{14}{5}}{-13}$$

$$\alpha_3^{novo} = \frac{14}{65}$$

Não há necessidade de realizar o ajuste pois α_3^{novo} está contido no intervalo $[0, \frac{2}{5}]$.

$$\alpha_1^{novo} = \alpha_1 + y_1 y_3 (\alpha_3 - \alpha_3^{novo})$$

$$\alpha_1^{novo} = \frac{2}{5} + 1(0 - \frac{14}{65})$$

$$\alpha_1^{novo} = \frac{12}{65}$$

– *Passo 04: Determinar o limiar b*

$$b_1 = E_1 + y_1(\alpha_1^{novo} - \alpha_1)(\mathbf{x}_1 \cdot \mathbf{x}_1) + y_3(\alpha_3^{novo} - \alpha_3)(\mathbf{x}_1 \cdot \mathbf{x}_3) + b$$

$$b_1 = \frac{6}{5} + (\frac{12}{65} - \frac{2}{5})5 + \frac{14}{65}(-3) + \frac{3}{5}$$

$$b_1 = \frac{5}{65}$$

$$b_3 = E_3 + y_1(\alpha_1^{novo} - \alpha_1)(\mathbf{x}_1 \cdot \mathbf{x}_3) + y_3(\alpha_3^{novo} - \alpha_3)(\mathbf{x}_3 \cdot \mathbf{x}_3) + b$$

$$b_3 = -\frac{8}{5} + (\frac{12}{65} - \frac{2}{5})(-3) + \frac{14}{65}(2) + \frac{3}{5}$$

$$b_3 = \frac{5}{65}$$

Como esperado, os dois limiares têm o mesmo valor já que tanto α_1 quanto α_3 são válidos.

– *Passo 05: Atualizar a função preditora.*

$$f(\mathbf{x}) = y_1 \alpha_1 (\mathbf{x}_1 \cdot \mathbf{x}) + y_2 \alpha_2 (\mathbf{x}_2 \cdot \mathbf{x}) + y_3 \alpha_3 (\mathbf{x}_3 \cdot \mathbf{x}) + y_4 \alpha_4 (\mathbf{x}_4 \cdot \mathbf{x}) + b$$

$$f(\mathbf{x}) = \frac{12}{65}(\mathbf{x}_1 \cdot \mathbf{x}) - \frac{26}{65}(\mathbf{x}_2 \cdot \mathbf{x}) + \frac{14}{65}(\mathbf{x}_3 \cdot \mathbf{x}) + \frac{5}{65}$$

Novamente com a finalidade de realizar uma análise de hiperplano obtido após esta primeira iteração calcula-se o vetor $\mathbf{w}$ conforme a Equação 3.16 e a variável independente b , conforme a Equação (3.18). Para o cálculo de b não é preciso considerar no somatório os rótulos y_i , uma vez que se anulam.

$$\mathbf{w} = \sum_{i=1}^N \alpha_i y_i \mathbf{x}_i = \alpha_1 \mathbf{x}_1 - \alpha_2 \mathbf{x}_2 + \alpha_3 \mathbf{x}_3 - \alpha_4 \mathbf{x}_4$$

$$\mathbf{w} = \frac{12}{65}(-1, 2) - \frac{26}{65}(1, 1) + \frac{14}{65}(1, -1) - 0(2, -1)$$

$$\mathbf{w} = (-\frac{24}{65}, -\frac{16}{65})$$

$$b = \frac{1}{N} \sum_{i=1}^4 (\frac{1}{y_i} - \mathbf{w} \cdot \mathbf{x}_i)$$

$$b = -\frac{\mathbf{w} \cdot (x_1 + x_2 + x_3 + x_4)}{4}$$

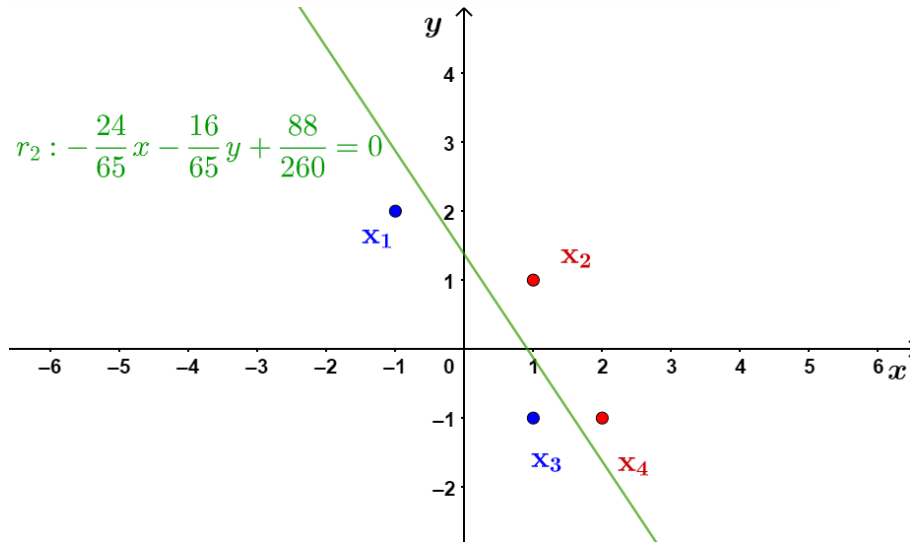
$$b = \frac{88}{260}$$

Calculado os valores de $\mathbf{w}$ e b prossegue-se para a determinação da Equação do hiperplano conforme (3.1). O resultado obtido está descrito na Equação (4.25) e o seu gráfico será a reta r_2

destacada na Figura 18.

$$\begin{aligned}
 f(\mathbf{x}) &= \mathbf{w} \cdot \mathbf{x} + b = 0 \\
 \left(-\frac{24}{65}, -\frac{16}{65}\right) \cdot (x, y) + \frac{88}{260} &= 0 \\
 -\frac{24}{65}x - \frac{16}{65}y + \frac{88}{260} &= 0
 \end{aligned} \tag{4.25}$$

Figura 18 – Hiperplano após 2 iterações



Fonte: Próprio autor.

Uma análise da Figura 18 permite concluir que já após a segunda iteração o algoritmo SMO retorna multiplicadores pelos quais é possível definir um hiperplano que se aproxima fortemente do hiperplano ótimo. Apesar de se estar trabalhando com um exemplo simples, este fato confirma as afirmações feitas por Platt (1998) de que o SMO é um algoritmo rápido na resolução deste tipo de problema de otimização.

Ainda para confirmar o bom resultado obtido após a segunda iteração pode ser avaliado os valores conseguidos por meio da função sinal (3.2). Percebe-se que os rótulos coincidem com os destacados na Tabela 3, mostrando que o hiperplano classifica corretamente todos os exemplos.

Para $\mathbf{x}_1$ temos:

$$\begin{aligned}
 f(\mathbf{x}_1) &= -\frac{24}{65}(-1) - \frac{16}{65}(2) + \frac{88}{260} = \frac{14}{65} = 0,2153... \\
 \text{sgn}(f(\mathbf{x}_1)) &= 1
 \end{aligned}$$

Para $\mathbf{x}_2$ temos:

$$\begin{aligned}
 f(\mathbf{x}_2) &= -\frac{24}{65}(1) - \frac{16}{65}(1) + \frac{88}{260} = -\frac{18}{65} = -0,2769... \\
 \text{sgn}(f(\mathbf{x}_2)) &= -1
 \end{aligned}$$

Para $\mathbf{x}_3$ temos:

$$f(\mathbf{x}_3) = -\frac{24}{65}(1) - \frac{16}{65}(-1) + \frac{88}{260} = \frac{14}{65} = 0,2153...$$

$$\text{sgn}(f(\mathbf{x}_3)) = 1$$

Para $\mathbf{x}_4$ temos:

$$f(\mathbf{x}_4) = -\frac{24}{65}(2) - \frac{16}{65}(-1) + \frac{88}{260} = -\frac{2}{13} = -0,1538...$$

$$\text{sgn}(f(\mathbf{x}_4)) = -1$$

É válido destacar que o SMO, continuaria a realizar às iterações até que as condições de KKT descritas pelas equivalências (4.17) fossem atendidas para todo o conjunto de treinamento. No entanto, entendendo que o objetivo, como contribuição didática para a compreensão da convergência do algoritmo foi alcançado não será feito mais iterações. Principalmente porque os cálculos são repetitivos e enfadonhos. Supondo, desse modo, que a equação encontrada na segunda iteração seja a ótima parte-se para a definição das margens. O cálculo para tanto é relativamente simples e baseado na Equação (3.3), bastando supor as duas possibilidades existentes para o módulo. Isto é, iguala-se a equação do hiperplano a 1 e a -1 , o que pode ser visualizado através da Figura 10. Com este procedimento chega-se nas equações (4.26) e (4.27) que definem respectivamente as margens H_1 e H_2 .

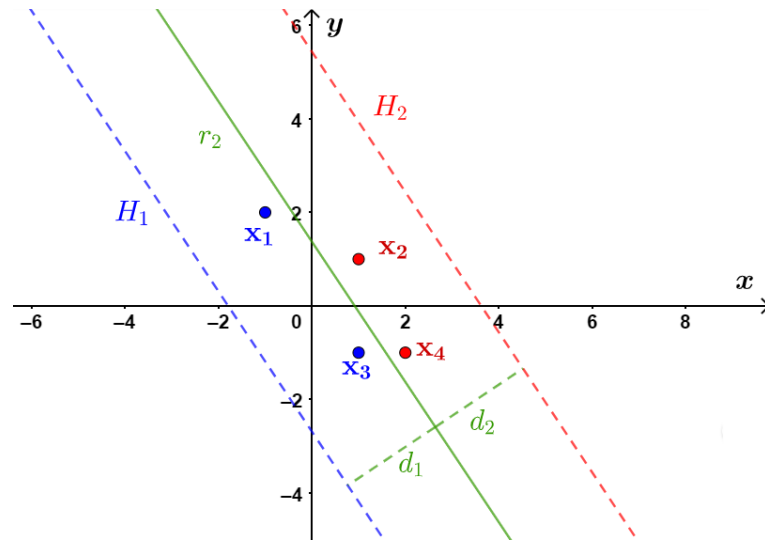
$$H_1 : -\frac{24}{65}x - \frac{16}{65}y + \frac{88}{260} = 1 \quad (4.26)$$

$$H_2 : -\frac{24}{65}x - \frac{16}{65}y + \frac{88}{260} = -1 \quad (4.27)$$

A Figura 19 destaca o hiperplano com suas respectivas margens. Chama-se atenção para o fato de que a distância d_1 e d_2 entre o hiperplano e qualquer uma de suas margens é, como discutido no início da Seção 3.2, $\frac{1}{\|w\|}$ (vide Equação (3.6)). O que pode ser aferido realizando algebricamente o cálculo da distância entre as retas, resultando em aproximadamente 2,253.

Ainda observando a Figura 19 percebe-se que todos os vetores que compuseram o conjunto de treinamento estão no interior das margens do hiperplano. O que justifica essa situação é a escolha do parâmetro de regularização $C = 1$. Como explicado na Seção 3.2, mais especificamente ao tratar do problema de otimização na forma primal. O valor de C , quanto mais próximo de zero implica em um modelo de predição mais flexível, por outro lado, quando tende para infinito positivo aproxima a função preditora para o caso de margens rígidas. Em particular,

Figura 19 – Hiperplano após 2 iterações (com as margens)



Fonte: Próprio autor.

foi escolhido um valor relativamente próximo de zero, pelo menos considerando o fato que poderia crescer infinitamente. Como consequência houve a flexibilização da função classificadora, perceptível graficamente pelo aumento da distância entre as margens e o hiperplano separador.

Fica assim encerrado o estudo da convergência algoritmo SMO para a resolução do problema de otimização gerado pelo treinamento de uma SVM tendo como base um exemplo artificial, que por ambições didáticas foi construído com simplicidade. Foi possível perceber a interlocução com a teoria matemática a qual a técnica das SVMs é fortemente embasada, o que realça a importância do pensamento matemático para a construção de soluções de problemas comuns ao cotidiano de uma sociedade cada vez mais tecnológica.

5 CONSIDERAÇÕES FINAIS

O objetivo deste trabalho consistiu em realizar um estudo a respeito do algoritmo *Sequential Minimal Optimization* (SMO) quando aplicado ao problema de otimização quadrático e convexo gerado a partir do treinamento de uma Máquina de Vetores de Suporte (SVM). Com este propósito, buscou-se compreender a conexão entre diferentes assuntos da matemática e o problema de otimização em evidência, cuja resolução pode ser feita usando o algoritmo SMO. Também foi descrito de forma detalhada o SMO e a obtenção do problema de otimização a ser resolvido pelo mesmo enfatizando como a otimização matemática esta relacionada ao treinamento de uma SVM. A fim de alcançar os objetivos traçados, a metodologia adotada foi uma extensa pesquisa bibliográfica de modo que o trabalho teve um caráter exploratório.

A dificuldade de compreensão da temática, tendo em vista a necessidade de assimilar diferentes conceitos matemáticos, justificou a elaboração deste relatório seguindo a perspectiva de obtenção de uma material didático para que futuros interessados no assunto encontrasse nele um apoio para os estudos iniciais. Outra justificativa se constituiu no desejo acadêmico em ter acesso a conhecimentos matemáticos não vistos durante o curso de Licenciatura em Matemática.

Ao longo do trabalho foi realizada uma revisão geral de conceitos matemáticos relacionados a aplicação estudada, detalhou-se os aspectos matemáticos inerentes a técnica das SVMs e também ao funcionamento do algoritmo SMO. Em relação à revisão dos conceitos matemáticos, abordou-se especificamente temas de quatro campos de estudo: Geometria Analítica, Álgebra Linear, Cálculo Diferencial e alguns tópicos sobre Otimização. Em se tratando das SVMs, buscou-se detalhar todo o processo de treinamento, enfatizando as modificações para os diferentes casos, conforme as características do conjunto de treinamento. Por fim, detalhou-se o algoritmo SMO, sendo realizado um estudo da convergência do mesmo à luz de uma aplicação com um conjunto de treinamento simples constituído de quatro vetores.

Os objetivos estabelecidos no início deste trabalho foram alcançados tendo em vista que houve a produção de um material em que foi destacado com detalhes os aspectos matemáticos relacionados ao treinamento das SVMs, assim como a descrição do processo de treinamento com todo embasamento matemático próprio desta técnica. A explicação das etapas que compõem as iterações feitas pelo algoritmo SMO também foi fundamental para a compreensão da temática estudada. A resolução de um exemplo artificial encerrou o estudo do assunto e mostrou como de fato teoria e prática estão intimamente relacionadas. O trabalho também permitiu visualizar o quanto é frequente, em nosso cotidiano, problemas de classificação assim como a complexidade

que estes carregam quando aplicados a uma técnica de aprendizagem de máquina.

Como sugestão para estudos posteriores recomenda-se um olhar aos aspectos teóricos relativos à convergência do algoritmo SMO, uma vez que se verificou certa dificuldade na compreensão matemática dos motivos que explicam sua maior rapidez quando comparado à outros algoritmos. Outra sugestão seria a implementação em software do SMO e a resolução de um problema real, haja vista que qualquer aplicação mais sofisticada torna a resolução de forma manual inviável e a aplicação trabalhada neste relatório teve pretensão unicamente didática.

REFERÊNCIAS

- ALES, V. T. **O Algoritmo Sequential Minimal Optimisation para Resolução do Problema de Support Vector Machine: Uma Técnica para Reconhecimento de Padrões**. Dissertação (Mestrado) — Universidade Federal do Paraná, Curitiba, 2008.
- ALVES, J. d. S. **Definitividade de Formas Quadráticas [manuscrito] : Uma Abordagem Polinomial**. Tese (Doutorado) — Universidade Federal de Goiás, Instituto de Informática (INF), 2016.
- BURGES, C. J. A tutorial on support vector machines for pattern recognition. **Data mining and knowledge discovery**, Springer, v. 2, n. 2, p. 121–167, 1998. Disponível em: <<https://doi.org/10.1023/A:1009715923555>>. Acesso em: 10 de out. de 2021.
- CARVALHO, R. L. d.; ALMEIDA, T. d. S.; ROCHA, M. L. **Introdução às Metaheurísticas**. 1. ed. Palmas: EDUFT, 2020.
- COELHO, F. U.; LOURENÇO, M. L. **Um Curso de Álgebra Linear**. 2. ed. São Paulo: Editora da Universidade de São Paulo, 2005.
- COLLAZO, R. A.; REIS, A. F. d.; PEREIRA, B. d. B.; LAURA, B. Aplicação do modelo support vector machine para a predição do risco de morte de pacientes internados com síndrome coronariana aguda. XLII SBPO, 2009. **Anais [...]**. Simpósio Brasileiro de Pesquisa Operacional.
- GIL, A. C. **Como Elaborar Projetos de Pesquisa**. 4. ed. São Paulo: Atlas, 2002.
- HAESER, G. **Condições de Otimalidade e Algoritmos em Otimização não Linear**. São Carlos - SP, 2016. v. 83. Disponível em: <<https://www.ime.usp.br/~ghaeser/tese.pdf>>. Acesso em: 10 de mar. 2021.
- HAMDY, H. A. **Pesquisa operacional: uma visão geral**. 8. ed. São Paulo: Pearson Prentice Hall, 2008. Tradução de Arlete Simille Marques, Revisão técnica de Rodrigo Arnaldo Scarpel.
- JUNIOR, A. C. M.; OLIVEIRA, V. A. de. Programação quadrática sequencial para problemas de otimização não-linear com restrições. **Proceeding Series of the Brazilian Society of Computational and Applied Mathematics**, v. 5, n. 1, 2017. Disponível em: <<https://proceedings.sbmoc.org.br/sbmoc/article/view/1649/1659>>. Acesso em: 20 de nov. de 2021.
- LORENA, A. C.; CARVALHO, A. C. de. Uma introdução às support vector machines. **Revista de Informática Teórica e Aplicada**, v. 14, n. 2, p. 43–67, 2007. Disponível em: <<https://doi.org/10.22456/2175-2745.5690>>. Acesso em: 01 de mar. de 2021.
- MAURI, G. R.; LORENA, L. A. N. Decomposições lagrangeanas para o problema de programação quadrática binária irrestrita. **Pesquisa Operacional**, SciELO Brasil, v. 29, n. 1, p. 111–127, 2009. Disponível em: <<https://www.scielo.br/j/pope/a/pLXFqFCgLZzxJd5DvxKmDXn/?lang=pt>>. Acesso em: 22 de nov. de 2021.
- NORONHA, D. H. **Proposta de Implementação em FPGA de Máquina de Vetores de Suporte (SVM) utilizando Otimização Sequencial Mínima (SMO)**. Dissertação (Mestrado) — Universidade Federal do Rio Grande do Norte, Centro de Tecnologia, Programa de Pós-graduação em Engenharia Elétrica e Computação, 2017.

PLATT, J. **Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines**. [S.l.], 1998. Disponível em: <<https://www.microsoft.com/en-us/research/publication/sequential-minimal-optimization-a-fast-algorithm-for-training-support-vector-machines/>>. Acesso em: 03 maio 2021.

SANTOS, C. E. d. S. **Seleção de Parâmetros de Máquinas de Vetores de Suporte Usando Otimização Multiobjetivo Baseada em Meta-Heurísticas**. Tese (Doutorado) — Universidade de Brasília, 2019.

SANTOS, E. M. d. **Teoria e Aplicação de Support Vector Machines à Aprendizagem e reconhecimento de Objetos Baseado na Aparência**. Dissertação (Mestrado) — Universidade Federal de Campina Grande, Pós-Graduação em Informática, 2002.

STEINBRUCH, A.; WINTERLE, P. **Um Curso de Álgebra Linear**. 2. ed. São Paulo: Pearson Makron Books, 1987.

THOMAS, G. B.; WEIR, M. D.; HASS, J. **Cálculo, volume 2**. 12. ed. São Paulo: Pearson Education do Brasil, 2012. Tradução de Carlos Scalici; revisão técnica de Claudio Hirofume Asano.

VAPNIK, V. N. **The nature of Statistical learning theory**. [S.l.]: Springer-Verlag, New York, 1995.

WINTERLE, P. **Vetores e Geometria Analítica**. 2. ed. São Paulo: Pearson Education do Brasil, 2014.