UNIVERSIDADE FEDERAL DE SANTA CATARINA
CAMPUS BLUMENAU
LICENCIATURA EM MATEMÁTICA

Paula Cristina Rohr Ertel

Análise teórica das máquinas de vetores suporte e aplicação à classificação de dados

Blumenau
2020

Paula Cristina Rohr Ertel

Análise teórica das máquinas de vetores suporte e aplicação à classificação de dados

Trabalho de Conclusão de Curso de Graduação em Licenciatura em Matemática do Campus Blumenau da Universidade Federal de Santa Catarina para a obtenção do título de Licenciada em Matemática.
Orientador:: Prof. Dr. Luiz Rafael dos Santos

Blumenau
2020

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Ertel, Paula Cristina Rohr
Análise teórica das máquinas de vetores suporte e aplicação à classificação de dados / Paula Cristina Rohr Ertel ; orientador, Luiz Rafael dos Santos, 2020.
158 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Campus Blumenau, Graduação em Matemática, Blumenau, 2020.

Inclui referências.

1. Matemática. 2. Máquinas de vetores suporte. 3. Aprendizagem de máquina. 4. Otimização. I. Santos, Luiz Rafael dos. II. Universidade Federal de Santa Catarina. Graduação em Matemática. III. Título.

Paula Cristina Rohr Ertel

Análise teórica das máquinas de vetores suporte e aplicação à classificação de dados

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do Título de Licenciada em Matemática e aprovado em sua forma final pelo Curso de Licenciatura em Matemática do Centro de Blumenau da Universidade Federal de Santa Catarina.

Blumenau, 26 de Novembro de 2020.

Prof. Dr. Júlio Faria Côrrea
Coordenador do Curso de Licenciatura em Matemática

Banca Examinadora:

Prof. Dr. Luiz Rafael dos Santos
Orientador:
Universidade Federal de Santa Catarina - UFSC

Prof. Dr. Ciro André Pitz
Avaliador
Universidade Federal de Santa Catarina - UFSC

Documento assinado digitalmente

Hugo Jose Lara Urdaneta
Data: 01/12/2020 12:22:30-0300
CPF: 051.656.117-01

Prof. Dr. Hugo José Lara Urdaneta
Avaliador
Universidade Federal de Santa Catarina - UFSC

Dedico este trabalho aos meus pais Enei e Roseli e ao meu irmão Antonio.

AGRADECIMENTOS

Primeiramente, agradeço aos meus pais, Enei e Roseli, e ao meu irmão Antonio, por todo o amor, suporte e incentivo durante essa caminhada. O apoio de vocês foi fundamental para que eu concluísse mais essa etapa.

Ao meu orientador, professor Luiz Rafael dos Santos, por todos os aprendizados que você me proporcionou. Sou imensamente grata pela sua paciência, disponibilidade e incentivo durante a orientação.

Aos docentes do curso de Licenciatura em Matemática da UFSC Blumenau, por todos os ensinamentos e dedicação. Em especial ao professor André Vanderlinde, por sempre se mostrar acessível e por toda a ajuda que tornou a minha permanência na UFSC possível.

Ao meu namorado, Bruno, pelo companheirismo e por todo o apoio que recebi durante a graduação.

Por fim, agradeço a todos que direta ou indiretamente contribuíram de alguma forma durante a minha caminhada na graduação.

RESUMO

Em problemas que exigem a análise de uma grande quantidade de dados para classificá-los, um processo manual torna-se inviável, motivando o desenvolvimento de técnicas computacionais capazes de reconhecer padrões para desempenhar tal tarefa. Assim, o objetivo central deste trabalho é desenvolver um estudo teórico, do ponto de vista da otimização, de uma técnica de aprendizagem de máquina supervisionada aplicada à classificação binária de dados denominada Máquinas de Vetores Suporte (SVMs). Para tanto, tendo em vista que a formulação matemática da técnica SVM consiste num problema de programação quadrática convexa com restrições lineares, abordamos aspectos da teoria de otimização, com e sem restrições, assim como apresentamos definições e resultados de otimização convexa, as quais fornecem propriedades importantes relacionadas aos problemas de otimização, como a garantia de existência de soluções. Desenvolvemos com detalhes a modelagem matemática da técnica SVM com margem rígida, demonstrando que o problema de otimização decorrente admite uma única solução, bem como construímos sua generalização para um dos casos em que os dados não são linearmente separáveis, denominada SVM com margem flexível. Por fim, utilizando a linguagem de programação Julia, realizamos uma implementação computacional da técnica SVM para classificar o conjunto de dados Flor Íris em relação às suas espécies e, posteriormente, para classificar um conjunto de dados sobre células de câncer de mama em tumor maligno ou benigno. Através desses experimentos numéricos foi possível analisar a eficiência da técnica SVM. Em particular, no caso em que aplicamos SVM com margem flexível, tal eficiência está relacionada com a escolha de um parâmetro de penalização adequado.

Palavras-chave: Aprendizagem de Máquina, Classificação, Máquinas de Vetores Suporte, Otimização com restrições, Otimização convexa.

ABSTRACT

In problems that require the classification and analysis of large scale data, a manual process becomes unfeasible, motivating the development of computational techniques capable of recognizing patterns for the task. Thus, the main objective of this work is to develop a theoretical study, using the point of view of Mathematical Optimization, of a machine learning technique applied to binary data classification: Support Vector Machines (SVM). For this purpose, considering that the mathematical formulation of the SVM technique consists of a convex quadratic programming problem with linear constraints, we study theoretical aspects of constrained and unconstrained optimization, as well as present definitions and results of convex optimization, which provide important properties of SVM related to optimization problems, such as existence and uniqueness of solutions. We developed in detail the SVM mathematical model with rigid margin, deriving that the resulting optimization problem admits a unique solution, and extend the model to one of the cases in which the data are not linearly separable, called soft margin SVM. Finally, using the Julia programming language, we implement the SVM techniques aforementioned and test our implementation in two cases: to classify the Iris flower dataset and to classify a dataset containing data from diagnosis of breast cancer cells. The numerical experiments show the efficiency of the SVM technique. In particular, in the case where we apply the soft margin SVM, such efficiency is related to the choice of an appropriate penalty parameter.

Keywords: Machine Learning, Classification, Support Vector Machines, Constrained Optimization, Convex Optimization.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A análise inteligente de dados tem se tornado cada vez mais importante para auxiliar na tomada de decisões em diversos campos da ciência. A Aprendizagem de Máquina (AM, do inglês *Machine Learning*) é um campo da inteligência computacional que estuda o uso de técnicas computacionais para automaticamente detectar padrões em dados e utilizá-los para fazer predições e tomar decisões. AM tem dado contribuições em diversos campos da ciência e também está presente em diferentes situações do cotidiano como na recomendação de filmes e séries em serviços de *Streaming*, nos mecanismos de pesquisa do Google e na detecção de spam em e-mails, por exemplo.

Veremos neste trabalho que a Aprendizagem de Máquina também está diretamente relacionada à Otimização, uma vez que contribui para a formulação matemática e para a implementação computacional de muitos de seus algoritmos. As técnicas de Aprendizagem de Máquina podem ser divididas em dois tipos principais, a aprendizagem supervisionada, em que através de um conjunto de dados, cujas saídas são previamente conhecidas, o algoritmo detecta padrões e produz um modelo capaz de deduzir as saídas corretas para novos dados, e a aprendizagem não-supervisionada, empregada em problemas que não possuem dados previamente rotulados. Nosso objetivo neste trabalho é realizar um estudo teórico, do ponto de vista da Otimização, de uma técnica de aprendizagem supervisionada: as Máquinas de Vetores Suporte (SVMs, do inglês *Support Vector Machines*).

A técnica SVM é fundamentada na Teoria de Aprendizagem Estatística e foi desenvolvida por Vladimir Vapnik, Bernhard Boser, Isabelle Guyon e Corrina Cortes [5, 7]. Conforme destacado por Krulikovski [15], as SVMs são amplamente empregadas em problemas de regressão e de classificação, possuem um embasamento teórico bem consolidado e apresentam uma boa capacidade

de generalização. São indicadas nos casos em que ocorrem dados de dimensões elevadas e com altos níveis de ruídos.

Aqui, será abordada a modelagem matemática da técnica SVM aplicada ao problema de classificação, que resulta no seguinte problema de programação quadrática convexa e com restrições lineares

$$\begin{aligned} \min_{w,b} \quad & f(w) \\ \text{s.a.} \quad & g(w,b) \leq 0, \end{aligned} \tag{1}$$

em que as funções $f : \mathbb{R}^{n+1} \to \mathbb{R}$ e $g : \mathbb{R}^{n+1} \to \mathbb{R}^m$ são continuamente diferenciáveis. A resolução do problema (1) fornece uma solução (w, b), a partir da qual obtém-se um classificador linear. Geometricamente, dado um conjunto de dados, a técnica SVM realiza a separação deste conjunto em diferentes classes através de um hiperplano definido pela equação $w^T x + b = 0$ [15], que é construído a partir da solução do problema (1).

Dependendo do conjunto de dados e da complexidade do problema, a técnica SVM apresenta três diferentes formulações: SVM com margem rígida (Figura 1a), SVM com margem flexível (Figura 1b) e SVM não-linear (Figura 1c).

O problema de classificação representado na Figura 1a é um exemplo que pode ser modelado pelo problema (1). Neste caso os dados são linearmente separáveis, sendo possível encontrar um hiperplano que os classifique corretamente. As Figuras 1b e 1c, no entanto, apresentam problemas cujos dados não são linearmente separáveis. Na Figura 1b temos um exemplo de SVM de margem flexível, em que promovendo um relaxamento das restrições através de variáveis de folga ξ_i associadas a cada atributo x^i ainda é possível obter um hiperplano que classifique os dados. Já no caso da Figura 1c é preciso aplicar a SVM não linear, na qual o hiperplano ótimo é obtido através de um mapeamento dos dados para um espaço de dimensão elevada [15]. Neste trabalho abordaremos somente os casos de SVM com margem rígida e margem flexível, pois para a

modelagem da técnica SVM não-linear são necessários alguns conceitos matemáticos mais complexos que não serão abordados neste trabalho e portanto, ela se constitui numa proposta de estudo a ser desenvolvida em projetos futuros.

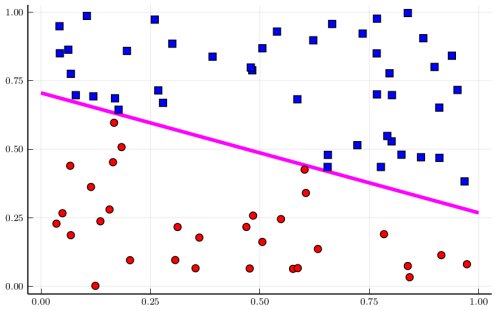
Como o problema de classificação (1) que desejamos formular trata-se de um problema de programação quadrática convexa com restrições lineares, o Capítulo 2 apresenta os principais conceitos e resultados de otimização irrestrita e com restrições lineares de igualdade e desigualdade. As principais referências consultadas para elaboração deste capítulo foram Friedlander [12], Izmailov e Solodov [14], Luenberger e Ye [19] e Ribeiro e Karas [22].

No Capítulo 3 são discutidos os conceitos de conjunto convexo e função convexa com o intuito de analisar a convexidade dos problemas de SVM. Para construção deste capítulo as principais bibliografias utilizadas foram Bertsekas [2], Izmailov e Solodov [14], Luenberger e Ye [19] e Ribeiro e Karas [22].

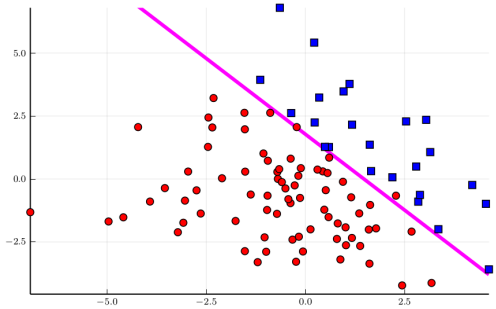
A modelagem matemática do problema de classificação através da técnica SVM, tanto para o caso de margem rígida quanto de margem flexível, é realizada no Capítulo 4. Para tanto, foram utilizadas como principais referências Deisenroth et al. [8] e Krulikovski [15].

Por fim, no Capítulo 5 realizamos a implementação computacional da técnica SVM para a classificação binária de dados. Num primeiro momento, aplicamos a técnica SVM para classificar o conjunto de dados de flores Íris [11] em relação às suas espécies, e posteriormente, para a classificação de amostras de células de câncer de mama em tumor maligno ou benigno [1, 27]. Para a realização de tais experimentos utilizamos a linguagem de programação `Julia` e os códigos foram escritos no *software* Jupyter Notebook. Assim, no Capítulo 5 apresentamos junto ao texto os principais códigos e resultados obtidos, com os comentários pertinentes.

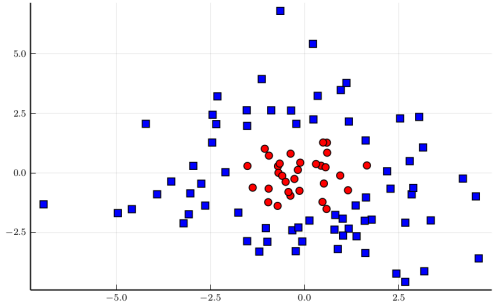
Das figuras presentes no trabalho, algumas foram retiradas de [8, 12, 15, 19], e outras foram elaboradas com o auxílio dos *softwares* `Julia` [3] e Ipe.



(a) Margem Rígida.



(b) Margem Flexível.



(c) Não-Linear.

Figura 1 – Conjunto de dados para classificação.

2 CONCEITOS DE OTIMIZAÇÃO

Neste capítulo pretendemos discutir alguns conceitos e resultados para o problema

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{s.a} \quad & x \in \Omega, \end{aligned} \tag{2}$$

em que $f : \mathbb{R}^n \to \mathbb{R}$ é chamada *função objetivo*, $\Omega \subset \mathbb{R}^n$ é chamado *conjunto factível* do problema (2) e os pontos de Ω são chamados de *pontos factíveis*.

Como o problema (1) que surge da formulação matemática da técnica SVM para classificação trata-se de um problema de otimização que possui restrições lineares apenas, toda a teoria de condições de otimalidade somente será estudada para problemas desse tipo: com restrições lineares. Para o desenvolvimento da teoria de otimização irrestrita, as principais referências utilizadas foram Izmailov e Solodov [14] e Ribeiro e Karas [22]. Já os conceitos e resultados da teoria de otimização com restrições foram desenvolvidos, principalmente, com base em Friedlander [12] e Luenberger e Ye [19].

Inicialmente, vamos caracterizar os pontos que são solução do problema (2).

Definição 2.1. Dizemos que um ponto $x^* \in \Omega$ é

(a) *minimizador local* de f em Ω se, e somente se, existe $\varepsilon > 0$ tal que $f(x^*) \leq f(x)$ para todo $x \in B(x^*, \varepsilon) \cap \Omega$.

(b) *minimizador global* de f em Ω se, e somente se, $f(x^*) \leq f(x)$ para todo $x \in \Omega$

Quando as desigualdades na Definição 2.1 forem estritas para $x \neq x^*$, diremos que x^* é minimizador estrito. Essas definições são representadas na Figura 2.

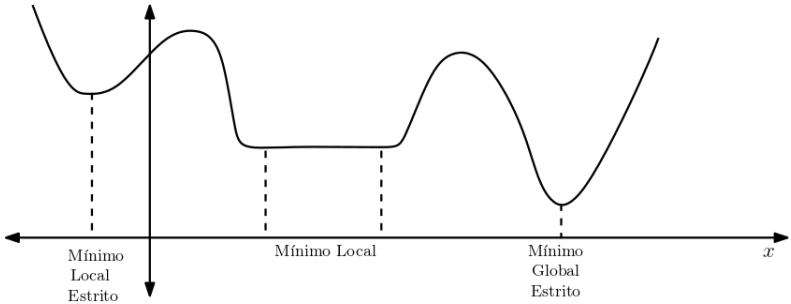


Figura 2 – Mínimo local e mínimo global em uma dimensão.

Pela Definição 2.1, todo minimizador global é também minimizador local, porém a recíproca não é verdadeira. É interessante salientar que em muitas circunstâncias iremos nos contentar com um ponto de mínimo local pois, de modo geral, condições globais e soluções globais só podem ser encontradas se o problema possuir certas propriedades que garantem, essencialmente, que qualquer mínimo local é global. Uma destas propriedades é a convexidade, a qual abordaremos mais adiante (veja Capítulo 3).

Observação 2.2. Todo problema de maximização

$$\begin{aligned} \max_{x} \quad & f(x) \\ \text{s.a} \quad & x \in \Omega \end{aligned}$$

pode ser transformado em um problema de minimização equivalente

$$\begin{aligned} \min_{x} \quad & -f(x) \\ \text{s.a} \quad & x \in \Omega. \end{aligned}$$

Em particular, as soluções locais e globais de ambos os problemas são as mesmas, com sinais opostos para os valores ótimos.

Ademais, quando o conjunto factível $\Omega = \mathbb{R}^n$ dizemos que o problema (2) é irrestrito. No caso em que Ω é definido por um sistema de igualdades ou desigualdades como

$$\Omega = \{x \in \mathbb{R}^n \mid h(x) = 0,\ g(x) \leq 0\},$$

falamos em otimização com restrições.

Frequentemente, a formulação de problemas mais complexos envolve restrições à função objetivo. No entanto, veremos mais adiante que muitos desses problemas podem ser convertidos em problemas irrestritos, utilizando as restrições para estabelecer relações entre as variáveis. Em vista disso, abordaremos primeiramente a teoria de otimização para o caso irrestrito para posteriormente obter as condições de otimalidade para problemas com restrições de igualdade e desigualdade, haja vista que o problema de classificação, o qual estamos interessados em resolver, possui tal formato.

Definição 2.3. Dizemos que $\bar{v} \in \mathbb{R} \cup \{-\infty\}$ definido por

$$\bar{v} = \inf_{x \in \Omega} f(x)$$

é o *valor ótimo* do problema (2).

No estudo do problema de otimização irrestrita uma das principais questões que surge diz respeito a existência da solução. Observe que se na Definição 2.3 temos $\bar{v} = -\infty$, o problema (2) não admite solução global, pois neste caso f é ilimitada inferiormente no conjunto factível. Outro caso em que também não existe minimizador global ocorre quando $\bar{v}$ não é atingido em nenhum ponto factível. Vejamos um exemplo.

Exemplo 2.4. Seja $f : \mathbb{R} \to \mathbb{R}$, definida por $f(x) = e^x$, $\Omega = \mathbb{R}$. Note que $\bar{v} = \inf_{x \in \mathbb{R}} e^x = 0$, contudo, não existe $x \in \mathbb{R}$ tal que $e^x = 0$. De modo análogo,

considere f definida como anteriormente e $\Omega = (0, 1]$. Temos que $\bar{v} = \inf_{x \in (0,1]} e^x = 1$ e novamente não existe $x \in (0, 1]$ tal que $e^x = 1$. Observe que a função f é contínua em Ω, porém no primeiro caso Ω não é limitado e no segundo, Ω não é fechado. Considere agora $\Omega = [0, 1]$, $f(x) = e^x$ para $x \in (0, 1]$ e $f(0) = 2$. Novamente, $\bar{v} = \inf_{x \in [0,1]} e^x = 1$, porém não existe $x \in \Omega$ tal que $f(x) = 1$. Neste exemplo, Ω é compacto mas f não é contínua.

Assim, a partir desses exemplos é possível perceber que a existência de soluções está relacionada à continuidade da função objetivo e à compacidade do conjunto factível. O principal resultado que garante a existência de soluções globais é o Teorema de Weierstrass.

Teorema 2.5. ***(Weierstrass)*** *Sejam $f : \mathbb{R}^n \to \mathbb{R}$ uma função contínua e $\Omega \in \mathbb{R}^n$ um conjunto compacto não-vazio. Então existe minimizador global de f em Ω.*

Demonstração. A imagem de um conjunto compacto por uma função contínua é compacta. Assim, $f(\Omega)$ é compacto, em particular, como $f(\Omega) \in \mathbb{R}$, este conjunto é limitado inferiormente, ou seja, existe $\bar{v} \in \mathbb{R}$ tal que

$$\bar{v} = \inf_{x \in \Omega} f(x).$$

Pela definição de ínfimo, temos que para todo $k \in \mathbb{N}$ existe $x^k \in \Omega$ tal que

$$\bar{v} \leq f(x^k) \leq \bar{v} + \frac{1}{k}.$$

Passando ao limite quando $k \to \infty$, e usando o Teorema do Sanduíche, concluímos que

$$\lim_{k \to \infty} f(x^k) = \bar{v}. \tag{3}$$

Como $\{x^k\} \in \Omega$ e Ω é compacto, temos que $\{x^k\}$ é uma sequência limitada. Logo, ela possui uma subsequência convergente em Ω, isto é, existe uma subsequência $\{x^{k_j}\}$ que converge para um ponto $\bar{x} \in \Omega$,

$$\lim_{j\to\infty} x^{k_j} = \bar{x} \in \Omega.$$

Como f é contínua, temos que

$$\lim_{j\to\infty} f(x^{k_j}) = f(\bar{x}).$$

Usando (3), temos que $f(\bar{x}) = \bar{v}$ e portanto, f assume valor mínimo no ponto $\bar{x} \in \Omega$. Em outras palavras, $\bar{x}$ é um minimizador global de f em Ω. □

2.1 CONDIÇÕES DE OTIMALIDADE PARA PROBLEMAS SEM RESTRIÇÕES

Considere o problema (2) irrestrito , isto é, com $\Omega = \mathbb{R}^n$,

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{s.a} \quad & x \in \mathbb{R}^n, \end{aligned} \tag{4}$$

em que $f : \mathbb{R}^n \to \mathbb{R}$. Nesta seção serão determinadas as condições que um ponto $x^* \in \mathbb{R}^n$ deve satisfazer quando é minimizador local do problema (4). Condições desse tipo são chamadas de *condições necessárias de otimalidade*. Determinaremos também as condições que garantem que um ponto dado é minimizador local do problema, que são denominadas *condições suficientes de otimalidade*.

Para os resultados que vem a seguir vamos utilizar os Fatos 2.6 e 2.7, que fornecem a Fórmula de Taylor de primeira e segunda ordem. As demonstrações destes fatos podem ser encontradas em Lima [17, p.194].

Fato 2.6. *(**Taylor de Primeira Ordem**) Seja $f : \mathbb{R}^n \to \mathbb{R}$ uma função diferenciável e $\bar{x} \in \mathbb{R}^n$. Então podemos escrever*

$$f(x) = f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x}) + o(x),$$

com $\lim_{x \to \bar{x}} \dfrac{o(x)}{\|x - \bar{x}\|} = 0.$

Fato 2.7. *(**Taylor de Segunda Ordem**) Seja $f : \mathbb{R}^n \to \mathbb{R}$ uma função duas vezes diferenciável e $\bar{x} \in \mathbb{R}^n$. Então*

$$f(x) = f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x}) + \frac{1}{2}(x - \bar{x})^T \nabla^2 f(\bar{x})(x - \bar{x}) + o(x),$$

com $\lim_{x \to \bar{x}} \dfrac{o(x)}{\|x - \bar{x}\|^2} = 0.$

Agora, nos teoremas que seguem, mostraremos as condições de otimalidade de primeira e de segunda ordem para o problema de minimização irrestrito.

Teorema 2.8. *(**Condição necessária de primeira ordem**) Seja $f : \mathbb{R}^n \to \mathbb{R}$ diferenciável no ponto $x^* \in \mathbb{R}^n$. Se x^* é um minimizador local de f, então*

$$\nabla f(x^*) = 0. \tag{5}$$

Demonstração. Considere $d \in \mathbb{R}^n$ arbitrário, porém fixo. Pela definição de minimizador local, existe $\varepsilon > 0$ tal que

$$f(x^*) \leq f(x^* + td), \tag{6}$$

para todo $t \in (0, \varepsilon)$. Pela diferenciabilidade de f em x^*, aplicando o Fato 2.6, podemos escrever

$$f(x^* + td) = f(x^*) + t\nabla f(x^*)^T d + o(t),$$

com $\lim_{t\to 0} \frac{o(t)}{t} = 0$. Utilizando (6), temos

$$0 \leq t\nabla f(x^*)^T d + o(t),$$

e dividindo por $t > 0$,

$$0 \leq \nabla f(x^*)^T d + \frac{o(t)}{t}.$$

Passando o limite quando $t \to 0$, obtemos

$$0 \leq \nabla f(x^*)^T d.$$

Supondo que $\nabla f(x^*)$ fosse não-nulo, como $d \in \mathbb{R}^n$ é arbitrário, poderíamos escolher $d = -\nabla f(x^*)$, resultando em

$$0 \leq \nabla f(x^*)^T d = -\|\nabla f(x^*)\|^2,$$

o que é uma contradição. Portanto, $\nabla f(x^*) = 0$. □

Definição 2.9. Um ponto $x^* \in \mathbb{R}^n$ que cumpre a condição (5) é dito *ponto crítico* ou *estacionário* da função f.

Portanto, a partir da definição acima temos que, se f é uma função diferenciável, minimizadores locais devem ser pontos estacionários. O teorema a seguir estabelece a condição necessária de segunda ordem.

Teorema 2.10. ***(Condição necessária de segunda ordem)*** *Seja $f : \mathbb{R}^n \to \mathbb{R}$ duas vezes diferenciável no ponto $x^* \in \mathbb{R}^n$. Se x^* é um minimizador local de f, então*

$$d^T \nabla^2 f(x^*) d \geq 0,$$

para todo $d \in \mathbb{R}^n$.

Demonstração. Seja novamente $d \in \mathbb{R}^n$ arbitrário, porém fixo. Se x^* é minimizador local de f, então para todo $t > 0$ suficientemente pequeno, temos que

$$0 \leq f(x^* + td) - f(x^*). \tag{7}$$

Aplicando o Fato 2.7,

$$f(x^* + td) = f(x^*) + t\nabla f(x^*)^T d + \frac{t^2}{2} d^T \nabla^2 f(x^*) d + o(t^2),$$

com $\lim_{t\to 0} \frac{o(t^2)}{t^2} = 0$, e (7) implica em

$$0 \leq \frac{t^2}{2} d^T \nabla^2 f(x^*) d + o(t^2), \tag{8}$$

pois como x^* é minimizador local, o Teorema 2.8 garante que $\nabla f(x^*) = 0$. Dividindo ambos os lados da desigualdade (8) por t^2 e tomando o limite quando $t \to 0$, obtemos

$$d^T \nabla^2 f(x^*) d \geq 0,$$

para todo $d \in \mathbb{R}^n$. □

Definição 2.11. Seja $A \in \mathbb{R}^{n\times n}$ uma matriz simétrica. Dizemos que A é *definida positiva* quando $x^T Ax > 0$, para todo $x \in \mathbb{R}^n$, $x \neq 0$. Tal propriedade é denotada por $A \succ 0$. Se $x^T Ax \geq 0$, para todo $x \in \mathbb{R}^n$, A é dita *semidefinida positiva*, fato este denotado por $A \succcurlyeq 0$.

Observação 2.12. A definição geral de positividade de uma matriz não exige que ela seja simétrica. Entretanto, neste trabalho sempre que dissermos que a matriz é definida ou semidefinida positiva assumimos implicitamente que a matriz é simétrica.

Assim, de acordo com a Definição 2.11, o Teorema 2.10 nos diz que se um ponto x^* é minimizador local de f, então a matriz Hessiana $\nabla^2 f(x^*)$ é semidefinida positiva.

É possível obter, a partir da condição necessária de segunda ordem, a condição suficiente de segunda ordem, isto é, condição que implica que x^* é minimizador local estrito, bastando para isso pedir a definição positiva da Hessiana.

Teorema 2.13. ***(Condição suficiente de segunda ordem)*** *Seja $f : \mathbb{R}^n \to \mathbb{R}$ duas vezes diferenciável no ponto $x^* \in \mathbb{R}^n$. Se x^* é um ponto estacionário da função f e $\nabla^2 f(x^*)$ é definida positiva, então x^* é minimizador local estrito de f.*

Demonstração. Seja $B \coloneqq \{h \in \mathbb{R}^n \mid \|h\| = 1\}$ e consideremos a função $\phi : B \to \mathbb{R}$ dada por

$$\phi(h) = h^T \nabla^2 f(x^*) h.$$

A função ϕ é contínua e B é um conjunto compacto, portanto, pelo Teorema 2.5, ϕ atinge um valor máximo e um valor mínimo em B. Considere $\phi_{\min} \coloneqq \phi(h^*)$ o valor mínimo de ϕ em B. Como $\nabla^2 f(x^*) \succ 0$, então $\phi(h) > 0$, para todo $h \in B$ e, em particular,

$$\phi(h) \geq \phi_{\min} > 0.$$

Agora, consideremos $d \in \mathbb{R}^n$, arbitrário não-nulo. Como $\dfrac{d}{\|d\|} \in B$, temos que

$$\frac{d}{\|d\|}^T \nabla^2 f(x^*) \frac{d}{\|d\|} \geq \phi_{\min}$$

resulta em

$$d^T \nabla^2 f(x^*) d \geq \phi_{\min} \|d\|^2. \tag{9}$$

Aplicando o Fato 2.7 em torno de x^*, temos

$$f(x^* + d) - f(x^*) = \nabla f(x^*)^T d + \frac{1}{2} d^T \nabla^2 f(x^*) d + o(\|d\|^2). \tag{10}$$

Como, por hipótese, $\nabla f(x^*) = 0$, (9) e (10) implicam que

$$f(x^* + d) - f(x^*) = \frac{1}{2} d^T \nabla^2 f(x^*) d + o(\|d\|^2),$$

e assim,

$$f(x^* + d) - f(x^*) \geq \frac{\phi_{\min}}{2} \|d\|^2 + o(\|d\|^2). \tag{11}$$

Então, para todo d tal que $\|d\|$ é suficientemente pequena, o primeiro termo do membro direito da desigualdade (11) define o sinal deste lado. Além disso,

$$\frac{\phi_{\min}}{2} \|d\|^2 > 0.$$

Portanto, para $\|d\|$ suficientemente pequena não-nula, temos que

$$f(x^* + d) - f(x^*) > 0,$$

e assim,

$$f(x^*) < f(x^* + d).$$

Portanto, para todo $x = x^* + d$, com d suficientemente pequeno não-nulo, temos que $f(x^*) < f(x)$. Logo, x^* é um minimizador local estrito de f. □

Vejamos um exemplo em que a condição suficiente de segunda ordem nos permite determinar uma solução para um problema de minimização irrestrito.

Exemplo 2.14. Considere o problema irrestrito

$$\begin{aligned} \min_x \quad & f(x) = x_1^2 - x_1 x_2 + 2x_2^2 - 2x_1 + \frac{2}{3} x_2 + e^{x_1 + x_2} \\ \text{s.a} \quad & x \in \mathbb{R}^2. \end{aligned} \tag{12}$$

Para um ponto x^* ser minimizador local estrito de (12) ele deve satisfazer as condições suficientes de segunda ordem do Teorema 2.13, que são

$$\nabla f(x^*) = \begin{bmatrix} 2x_1^* - x_2^* - 2 + e^{x_1^*+x_2^*} \\ -x_1^* + 4x_2^* + \dfrac{2}{3} + e^{x_1^*+x_2^*} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \tag{13}$$

e

$$\nabla^2 f(x^*) = \begin{bmatrix} 2 + e^{x_1^*+x_2^*} & -1 + e^{x_1^*+x_2^*} \\ -1 + e^{x_1^*+x_2^*} & 4 + e^{x_1^*+x_2^*} \end{bmatrix} \succ 0.$$

Com efeito, o ponto $x^* = (1/3, -1/3)$ satisfaz (13), pois

$$\nabla f(x^*) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Além disso, $\nabla^2 f(x^*)$ é definida positiva, pois

$$\nabla^2 f(x^*) = \begin{bmatrix} 3 & 0 \\ 0 & 5 \end{bmatrix},$$

e, para todo $x \in \mathbb{R}^2 \backslash \{0\}$, temos que

$$\begin{bmatrix} x_1 & x_2 \end{bmatrix}^T \begin{bmatrix} 3 & 0 \\ 0 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 3x_1^2 + 5x_2^2 > 0.$$

Portanto, $x^* = (1/3, -1/3)$ é minimizador local estrito do problema (12).

Tendo em vista que o problema de classificação (1) que estamos interessados em resolver apresenta restrições lineares e seu conjunto factível é poliedral, abordaremos nas seções seguintes o problema de minimização com restrições nesse contexto. Para tanto, iniciamos analisando o problema de minimização com restrições lineares de igualdade.

2.2 MINIMIZAÇÃO COM RESTRIÇÕES LINEARES DE IGUALDADE

Nosso objetivo nesta seção será analisar o seguinte problema de minimização com restrições lineares de igualdade

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{s.a} \quad & Ax = b, \end{aligned} \tag{14}$$

em que $A \in \mathbb{R}^{m\times n}$, $1 \leq m < n$ e posto$(A) = m$. Assim como no caso irrestrito, estamos interessados em obter as condições que caracterizam as soluções do problema (14).

O conjunto

$$\Omega := \{x \in \mathbb{R}^n \mid Ax = b\}$$

é chamado *conjunto de factibilidade* de (14). Este conjunto é a variedade afim de soluções do sistema linear

$$Ax = b. \tag{15}$$

Note que Ω é uma reta se $m = n - 1$, um plano se $m = n - 2$ e uma variedade de dimensão $n - m$ para m genérico. No caso em que $n > 2$ e $m = 1$ dizemos que Ω é um hiperplano.

Para obter as condições de otimalidade para um ponto de mínimo x^* do problema (14), a ideia central é considerar o movimento para longe do ponto x^* em uma determinada direção, desde que possamos continuar, pelo menos inicialmente, no conjunto Ω. Desse modo, dado $x \in \Omega$ dizemos que um vetor d é uma *direção factível* de x se existe $\bar{\alpha} > 0$ tal que $x + \alpha d \in \Omega$, para todo α com $0 \leq \alpha \leq \bar{\alpha}$. Em vista disso, o primeiro passo será determinar o conjunto de direções factíveis em Ω.

Para tanto, associado a Ω temos o conjunto de soluções do sistema homogêneo $Ax = 0$, que é chamado de Núcleo de A e denotado por $\mathcal{N}(A)$. Este

conjunto é um subespaço de $\mathbb{R}^n$ de dimensão $n - m$, pois posto$(A) = m$. É interessante observar que $\mathcal{N}(A)$ é um subespaço paralelo à variedade afim Ω e passa pela origem. Esta noção é representada na Figura 3.

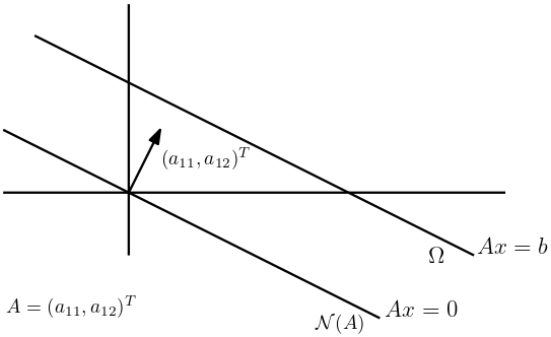


Figura 3 – Núcleo de uma matriz.

Fonte – Friedlander [12]

Como posto$(A) = m$, as m linhas de A formam um conjunto de vetores linearmente independentes que geram o subespaço chamado imagem de A^T de dimensão m, que é denotado por $Im(A^T)$. Dessa forma, como é possível observar na Figura 3, as linhas de A são ortogonais ao $\mathcal{N}(A)$, ou, em outras palavras, $\mathcal{N}(A)$ é o complemento ortogonal de $Im(A^T)$. Vamos demonstrar este resultado.

Proposição 2.15. *Seja $A \in \mathbb{R}^{m\times n}$. $\mathcal{N}(A) = Im(A^T)^{\perp}$.*

Demonstração. Um vetor $v \in \mathcal{N}(A)$ se, e somente se, $Av = 0$. Mas isso ocorre se, e somente se, Av é ortogonal a todo vetor $u \in \mathbb{R}^m$, isto é, $u^T Av = 0$, para todo $u \in \mathbb{R}^m$. Assim, $(u^T Av)^T = 0^T$ implica que $v^T(A^T u) = 0$. Agora, variando $u \in \mathbb{R}^m$, temos que $A^T u$ fornece o conjunto $Im(A^T)$ e desse modo,

$v^T(A^T v) = 0$, para todo $u \in \mathbb{R}^m$, se, e somente se, $v \in Im(A^T)^\perp$. Portanto, $v \in \mathcal{N}(A)$ se, e somente se, $v \in Im(A^T)^\perp$. □

Assim, $\mathcal{N}(A)$ e $Im(A^T)$ são subespaços vetoriais ortogonais e verificam

$$\mathcal{N}(A) \cap Im(A^T) = \{0\} \quad \text{e} \quad \mathbb{R}^n = \mathcal{N}(A) \oplus Im(A^T). \tag{16}$$

Observação 2.16. A demonstração de (16) pode ser encontrada em Lima [16].

Em vista disso, o teorema a seguir nos dá uma caracterização para o conjunto das direções factíveis em Ω

Teorema 2.17. *Seja $\tilde{x} \in \Omega$. Um vetor $d \in \mathbb{R}^n$ é uma direção factível a partir de $\tilde{x}$ se, e somente se, $d \in \mathcal{N}(A)$.*

Demonstração. Seja $d \in \mathbb{R}^n$ uma direção factível em $\tilde{x}$, então $x := \tilde{x} + \alpha d \in \Omega$, isto é, $Ax = b$. Assim,

$$A(\tilde{x} + \alpha d) = b,$$

e como $A\tilde{x} = b$ e $\alpha > 0$, temos que

$$Ad = 0,$$

e portanto, $d \in \mathcal{N}(A)$.

A recíproca também é verdadeira, pois se $d \in \mathcal{N}(A)$ e $\tilde{x} \in \Omega$, então $x = \tilde{x} + \alpha d \in \Omega$, pois

$$A(\tilde{x} + \alpha d) = A\tilde{x} + \alpha Ad = A\tilde{x} = b,$$

e portanto, qualquer $d \in \mathcal{N}(A)$ é uma direção factível a partir de $\tilde{x}$. □

Dessa forma, concluímos que $\mathcal{N}(A)$ é o conjunto de direções factíveis em Ω, ou seja, qualquer $d \in \mathcal{N}(A)$ é uma direção no espaço na qual podemos nos

deslocar a partir de uma solução factível sem correr o risco de abandonar a região de factibilidade.

A partir disso, é possível construir uma parametrização que caracterize o conjunto factível. Se $\{z_1, z_2, \ldots, z_{n-m}\}$ é uma base de $\mathcal{N}(A)$ e Z a matriz de dimensão $n \times (n-m)$ cujas colunas são os vetores z_i, então para todo $d \in \mathcal{N}(A)$, existe $\boldsymbol{\gamma} \in \mathbb{R}^{n-m}$ tal que $d = Z\boldsymbol{\gamma}$, ou seja, d é escrito como combinação linear dos vetores da base do núcleo de A. Assim, se $\tilde{x}$ é uma solução de (15), podemos caracterizar o conjunto factível da seguinte forma

$$\Omega = \{x \in \mathbb{R}^n \mid x = \tilde{x} + Z\boldsymbol{\gamma},\ \boldsymbol{\gamma} \in \mathbb{R}^{n-m} \text{ e } A\tilde{x} = b\}. \tag{17}$$

2.2.1 Condições necessárias de primeira ordem

Como vimos anteriormente, de modo geral, se um ponto é solução de um problema de otimização então deve satisfazer determinadas propriedades, que são chamadas de condições de otimalidade. Nesta seção abordaremos a condição necessária de primeira ordem para o problema de minimização com restrições de igualdade. Para obter esta condição utilizaremos a parametrização do conjunto factível proposta em (17), transferindo as restrições de (14) para sua função objetivo. Com isso obtemos um novo problema de minimização irrestrita, para o qual as condições necessárias de primeira e segunda ordem já são conhecidas.

Assim sendo, a caracterização de Ω dada em (17) permite definir a seguinte função $\varphi : \mathbb{R}^{n-m} \to \mathbb{R}$ dada por

$$\varphi(\boldsymbol{\gamma}) = f(\tilde{x} + Z\boldsymbol{\gamma}), \tag{18}$$

e a partir dela podemos considerar o seguinte problema de minimização sem restrições

$$\min_{\boldsymbol{\gamma}} \varphi(\boldsymbol{\gamma}). \tag{19}$$

Vejamos um exemplo de como determinar uma parametrização para o conjunto factível e a partir dela obter a função φ.

Exemplo 2.18. Considere o problema

$$\begin{aligned} \min_{x} \quad & x_1^2 + 2x_2^2 - 2x_1 - 2x_1x_2 \\ \text{s.a} \quad & 2x_1 + x_2 = 1. \end{aligned} \tag{20}$$

Podemos reescrever a restrição na forma matricial

$$\begin{bmatrix} 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \end{bmatrix}.$$

Assim, denotando $A = \begin{bmatrix} 2 & 1 \end{bmatrix}$, é preciso primeiramente determinar uma base para $\mathcal{N}(A)$, pois ele é o conjunto das direções factíveis. Para tanto, basta resolver o sistema $Ad = 0$ e a partir disso obtemos

$$Z = \begin{bmatrix} 1 \\ -2 \end{bmatrix},$$

em que a coluna da matriz Z é uma base do $\mathcal{N}(A)$. Por conseguinte, é necessário determinar um ponto $\tilde{x}$ factível. Então, resolvendo

$$2x_1 + x_2 = 1,$$

temos que

$$x_2 = 1 - 2x_1,$$

e escolhendo $x_1 = 1$ obtemos $x_2 = -1$, isto é, $\tilde{x} = (1, -1)$. Assim, podemos reescrever o conjunto factível Ω como proposto em (17):

$$\Omega = \left\{ x \in \mathbb{R}^2 \mid x = \begin{bmatrix} 1 \\ -1 \end{bmatrix} + \gamma \begin{bmatrix} 1 \\ -2 \end{bmatrix},\ \gamma \in \mathbb{R} \right\}.$$

A partir dessa caracterização de Ω definimos a função $\varphi : \mathbb{R} \to \mathbb{R}$ dada por

$$\varphi(\gamma) = f(\tilde{x} + Z\gamma) = 13\gamma^2 + 14\gamma + 3,$$

e o problema (20) pode ser reformulado para um problema de minimização sem restrições dado por

$$\min_{\boldsymbol{\gamma}} \varphi(\boldsymbol{\gamma}) = 13\gamma^2 + 14\gamma + 3. \tag{21}$$

Dessa forma, para determinar uma solução do problema (21) utilizamos a condição suficiente de segunda ordem para problemas irrestritos dada pelo Teorema 2.13. Ou seja, dado $\gamma^* \in \mathbb{R}$, devemos ter

$$\nabla\varphi(\gamma^*) = \Big[26\gamma^* + 14\Big] = 0,$$

e resolvendo para γ^*, obtemos

$$\gamma^* = -\frac{7}{13}.$$

Além disso, como

$$\nabla^2\varphi(\gamma^*) = \Big[26\Big] \succ 0,$$

pois para todo $x \in \mathbb{R}$, com $x \neq 0$, temos que $x^T\nabla^2\varphi(\gamma^*)x = 26x^2 \geq 0$, concluímos que $\gamma^* = -\dfrac{7}{13}$ é minimizador local de (21). E definindo $x^* \coloneqq \tilde{x} + Z\boldsymbol{\gamma}^*$ temos que

$$x^* = \begin{bmatrix} 1 \\ -1 \end{bmatrix} - \frac{7}{13}\begin{bmatrix} 1 \\ -2 \end{bmatrix} = \begin{bmatrix} \frac{6}{13} \\ \frac{1}{13} \end{bmatrix},$$

é minimizador local do problema (20).

Com efeito, para determinar a solução do problema (20) utilizamos o fato de haver uma equivalência entre sua solução e a solução do problema irrestrito (21). Essa equivalência entre as soluções é garantida pela proposição a seguir.

Proposição 2.19. *O vetor $\boldsymbol{\gamma}^*$ é um minimizador local (global) de φ em $\mathbb{R}^{n-m}$ se, e somente se, $x^* \coloneqq \tilde{x} + Z\boldsymbol{\gamma}^*$ é um minimizador local (global) de* (14).

Demonstração. Seja $\boldsymbol{\gamma}^*$ um minimizador local (global) de φ em $\mathbb{R}^{n-m}$. Então, existe $\varepsilon > 0$ tal que

$$\varphi(\boldsymbol{\gamma}^*) \leq \varphi(\boldsymbol{\gamma}),$$

para todo $\boldsymbol{\gamma} \in B(\boldsymbol{\gamma}^*, \delta) \cap \mathbb{R}^{n-m}$, em que $\delta \coloneqq \dfrac{\varepsilon}{\|Z\|}$. Logo, definindo $x^* \coloneqq \tilde{x} + Z\boldsymbol{\gamma}^* \in \Omega$ temos, para $\boldsymbol{\gamma} \in B(\boldsymbol{\gamma}^*, \delta) \cap \mathbb{R}^{n-m}$, que

$$f(x^*) = f(\tilde{x} + Z\boldsymbol{\gamma}^*) = \varphi(\boldsymbol{\gamma}^*) \leq \varphi(\boldsymbol{\gamma}) = f(\tilde{x} + Z\boldsymbol{\gamma}) = f(x),$$

em que $x = \tilde{x} + Z\boldsymbol{\gamma} \in \Omega$. Note que

$$\begin{aligned} \|x - x^*\| &= \|(\tilde{x} + Z\boldsymbol{\gamma}) - (\tilde{x} + Z\boldsymbol{\gamma}^*)\| \\ &= \|Z(\boldsymbol{\gamma} - \boldsymbol{\gamma}^*)\| \\ &\leq \|Z\|\|\boldsymbol{\gamma} - \boldsymbol{\gamma}^*\| \\ &< \|Z\|\frac{\varepsilon}{\|Z\|} = \varepsilon, \end{aligned}$$

em que na primeira desigualdade usamos a consistência da norma-2. Portanto, $x \in B(x^*, \varepsilon) \cap \Omega$, e assim, x^* é um minimizador local (global) de (14).

Reciprocamente, suponhamos que $\boldsymbol{\gamma}^* \in \mathbb{R}^{n-m}$ não é mínimo local (global) de φ. Então, dado $\varepsilon > 0$ existe $\bar{\boldsymbol{\gamma}} \in B(\boldsymbol{\gamma}^*, \delta)$, com $\delta \coloneqq \dfrac{\varepsilon}{\|Z\|}$, tal que

$$\varphi(\bar{\boldsymbol{\gamma}}) < \varphi(\boldsymbol{\gamma}^*).$$

Neste caso, $f(\bar{x}) < f(x^*)$, em que $x^* \coloneqq \tilde{x} + Z\boldsymbol{\gamma}^*, \bar{x} \coloneqq \tilde{x} + Z\bar{\boldsymbol{\gamma}} \in \Omega$. Além disso,

$$\begin{aligned}\|\bar{x} - x^*\| &= \|(\tilde{x} + Z\bar{\boldsymbol{\gamma}}) - (\tilde{x} + Z\boldsymbol{\gamma}^*)\| \\ &= \|Z(\bar{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)\| \\ &\leq \|Z\|\|\bar{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\| \\ &< \|Z\|\delta \\ &= \|Z\|\frac{\varepsilon}{\|Z\|} = \varepsilon,\end{aligned}$$

isto é, $\bar{x} \in B(x^*, \varepsilon)$. Portanto, x^* não é minimizador local (global) de (14). □

Agora, como (19) é um problema de minimização irrestrito, o Teorema 2.8 nos diz que a condição necessária de primeira ordem para algum $\boldsymbol{\gamma}^* \in \mathbb{R}^{n-m}$ é

$$\nabla\varphi(\boldsymbol{\gamma}^*) = 0. \tag{22}$$

Por (18) e definindo a função $g : \mathbb{R}^{n-m} \to \mathbb{R}^n$, com $g(\boldsymbol{\gamma}) = \tilde{x} + Z\boldsymbol{\gamma}$, temos que $\varphi(\boldsymbol{\gamma}) = f(g(\boldsymbol{\gamma}))$. Aplicando a regra da cadeia para calcular sua derivada, obtemos

$$\nabla\varphi(\boldsymbol{\gamma})^T = \nabla f(g(\boldsymbol{\gamma}))J_g(\boldsymbol{\gamma}),$$

em que $J_g(\boldsymbol{\gamma})$ é a matriz Jacobiana de g em $\boldsymbol{\gamma}$. Note que $J_g(\boldsymbol{\gamma}) = Z$, portanto

$$\nabla\varphi(\boldsymbol{\gamma}) = Z^T\nabla f(g(\boldsymbol{\gamma})).$$

Desse modo, da condição (22), resulta que

$$0 = \nabla\varphi(\boldsymbol{\gamma}^*) = Z^T\nabla f(\tilde{x} + Z\boldsymbol{\gamma}^*) = Z^T\nabla f(x^*).$$

Consequentemente, uma condição necessária de primeira ordem para que x^* seja minimizador local de (14) é que

$$Z^T\nabla f(x^*) = 0, \tag{23}$$

ou seja, que $(z_i)^T \nabla f(x^*) = 0$, para todo $i = 1, \ldots, n-m$. Como $\{z_1, \ldots, z_{n-m}\}$ é uma base de $\mathcal{N}(A)$, tal condição implica que $\nabla f(x^*)$ seja ortogonal a $\mathcal{N}(A)$. Logo, pelas considerações feitas anteriormente (Proposição 2.15), temos que $\nabla f(x^*) \in Im(A^T)$, em outras palavras, $\nabla f(x^*)$ deve ser uma combinação linear das linhas de A. Portanto, existe $\boldsymbol{\lambda}^* \in \mathbb{R}^m$ tal que

$$\nabla f(x^*) = A^T \boldsymbol{\lambda}^*. \tag{24}$$

Observe que as condições propostas em (23) e (24) são equivalentes. Com efeito, se (23) se verifica, isso implica que $\nabla f(x^*) \in \mathcal{N}(A)^\perp = Im(A^T)$ e portanto (24) é verdadeiro. Reciprocamente, se (24) ocorre, temos que, pela Proposição 2.15, $Z^T \nabla f(x^*) = Z^T (A^T \boldsymbol{\lambda}^*) = 0$.

Assim, em vista do que segue acima temos o seguinte teorema.

Teorema 2.20. *(**Multiplicadores de Lagrange**) Se x^* é um minimizador local de (14), então existe $\boldsymbol{\lambda}^* \in \mathbb{R}^m$ tal que $(x^*, \boldsymbol{\lambda}^*)$ é solução do seguinte sistema de $(n+m)$ equações*

$$\begin{aligned} \nabla f(x^*) &= A^T \boldsymbol{\lambda}^* \\ Ax^* &= b. \end{aligned} \tag{25}$$

O vetor $\boldsymbol{\lambda}^* \in \mathbb{R}^m$ é chamado vetor de *multiplicadores de Lagrange* associado a x^*.

A solução de (14) é necessariamente solução de (25), porém para que a recíproca seja verdadeira é necessária informação de segunda ordem.

Vejamos a seguir um exemplo em que as condições de otimalidade de primeira ordem nos permitem determinar a expressão geral para uma solução de um problema.

Exemplo 2.21. Queremos encontrar uma solução do sistema linear $Ax = b$, em que $m < n$, (portanto um sistema linear com infinitas soluções), com a

menor norma possível. Podemos descrever matematicamente este problema da seguinte forma

$$\begin{aligned} \min_{x} \quad & \frac{1}{2}\|x\|^2 \\ \text{s.a} \quad & Ax = b, \end{aligned} \tag{26}$$

com $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $m < n$ e posto$(A) = m$. Com efeito, seja $\tilde{x}$ solução de (26). Então, $\tilde{x}$ também minimiza $\|x\|^2$ que por sua vez minimiza $\|x\|$, pois os problemas são equivalentes. Além disso, veremos que a solução $\tilde{x}$ desse problema pode ser escrita como $\tilde{x} = A^{\dagger}b$, em que $A^{\dagger} \in \mathbb{R}^{n\times m}$ e $AA^{\dagger} = I$.

Assim, inicialmente $\tilde{x}$ deve satisfazer a restrição de igualdade, isto é,

$$A\tilde{x} = b. \tag{27}$$

Além disso, se $\tilde{x}$ é solução então,

$$\nabla f(\tilde{x}) = A^T\tilde{\boldsymbol{\lambda}},$$

e como $\nabla f(\tilde{x}) = \tilde{x}$, obtemos

$$\tilde{x} = A^T\tilde{\boldsymbol{\lambda}}. \tag{28}$$

Substituindo (28) em (27), temos

$$AA^T\tilde{\boldsymbol{\lambda}} = b,$$

e como por hipótese A é posto completo, temos que AA^T é não singular. Desse modo,

$$(AA^T)^{-1}AA^T\tilde{\boldsymbol{\lambda}} = (AA^T)^{-1}b,$$

o que implica em

$$\tilde{\boldsymbol{\lambda}} = (AA^T)^{-1}b. \tag{29}$$

Agora, substituindo (29) em (28), concluímos que

$$\tilde{x} = A^T(AA^T)^{-1}b = A^{\dagger}b,$$

em que $A^{\dagger} = A^T(AA^T)^{-1} \in \mathbb{R}^{n\times m}$ e $AA^{\dagger} = AA^T(AA^T)^{-1} = I$.

Observação 2.22. A matriz $A^{\dagger}$ é também chamada de *pseudo-inversa* de Moore-Penrose. Para mais informações consulte Meyer [21, p. 423].

2.2.2 Condições necessárias e suficientes de segunda ordem

A condição necessária de segunda ordem para uma solução $\boldsymbol{\gamma}^*$ do problema (19) é dada pelo Teorema 2.10 e toma a seguinte forma

$$\nabla^2\varphi(\boldsymbol{\gamma}^*) \succcurlyeq 0. \tag{30}$$

Derivando $\nabla\varphi(\boldsymbol{\gamma}) = Z^T\nabla f(\tilde{x} + Z\boldsymbol{\gamma})$ através da regra da cadeia, obtemos

$$\nabla^2\varphi(\boldsymbol{\gamma}) = Z^T\nabla^2 f(\tilde{x} + Z\boldsymbol{\gamma})Z.$$

Assim, por (30) temos que a condição necessária de segunda ordem para que x^* seja minimizador local de (14) é

$$Z^T\nabla^2 f(x^*)Z \succcurlyeq 0.$$

Ademais, observe que $Z^T\nabla^2 f(x^*)Z$ é uma matriz de ordem $(n-m)\times(n-m)$, e o fato de ser semidefinida positiva significa que

$$y^T\nabla^2 f(x^*)y \geq 0 \text{ para todo } y \in \mathcal{N}(A).$$

Isso significa que $\nabla^2 f(x^*)$ é semidefinida positiva para todo $y \in \mathcal{N}(A)$. Desse modo, com base no que foi desenvolvido acima temos o seguinte teorema da condição necessária de segunda ordem.

Teorema 2.23. ***(Condição necessária de segunda ordem)*** *Se $x^* \in \mathbb{R}^n$ é minimizador local de* (14)*, então*

(i) existe $\boldsymbol{\lambda}^* \in \mathbb{R}^m$ tal que $\nabla f(x^*) = A^T\boldsymbol{\lambda}^*$;

(ii) para todo $y \in \mathcal{N}(A)$ temos que $y^T\nabla^2 f(x^*)y \geq 0$.

Analogamente, aplicando o Teorema 2.13 para o problema (19) podemos determinar as condições suficientes de segunda ordem para o problema (14). De fato, pelo Teorema 2.13, se $\boldsymbol{\gamma}^* \in \mathbb{R}^{n-m}$ satisfaz

$$\nabla\varphi(\boldsymbol{\gamma}^*) = 0 \tag{31}$$

e

$$\nabla^2\varphi(\boldsymbol{\gamma}^*) \succ 0, \tag{32}$$

então $\boldsymbol{\gamma}^*$ é minimizador local de (19). Então, com base no que foi desenvolvido até o momento, temos que de (31) decorre que existe $\boldsymbol{\lambda}^* \in \mathbb{R}^m$ tal que

$$\nabla f(x^*) = A^T\boldsymbol{\lambda}^*, \tag{33}$$

e de (32) resulta que para todo $y \in \mathcal{N}(A)$, com $y \neq 0$,

$$y^T\nabla^2 f(x^*)y > 0. \tag{34}$$

Portanto, se $x^* \in \Omega$ satisfaz (33) e (34), então ele é um minimizador local de (14). Assim, usando os mesmos argumentos dos Teoremas 2.20 e 2.23, segue o Teorema 2.24.

Teorema 2.24. ***(Condição suficiente de segunda ordem)*** *Se $x^* \in \mathbb{R}^n$ verifica $Ax^* = b$ e*

(i) existe $\boldsymbol{\lambda}^* \in \mathbb{R}^m$ tal que $\nabla f(x^*) = A^T\boldsymbol{\lambda}^*$;

(ii) para todo $y \in \mathcal{N}(A)$*, com* $y \neq 0$*, temos que* $y^T \nabla^2 f(x^*) y > 0$*;*

então x^* *é um minimizador local de* (14)*.*

Vejamos um exemplo em que as condições suficientes nos permitem determinar uma solução para o problema de minimizar uma função quadrática sujeita a restrições de igualdade.

Exemplo 2.25. Considere o problema

$$\begin{aligned} \min_{x} \quad & \frac{1}{2} x^T Q x + p^T x + q \\ \text{s.a} \quad & Ax = b, \end{aligned} \tag{35}$$

em que $Q \in \mathbb{R}^{n \times n}$ é simétrica, $x, p \in \mathbb{R}^n$, $q \in \mathbb{R}$, $A \in \mathbb{R}^{m \times n}$ e $b \in \mathbb{R}^m$. Seja Z uma base do $\mathcal{N}(A)$ e suponha que $Z^T Q Z$ é definida positiva. Seja x^0 tal que $Ax^0 = b$. Então a solução $\tilde{x}$ é dada por

$$\tilde{x} = x^0 - Z(Z^T Q Z)^{-1} Z^T (Qx^0 + p). \tag{36}$$

Para provar que (36) é solução é preciso verificar se $\tilde{x}$ cumpre as condições suficientes de segunda ordem. Para tanto, devemos ter:

(i) $\tilde{x}$ cumpre a restrição de igualdade. De fato,

$$\begin{aligned} A\tilde{x} &= A(x^0 - Z(Z^T Q Z)^{-1} Z^T (Qx^0 + p)) \\ &= Ax^0 - AZ(Z^T Q Z)^{-1} Z^T (Qx^0 + p) \\ &= b, \end{aligned}$$

pois $AZ = 0$ pelo fato de cada coluna de Z pertencer ao $\mathcal{N}(A)$ e portanto, serem ortogonais as linhas de A.

(ii) É preciso verificar se $Z^T \nabla f(\tilde{x}) = 0$. Como $\nabla f(\tilde{x}) = Q\tilde{x} + p$, temos que

$$\begin{aligned} Z^T(Q\tilde{x} + p) &= Z^T Q(x^0 - Z(Z^T QZ)^{-1} Z^T(Qx^0 + p)) + Z^T p \\ &= Z^T Qx^0 - Z^T QZ(Z^T QZ)^{-1} Z^T(Qx^0 + p) + Z^T p \\ &= Z^T Qx^0 - Z^T Qx^0 - Z^T p + Z^T p \\ &= 0. \end{aligned}$$

(iii) Por fim, $Z^T \nabla^2 f(\tilde{x}) Z$ deve ser definida positiva. Assim, como $\nabla^2 f(\tilde{x}) = Q$, temos que

$$Z^T QZ \succ 0.$$

Portanto, $\tilde{x}$ é solução do problema (35).

Observação 2.26. Veremos mais a frente que no caso em que Q é uma matriz semidefinida positiva, os problemas quadráticos são convexos.

2.3 MINIMIZAÇÃO COM RESTRIÇÕES LINEARES DE DESIGUALDADE

Vamos considerar agora problemas da forma

$$\begin{aligned} \min_x \quad & f(x) \\ \text{s.a} \quad & Wx \leq c, \end{aligned} \tag{37}$$

em que $x \in \mathbb{R}^n$ e $W \in \mathbb{R}^{m \times n}$. Como na seção anterior, estamos interessados nesse primeiro momento em analisar a região de factibilidade e determinar as direções factíveis, que são aquelas em que há espaço para se movimentar dentro da região Ω.

Denotando cada linha da matriz W da forma $w_i^T = (w_{i1}, w_{i2}, \ldots, w_{in})$, podemos caracterizar o conjunto factível da seguinte forma

$$\Omega = \{x \in \mathbb{R}^n \mid w_i^T x \leq c_i \text{ para todo } i = 1, 2, \ldots, m\}.$$

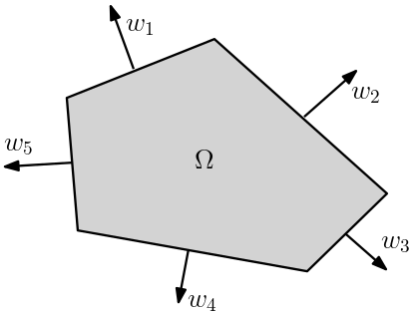
Cada uma das m desigualdades $w_i^T x \leq c_i$ define em $\mathbb{R}^n$ um *semiespaço*. O hiperplano divisor é dado por $\mathcal{H}(w_i, c_i) = \{x \mid w_i^T x = c_i\}$ e o semiespaço definido é aquele que está do lado contrário à direção apontada pelo vetor w_i. Desta forma, a região Ω consiste na intersecção dos m semiespaços. Este tipo de conjunto de $\mathbb{R}^n$ é chamado *poliedro* ou conjunto *poliedral*. Vamos definir formalmente este conceito.

Definição 2.27. Um *poliedro* ou um conjunto *poliedral* $\mathcal{P} \subset \mathbb{R}^n$ é definido como o conjunto de soluções de um sistema finito de equações e inequações lineares:

$$\mathcal{P} = \{x \in \mathbb{R}^n \mid Ax = b,\ Wx \leq c\},$$

em que $A \in \mathbb{R}^{m \times n}$, $W \in \mathbb{R}^{p \times n}$, $b \in \mathbb{R}^m$ e $c \in \mathbb{R}^p$. Neste contexto, dizemos que $h : \mathbb{R}^n \to \mathbb{R}^m$, dada por $h(x) = Ax - b$, e $g : \mathbb{R}^n \to \mathbb{R}^p$, dada por $g(x) = Wx - c$, são funções *afim*.

A Figura 4 representa um poliedro em $\mathbb{R}^2$.



Fonte – Friedlander [12]

Figura 4 – Conjunto factível poliedral.

Um conceito fundamental para formular as condições de otimalidade para o problema (37) e que fornece uma grande quantidade de informações é o de *restrição ativa*. Portanto, consideramos, no que segue, que Ω é o conjunto factível deste problema.

Definição 2.28. Dado $\bar{x} \in \Omega$, dizemos que a restrição de desigualdade $w_i^T \bar{x} \leq c_i$ que corresponde ao índice $i \in \{1, 2, \ldots, m\}$ é *ativa* no ponto $\bar{x}$ quando $w_i^T \bar{x} = c_i$. Caso $w_i^T \bar{x} < c_i$ dizemos que a restrição é *inativa* em $\bar{x}$. O conjunto dos índices das restrições de desigualdade ativas no ponto $\bar{x} \in \Omega$ é denotado por

$$\mathcal{I}(\bar{x}) = \{i = 1, 2, \ldots, m \mid w_i^T \bar{x} = c_i\}.$$

A cada $x \in \Omega$ podemos associar um número $r(x)$, com $0 \leq r(x) \leq m$, que representa a quantidade de restrições ativas em x. Assim, na Figura 5 temos que $r(x^1) = 1$, $r(x^2) = 2$, $r(x^3) = 1$, $r(x^4) = 0$ e $r(x^5) = 2$.

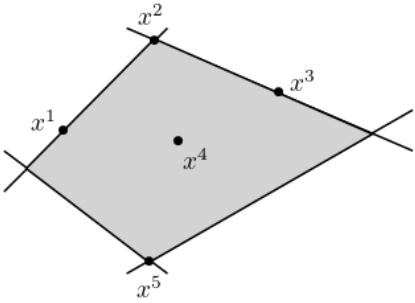


Figura 5 – Restrições ativas e inativas.

Para $x^* \in \Omega$, com $1 \leq r(x^*) \leq m$, vamos denotar por $\mathcal{I}(x^*) = \{i_1, i_2, \ldots, i_{r(x^*)}\}$ o conjunto dos índices das restrições ativas em x^*, e por $W_{\mathcal{I}} \in \mathbb{R}^{r(x^*) \times n}$ a submatriz de W cujas linhas são as que tem índices em

$\mathcal{I}(x^*)$ e, consequentemente,

$$c_{\mathcal{I}} = \begin{bmatrix} c_{i_1} \\ c_{i_2} \\ \vdots \\ c_{i_{r(x^*)}} \end{bmatrix}.$$

Ademais, o conjunto das direções factíveis a partir de x depende das restrições ativas nesse ponto, pois são elas que restringem o domínio de factibilidade na vizinhança de x, enquanto que as restrições inativas não possuem influência na vizinhança de x. O Teorema 2.29 nos fornece uma caracterização das direções factíveis em x.

Teorema 2.29. *Considere $x \in \Omega$ tal que $r(x) = p$ com $0 < p \leq m$. Um vetor $d \in \mathbb{R}^n$ é uma direção factível a partir de x se, e somente se, $w_i^T d \leq 0$ para todo $i \in \mathcal{I}(x)$.*

Demonstração. Suponhamos que $d \in \mathbb{R}^n$ é uma direção factível em x. Então existe $\tilde{\alpha} > 0$ tal que $x + \alpha d \in \Omega$ para todo $\alpha \in (0, \tilde{\alpha}]$, o que ocorre se, e somente se, $W(x + \alpha d) \leq c$, ou seja, $w_i^T(x + \alpha d) \leq c_i$, para todo $i \in \mathcal{M} \coloneqq \{1, 2, \ldots, m\}$. Em particular, se $i \in \mathcal{I}(x)$ temos que

$$w_i^T(x + \alpha d) = c_i + \alpha w_i^T d \leq c_i,$$

e consequentemente, temos

$$w_i^T d \leq 0.$$

Reciprocamente, seja $d \in \mathbb{R}^n$ e suponhamos que $w_i^T d \leq 0$ para todo $i \in \mathcal{I}(x)$. Assim, se $i \in \mathcal{I}(x)$, temos que $w_i^T x = c_i$, implicando que, para qualquer $\alpha > 0$, $w_i^T(x + \alpha d) \leq c_i$. Por outro lado, se $i \notin \mathcal{I}(x)$, temos que $w_i^T x < c_i$. Neste caso, é preciso analisar duas situações:

(i) Se $w_i^T d \leq 0$, então $w_i^T(x + \alpha d) \leq c_i$, para qualquer $\alpha > 0$;

(ii) Se $w_i^T d > 0$ podemos encontrar $\alpha_i > 0$ de modo que $w_i^T x + w_i^T(\alpha_i d) = c_i$. Considere que como $w_i^T x < c_i$ existe $\beta_i > 0$, tal que $w_i^T x + \beta_i = c_i$. Desse modo, temos que escolher α_i tal que $\beta_i = \alpha_i w_i^T d$, isto é,

$$w_i^T x + \alpha_i w_i^T d = c_i,$$

para todo $i \notin \mathcal{I}(x)$. Podemos fazer isso resolvendo a equação anterior para α_i, encontrando

$$\alpha_i = \frac{c_i - w_i^T x}{w_i^T d}. \tag{38}$$

Assim, se tomarmos o mínimo de (38), para todo $i \in \mathcal{M}\backslash\mathcal{I}(x)$, isto é, se definirmos

$$\tilde{\alpha} = \min_{i \in \mathcal{M}\backslash\mathcal{I}(x)} \left(\frac{c_i - w_i^T x}{w_i^T d} \right), \tag{39}$$

teremos que $w_i^T(x + \alpha d) \leq c_i$ para todo $i \in \mathcal{M}$ e $\alpha \in (0, \tilde{\alpha}]$.

Portanto, de (i) e (ii), d será uma direção factível a partir de x. □

A demonstração do Teorema 2.29 nos dá, na equação (39), o quanto podemos andar em uma direção factível d a partir de $x \in \Omega$ e ainda continuarmos sobre o conjunto factível. Este critério é também chamado *Teste da razão*.

2.3.1 Condições necessárias de primeira ordem

No intuito de determinar as condições de otimalidade no caso de problemas com restrições de desigualdade, acabamos de desenvolver a noção de direções factíveis para avaliar a estrutura do conjunto factível na vizinhança de uma solução. Agora, é necessário definir o conceito de *direção de descida*.

Definição 2.30. Dizemos que $d \in \mathbb{R}^n$ é uma *direção de descida* de $f : \mathbb{R}^n \to \mathbb{R}$ no ponto $\bar{x} \in \mathbb{R}^n$, se existe $\varepsilon > 0$ tal que

$$f(\bar{x} + td) < f(\bar{x})$$

para todo $t \in (0, \varepsilon]$.

Denotamos por $\mathcal{D}_f(\bar{x})$ o conjunto de todas as direções de descida da função f no ponto $\bar{x}$.

Lema 2.31. *Seja $f : \mathbb{R}^n \to \mathbb{R}$ uma função diferenciável no ponto $\bar{x} \in \mathbb{R}^n$. Então:*

(i) Para todo $d \in \mathcal{D}_f(\bar{x})$, tem-se que $\nabla f(\bar{x})^T d \leq 0$.

(ii) Se $d \in \mathbb{R}^n$ satisfaz $\nabla f(\bar{x})^T d < 0$, tem-se que $d \in \mathcal{D}_f(\bar{x})$.

Demonstração. Seja $d \in \mathcal{D}_f(\bar{x})$. Assim, para todo $t > 0$ suficientemente pequeno,

$$f(\bar{x} + td) < f(\bar{x}). \tag{40}$$

Pelo Fato 2.6, temos

$$f(\bar{x} + td) - f(\bar{x}) = t\nabla f(\bar{x})^T d + o(t), \tag{41}$$

com $\lim\limits_{t \to 0} \dfrac{o(t)}{t} = 0$, e de (40) e (41), obtemos

$$t\nabla f(\bar{x})^T d + o(t) < 0. \tag{42}$$

Dividindo ambos os lados da desigualdade (42) por $t > 0$ e passando o limite quando $t \to 0^+$, obtemos

$$\nabla f(\bar{x})^T d \leq 0,$$

provando o item (i).

Suponhamos agora que $\nabla f(\bar{x})^T d < 0$. Aplicando o Fato 2.6, podemos escrever

$$f(\bar{x} + td) - f(\bar{x}) = t\left(\nabla f(\bar{x})^T d + \frac{o(t)}{t}\right).$$

Em particular, para todo $t > 0$ suficientemente pequeno, temos

$$\nabla f(\bar{x})^T d + \frac{o(t)}{t} \leq \frac{1}{2}\nabla f(\bar{x})^T d < 0,$$

o que implica em

$$f(\bar{x} + td) - f(\bar{x}) < 0.$$

Portanto, $d \in \mathcal{D}_f(\bar{x})$. □

Assim, a partir de um ponto $\bar{x} \in \Omega$, estamos interessados em saber se existem direções de descida factíveis, isto é, direções factíveis tais que

$$\nabla f(\bar{x})^T d < 0, \tag{43}$$

pois se existe uma direção d que satisfaz (43), decorre do Lema 2.31 e da Definição 2.30 que $\bar{x}$ não é minimizador local do problema (37).

Observação 2.32. Novamente, a análise dependerá das restrições ativas em um dado ponto $\bar{x}$. Em particular, se $r(\bar{x}) = 0$, o ponto está no interior de Ω e as condições necessárias e suficientes são as mesmas do caso em que o problema é irrestrito, pois qualquer direção é factível nesse ponto.

Agora, já possuímos os ferramentais necessários para determinar a condição necessária de primeira ordem a ser satisfeita por uma solução do problema (37).

Teorema 2.33. ***(Condição necessária de primeira ordem)*** *Considere o problema* (37) *com* $f \in \mathcal{C}^1$. *Se* x^* *é minimizador local de* (37) *e* $\text{posto}(W_{\mathcal{I}}) = r(x^*)$, *então existe* $\boldsymbol{\mu} \in \mathbb{R}^{r(x^*)}$ *tal que*

$$\nabla f(x^*) + W_{\mathcal{I}}^T \boldsymbol{\mu} = 0 \quad e \quad \mu_k \geq 0,\ 1 \leq k \leq r(x^*). \tag{44}$$

Demonstração. Suponhamos, por contradição, que (44) não ocorre. Isso pode acontecer por dois motivos:

(i) $\nabla f(x^*) + W_{\mathcal{I}}^T \boldsymbol{\mu} \neq 0$ para todo $\boldsymbol{\mu} \in \mathbb{R}^{r(x^*)}$.

Assim, temos que $\nabla f(x^*)$ não pode ser escrito como combinação linear das linhas de $W_{\mathcal{I}}$. Em outras palavras, x^* não é minimizador local do seguinte problema

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{s.a} \quad & W_{\mathcal{I}} x = c_{\mathcal{I}}, \end{aligned} \tag{45}$$

pois x^* não satisfaz a condição necessária de primeira ordem para o problema (45) estabelecida pelo Teorema 2.20. Logo, x^* não pode ser minimizador local do problema (37), o que é uma contradição.

(ii) $\nabla f(x^*) + W_{\mathcal{I}}^T \boldsymbol{\mu} = 0$, mas existe j tal que $\mu_j < 0$.

Primeiramente, se $r(x^*) = 1$ temos $\mathcal{I} = \{i_1\}$ e então,

$$\nabla f(x^*) + \mu_1 w_{i_1} = 0,$$

com $\mu_1 < 0$, ou equivalentemente,

$$\nabla f(x^*) = -\mu_1 w_{i_1}.$$

Assim, tomando $d = -\nabla f(x^*) = \mu_1 w_{i_1}$, com $\mu_1 < 0$, temos que d é uma direção de descida, pois

$$\nabla f(x^*)^T d = -\nabla f(x^*)^T \nabla f(x^*) = -\|\nabla f(x^*)\|^2 < 0,$$

e além disso, d é uma direção factível, pois

$$w_{i_1}^T d = \mu_1 w_{i_1}^T w_{i_1} = \mu_1 \|w_{i_1}\|^2 < 0,$$

devido o fato de $\mu_1 < 0$. Desse modo, d é uma direção de descida factível no ponto x^*, e portanto x^* não pode ser minimizador local de (37), contradizendo a hipótese.

Agora, considere o caso em que $2 \leq r(x^*) \leq n$. Vamos denotar por $\tilde{W_{\mathcal{I}}}$ a matriz obtida retirando a linha w_{i_j} que corresponde ao multiplicador $\mu_j < 0$. Tomando $d = \text{proj}_{\mathcal{N}(\tilde{W_{\mathcal{I}}})}(-\nabla f(x^*))$, temos que

$$(-\nabla f(x^*) - d)^T d = 0,$$

pois eles são ortogonais (veja Figura 6), o que implica em

$$\nabla f(x^*)^T d = -d^T d = -\|d\|^2 < 0. \tag{46}$$

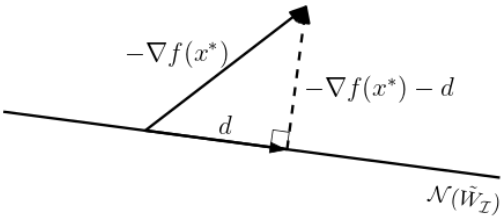


Figura 6 – Projeção de $-\nabla f(x)$ sobre o $\mathcal{N}(\tilde{W_{\mathcal{I}}})$.

Fonte – Friedlander [12]

Logo, d é uma direção de descida. Além disso, d é uma direção factível. Com efeito, temos que

$$\nabla f(x^*) + \mu_1 w_{i_1} + \mu_2 w_{i_2} + \ldots + \mu_j w_{i_j} + \ldots + \mu_{r(x^*)} w_{i_{r(x^*)}} = 0$$

com $\mu_j < 0$, e como tomamos $d \in \mathcal{N}(\tilde{W_{\mathcal{I}}})$ temos que

$$w_{i_k}^T d = 0, \tag{47}$$

para todo $k \neq j$. Então,

$$\nabla f(x^*)^T d + \mu_j w_{i_j}^T d = 0,$$

e reescrevendo, obtemos

$$\nabla f(x^*)^T d = -\mu_j w_{i_j}^T d.$$

Agora, utilizando (46) e o fato de $\mu_j < 0$, obtemos

$$w_{i_j}^T d < 0. \tag{48}$$

Assim, de (47) e (48), concluímos que

$$w_{i_k}^T d \leq 0,$$

para todo k tal que $1 \leq k \leq r(x^*)$, ou seja, d é uma direção factível. Portanto, d é uma direção factível e de descida a partir de x^*, o que contradiz o fato de x^* ser minimizador local de (37).

□

2.3.2 Condições necessárias e suficientes de segunda ordem

Desenvolveremos a seguir as condições necessárias e as condições suficientes de segunda ordem para o problema (37) com restrições de desigualdade.

Teorema 2.34. ***(Condição necessária de segunda ordem)*** *Considere o problema* (37) *com* $f \in \mathcal{C}^2$ *e* $r(x^*)$ *e* $\mathcal{I}(x^*)$ *definidos como anteriormente. Se* x^* *é um minimizador local do problema* (37)*, então*

(i) existe $\boldsymbol{\mu} \in \mathbb{R}^{r(x^*)}$ *tal que* $\nabla f(x^*) + W_{\mathcal{I}}^T \boldsymbol{\mu} = 0$ *e* $\mu_k \geq 0$ *para todo* $i_k \in \mathcal{I}(x^*)$*;*

(ii) para todo $y \in \mathcal{N}(W_{\mathcal{I}})$ *temos que* $y^T \nabla^2 f(x^*) y \geq 0$.

Demonstração. Perceba que se x^* é minimizador local de (37), então pelo Teorema 2.33, existe $\boldsymbol{\mu} \in \mathbb{R}^{r(x^*)}$ tal que

$$\nabla f(x^*) + W_{\mathcal{I}}^T \boldsymbol{\mu} = 0 \quad \text{e} \quad \mu_k \geq 0,$$

para todo $i_k \in \mathcal{I}(x^*)$.

Por conseguinte, se x^* é minimizador local de (37), em particular, ele é solução do seguinte problema com restrições de igualdade

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{s.a} \quad & W_{\mathcal{I}} x = c_{\mathcal{I}}, \end{aligned}$$

em que $W_{\mathcal{I}}$ é a submatriz de W cujas linhas correspondem às restrições ativas em x^*. Então, pelo Teorema 2.23, concluímos que

$$y^T \nabla^2 f(x^*) y \geq 0,$$

para todo $y \in \mathcal{N}(W_{\mathcal{I}})$. □

Teorema 2.35. ***(Condição suficiente de segunda ordem)*** *Considere o problema* (37) *com* $f \in \mathcal{C}^2$ *e* $r(x^*)$ *e* $\mathcal{I}(x^*)$ *definidos como anteriormente. Se* $x^* \in \Omega$ *satisfaz*

(i) $\nabla f(x^*) + W_{\mathcal{I}}^T \boldsymbol{\mu} = 0$ *com* $\mu_k \geq 0$ *para todo* $i_k \in \mathcal{I}(x^*)$*;*

(ii) $y^T \nabla^2 f(x^*) y > 0$ *para todo* $y \in \mathcal{N}(W_{\mathcal{J}})$, $y \neq 0$, *em que* $\mathcal{J} = \{k \in \{1, 2, \ldots, r(x^*)\} \mid \mu_k > 0\}$*;*

então x^* *é um minimizador local de* (37).

Demonstração. Primeiramente, se $\mu_k > 0$ para todo $i_k \in \mathcal{I}(x^*)$, então

$$W_{\mathcal{I}} = W_{\mathcal{J}}.$$

Assim, como $x^* \in \Omega$ satisfaz

$$\nabla f(x^*) + W_{\mathcal{J}}^T \boldsymbol{\mu} = 0$$

e

$$y^T \nabla^2 f(x^*) y > 0,$$

para todo $y \in \mathcal{N}(W_{\mathcal{J}})$ e $y \neq 0$, então, pelo Teorema 2.24, x^* é minimizador local do problema

$$\begin{aligned} \min_x \quad & f(x) \\ \text{s.a} \quad & W_{\mathcal{J}} x = c_{\mathcal{J}}, \end{aligned}$$

e consequentemente, é minimizador local do problema (37).

Agora, suponha que existe pelo menos um k tal que $\mu_k = 0$. Neste caso, observe que

$$W_{\mathcal{I}}^T \boldsymbol{\mu} = W_{\mathcal{J}}^T \boldsymbol{\mu}_{\mathcal{J}},$$

em que $\boldsymbol{\mu}_{\mathcal{J}}$ é vetor formado pelas componentes μ_k que possuem índice em $\mathcal{J} = \{k \in \{1, 2, \ldots, r(x^*)\} \mid \mu_k > 0\}$. Desse modo, temos que $x^* \in \Omega$ satisfaz

$$\nabla f(x^*) + W_{\mathcal{J}}^T \boldsymbol{\mu}_{\mathcal{J}} = 0$$

e

$$y^T \nabla^2 f(x^*) y > 0,$$

para todo $y \in \mathcal{N}(W_{\mathcal{J}})$ e $y \neq 0$. Assim, o Teorema 2.24 nos garante que x^* é minimizador local do problema

$$\begin{aligned} \min_x \quad & f(x) \\ \text{s.a} \quad & W_{\mathcal{J}} x = c_{\mathcal{J}}, \end{aligned}$$

e portanto, x^* é minimizador local do problema (37). □

Vejamos no exemplo a seguir as condições de otimalidade de segunda ordem para um problema de minimização em que a função objetivo é uma quadrática.

Exemplo 2.36. Considere o problema de programação quadrática

$$\begin{aligned} \min_x \quad & \frac{1}{2}x^T Q x + p^T x \\ \text{s.a} \quad & Wx \leq c, \end{aligned} \tag{49}$$

em que $Q \in \mathbb{R}^{n\times n}$ é simétrica, $p \in \mathbb{R}^n$, $W \in \mathbb{R}^{m\times n}$ e $c \in \mathbb{R}^m$. De acordo com o Teorema 2.34, as condições necessárias de segunda ordem a serem satisfeitas por um ponto x^* que é um minimizador local do problema (49) são

(i) existe $\boldsymbol{\mu} \in \mathbb{R}^{r(x^*)}$ tal que

$$(Qx^* + p) + W_{\mathcal{I}}^T \boldsymbol{\mu} = 0 \quad \text{e} \quad \mu_k \geq 0,$$

para todo $i_k \in \mathcal{I}(x^*)$, em que $\mathcal{I}(x^*)$ é o conjunto dos índices das restrições ativas em x^* já definido anteriormente, e $\nabla f(x^*) = Qx^* + p$;

(ii) para todo $y \in \mathcal{N}(W_{\mathcal{I}})$, isto é, tal que $W_{\mathcal{I}} y = 0$, devemos ter

$$y^T Q y \geq 0,$$

em que $\nabla^2 f(x^*) = Q$.

Por outro lado, para que x^* seja minimizador local do problema (49), pelo Teorema 2.35 é suficiente que x^* satisfaça a restrição de desigualdade $Wx \leq b$ e também

(i) existe $\boldsymbol{\mu} \in \mathbb{R}^{r(x^*)}$ tal que

$$(Qx^* + p) + W_{\mathcal{I}}^T \boldsymbol{\mu} = 0 \quad \text{e} \quad \mu_k \geq 0,$$

para todo $i_k \in \mathcal{I}(x^*)$;

(ii) para todo $y \in \mathcal{N}(W_{\mathcal{J}})$, com $y \neq 0$, temos que

$$y^T Q y > 0,$$

em que $\mathcal{J} = \{k \in \{1, 2, \ldots, r(x^*)\} \mid \mu_k > 0\}$.

Além disso, é interessante observar que se tomarmos $Q = I$ e $p = 0$, o problema (49) pode ser reescrito da seguinte forma

$$\begin{aligned} \min_{x} \quad & \frac{1}{2}\|x\|^2 \\ \text{s.a} \quad & Wx \leq b. \end{aligned}$$

Tal problema consiste em encontrar uma solução do sistema linear $Wx \leq b$ com a menor norma possível, e como as condições de otimalidade para problemas com restrições de desigualdade consideram, essencialmente, as restrições que são ativas em uma solução do problema, sua solução pode ser caracterizada de modo análogo ao realizado no Exemplo 2.21.

2.4 MINIMIZAÇÃO COM RESTRIÇÕES LINEARES DE IGUALDADE E DESIGUALDADE

Nesta seção abordaremos o problema geral de otimização, isto é, com restrições de igualdade e desigualdade. Para tanto, utilizando os argumentos das Seções 2.2 e 2.3, pretendemos obter as condições de otimalidade de Karush-Kuhn-Tucker (KKT) e as condições necessárias e suficientes de segunda ordem

para o problema

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{s.a} \quad & Ax = b,\ Wx \leq c, \end{aligned} \tag{50}$$

em que $A \in \mathbb{R}^{m\times n}$, com $m < n$ e posto$(A) = m$, $W \in \mathbb{R}^{p\times n}$, $b \in \mathbb{R}^m$ e $c \in \mathbb{R}^p$.

Para o problema (50), o conjunto de factibilidade Ω é um poliedro em $\mathbb{R}^n$ definido por

$$\Omega := \{x \in \mathbb{R}^n \mid Ax = b \text{ e } Wx \leq c\}.$$

Por convenção, as restrições de igualdade são consideradas ativas em todo ponto factível, e em decorrência disso, as restrições correspondentes às linhas de A são sempre ativas. Assim, dado $x^* \in \Omega$, o conjunto dos índices das restrições ativas em x^* é dado por

$$\mathcal{I}(x^*) = \{1, 2, \ldots, m, i_1, i_2, \ldots, i_{s(x)}\},$$

em que $\mathcal{J}(x^*) := \{i_1, i_2, \ldots, i_{s(x)}\}$ será o conjunto de índices que correspondem às restrições de desigualdade que são ativas em x^*. Ademais, temos que $s(x^*)$, com $0 \leq s(x^*) \leq p$, é a quantidade de restrições de desigualdade ativas em x^*, enquanto que $r(x^*)$, com $m \leq r(x^*) \leq m + p$, é o número total de restrições ativas em x^*.

Por conseguinte, as caracterizações das direções factíveis estabelecidas pelos Teoremas 2.17 e 2.29 implicam no seguinte teorema.

Teorema 2.37. *Dado um ponto $x^* \in \Omega$, o vetor $d \in \mathbb{R}^n$ é uma direção factível em x^* se, e somente se, $Ad = 0$ e $w_j^T d \leq 0$ para todo $j \in \mathcal{J}(x^*)$.*

2.4.1 Condições necessárias de primeira ordem

Para enunciar os teoremas a seguir vamos considerar os conjuntos $\mathcal{I}(x^*)$ e $\mathcal{J}(x^*)$ como definidos anteriormente. Além disso, denotaremos por $W_{\mathcal{J}}$ a submatriz

de W cujas linhas são as que têm os índices em $\mathcal{J}(x^*)$, por $c_{\mathcal{J}} \in \mathbb{R}^{s(x^*)}$ o vetor formado pelas componentes de c correspondentes a $\mathcal{J}(x^*)$ e por $B \in \mathbb{R}^{r(x^*)\times n}$ a matriz dada por

$$B = \begin{bmatrix} A \\ W_{\mathcal{J}} \end{bmatrix},$$

em que $\text{posto}(B) = r(x^*)$.

Teorema 2.38. ***(Condição necessária de primeira ordem)*** *Considere o problema* (50) *com* $f \in \mathcal{C}^1$ *e* $x^* \in \Omega$ *tal que* $m \leq r(x^*) \leq n$ *e* $s(x^*) \geq 1$. *Se* x^* *é minimizador local de* (50), *e* $\text{posto}(B) = r(x^*)$, *então existem* $\boldsymbol{\lambda} \in \mathbb{R}^m$ *e* $\boldsymbol{\mu} \in \mathbb{R}^{s(x^*)}$ *tais que*

$$\nabla f(x^*) + A^T\boldsymbol{\lambda} + W_{\mathcal{J}}^T\boldsymbol{\mu} = 0 \quad e \quad \mu_k \geq 0,\ 1 \leq k \leq s(x^*). \tag{51}$$

Demonstração. Suponhamos por contradição que (51) não ocorre. Isso pode acontecer por dois motivos:

(i) $\nabla f(x^*) + A^T\boldsymbol{\lambda} + W_{\mathcal{J}}^T\boldsymbol{\mu} \neq 0$ para todo $\boldsymbol{\mu} \in \mathbb{R}^{s(x^*)}$.

Assim, temos que x^* não é minimizador local do problema

$$\begin{aligned} \min_x \quad & f(x) \\ \text{s.a} \quad & Ax = b, W_{\mathcal{J}}x = c_{\mathcal{J}}, \end{aligned} \tag{52}$$

pois x^* não satisfaz a condição necessária de primeira ordem para o problema (52) dada pelo Teorema 2.20. Em virtude disso, x^* não é minimizador local do problema (50), o que é uma contradição.

(ii) $\nabla f(x^*) + A^T\boldsymbol{\lambda} + W_{\mathcal{J}}^T\boldsymbol{\mu} = 0$, mas existe j tal que $\mu_j < 0$.

Considere $\tilde{W}_{\mathcal{J}}$ a matriz obtida retirando a linha w_{i_j} que corresponde ao multiplicador μ_j, e a matriz $\tilde{B}$ dada por $\tilde{B} = \begin{bmatrix} A \\ \tilde{W}_{\mathcal{J}} \end{bmatrix}$. Então, tomando

$d = \text{proj}_{\mathcal{N}(\tilde{B})}(-\nabla f(x^*))$, temos que

$$(-\nabla f(x^*) - d)^T d = 0,$$

pois são ortogonais, e isso implica que

$$\nabla f(x^*)^T d = -d^T d = -\|d\|^2 < 0. \tag{53}$$

Portanto, d é uma direção de descida. Agora, vejamos que d é uma direção factível. De (51) temos que

$$\nabla f(x^*) + \lambda_1 a_1 + \ldots + \lambda_m a_m + \mu_1 w_{i_1} + \ldots + \mu_j w_{i_j} + \ldots + \mu_{s(x^*)} w_{i_{s(x^*)}} = 0,$$

com $\mu_j < 0$. Como tomamos $d \in \mathcal{N}(\tilde{B})$, temos que

$$Ad = 0 \quad \text{e} \quad w_{i_k}^T d = 0, \tag{54}$$

para todo $k \neq j$. Então,

$$\nabla f(x^*)^T d + \mu_j w_{i_j}^T d = 0,$$

e reescrevendo obtemos

$$\nabla f(x^*)^T d = -\mu_j w_{i_j}^T d.$$

Agora, utilizando (53) e o fato de $\mu_j < 0$, concluímos que

$$w_{i_j}^T d < 0. \tag{55}$$

Desse modo, de (54) e (55) resulta que

$$Ad = 0 \quad \text{e} \quad w_{i_k}^T d \leq 0,$$

para todo k tal que $1 \leq k \leq s(x^*)$. Logo, d é uma direção factível. Portanto, d é uma direção factível e de descida a partir de x^*, contradizendo a hipótese de x^* ser minimizador local de (50).

□

As condições em (51) são conhecidas como condições de Karush-Kuhn-Tucker (KKT). A partir delas podemos definir *ponto estacionário*.

Definição 2.39. (Ponto estacionário ou KKT) Um ponto $x^* \in \mathbb{R}^n$ é um *ponto estacionário*, ou ainda, um *ponto KKT* do problema (50) quando $x^* \in \Omega$ e existem $\boldsymbol{\lambda}^* \in \mathbb{R}^m$ e $\boldsymbol{\mu}^* \in \mathbb{R}^p$, tais que as condições (51) são satisfeitas. Os elementos $\boldsymbol{\lambda}^*$ e $\boldsymbol{\mu}^*$ são chamados de *multiplicadores de Lagrange* associados ao ponto estacionário x^*.

Em outras palavras, x^* é dito estacionário quando satisfaz as condições KKT.

2.4.2 Condições necessárias e suficientes de segunda ordem

As condições necessárias e suficientes de segunda ordem para problemas com restrições de igualdade e desigualdade podem ser obtidas considerando-se essencialmente as restrições ativas.

Teorema 2.40. ***(Condição necessária de segunda ordem)*** *Considere o problema* (50) *com* $f \in \mathcal{C}^2$ *e* $r(x^*)$, $s(x^*)$, $\mathcal{J}(x^*)$ *e* B *definidos como anteriormente. Se* x^* *é um minimizador local do problema* (50)*, então*

(i) existem $\boldsymbol{\lambda} \in \mathbb{R}^m$ *e* $\boldsymbol{\mu} \in \mathbb{R}^{s(x^*)}$ *tais que* $\nabla f(x^*) + A^T\boldsymbol{\lambda} + W_{\mathcal{J}}^T\boldsymbol{\mu} = 0$ *e* $\mu_k \geq 0$ *para todo* $k \in \{1, 2, \ldots, s(x^*)\}$*;*

(ii) para todo $y \in \mathcal{N}(B)$ *temos que* $y^T\nabla^2 f(x^*)y \geq 0$ *.*

Demonstração. Primeiramente, se x^* é minimizador local do problema (50) então, pelo Teorema 2.38, existem $\boldsymbol{\lambda} \in \mathbb{R}^m$ e $\boldsymbol{\mu} \in \mathbb{R}^{s(x^*)}$ tais que

$$\nabla f(x^*) + A^T\boldsymbol{\lambda} + W_{\mathcal{J}}^T\boldsymbol{\mu} = 0 \;\text{ e }\; \mu_k \geq 0,\; k \in \{1, 2, \ldots, s(x^*)\}.$$

Ademais, se x^* é minimizador local de (50) então ele é solução do seguinte problema obtido considerando apenas as restrições ativas em x^*

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{s.a} \quad & Ax = b, W_{\mathcal{J}} x = c_{\mathcal{J}}. \end{aligned}$$

Assim, como $B = \begin{bmatrix} A \\ W_{\mathcal{J}} \end{bmatrix}$, concluímos pelo Teorema 2.23 que

$$y^T \nabla^2 f(x^*) y \geq 0,$$

para todo $y \in \mathcal{N}(B)$. □

Agora, para saber se um determinado ponto é um mínimo local do problema (50), é necessário que ele cumpra certas condições suficientes que são estabelecidas pelo teorema a seguir.

Teorema 2.41. ***(Condição suficiente de segunda ordem)*** *Considere o problema* (50) *com* $f \in \mathcal{C}^2$ *e* $r(x^*)$, $s(x^*)$, $\mathcal{J}(x^*)$ *e* B *definidos como anteriormente. Se* $x^* \in \Omega$ *verifica*

(i) existem $\boldsymbol{\lambda} \in \mathbb{R}^m$ *e* $\boldsymbol{\mu} \in \mathbb{R}^{s(x^*)}$ *tais que* $\nabla f(x^*) + A^T \boldsymbol{\lambda} + W_{\mathcal{J}}^T \boldsymbol{\mu} = 0$ *e* $\mu_k \geq 0$ *para todo* $k \in \{1, 2, \ldots, s(x^*)\}$*;*

(ii) se $y^T \nabla^2 f(x^*) y > 0$ *para todo* $y \in \mathcal{N}(\tilde{B})$*, em que*

$$\tilde{B} = \begin{bmatrix} A \\ W_{\mathcal{K}} \end{bmatrix}$$

e

$$\mathcal{K} = \{j \in \mathcal{J} \mid \mu_j > 0\};$$

então x^* *é um minimizador local de* (50).

Demonstração. Suponhamos, inicialmente, que $\mu_j > 0$ para todo $j \in \mathcal{J}(x^*)$, então $W_{\mathcal{J}} = W_{\mathcal{K}}$ e, consequentemente,

$$B = \tilde{B}.$$

Desse modo, se $x^* \in \Omega$ satisfaz

$$\nabla f(x^*) + A^T \boldsymbol{\lambda} + W_{\mathcal{K}}^T \boldsymbol{\mu} = 0$$

e

$$y^T \nabla^2 f(x^*) y > 0,$$

para todo $y \in \mathcal{N}(\tilde{B})$ tal que $y \neq 0$, então, pelo Teorema 2.24, x^* é minimizador local do problema

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{s.a} \quad & Ax = b,\ W_{\mathcal{K}} x = c_{\mathcal{K}}, \end{aligned}$$

e consequentemente, é minimizador local do problema (50).

Agora, suponhamos que existe pelo menos um j tal que $\mu_j = 0$. Dessa forma,

$$W_{\mathcal{J}}^T \boldsymbol{\mu} = W_{\mathcal{K}}^T \boldsymbol{\mu}_{\mathcal{K}},$$

em que $\boldsymbol{\mu}_{\mathcal{K}}$ é o vetor formado pelas componentes μ_j que possuem índices em $\mathcal{K}$. Então, se $x^* \in \Omega$ satisfaz

$$\nabla f(x^*) + A^T \boldsymbol{\lambda} + W_{\mathcal{K}}^T \boldsymbol{\mu} = 0$$

e

$$y^T \nabla^2 f(x^*) y > 0,$$

para todo $y \in \mathcal{N}(\tilde{B})$ e $y \neq 0$, então, pelo Teorema 2.24, temos que x^* é minimizador local do problema

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{s.a} \quad & Ax = b,\ W_{\mathcal{K}} x = c_{\mathcal{K}}. \end{aligned}$$

Assim, concluímos que x^* é minimizador local do problema (50). □

Portanto, com este teorema finalizamos o Capítulo 2, no qual discutimos elementos da teoria de otimização irrestrita e com restrições lineares, abordando as condições de otimalidade para cada caso. Tendo em vista que o problema decorrente da modelagem da técnica SVM, que será formulado no Capítulo 4, consiste num problema de programação quadrática convexa com restrições lineares, no próximo capítulo daremos continuidade aos estudos da teoria de otimização, abordando alguns conceitos de otimização convexa, que fornece resultados importantes aos problemas de otimização.

3 OTIMIZAÇÃO CONVEXA

Neste capítulo o objetivo principal é apresentar alguns resultados e definições relacionados aos problemas de otimização convexa, isto é, quando a função objetivo e o conjunto factível são convexos. A noção de convexidade é muito importante na teoria de otimização, pois ela garante que pontos estacionários são minimizadores e permite concluir que minimizadores locais são globais. Além disso, as condições necessárias de otimalidade tornam-se suficientes sob a hipótese de convexidade. A partir disso, os problemas de classificação podem ser formulados em termos de otimização convexa. Os conceitos e resultados abordados neste capítulo foram desenvolvidos, principalmente, com base nas seguintes referências: Izmailov e Solodov [14], Krulikovski [15], Luenberger e Ye [19] e Ribeiro e Karas [22].

3.1 CONJUNTOS CONVEXOS

Geometricamente, um conjunto é convexo se, dados dois pontos no conjunto, cada ponto no segmento de linha que une esses dois pontos também pertence conjunto.

Definição 3.1. Um *conjunto* $C \subset \mathbb{R}^n$ é dito *convexo* quando dados $x, y \in C$, o segmento $[x, y] \coloneqq \{(1-t)x + ty \mid t \in [0,1]\}$ estiver inteiramente contido em C.

A Figura 7 ilustra a noção de convexidade de conjuntos. Alguns exemplos de conjuntos convexos são o conjunto vazio, o espaço $\mathbb{R}^n$, e um conjunto que contém um ponto só. Além desses conjuntos, as duas proposições a seguir garantem que qualquer hiperplano e qualquer semiespaço em $\mathbb{R}^n$ também são conjuntos convexos.

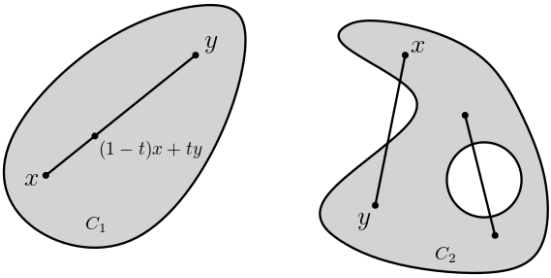


Figura 7 – O conjunto C_1 é convexo, o conjunto C_2 não é convexo.

Proposição 3.2. *Qualquer hiperplano em $\mathbb{R}^n$, isto é, um conjunto da forma $\{x \in \mathbb{R}^n \mid w^T x = b\}$ em que $w \in \mathbb{R}^n$ e $b \in \mathbb{R}$, é um conjunto convexo.*

Demonstração. Considere o hiperplano $\mathcal{H} = \{x \in \mathbb{R}^n \mid w^T x = b\}$, com $w \in \mathbb{R}^n$ e $b \in \mathbb{R}$, e sejam $x, y \in \mathcal{H}$ arbitrários. Além disso, tome a combinação convexa de x e y

$$(1-t)x + ty,$$

com $t \in [0,1]$. Assim, como $w^T x = b$ e $w^T y = b$, temos que

$$\begin{aligned} w^T[(1-t)x + ty] &= w^T[(1-t)x] + w^T(ty) \\ &= (1-t)w^T x + tw^T y \\ &= (1-t)b + tb \\ &= b. \end{aligned}$$

Portanto, $(1-t)x + ty \in \mathcal{H}$ para todo $t \in [0,1]$, e com isso concluímos que $\mathcal{H}$ é um conjunto convexo. □

Proposição 3.3. *Qualquer semiespaço em $\mathbb{R}^n$, isto é, um conjunto da forma $\{x \in \mathbb{R}^n \mid w^T x \leq b\}$ em que $w \in \mathbb{R}^n$ e $b \in \mathbb{R}$, é um conjunto convexo.*

Demonstração. Considere o semiespaço $\mathcal{S} = \{x \in \mathbb{R}^n \mid w^T x \leq b\}$, com $w \in \mathbb{R}^n$ e $b \in \mathbb{R}$, e tome $x, y \in \mathcal{S}$ arbitrários. Vamos provar que a combinação convexa de x e y, $(1-t)x + ty$, pertence ao semiespaço $\mathcal{S}$ para todo $t \in [0,1]$. De fato, como $w^T x \leq b$ e $w^T y \leq b$, temos que

$$\begin{aligned} w^T[(1-t)x + ty] &= w^T[(1-t)x] + w^T(ty) \\ &= (1-t)w^T x + tw^T y \\ &\leq (1-t)b + tb \\ &= b. \end{aligned}$$

Portanto, $\mathcal{S}$ é um conjunto convexo. □

Outro resultado interessante acerca da convexidade de conjuntos é dado a seguir, o qual garante que a interseção de conjuntos convexos preserva a convexidade.

Proposição 3.4. *Sejam $C_i \subset \mathbb{R}^n$, $i \in I$, conjuntos convexos, em que I é um conjunto qualquer (possivelmente infinito). Então a interseção $C = \bigcap_{i \in I} C_i$ também é um conjunto convexo.*

Demonstração. Sejam $x, y \in C$ arbitrários. Pela definição de interseção de conjuntos, $x, y \in C_i$, para todo $i \in I$. Como cada conjunto C_i, $i \in I$, é convexo por hipótese, temos que

$$(1-t)x + ty \in C_i,$$

para qualquer $t \in [0,1]$ e todo $i \in I$. Dessa forma, pela definição de interseção, segue que $(1-t)x + ty \in C$, para qualquer $t \in [0,1]$. Portanto, C é convexo. □

Utilizando as Proposições 3.2 a 3.4, podemos verificar que um conjunto poliedral em $\mathbb{R}^n$ também é convexo.

Corolário 3.5. *Um conjunto poliedral em $\mathbb{R}^n$ é convexo.*

Demonstração. Pela Definição 2.27 temos que um conjunto poliedral é o conjunto de soluções de um sistema finito de equações e inequações lineares. Em outras palavras, é uma interseção de semiespaços e hiperplanos que, pelas Proposições 3.2 e 3.3, são conjuntos convexos. Portanto, pela Proposição 3.4 concluímos que um conjunto poliedral é convexo. □

No próximo lema veremos que se dois vetores têm a mesma norma euclidiana, então qualquer combinação linear que não seja a trivial ($t = 0$ ou $t = 1$) terá norma estritamente menor.

Lema 3.6. *Considere $\|\cdot\|$ a norma euclidiana em $\mathbb{R}^n$. Sejam $u, v \in \mathbb{R}^n$ com $u \neq v$. Se $\|u\| = \|v\| = r$, então $\|(1-t)u + tv\| < r$, para todo $t \in (0,1)$.*

Demonstração. Seja $t \in (0,1)$ e suponha que $\|u\| = \|v\| = r$. Aplicando a desigualdade triangular, temos

$$\|(1-t)u + tv\| \leq (1-t)\|u\| + t\|v\| = (1-t)r + tr = r.$$

Agora, suponha por absurdo que $\|(1-t)u + tv\| = r$. Então

$$(1-t)^2 u^T u + 2t(1-t)u^T v + t^2 v^T v = \|(1-t)u + tv\|^2 = r^2. \tag{56}$$

Como $u^T u = v^T v = r^2$ e $t \in (0,1)$, substituindo em (56) e desenvolvendo, obtemos

$$\begin{aligned} r^2 &= (1 - 2t + t^2)u^T u + (2t - 2t^2)u^T v + t^2 v^T v \\ &= (1 - 2t + t^2)r^2 + (2t - 2t^2)u^T v + t^2 r^2 \\ &= r^2 - 2tr^2 + t^2 r^2 + (2t - 2t^2)u^T v + t^2 r^2. \end{aligned}$$

Evidenciando r^2, temos

$$(2t - 2t^2)r^2 = (2t - 2t^2)u^T v,$$

e, portanto,

$$r^2 = u^T v. \tag{57}$$

Assim, por (57),

$$\|u - v\|^2 = u^T u - 2u^T v + v^T v = r^2 - 2r^2 + r^2 = 0,$$

o que é uma contradição, pois por hipótese $u \neq v$. Portanto, concluímos que $\|(1-t)u + tv\| < r$, para todo $t \in (0,1)$. □

Agora, dado um conjunto $S \subset \mathbb{R}^n$ e um ponto $z \in \mathbb{R}^n$, considere o problema de encontrar um ponto de S mais próximo de z, em outras palavras, queremos minimizar a distância de um ponto a um conjunto. Assim, os próximos resultados garantem a existência da solução no caso de S ser um conjunto fechado e sua unicidade se, além de fechado, S for convexo. Tal solução é chamada de projeção de z sobre S, e denotada por $\text{proj}_S(z)$.

Lema 3.7. *Seja $S \subset \mathbb{R}^n$ um conjunto fechado não vazio. Dado $z \in \mathbb{R}^n$, existe $\bar{z} \in S$ tal que*

$$\|z - \bar{z}\| \leq \|z - x\|,$$

para todo $x \in S$.

Demonstração. Seja $\alpha = \inf\{\|z - x\| \mid x \in S\}$. Então, para cada $n \in \mathbb{N}$, existe $x^n \in S$ tal que

$$\alpha \leq \|z - x^n\| \leq \alpha + \frac{1}{n}. \tag{58}$$

Em particular, $\|z - x^n\| \leq \alpha + 1$, para todo $n \in \mathbb{N}$. Logo, pelo Teorema de Bolzano-Weierstrass, existe uma subsequência (x^{n^k}) convergente, com $k \in \mathbb{N}'$, tal que $x^{n^k} \longrightarrow \bar{z}$. Como S é fechado temos que $\bar{z} \in S$. Além disso,

$$\|z - x^n\| \longrightarrow \|z - \bar{z}\|.$$

Mas, por (58), temos que $\|z - x^n\| \longrightarrow \alpha$, e portanto, concluímos que $\|z - \bar{z}\| = \alpha$.

□

Lema 3.8. *Seja $S \subset \mathbb{R}^n$ um conjunto não vazio, convexo e fechado. Dado $z \in \mathbb{R}^n$, existe um único $\bar{z} \in S$ tal que*

$$\|z - \bar{z}\| \leq \|z - x\|$$

para todo $x \in S$.

Demonstração. A existência é garantida pelo Lema 3.7. Para provar a unicidade suponha que existam $\bar{z}, \tilde{z} \in S$, com $\bar{z} \neq \tilde{z}$, tais que

$$\|z - \bar{z}\| \leq \|z - x\| \quad \text{e} \quad \|z - \tilde{z}\| \leq \|z - x\|, \tag{59}$$

para todo $x \in S$. Tomando $x = \tilde{z}$ na primeira desigualdade e $x = \bar{z}$ na segunda, obtemos

$$\|z - \bar{z}\| = \|z - \tilde{z}\|.$$

Por outro lado, o ponto $z^* = \dfrac{\bar{z} - \tilde{z}}{2}$ pertence ao conjunto convexo S. Além disso, pelo Lema 3.6, com $r = \|z - \bar{z}\| = \|z - \tilde{z}\|$ e $t = 1/2$, temos

$$\begin{aligned} \|z - z^*\| &= \|z - t(\bar{z} + \tilde{z})\| \\ &= \|z - t\bar{z} - t\tilde{z}\| \\ &= \|(1-t)(z - \bar{z}) + t(z - \tilde{z})\| \\ &< r, \end{aligned}$$

o que é uma contradição, pois por (59) teríamos

$$r = \|z - \bar{z}\| = \|z - \tilde{z}\| \leq \|z - z^*\| < r.$$

Portanto, $\bar{z} = \tilde{z}$. □

No Lema 3.8 denotamos $\bar{z} = \text{proj}_S(z)$.

Teorema 3.9. *Sejam $S \subset \mathbb{R}^n$ um conjunto não vazio, convexo e fechado, $z \in \mathbb{R}^n$ e $\bar{z} = \text{proj}_S(z)$. Então,*

$$(z - \bar{z})^T (x - \bar{z}) \leq 0,$$

para todo $x \in S$.

Demonstração. Sejam $x \in S$ um ponto arbitrário e $\bar{z} = \text{proj}_S(z)$. Pelo Lema 3.7 $\bar{z} \in S$ e, dado $t \in (0, 1)$, pela convexidade de S, temos que $(1 - t)\bar{z} + tx \in S$. Assim,

$$\|z - \bar{z}\| \leq \|z - (1 - t)\bar{z} - tx\| = \|(z - \bar{z}) - t(x - \bar{z})\|.$$

Então,

$$\|z - \bar{z}\|^2 \leq \|(z - \bar{z}) - t(x - \bar{z})\|^2 = \|z - \bar{z}\|^2 - 2t(z - \bar{z})^T (x - \bar{z}) + t^2 \|x - \bar{z}\|^2,$$

e como $t > 0$, temos

$$2(z - \bar{z})^T (x - \bar{z}) \leq t\|x - \bar{z}\|^2. \tag{60}$$

Passando o limite em (60) quando $t \to 0$, obtemos

$$(z - \bar{z})^T (x - \bar{z}) \leq 0,$$

completando a demonstração. □

O Teorema 3.9 estabelece uma condição necessária e suficiente para caracterizar a projeção. Este resultado é provado no Lema 3.10.

Lema 3.10. *Sejam $S \subset \mathbb{R}^n$ um conjunto não vazio, convexo e fechado e $z \in \mathbb{R}^n$. Se $\bar{z} \in S$ satisfaz*

$$(z - \bar{z})^T (x - \bar{z}) \leq 0,$$

para todo $x \in S$, então $\bar{z} = \text{proj}_S(z)$.

Demonstração. Dado $x \in S$ arbitrário, temos

$$\begin{aligned}
\|z - \bar{z}\|^2 - \|z - x\|^2 &= z^T z - 2z^T \bar{z} + \bar{z}^T \bar{z} - z^T z + 2z^T x - x^T x \\
&= -2z^T \bar{z} + \bar{z}^T \bar{z} + 2z^T x - x^T x \\
&= (x - \bar{z})^T (2z - x - \bar{z}) \\
&= (x - \bar{z})^T (2(z - \bar{z}) - (x - \bar{z})) \\
&= 2(x - \bar{z})^T (z - \bar{z}) - (x - \bar{z})^T (x - \bar{z}) \\
&\leq 0.
\end{aligned}$$

pois $(x - \bar{z})^T (z - \bar{z}) \leq 0$ por hipótese, e $(x - \bar{z})^T (x - \bar{z}) = \|(x - \bar{z})\| \geq 0$. Logo,

$$\|z - \bar{z}\|^2 - \|z - x\|^2 \leq 0,$$

e, então

$$\|z - \bar{z}\|^2 \leq \|z - x\|^2,$$

para todo $x \in S$. Portanto, $\bar{z} = \text{proj}_S(z)$. □

3.2 FUNÇÕES CONVEXAS

Definição 3.11. Seja $C \subset \mathbb{R}^n$ um conjunto convexo. Dizemos que a *função* $f : \mathbb{R}^n \to \mathbb{R}$ é *convexa* em C quando

$$f((1 - t)x + ty) \leq (1 - t)f(x) + tf(y),$$

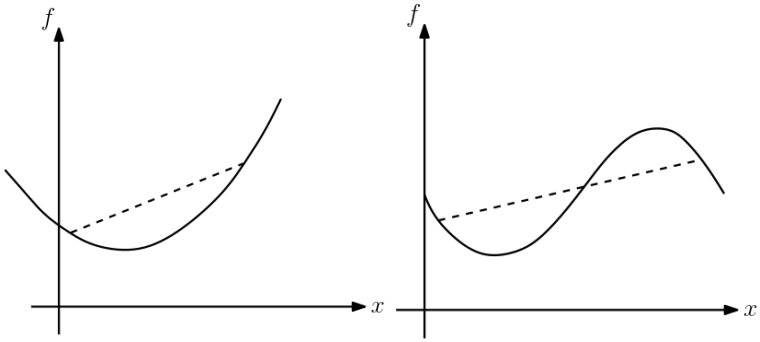
para todos $x, y \in C$ e $t \in [0,1]$.

Se para todo $t \in (0,1)$ e $x \neq y$ vale que

$$f((1-t)x + ty) < (1-t)f(x) + tf(y),$$

dizemos que f é *estritamente convexa*.

Na Figura 8a temos um exemplo de função convexa, enquanto que na Figura 8b podemos visualizar uma função não-convexa.



(a) Função Convexa (b) Função não-convexa.

Figura 8 – Funções convexas e não-convexas.

Geometricamente, uma função é convexa se qualquer arco do seu gráfico está sempre abaixo do segmento que liga as extremidades. Tal noção é representada na Figura 9.

A partir da definição de função convexa podemos definir o que é uma *função côncava*.

Definição 3.12. Seja $C \subset \mathbb{R}^n$ um conjunto convexo. Uma função $f : C \to \mathbb{R}$ é uma *função côncava* em C se a função $-f$ é convexa em C. A função f é *estritamente côncava* se $-f$ é estritamente convexa.

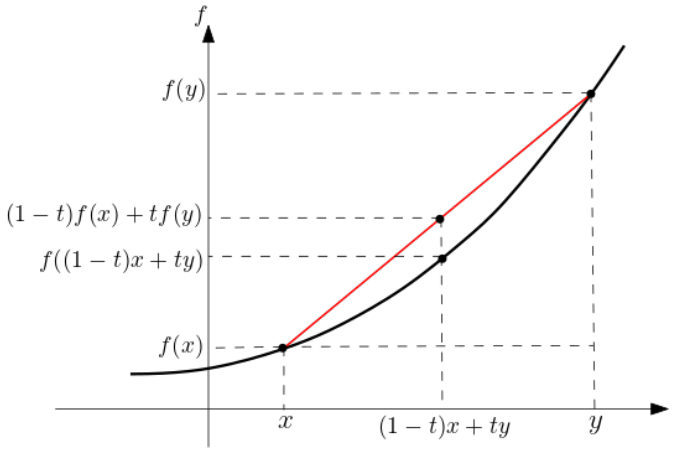


Figura 9 – Noção geométrica de uma função convexa.

Dessa forma, maximizar uma função côncava num conjunto convexo equivale a minimizar uma função convexa num conjunto convexo.

Por conseguinte, veremos na proposição a seguir que a soma de duas funções convexas é uma função convexa e ao multiplicar uma função convexa por um escalar, a nova função obtida também será convexa.

Proposição 3.13. *Sejam f e g funções convexas no conjunto convexo C, $\alpha \geq 0$ e c um número real qualquer. Então*

(i) A função $f + g$ é convexa em C;

(ii) A função αf é convexa para qualquer $\alpha \geq 0$.

Demonstração. Sejam $x_1, x_2 \in C$ e $t \in (0, 1)$. Então

(i)

$$\begin{aligned}(f+g)((1-t)x_1+tx_2) &= f((1-t)x_1+tx_2)+g((1-t)x_1+tx_2)\\ &\leq (1-t)f(x_1)+tf(x_2)+(1-t)g(x_1)+tg(x_2)\\ &= (1-t)[f(x_1)+g(x_1)]+t[f(x_2)+g(x_2)]\\ &= (1-t)(f+g)(x_1)+t(f+g)(x_2).\end{aligned}$$

Portanto, $f+g$ é convexa em C.

(ii)

$$\begin{aligned}\alpha f((1-t)x_1+tx_2) &\leq \alpha[(1-t)f(x_1)+tf(x_2)]\\ &= (1-t)(\alpha f)(x_1)+t(\alpha f)(x_2).\end{aligned}$$

Portanto, αf é convexa em C.

□

A proposição anterior também nos garante que a combinação de funções convexas é convexa, pois observe que, se aplicarmos os itens (i) e (ii) da Proposição 3.13 de maneira indutiva, a combinação positiva de funções convexas $\alpha_1 f_1 + \alpha_2 f_2 + \ldots + \alpha_m f_m$ também é convexa.

Agora, vamos analisar conjuntos definidos por restrições de desigualdade. Neste caso, quando as funções da restrição de um problema de otimização são convexas, o conjunto de restrições resultante é convexo.

Proposição 3.14. *Seja g uma função convexa em C. O conjunto $S_c = \{x \mid x \in C, g(x) \leq c\}$ é convexo para todo número real c.*

Demonstração. Sejam $x_1, x_2 \in S_c$. Então, como $g(x_1) \leq c$ e $g(x_2) \leq c$, temos que

$$g((1-t)x_1+tx_2) \leq (1-t)g(x_1)+tg(x_2) \leq (1-t)c+tc = c,$$

para todo $t \in (0,1)$. Portanto, $(1-t)x_1 + tx_2 \in S_c$. □

Portanto, a partir das Proposições 3.4 e 3.14, concluímos que o conjunto de pontos que satisfaz simultaneamente

$$g_1(x) \leq c_1,\ g_2(x) \leq c_2,\ \ldots,\ g_m(x) \leq c_m,$$

em que cada g_i é uma função convexa, define um conjunto convexo. Esta consideração é importante, pois o conjunto de restrições dos problemas de otimização que estamos interessados em resolver são definidos dessa forma.

Ademais, quando uma função é diferenciável, a convexidade pode ser caracterizada de diferentes formas. Tais caracterizações são úteis para determinar se uma função é convexa, de modo a não depender apenas da definição usual de função convexa. O teorema a seguir apresenta outra forma de caracterizar a convexidade de uma função quando temos hipótese de diferenciabilidade.

Teorema 3.15. *Sejam $f : \mathbb{R}^n \to \mathbb{R}$ uma função diferenciável e $C \in \mathbb{R}^n$ um conjunto convexo. A função f é convexa em C se, e somente se,*

$$f(y) \geq f(x) + \nabla f(x)^T (y - x),$$

para todos $x, y \in C$.

Demonstração. Primeiramente, suponha que f é uma função convexa em C. Para quaisquer $x, y \in C$ e $t \in (0,1]$ temos que $(1-t)x + ty = x + t(y-x)$. Então, tomando $d = y - x$, temos que $x + td \in C$ e, pela convexidade de f,

$$f(x + td) = f((1-t)x + ty) \leq (1-t)f(x) + tf(y). \tag{61}$$

Reescrevendo (61), obtemos

$$f(x + td) \leq f(x) + t(f(y) - f(x)),$$

que, para $t \in (0, 1]$, implica em

$$f(y) - f(x) \geq \frac{f(x + td) - f(x)}{t}.$$

Passando o limite quando $t \to 0^+$, temos

$$f(y) - f(x) \geq \lim_{t \to 0^+} \frac{f(x + td) - f(x)}{t} = \nabla f(x)^T d = \nabla f(x)^T (y - x),$$

e, portanto,

$$f(y) \geq f(x) + \nabla f(x)^T (y - x).$$

Reciprocamente, fixe $x, y \in C$ e $t \in (0, 1)$. Tomando $z = (1 - t)x + ty$, observe que

$$f(x) \geq f(z) + \nabla f(z)^T (x - z) \quad \text{e} \quad f(y) \geq f(z) + \nabla f(z)^T (y - z).$$

Agora, multiplicando a primeira desigualdade por $(1 - t)$, a segunda por t e somando, obtemos

$$\begin{aligned}
(1 - t)f(x) + tf(y) &\geq (1 - t)(f(z) + \nabla f(z)^T (x - z)) + t(f(z) + \nabla f(z)^T (y - z)) \\
&\geq f(z) + \nabla f(z)^T (x - z) + t\nabla f(z)^T (-x + z + y - z) \\
&\geq f(z) + \nabla f(z)^T (x - z) + t\nabla f(z)^T (y - x) \\
&\geq f(z) + \nabla f(z)^T (x - z) + t\nabla f(z)^T \frac{(z - x)}{t} \\
&\geq f(z) + \nabla f(z)^T (x - z) - \nabla f(z)^T (x - z) \\
&= f(z) \\
&= f((1 - t)x + ty).
\end{aligned}$$

Portanto, a função f é convexa em C. □

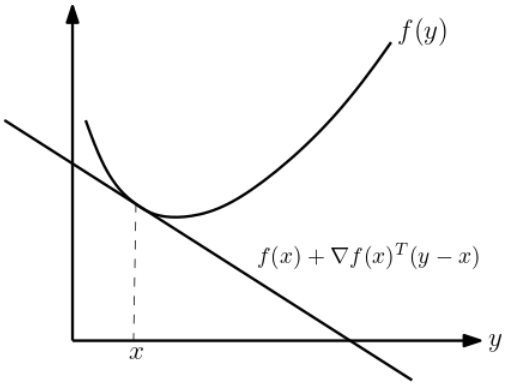


Figura 10 – Interpretação geométrica do Teorema 3.15.

Fonte – Luenberger e Ye [19]

Uma interpretação geométrica do Teorema 3.15 é que a aproximação linear através da derivada local de uma função convexa diferenciável está sempre abaixo do gráfico da função. Tal resultado é ilustrado na Figura 10.

Para funções convexas duas vezes diferenciáveis, existe outra forma de caracterizá-las que é apresentada no Teorema 3.18. Para demonstrar esse resultado vamos utilizar o Fato 3.16, que fornece a fórmula de Taylor com resto de Lagrange. Sua demonstração pode ser encontrada em Lima [17, p.196].

Fato 3.16. *(**Taylor com Resto de Lagrange**) Considere $f : \mathbb{R}^n \to \mathbb{R}$ uma função de classe $\mathcal{C}^1$ e $\bar{x}, d \in \mathbb{R}^n$. Se f é duas vezes diferenciável no segmento $(\bar{x}, \bar{x} + d)$, então existe $t \in (0, 1)$ tal que*

$$f(\bar{x} + d) = f(\bar{x}) + \nabla f(\bar{x})^T d + \frac{1}{2} d^T \nabla^2 f(\bar{x} + td) d.$$

Além do Fato 3.16, também será necessário o seguinte lema para demonstrar o Teorema 3.18.

Lema 3.17. *Sejam $C \subset \mathbb{R}^n$ convexo, $x \in \bar{C}$ e $y \in \operatorname{int} C$. Então, $(x, y] \subset \operatorname{int} C$.*

Demonstração. Suponhamos que $(x, y] \nsubseteq \operatorname{int} C$, ou seja, existe $t \in (0, 1)$ tal que $(1-t)x + ty \notin \operatorname{int} C$. Isso significa que para todo $\varepsilon > 0$, $B((1-t)x + ty, \varepsilon) \nsubseteq C$, pois para cada ε existe pelo menos um ponto $z \in B((1-t)x + ty, \varepsilon)$, tal que $z \notin C$. Mas neste caso existiria pelo menos um ponto $w \in C$, de modo que $z \in [w, y]$. Logo, o segmento $[w, y] \nsubseteq C$, contradizendo a hipótese do conjunto C ser convexo. □

Teorema 3.18. *Sejam $f : \mathbb{R}^n \to \mathbb{R}$ uma função de classe $\mathcal{C}^2$ e $C \subset \mathbb{R}^n$ um conjunto convexo com interior não vazio. A função f é convexa em C se, e somente se, a Hessiana $\nabla^2 f(x)$ é semidefinida positiva para todo $x \in C$.*

Demonstração. Primeiro, suponha que f é convexa e considere $x \in \operatorname{int} C$. Então, dado $d \in \mathbb{R}^n$ temos que $x + td \in C$, para t suficientemente pequeno. Assim, pelo Teorema 3.15, resulta que

$$f(x + td) \geq f(x) + t\nabla f(x)^T d,$$

e reescrevendo, obtemos

$$f(x + td) - f(x) - t\nabla f(x)^T d \geq 0. \tag{62}$$

Aplicando o Fato 2.7 para $f(x + td)$, temos que

$$f(x + td) = f(x) + t\nabla f(x)^T d + \frac{t^2}{2} d^T \nabla^2 f(x) d + o(t^2),$$

e substituindo em (62), obtemos

$$\frac{t^2}{2} d^T \nabla^2 f(x) d + o(t^2) \geq 0,$$

com $\lim_{t\to 0} \frac{o(t^2)}{t^2} = 0$. Dividindo ambos os lados da desigualdade por t^2 e passando o limite com $t \to 0$, temos que

$$d^T \nabla^2 f(x) d \geq 0.$$

Agora, considere $x \in C$ arbitrário. Como existe $y \in \operatorname{int} C$, o Lema 3.17 garante que todos os pontos do segmento $(x, y]$ pertencem ao $\operatorname{int} C$. Então, pelo que acabamos de provar, dados $d \in \mathbb{R}^n$ e $t \in (0, 1]$, vale

$$d^T \nabla^2 f((1-t)x + ty) d \geq 0.$$

Fazendo $t \to 0^+$ e usando a continuidade de $\nabla^2 f$, obtemos

$$d^T \nabla^2 f(x) d \geq 0,$$

para todo $x \in C$. Portanto, a Hessiana $\nabla^2 f(x)$ é semidefinida positiva para todo $x \in C$.

Reciprocamente, dados $x \in C$ e $d \in \mathbb{R}^n$ tal que $x + d \in C$, pelo Fato 3.16,

$$f(x+d) = f(x) + \nabla f(x)^T d + \frac{1}{2} d^T \nabla^2 f(x + td) d,$$

para algum $t \in (0, 1)$. Como $\nabla^2 f(x + td) \succcurlyeq 0$, concluímos que

$$f(x+d) \geq f(x) + \nabla f(x)^T d.$$

Logo, pelo Teorema 3.15, a função f é convexa. □

O próximo teorema estabelece que no caso em que a função objetivo é convexa, então todo minimizador local é global. Esta afirmação é muito importante e justifica o fato da convexidade ser uma propriedade tão importante para os problemas de otimização.

Teorema 3.19. *Sejam $C \subset \mathbb{R}^n$ um conjunto convexo e $f : C \to \mathbb{R}$ uma função convexa. Se $x^* \in C$ é minimizador local de f, então x^* é minimizador global de f.*

Demonstração. Seja x^* um minimizador local de f. Então, existe $\delta > 0$ tal que

$$f(x^*) \leq f(x),$$

para todo $x \in B(x^*, \delta) \cap C$. Considere $y \in C$, tal que $y \notin B(x^*, \delta)$, e tome $t \in (0, 1]$ de modo que $t\|y - x^*\| < \delta$. Assim, o ponto $x = (1-t)x^* + ty$ satisfaz

$$\|x - x^*\| = \|(1-t)x^* + ty - x^*\| = \|x^* - tx^* + ty - x^*\| = t\|y - x^*\| < \delta,$$

e, portanto, $x \in B(x^*, \delta) \cap C$. Desse modo, como f é uma função convexa, temos

$$\begin{aligned} f(x^*) &\leq f(x) \\ &= f((1-t)x^* + ty) \\ &\leq (1-t)f(x^*) + tf(y) \\ &= f(x^*) + t(f(y) - f(x^*)), \end{aligned}$$

donde segue que $f(x^*) \leq f(y)$. Portanto, x^* é minimizador global de f. □

O resultado a seguir também é muito importante, pois garante que pontos estacionários de problemas convexos são minimizadores globais. Para tanto, considere o problema geral de otimização dado em (50), em que as restrições são definidas pelas funções $g(x) = Wx - c \leq 0$ e $h(x) = Ax - b = 0$.

Teorema 3.20. *Seja $x^* \in \mathbb{R}^n$ um ponto estacionário do problema* (50)*, no sentido da Definição 2.39. Sejam f e g funções convexas e seja h uma função afim. Então, x^* é minimizador global do problema* (50)*.*

Demonstração. Sejam $\boldsymbol{\lambda}^* \in \mathbb{R}^m$ e $\boldsymbol{\mu}^* \in \mathbb{R}^p_+$ um par de multiplicadores associados ao ponto estacionário x^*. Além disso, para todo $x \in \Omega$, temos que $h(x) = 0$ e $g(x) \leq 0$. Pela convexidade de g_i e usando o Teorema 3.15, obtemos

$$\nabla g_i(x^*)^T(x - x^*) = g_i(x^*) + \nabla g_i(x^*)^T(x - x^*) \leq g_i(x) \leq 0,$$

para todo $i \in \mathcal{I}(x^*)$ e $x \in \Omega$. Ademais, como h é afim podemos escrever $h_i(x) = a_i^T x - b$, e disso segue que

$$\nabla h_i(x^*)^T(x - x^*) = a_i^T(x - x^*) = (a_i^T x - b) - (a_i^T x^* - b) = 0,$$

para todo $i = 1, \ldots, m$ e $x \in \Omega$. Como x^* é um ponto estacionário,

$$\nabla f(x^*) = -\sum_{i \in \mathcal{I}(x^*)} \mu_i^* \nabla g_i(x^*) - \sum_{i=1}^{m} \lambda_i^* \nabla h_i(x^*),$$

e fazendo o produto interno por $(x - x^*)$ em ambos os lados da igualdade, obtemos

$$\nabla f(x^*)^T(x-x^*) = -\sum_{i \in \mathcal{I}(x^*)} \mu_i^* \nabla g_i(x^*)^T(x-x^*) - \sum_{i=1}^{m} \lambda_i^* \nabla h_i(x^*)^T(x-x^*) \geq 0.$$

Desse modo, pela convexidade de f e usando o Teorema 3.15, temos que

$$f(x^*) \leq f(x^*) + \nabla f(x^*)^T(x - x^*) \leq f(x),$$

e portanto,

$$f(x^*) \leq f(x),$$

para todo $x \in \Omega$. □

O Teorema 3.20 se constitui num resultado de grande relevância, pois no caso em que o problema de otimização é convexo, isto é, a função objetivo

é convexa e o conjunto factível determinado pelas restrições é convexo, temos que as condições de KKT são necessárias e suficientes para x^* ser minimizador global.

Por fim, tendo em vista que o problema decorrente da formulação matemática da técnica SVM é um problema de minimização cuja função objetivo é quadrática, é necessário então avaliar sob quais condições uma função quadrática é convexa. Tal resultado é apresentado a seguir e sua demonstração decorre diretamente do Teorema 3.18.

Teorema 3.21. *Seja $C \in \mathbb{R}^n$ um conjunto convexo e $Q \in \mathbb{R}^{n\times n}$ uma matriz quadrada e simétrica. Seja $f : C \to \mathbb{R}$ tal que $f(x) = \frac{1}{2}x^TQx - b^Tx$ é uma função quadrática. Então, f é convexa se, e somente se, Q é semidefinida positiva.*

Demonstração. Primeiramente, suponhamos que f é convexa em C. Então, pelo Teorema 3.18, a Hessiana $\nabla^2 f(x)$ é semidefinida positiva para todo $x \in C$, e como $\nabla^2 f(x) = Q$, concluímos que Q é semidefinida positiva. Reciprocamente, se $\nabla^2 f(x) = Q$ é semidefinida positiva, então, pelo Teorema 3.18, a função f é convexa em C. □

Portanto, apresentamos neste capítulo alguns conceitos e resultados de otimização convexa que foram considerados importantes para a formulação matemática da técnica Máquinas de Vetores Suporte que será tratada no próximo capítulo.

4 MÁQUINAS DE VETORES SUPORTE

A Aprendizagem de Máquina desempenha um papel importante dentro do campo de reconhecimento de padrões, o qual, de acordo com Bishop [4, p. 1], "(...) se preocupa com a descoberta automática de regularidades em dados através do uso de algoritmos de computador e com o uso destas regularidades para executar ações, como classificar dados em diferentes categorias" (tradução nossa). Assim, em problemas que exigem a análise de uma grande quantidade de dados para classificá-los, um processo manual torna-se inviável, motivando o desenvolvimento de técnicas computacionais capazes de reconhecer padrões para desempenhar tal tarefa. Neste capítulo desenvolvemos a modelagem matemática de uma técnica de aprendizagem supervisionada aplicada ao problema de classificação de dados, as Máquinas de Vetores Suporte (SVMs, do inglês *Support Vector Machines*).

Em um primeiro momento discutimos a formulação matemática de tal técnica para a classificação de dados linearmente separáveis, característica que qualifica a SVM com margem rígida. Posteriormente, estenderemos tal formulação para o caso em que os dados não são linearmente separáveis, obtendo assim a modelagem do problema de classificação com margem flexível. Para o desenvolvimento deste capítulo, as principais bibliografias utilizadas foram Deisenroth et al. [8] e Krulikovski [15].

4.1 CONCEITOS BÁSICOS DE APRENDIZAGEM DE MÁQUINA

Nas palavras de Deisenroth et al. [8, p. 11], a área da Aprendizagem de Máquina está interessada em

> (...) projetar algoritmos que extraem automaticamente informações valiosas dos dados. A ênfase aqui está em "automático", *i.e.*, a aprendizagem de máquina está preocupada com metodologias de uso geral que possam ser aplicadas em muitos con-

> juntos de dados, enquanto produz algo significativo (tradução nossa).

O aprendizado da máquina pode ser comparado ao do cérebro humano, pois os algoritmos de aprendizagem de máquina buscam "aprender" com a experiência, o que ocorre através do reconhecimento de padrões em dados [15].

Considere o exemplo de reconhecimento de dígitos numéricos escritos à mão. Para que uma técnica computacional seja capaz de diferenciá-los entre $0, 1, ..., 9$, é preciso primeiro fornecer à máquina imagens dos diferentes dígitos com suas respectivas classificações, isto é, informando quais dígitos são 0, quais são 1 e assim sucessivamente. A partir disso, esta técnica detecta padrões entre as características de cada dígito e sua classificação, e cria um modelo para deduzir a classificação de novos dígitos. Este é um exemplo de *aprendizado supervisionado*, em que através de um conjunto de dados de entrada fornecidos na forma (entrada, saída desejada), a máquina detecta padrões e produz um modelo capaz de deduzir as saídas corretas para novos dados [18]. Algumas técnicas para aprendizagem supervisionada são as Máquinas de Vetores Suporte (SVMs), Regressão Linear, Regressão Logística e Redes Neurais.

A *aprendizagem não-supervisionada* por sua vez é empregada em problemas que não possuem dados previamente rotulados. Em vista disso, tais técnicas são geralmente utilizadas com o intuito de auxiliar no entendimentos dos dados e obter informações acerca destes [18]. A Decomposição em Valores Singulares (SVD), Clusterização e Análise de Componentes Principais [15] são exemplos de técnicas de aprendizagem não-supervisionada.

A aprendizagem supervisionada é composta por uma etapa denominada *fase de treinamento*, na qual um *conjunto de treinamento* formado por vários pares na forma (x^i, y_i), em que x^i representa o vetor de características (ou atributos) e y_i corresponde à saída (ou rótulo), é fornecido à máquina e a partir do qual ela detecta padrões e cria um modelo para deduzir a saída de novos

dados não rotulados. Após este processo ocorre a *fase de testes*, na qual novas entradas, que compõem o *conjunto de teste*, serão testadas no intuito de analisar se a máquina está gerando as saídas corretas. Na aprendizagem supervisionada, assim como o nome já salienta, há a presença de um supervisor externo, o qual será responsável por fornecer à máquina os dados devidamente rotulados e averiguar se ela está gerando as saídas corretas.

Em muitas situações a aprendizagem supervisionada é aplicada a problemas cujo objetivo é obter uma classificação dos dados. Alguns exemplos são a detecção de *spam* em e-mails, o reconhecimento facial e o reconhecimento de dígitos escritos à mão. Em todos estes casos a técnica SVM pode ser aplicada, pois além de apresentar bons resultados quando comparada a outras técnicas de classificação, ela também possui uma boa capacidade de generalização, sendo indicada nos casos em que ocorrem dados de dimensões elevadas e com altos níveis de ruídos [15].

Assim, inicialmente mostraremos que a modelagem matemática do problema de classificação pela técnica SVM com margem rígida resulta num problema de programação quadrática, convexa e com restrições lineares, em que o objetivo é encontrar o hiperplano de máxima margem, isto é, que maximiza a distância entre os hiperplanos $w^T x + b = 1$ e $w^T x + b = -1$, os quais definem a faixa que separa os dados [15].

No entanto, problemas de classificação que envolvem situações reais dificilmente apresentam um conjunto de dados linearmente separável. Neste contexto faz-se necessário generalizar os resultados obtidos na modelagem da técnica SVM com margem rígida para o caso não-linear. Em vista disso, estudamos também SVM com margem flexível, em que os dados não são linearmente separáveis e para garantir a classificação correta através de um hiperplano no espaço de característica utilizamos regularização, introduzindo variáveis de folga ξ_i às restrições. No intuito de evitar classificações incorretas, tais variáveis são pena-

lizadas na função objetivo através de um parâmetro $C > 0$, por isso a técnica é também denominada CSVM.

É importante salientar que neste trabalho abordamos unicamente o problema de classificação binária, isto é, quando existem apenas duas saídas possíveis $\{-1, 1\}$. No entanto é possível generalizar para problemas que envolvam mais saídas, se configurando neste caso num problema multi-classes.

O nome da técnica SVM se refere aos vetores do conjunto de treinamento que estão sobre os hiperplanos que determinam a margem ótima. Tais vetores são os que apresentam maior relevância na hora de determinar o hiperplano ótimo, de modo que os demais vetores poderiam ser descartados sem interferir na determinação do classificador.

A seguir, desenvolveremos a modelagem matemática da técnica SVM com margem rígida e posteriormente sua generalização para margem flexível.

4.2 MÁQUINAS DE VETORES SUPORTE - MARGEM RÍGIDA

Nesta seção formularemos matematicamente a técnica SVM para a classificação binária de dados linearmente separáveis, que caracteriza o problema com margem rígida. Para tanto, abordaremos alguns conceitos importantes para a modelagem do problema, como a definição de hiperplano, conjuntos linearmente separáveis, além de discutir a noção de margem. Veremos também que a modelagem de tal problema recai num problema de otimização.

No caso da classificação binária, cada dado (atributo) será representado por um vetor de características x^i pertencente ao conjunto de entrada e a saída será representada por $y_i \in \{-1, 1\}$, de modo que diremos que x_i pertence à classe positiva se $y_i = 1$, e x^i pertence à classe negativa se $y_i = -1$. Sucintamente, a técnica SVM se concentra em obter um hiperplano ótimo $(w^*)^T x + b^* = 0$ que separe os dados de entrada x^i em duas saídas (classes) y_i

através de uma função de decisão.

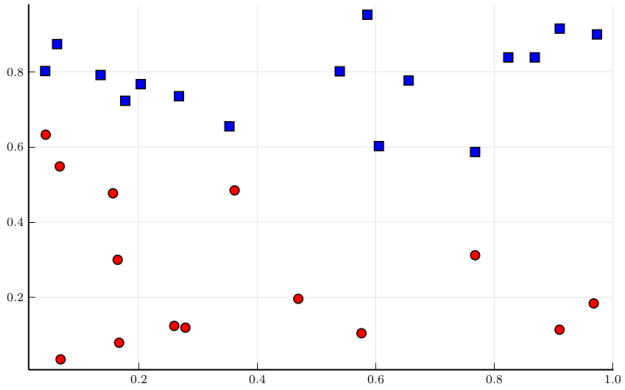
Considere um conjunto de dados de treinamento representado no $\mathbb{R}^2$ como na Figura 11a, em que os pontos em azul representam a classe positiva, e os pontos em vermelho a classe negativa. Para obter a classificação desse conjunto em duas classes, azul ou vermelho, a técnica SVM se concentra em determinar o hiperplano que melhor separa os dados corretamente. No entanto, para dados linearmente separáveis podem existir infinitos hiperplanos que separam corretamente os dados de treinamento; veja Figura 11b. Assim, a escolha do hiperplano ideal se baseia no conceito de distância entre os pontos que representam os dados e o hiperplano. Isto é, o *hiperplano ótimo*, como será denominado o hiperplano que melhor separa os dados, será aquele que possibilita a maior margem entre os dados positivos e negativos. Tal hiperplano é representado na Figura 12a pela cor rosa. Em outras palavras, estamos interessados em encontrar o hiperplano que maximiza a distância entre ele e os dados do conjunto de treinamento mais próximos. Note que, caso a margem seja muito estreita pequenas perturbações no hiperplano ou no conjunto de dados podem resultar uma classificação incorreta.

Inicialmente, para formular matematicamente o problema de classificação, é preciso estabelecer o conjunto de treinamento. Assim, considere o conjunto de treinamento $\mathcal{X} = \{(x^1, y_1), \ldots, (x^m, y_m) \mid x^i \in \mathbb{R}^n \text{ e } y_i \in \{-1, 1\}\}$, formado por pares de entrada x^i e saída y_i, com a seguinte partição

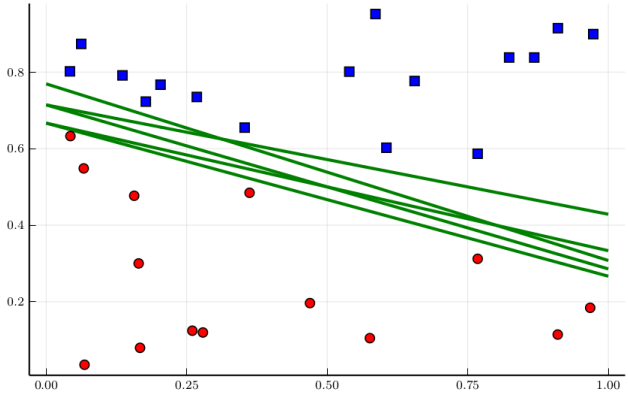
$$\mathcal{X}^+ = \{x^i \in \mathcal{X} \mid y_i = 1\} \quad \text{e} \quad \mathcal{X}^- = \{x^i \in \mathcal{X} \mid y_i = -1\}.$$

Diremos que $\mathcal{X}^+$ e $\mathcal{X}^-$ são os conjuntos formados pelos atributos pertencentes às classes positiva e negativa, respectivamente.

Definição 4.1. Considere um vetor não nulo $w \in \mathbb{R}^n$ e um escalar $b \in \mathbb{R}$. Um *hiperplano* com vetor normal w e constante b é um conjunto da forma $\mathcal{H}(w, b) = \{x \in \mathbb{R}^n \mid w^T x + b = 0\}$.

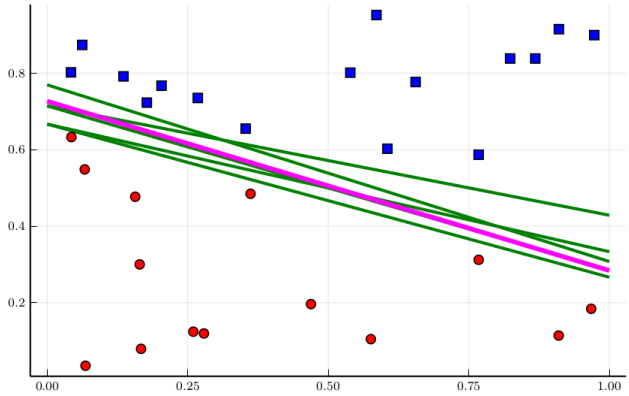


(a) Dados de treinamento.

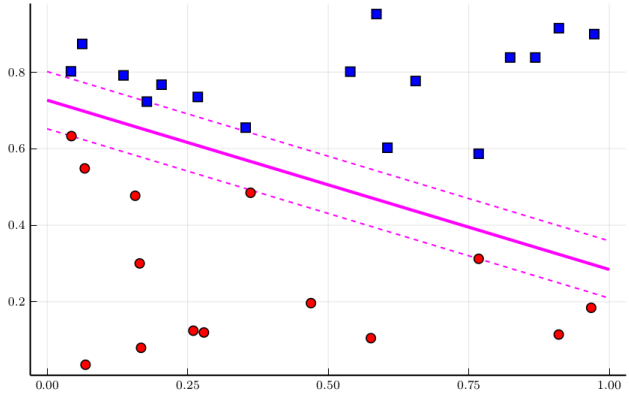


(b) Hiperplanos separadores.

Figura 11 – Conjunto de dados e hiperplanos.



(a) Hiperplano ótimo.



(b) Máxima margem.

Figura 12 – Hiperplano ótimo.

O hiperplano $\mathcal{H}(w,b)$ divide o espaço $\mathbb{R}^n$ em dois semiespaços, dados por

$$\mathcal{S}^+ = \{x \in \mathbb{R}^n \mid w^T x + b \geq 0\} \quad e \quad \mathcal{S}^- = \{x \in \mathbb{R}^n \mid w^T x + b \leq 0\}.$$

Definição 4.2. Os conjuntos $\mathcal{X}^+, \mathcal{X}^- \subset \mathbb{R}^n$ são ditos *linearmente separáveis* quando existem $w \in \mathbb{R}^n$ e $b \in \mathbb{R}$ tais que $w^T x + b > 0$ para todo $x \in \mathcal{X}^+$ e $w^T x + b < 0$ para todo $x \in \mathcal{X}^-$. O hiperplano $\mathcal{H}(w,b)$ é chamado hiperplano separador dos conjuntos $\mathcal{X}^+$ e $\mathcal{X}^-$.

Lema 4.3. *Suponha que os conjuntos $\mathcal{X}^+, \mathcal{X}^- \subset \mathbb{R}^n$ são finitos e linearmente separáveis, com hiperplano separador $\mathcal{H}(w,b)$. Então, existem $\overline{w} \in \mathbb{R}^n$ e $\overline{b} \in \mathbb{R}$ tais que $\mathcal{H}(w,b)$ pode ser descrito por*

$$\bar{w}^T x + \bar{b} = 0,$$

satisfazendo

$$\bar{w}^T x + \bar{b} \geq 1, \text{ para todo } x \in \mathcal{X}^+, \tag{63}$$

$$\bar{w}^T x + \bar{b} \leq -1, \text{ para todo } x \in \mathcal{X}^-. \tag{64}$$

Demonstração. Pela Definição 4.2, temos que existem $w \in \mathbb{R}^n$ e $b \in \mathbb{R}$ tais que

$$w^T x + b > 0, \text{ para todo } x \in \mathcal{X}^+,$$
$$w^T x + b < 0, \text{ para todo } x \in \mathcal{X}^-.$$

Como $\mathcal{X}^+ \cup \mathcal{X}^-$ é um conjunto finito, podemos definir

$$\gamma := \min_{x \in \mathcal{X}^+ \cup \mathcal{X}^-} |w^T x + b| > 0.$$

Portanto, para todo $x \in \mathcal{X}^+ \cup \mathcal{X}^-$, $\gamma \leq |w^T x + b|$ e consequentemente, $\dfrac{|w^T x + b|}{\gamma} \geq 1$. Assim, para $x \in \mathcal{X}^+$ temos

$$\frac{w^T x + b}{\gamma} = \frac{|w^T x + b|}{\gamma} \geq 1,$$

e para $x \in \mathcal{X}^-$, temos

$$-\frac{w^T x + b}{\gamma} = \frac{|w^T x + b|}{\gamma} \geq 1.$$

Logo, definindo $\bar{w} \coloneqq \dfrac{w}{\gamma}$ e $\bar{b} \coloneqq \dfrac{b}{\gamma}$, obtemos as desigualdades (63) e (64).

□

Sem perda de generalidade, a partir do Lema 4.3 podemos considerar o hiperplano $\mathcal{H}(w, b)$ com w e b satisfazendo (63) e (64). Logo, temos que $\mathcal{H}^+(w, b) \coloneqq \{x \in \mathbb{R}^n \mid w^T x + b = 1\}$ e $\mathcal{H}^-(w, b) \coloneqq \{x \in \mathbb{R}^n \mid w^T x + b = -1\}$ são os hiperplanos que definem a faixa que separa os conjuntos $\mathcal{X}^+$ e $\mathcal{X}^-$.

Como mencionado anteriormente, estamos interessados em obter o hiperplano que maximiza a margem entre os dados das classes positiva e negativa. Para tanto, precisamos calcular a distância entre um vetor x^i e o hiperplano $\mathcal{H}(w, b)$, para a partir disso obter a medida da margem. Tal distância será dada pela projeção ortogonal. Os dois resultados a seguir estabelecem, respectivamente, a projeção de um ponto $x \in \mathbb{R}^n$ qualquer sobre um hiperplano e a distância entre os hiperplanos $\mathcal{H}^+$ e $\mathcal{H}^-$.

Proposição 4.4. *A projeção ortogonal de um vetor $\bar{x} \in \mathbb{R}^n$ sobre um hiperplano afim $\mathcal{H}(w, b)$, é dada por*

$$\text{proj}_{\mathcal{H}(w,b)}(\bar{x}) = \bar{x} - \frac{w^T \bar{x} + b}{w^T w} w.$$

Além disso, $\text{proj}_{\mathcal{H}(w,b)}(\bar{x})$ *satisfaz a menor distância entre $\bar{x}$ e $\mathcal{H}(w, b)$.*

Demonstração. Sejam $w \in \mathbb{R}^n$ o vetor normal ao hiperplano $\mathcal{H}(w, b)$, $\bar{z} \in \mathcal{H}(w, b)$ e $x^* \coloneqq \text{proj}_{\mathcal{H}(w,b)}(\bar{x})$ a projeção ortogonal de $\bar{x}$ sobre $\mathcal{H}(w, b)$. Assim, temos que

$$w^T(x^* - \bar{z}) = 0 \tag{65}$$

e

$$\bar{x} - x^* = \lambda w \Longrightarrow x^* = \bar{x} - \lambda w. \tag{66}$$

Substituindo (66) em (65), obtemos

$$\begin{aligned} 0 &= w^T(\bar{x} - \lambda w - \overline{z}) \\ &= w^T\bar{x} - \lambda w^T w - w^T\bar{z}. \end{aligned}$$

Resolvendo para λ e como $w^T\bar{z} = -b$, temos

$$\lambda = \frac{w^T\bar{x} - w^T\bar{z}}{w^T w} = \frac{w^T\bar{x} + b}{w^T w}.$$

Portanto,

$$x^* = \bar{x} - \frac{w^T\bar{x} + b}{w^T w} w.$$

Ademais, vamos provar que a $\text{proj}_{\mathcal{H}(w,b)}(\bar{x})$ satisfaz a menor distância, isto é,

$$\|\bar{x} - x^*\|_2 \leq \|\bar{x} - x\|_2,$$

para todo $x \in \mathcal{H}(w, b)$. De fato, tomando $u = \bar{x} - x^*$ e $v = x^* - x$ observe que estes vetores são ortogonais, pois

$$\begin{aligned} u^T v &= (\bar{x} - x^*)^T(x^* - x) \\ &= (\bar{x} - \bar{x} + \lambda w)^T(x^* - x) \\ &= \lambda w^T(x^* - x) \\ &= \lambda(w^T x^* - w^T x) \\ &= \lambda(-b - (-b)) \\ &= 0. \end{aligned}$$

Assim, podemos utilizar Pitágoras, isto é, obtemos

$$\|u + v\|^2 = \|u\|^2 + 2u^T v + \|v\|^2 = \|u\|^2 + \|v\|^2,$$

e utilizamos as definições de u e v para obter

$$\|\bar{x} - x\|^2 = \|\bar{x} - x^*\|^2 + \|x^* - x\|^2 \geq \|x^* - x\|^2.$$

□

Agora, utilizando a Proposição 4.4 podemos demonstrar o Lema 4.5, o qual estabelece a largura da faixa entre os hiperplanos separadores $\mathcal{H}^+(w,b)$ e $\mathcal{H}^-(w,b)$.

Lema 4.5. *A distância entre os hiperplanos $\mathcal{H}^+(w,b)$ e $\mathcal{H}^-(w,b)$ é dada por*

$$\text{dist}(\mathcal{H}^+, \mathcal{H}^-) = \frac{2}{\|w\|}.$$

Demonstração. Considere um ponto arbitrário $\bar{x} \in \mathcal{H}^+(w,b)$ e seja $x^* \in \mathcal{H}^-(w,b)$ a projeção ortogonal de $\bar{x}$ sobre $\mathcal{H}^-(w,b)$. Usando a Proposição 4.4, temos

$$x^* = \text{proj}_{\mathcal{H}^-(w,b)}(\bar{x}) = \bar{x} - \frac{w^T\bar{x} + b + 1}{\|w\|^2} w. \tag{67}$$

Além disso, a distância entre dois conjuntos é definida por

$$\text{dist}(\mathcal{H}^+, \mathcal{H}^-) \coloneqq \inf\{\|x^+ - x^-\| : x^+ \in \mathcal{H}^+(w,b) \text{ e } x^- \in \mathcal{H}^-(w,b)\},$$

e como a $\text{proj}_{\mathcal{H}^-(w,b)}(\bar{x})$ satisfaz a menor distância entre $\bar{x}$ e $\mathcal{H}^-(w,b)$ e $\mathcal{H}^+(w,b)$ é paralelo a $\mathcal{H}^-(w,b)$, temos que

$$\text{dist}(\mathcal{H}^+, \mathcal{H}^-) = \|\bar{x} - x^*\|. \tag{68}$$

Substituindo (67) em (68), obtemos

$$\begin{aligned}\operatorname{dist}(\mathcal{H}^+, \mathcal{H}^-) &= \|\bar{x} - x^*\| \\ &= \left\| \bar{x} - \bar{x} + \frac{w^T\bar{x} + b + 1}{\|w\|^2} w \right\| \\ &= \frac{|w^T\bar{x} + b + 1|}{\|w\|^2}\|w\| \\ &= \frac{|w^T\bar{x} + b + 1|}{\|w\|},\end{aligned}$$

e como $\bar{x} \in \mathcal{H}^+(w, b)$, $w^T\bar{x} + b = 1$ implica

$$w^T\bar{x} = 1 - b,$$

concluimos que

$$\begin{aligned}\operatorname{dist}(\mathcal{H}^+, \mathcal{H}^-) &= \frac{|1 - b + b + 1|}{\|w\|} \\ &= \frac{2}{\|w\|}.\end{aligned}$$

□

Portanto, encontrar o hiperplano que melhor separa os dados implica maximizar a largura da margem, isto é, maximizar $\operatorname{dist}(\mathcal{H}^+, \mathcal{H}^-) = \dfrac{2}{\|w\|}$. Isso equivale a minimizar seu inverso $\dfrac{1}{2}\|w\|$ ou ainda minimizar $\dfrac{1}{2}\|w\|^2$. De fato, seja $w^* = \arg\max \dfrac{2}{\|w\|}$. Então, para todo $w \in \mathbb{R}^n$,

$$\frac{2}{\|w^*\|} \geq \frac{2}{\|w\|}$$

implica

$$\|w\| \geq \|w^*\|. \tag{69}$$

Logo, $w^* \in \arg\min \|w\|$. Além disso, como $\|\cdot\|$ é não negativa, elevando ao quadrado ambos os lados da desigualdade (69) temos que $\|w\|^2 \geq \|w^*\|^2$ implica

$$\frac{1}{2}\|w\|^2 \geq \frac{1}{2}\|w^*\|^2.$$

Portanto,

$$\arg\max \frac{2}{\|w\|} = \arg\min \frac{1}{2}\|w\|^2.$$

Ademais, como a faixa deve separar os dados das duas classes, as seguintes restrições devem ser satisfeitas

$$w^T x + b \geq 1, \text{ para todo } x \in \mathcal{X}^+,$$
$$w^T x + b \leq -1, \text{ para todo } x \in \mathcal{X}^-.$$

Considerando que $\mathcal{X}^+ = \{x^i \in \mathcal{X} \mid y_i = 1\}$ e $\mathcal{X}^- = \{x^i \in \mathcal{X} \mid y_i = -1\}$, podemos reescrever as restrições acima de uma forma mais compacta

$$y_i(w^T x^i + b) \geq 1, \quad i = 1, \ldots, m.$$

Portanto, o problema de encontrar o hiperplano ótimo pode ser formulado da seguinte maneira

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}\|w\|^2 \\ \text{s.a.} \quad & y_i(w^T x^i + b) \geq 1, \quad i = 1, \ldots, m, \end{aligned} \tag{70}$$

em que $w \in \mathbb{R}^n$ e $b \in \mathbb{R}$.

O problema (70) possui função objetivo

$$f(w, b) = \frac{1}{2}\|w\|^2,$$

a qual é convexa de acordo com o Teorema 3.21 (tomando $Q = I$ e $b = 0$, no teorema), e restrições lineares

$$g_i(w, b) = 1 - y_i(w^T x^i + b) \leq 0, \quad i = 1, \ldots, m,$$

em que a função $g : \mathbb{R}^{n+1} \to \mathbb{R}^m$ pode ser escrita da forma

$$g(w, b) = e - (YX^T w + by) \leq 0,$$

com e sendo o vetor cujas m componentes são todas iguais a 1, $Y = \text{diag}(y_i)$, $X = \text{diag}(x^i)$, $y^T = [y_1 \ \dots \ y_m]$, $w \in \mathbb{R}^n$ e $b \in \mathbb{R}$.

Como os conjuntos $\mathcal{X}^+$ e $\mathcal{X}^-$ são linearmente separáveis, concluímos que o conjunto factível $\Omega = \{(w, b) \in \mathbb{R}^{n+1} \mid g_i(w, b) \leq 0, i = 1, \dots, m\}$ é um poliedro não-vazio segundo a Definição 2.27. Logo, o Corolário 3.5 nos garante que o conjunto factível Ω do problema (70) é um conjunto convexo. Além disso, a função objetivo é limitada inferiormente, pois $\frac{1}{2}\|w\|^2 > 0$, e sua Hessiana $\nabla^2 f(x) = I$, e portanto, é semidefinida positiva para todo $(w, b) \in \Omega$. Assim, a função objetivo é uma função convexa. Portanto, o problema (70) é um problema de otimização convexa e valem os Teoremas 3.19 e 3.20, os quais garantem que todo minimizador local é global e todo ponto estacionário é minimizador global.

Agora, com base em Krulikovski [15, Teo. 2.5 e 2.7], nosso objetivo é garantir a existência e unicidade de um minimizador para o problema (70).

Teorema 4.6. *O problema* (70) *possui minimizador global.*

Demonstração. Primeiramente, pelas considerações acima temos que a função objetivo do problema (70) é limitada inferiormente e o conjunto $\Omega \neq \emptyset$. Logo, podemos definir

$$f^* := \inf_{(w,b)\in\Omega} f(w, b) > -\infty.$$

Pela definição de ínfimo, temos que para todo $k \in \mathbb{N}$, existe $(w_k, b_k)_{k\in\mathbb{N}} \subset \Omega$ tal que

$$\frac{1}{2}\|w_k\|^2 = f(w_k, b_k) \to f^*, \tag{71}$$

e portanto, a sequência $(\|w_k\|)_{k\in\mathbb{N}}$ é convergente. Consequentemente, como toda sequência convergente é limitada, concluímos que $(w_k)_{k\in\mathbb{N}}$ é limitada.

Mostraremos agora que $(b_k)_{k\in\mathbb{N}}$ também é limitada. Para tanto, considere $\bar{x} \in \mathcal{X}^+$ e $\tilde{x} \in \mathcal{X}^-$ arbitrários. Temos que

$$w_k^T \bar{x} + b_k \geq 1$$

e

$$w_k^T \tilde{x} + b_k \leq -1,$$

e portanto,

$$1 - w_k^T \bar{x} \leq b_k \leq -1 - w_k^T \tilde{x}.$$

Assim, como (w_k) é uma sequência limitada segue que (b_k) é limitada também.

Como a sequência (w_k, b_k) é limitada, pelo Teorema de Bolzano-Weierstrass, ela possui uma subsequência convergente $(w_{k_j}, b_{k_j})_{j\in\mathbb{N}}$, isto é,

$$(w_{k_j}, b_{k_j}) \to (w^*, b^*).$$

Logo, pela continuidade de f obtemos

$$f(w_{k_j}, b_{k_j}) \to f(w^*, b^*),$$

e por (71) temos que $f(w^*, b^*) = f^*$. Portanto, existe $(w^*, b^*) \in \Omega$ tal que

$$f(w^*, b^*) = f^* \leq f(w, b),$$

para todo $(w, b) \in \Omega$. □

Além da existência também podemos garantir a unicidade da solução do problema (70). Para sua demonstração utilizaremos o seguinte lema.

Lema 4.7. *Se (w^*, b^*) é uma solução ótima para o problema* (70) *então existem $\bar{x} \in \mathcal{X}^+$ e $\tilde{x} \in \mathcal{X}^-$ tais que $(w^*)^T \bar{x} + b^* = 1$ e $(w^*)^T \tilde{x} + b^* = -1$.*

Demonstração. Como (w^*, b^*) é uma solução ótima para o problema (70), temos que

$$f(w^*, b^*) \leq f(w, b),$$

para todo $(w, b) \in \Omega$. Suponhamos que não existe $\bar{x} \in \mathcal{X}^+$ satisfazendo $(w^*)^T \bar{x} + b^* = 1$. Então,

$$(w^*)^T x + b^* \geq 1 + \delta,$$

para todo $x \in \mathcal{X}^+$, em que $\delta \coloneqq \min_{x \in \mathcal{X}^+}[(w^*)^T x + b^* - 1] > 0$. Desse modo,

$$(w^*)^T x + b^* - \frac{\delta}{2} \geq 1 + \frac{\delta}{2},$$

e dividindo ambos os lados da desigualdade por $1 + \dfrac{\delta}{2}$, obtemos

$$\frac{(w^*)^T x + b^* - \dfrac{\delta}{2}}{1 + \dfrac{\delta}{2}} \geq 1.$$

Agora, tomando $\bar{w} = \dfrac{w^*}{1 + \dfrac{\delta}{2}}$ e $\bar{b} = \dfrac{b^* - \dfrac{\delta}{2}}{1 + \dfrac{\delta}{2}}$, temos

$$\bar{w}^T x + \bar{b} \geq 1,$$

para todo $x \in \mathcal{X}^+$.

Por conseguinte, para $x \in \mathcal{X}^-$ temos $(w^*)^T x + b^* \leq -1$, o que equivale a

$$(w^*)^T x + b^* - \frac{\delta}{2} \leq -1 - \frac{\delta}{2}.$$

Assim, temos que $\bar{w}^T x + \bar{b} \leq -1$, para todo $x \in \mathcal{X}^-$. Portanto, encontramos $(\bar{w}, \bar{b})$ satisfazendo as restrições do problema (70) e

$$f(\bar{w}, \bar{b}) = \frac{1}{2}\|\bar{w}\|^2 = \frac{1}{2}\frac{\|w^*\|^2}{\left(1 + \frac{\delta}{2}\right)^2} < \frac{1}{2}\|w^*\|^2 = f(w^*, b^*),$$

contradizendo a hipótese de que (w^*, b^*) é ótimo. □

Utilizando o Lema 4.7, podemos demonstrar a unicidade da solução do problema (70).

Teorema 4.8. *O minimizador global para o problema* (70) *é único.*

Demonstração. Suponhamos que $(\bar{w}, \bar{b})$ e $(\tilde{w}, \tilde{b})$ são soluções ótimas para o problema (70). Então, temos que $f(\bar{w}, \bar{b}) = f(\tilde{w}, \tilde{b})$, o que implica

$$\frac{1}{2}\|\bar{w}\|^2 = \frac{1}{2}\|\tilde{w}\|^2,$$

e consequentemente,

$$\|\bar{w}\| = \|\tilde{w}\|.$$

Como Ω é um conjunto convexo, temos que o segmento $[(1-t)(\bar{w}, \bar{b}) + t(\tilde{w}, \tilde{b})]$, tal que $t \in [0, 1]$, está contido em Ω. Em particular, tomando $t = \frac{1}{2}$,

$$(\hat{w}, \hat{b}) = \frac{1}{2}(\bar{w}, \bar{b}) + \frac{1}{2}(\tilde{w}, \tilde{b}) \in \Omega. \tag{72}$$

Pela igualdade de vetores em (72), temos que o vetor $\hat{w} = \frac{1}{2}(\bar{w} + \tilde{w})$. Ademais, pelo Lema 3.6 temos que, se $\|\bar{w}\| = \|\tilde{w}\| = r$, então $\|\frac{1}{2}(\bar{w} + \tilde{w})\| < r$. Portanto,

$$\|\hat{w}\| = \|\frac{1}{2}(\bar{w} + \tilde{w})\| < r = \|\bar{w}\|,$$

o que implica

$$f(\hat{w}, \hat{b}) < f(\bar{w}, \bar{b}),$$

contradizendo o fato de $(\bar{w}, \bar{b})$ ser uma solução ótima. Portanto, temos a unicidade de w.

Agora, vamos mostrar a unicidade de b. Suponha que $(w, \bar{b})$ e $(w, \tilde{b})$ são soluções ótimas para o problema (70) com, por exemplo, $\tilde{b} < \bar{b}$. Pelo Lema 4.7, existe $x \in \mathcal{X}^+$ tal que $w^T x + \bar{b} = 1$. Por outro lado,

$$1 \leq w^T x + \tilde{b} < w^T x + \bar{b} = 1,$$

o que é uma contradição. □

A seguir apresentamos a definição de vetores suporte, conceito que nomeia a técnica SVM, e comentamos brevemente sua relevância com base nas considerações feitas em [4, 15].

Definição 4.9. Considere um conjunto $\mathcal{X}$ de vetores linearmente separáveis e (w^*, b^*) a solução do problema (70). Os *vetores suporte* são os vetores $x^i \in \mathcal{X}$ tais que $y_i((w^*)^T x^i + b^*) = 1$.

Desse modo, os vetores suporte são aqueles localizados sobre os hiperplanos que definem a máxima margem que separa os dados, ou, em outras palavras, são os vetores que correspondem às restrições ativas na solução (w^*, b^*) do problema (70). Tal propriedade detém grande importância na aplicação prática da técnica SVM, pois de acordo com o Teorema 4.10 a seguir, temos que após o modelo ser treinado uma significativa quantidade de dados pode ser descartada e apenas os vetores suporte mantidos de modo que o classificador se mantém o mesmo.

Teorema 4.10. *Considere* (w^*, b^*) *a solução do problema* (70) *e* $\mathcal{I}^* = \mathcal{I}(w^*, b^*)$ *o conjunto dos índices das restrições ativas na solução. Então, a solução do problema*

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}\|w\|^2 \\ s.a. \quad & y_i(w^T x^i + b) \geq 1, \quad i \in \mathcal{I}^* \end{aligned} \tag{73}$$

é (w^*, b^*).

Demonstração. Seja $(\bar{w}, \bar{b})$ solução do problema (73). Como o conjunto factível do problema (70) está contido no conjunto factível do problema (73), temos que

$$f(\bar{w}, \bar{b}) \leq f(w^*, b^*).$$

Agora, considere para cada $t \in (0, 1)$

$$(w_t, b_t) = (1-t)(w^*, b^*) + t(\bar{w}, \bar{b}).$$

Para $i \in \mathcal{I}^*$, a convexidade da restrição g_i e a factibilidade das soluções garantem que

$$g_i(w_t, b_t) \leq (1-t)g_i(w^*, b^*) + tg_i(\bar{w}, \bar{b}) \leq 0.$$

Por outro lado, para $i \notin \mathcal{I}^*$, temos que $g_i(w^*, b^*) < 0$ e portanto, pela continuidade de g_i, vale $g_i(w_t, b_t) < 0$ para t suficientemente pequeno. Portanto, (w_t, b_t) é factível para o problema (70).

Por conseguinte, afirmamos que $f(\bar{w}, \bar{b}) = f(w^*, b^*)$. Com efeito, se $f(\bar{w}, \bar{b}) < f(w^*, b^*)$, então

$$\|w_t\| = \|(1-t)w^* + t\bar{w}\| \leq (1-t)\|w^*\| + \|\bar{w}\| < \|w^*\|,$$

implicando $f(w_t, b_t) < f(w^*, b^*)$, o que é uma contradição. Concluímos então que (w_t, b_t) e (w^*, b^*) são factíveis para o problema (73) com o mesmo valor

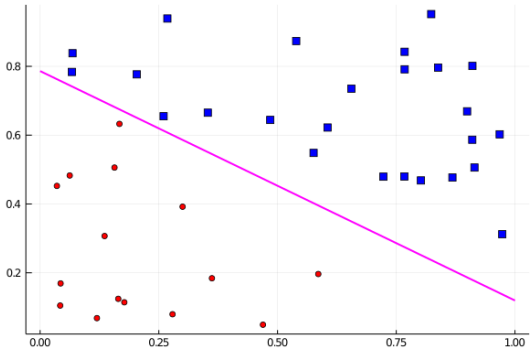
ótimo da função objetivo, ou seja, ambos são minimizadores globais do problema (70). Pelo Teorema 4.8, segue que $(\bar{w}, \bar{b}) = (w^*, b^*)$. □

Ademais, conforme mencionado por Krulikovski [15], como somente os vetores suporte possuem um papel relevante na determinação do hiperplano ótimo, o cálculo de w^* torna-se mais barato computacionalmente.

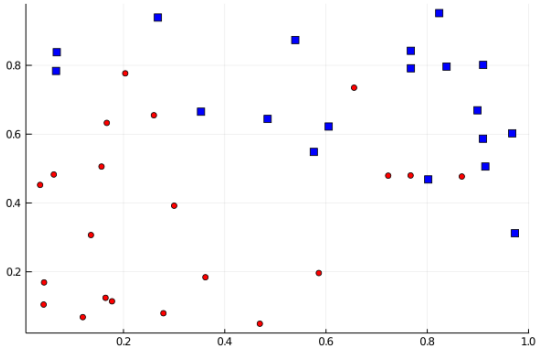
Portanto, nesta seção desenvolvemos a formulação matemática da técnica SVM com margem rígida para o problema de classificação. Concluímos que tal formulação resulta em um problema de otimização e demonstramos a existência de solução e sua unicidade. No entanto, como já mencionado em alguns momentos no decorrer do texto, problemas que envolvam situações reais dificilmente apresentam dados linearmente separáveis. Em vista disso, na seção a seguir trataremos da modelagem do problema com margem flexível, caso em que os dados não são linearmente separáveis.

4.3 MÁQUINAS DE VETORES SUPORTE - MARGEM FLEXÍVEL (CSVM)

No caso em que os dados não são linearmente separáveis não é possível determinar pela técnica SVM com margem rígida um hiperplano no espaço de entrada que separa corretamente todos os dados do conjunto de treinamento como representado na Figura 13a. Assim, como classificar um conjunto de dados como o representado na Figura 13b através de um hiperplano, sendo que alguns dados poderiam estar localizados na região da margem ou até mesmo do lado errado do hiperplano? É pensando nesse contexto que estenderemos agora os conceitos desenvolvidos na modelagem da técnica SVM com margem rígida para formular a técnica SVM com margem flexível, aplicada no caso em que desejamos classificar dados que não são linearmente separáveis mas que ainda é possível encontrar um hiperplano separador no espaço de entrada ao promover uma flexibilização da margem.



(a) Dados linearmente separáveis.



(b) Dados não-linearmente separáveis.

Figura 13 – Conjunto de dados.

Considerando um conjunto de dados não-linearmente separável, como na Figura 13b por exemplo, não existe um hiperplano separador como aquele estabelecido pela Definição 4.2 para conjuntos linearmente separáveis. Neste

caso, temos que o conjunto factível

$$\Omega = \{(w, b) \in \mathbb{R}^{n+1} \mid 1 - y_i(w^T x^i + b) \leq 0, \quad i = 1, \ldots, m\}$$

é vazio, comprovando que a formulação dada pelo problema (70) não fornece um classificador nesta situação.

A ideia central por trás da técnica SVM com margem flexível é introduzir variáveis de folga $\xi_i \geq 0$ associadas aos dados de treinamento x^i, com $i = 1, \ldots, m$, suavizando a margem e permitindo desse modo que o problema de estimar as variáveis w e b torne-se mais flexível. Em outras palavras, as restrições $1 - y_i(w^T x^i + b) \leq 0$ são relaxadas e substituídas por $1 - y_i(w^T x^i + b) \leq \xi_i$, em que $\xi_i \geq 0$. Tal processo é também denominado de regularização.

Cada variável de folga ξ_i corresponde à distância que determinado dado x^i está do hiperplano que delimita a margem, isto é, $\mathcal{H}^+$ ou $\mathcal{H}^-$. Dessa forma, caso um dado x^i esteja localizado no semiespaço correto a variável de folga $\xi_i = 0$, se estiver localizado no lado correto em relação ao hiperplano separador, porém na região da margem, temos que $0 < \xi_i < 1$, e se tal dado estiver no lado errado, então $\xi_i > 1$. Portanto, este processo de regularização flexibiliza a margem, permitindo que pontos da classe positiva permaneçam fora do semiespaço $\mathcal{S}^+ = \{x \in \mathbb{R}^n \mid w^T x + b \geq 1\}$ ou pontos da classe negativa fora do semiespaço $\mathcal{S}^- = \{x \in \mathbb{R}^n \mid w^T x + b \leq -1\}$.

Nesta formulação, o hiperplano separador $\mathcal{H}(w, b)$ é denominado hiperplano de margem flexível e as restrições que correspondem aos hiperplanos $\mathcal{H}^+(w, b)$ e $\mathcal{H}^-(w, b)$, que delimitam a margem, são reformuladas da seguinte maneira

$$w^T x^i + b \geq 1 - \xi_i, \text{ para todo } x^i \in \mathcal{X}^+, \tag{74}$$

$$w^T x^i + b \leq -1 + \xi_i, \text{ para todo } x^i \in \mathcal{X}^-. \tag{75}$$

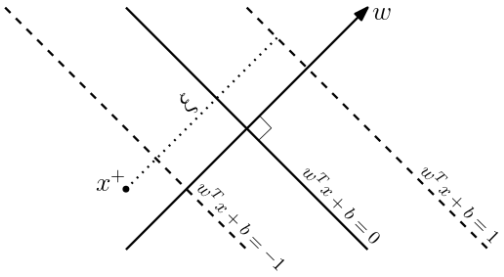


Figura 14 – Variáveis de folga.

Observe que, dados w e b arbitrários, podemos escolher $\xi_i \geq 0$ de modo que as restrições (74) e (75) sejam satisfeitas. Para tanto, basta definir

$$\xi_i = \begin{cases} \max\{0, 1 - w^T x^i - b\}, & \text{se} \quad x^i \in \mathcal{X}^+, \\ \max\{0, 1 + w^T x^i + b\}, & \text{se} \quad x^i \in \mathcal{X}^-. \end{cases}$$

No entanto, para obter um bom classificador não basta apenas maximizar a margem definida pelos hiperplanos $\mathcal{H}^+(w, b)$ e $\mathcal{H}^-(w, b)$ e introduzir as variáveis de folga nas restrições, mantendo a mesma função objetivo. Como exemplificado por Krulikovski [15, p. 45], a Figura 15 ilustra o hiperplano dado por $w_0^T x + b_0 = 0$, o qual satisfaz as restrições (74) e (75) mas não serve para classificar os dados.

Portanto, além de acrescentar as variáveis de folga às restrições, é preciso controlar seu valor para encorajar uma correta classificação. Em vista disso, acrescentamos à função objetivo o somatório dessas variáveis de folga ξ_i multiplicado a um parâmetro de penalização $C > 0$. Assim, o problema (70) é

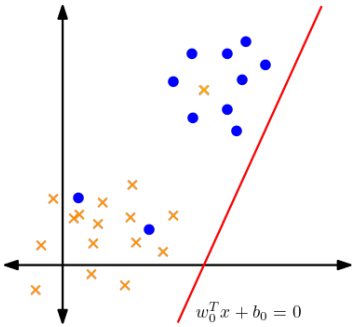


Figura 15 – Exemplo de hiperplano que satisfaz as restrições mas não classifica os dados.

reformulado da seguinte forma

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i \\ \text{s.a.} \quad & y_i(w^T x^i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, m, \\ & \xi_i \geq 0, \quad i = 1, \ldots, m. \end{aligned} \tag{76}$$

O parâmetro de penalização C tem como objetivo controlar a importância das variáveis de folga ao minimizar a função objetivo. Note que, caso seja atribuído a ele um valor muito alto o problema se concentra em minimizar as variáveis de folga, deixando de lado seu principal objetivo que é minimizar o termo $\frac{1}{2}\|w\|^2$, que implica na maximização da margem. Por outro lado, caso C assuma valores muito pequenos uma maior quantidade de vetores recebe folga, o que pode resultar numa classificação incorreta. Dessa forma, Krulikovski [15] salienta que o valor do parâmetro C que fornece uma boa classificação dos dados é escolhido de maneira heurística nas fases de treinamento e teste, sendo necessário considerar também natureza do problema.

É devido a presença do parâmetro C no problema (76) formulado para a técnica SVM com margem flexível que ela também é denominada CSVM.

O problema com margem flexível (76) também possui restrições lineares

$$g_i(w, b, \xi) = 1 - \xi_i - y_i(w^T x^i + b) \leq 0$$

e

$$h_i(w, b, \xi) = -\xi_i \leq 0,$$

para $i = 1, \ldots, m$, e assim, o conjunto factível

$$\Omega = \{(w, b, \xi) \in \mathbb{R}^{n+1+m} \mid g_i(w, b, \xi) \leq 0,\ h_i(w, b, \xi) \leq 0,\ i = 1, \ldots, m\} \quad (77)$$

é um poliedro não-vazio. Ademais, a função objetivo é quadrática e limitada inferiormente em Ω, pois

$$f(w, b, \xi) = \frac{1}{2}\|w\|^2 + \underbrace{C}_{>0} \sum_{i=1}^{m} \underbrace{\xi_i}_{\geq 0} \geq 0. \quad (78)$$

Novamente, baseando-se em Krulikovski [15, Teo. 2.13], apresentamos a seguir o teorema que garante a existência de um minimizador global para o problema (76).

Teorema 4.11. *O problema* (76) *possui minimizador global.*

Demonstração. Como visto em (77) e (78), o conjunto factível do problema (76) é não-vazio e a função objetivo é limitada inferiormente. Logo, definimos

$$f^* := \inf_{(w,b,\xi)\in\Omega} f(w, b, \xi) > -\infty.$$

Pela definição de ínfimo, podemos concluir que para todo $k \in \mathbb{N}$, existe $(w_k, b_k, \xi_k)_{k\in\mathbb{N}} \subset \Omega$ tal que

$$f^* \leq f(w_k, b_k, \xi_k) < f^* + \frac{1}{k}.$$

Desse modo, pelo Teorema do Sanduíche,

$$\frac{1}{2}\|w_k\|^2 + C\sum_{i=1}^{m}(\xi_k)_i = f(w_k, b_k, \xi_k) \to f^*. \tag{79}$$

Como

$$0 \leq \frac{1}{2}\|w_k\|^2 \leq f(w_k, b_k, \xi_k) \quad \text{e} \quad 0 \leq C\sum_{i=1}^{m}(\xi_k)_i \leq f(w_k, b_k, \xi_k),$$

temos que as sequências $(w_k)_{k\in\mathbb{N}}$ e $(\xi_k)_{k\in\mathbb{N}}$ são limitadas. Vejamos que a sequência $(b_k)_{k\in\mathbb{N}}$ também é limitada. Com efeito, dados $x^j \in \mathcal{X}^+$ e $x^l \in \mathcal{X}^-$ arbitrários, de (74) e (75) resulta que

$$1 - (\xi_k)_j - w_k^T x^j \leq b_k \leq -1 + (\xi_k)_l - w_k^T x^l.$$

Dessa forma, a sequência (w_k, b_k, ξ_k) é limitada e, pelo Teorema de Bolzano-Weierstrass, possui uma subsequência convergente

$$(w_{k_j}, b_{k_j}, \xi_{k_j}) \to (w^*, b^*, \xi^*).$$

Logo, pela continuidade de f, obtemos

$$f(w_{k_j}, b_{k_j}, \xi_{k_j}) \to f(w^*, b^*, \xi^*),$$

e de (79) decorre que $f(w^*, b^*, \xi^*) = f^*$. Assim, provamos que existe $(w^*, b^*, \xi^*) \in \Omega$ tal que

$$f(w^*, b^*, \xi^*) = f^* \leq f(w, b, \xi)$$

para todo $(w, b, \xi) \in \Omega$. □

Portanto, para o caso em que os dados não são linearmente separáveis, mas é possível determinar um hiperplano com margem flexível no espaço de

entrada que os separe, o problema de classificação pode ser resolvido através do problema (76), e este admite solução.

Em síntese, com base na formulação matemática da técnica SVM desenvolvida neste capítulo, sua atuação consiste em resolver, para o conjunto de treino, o problema (70) caso o conjunto de dados seja linearmente separável ou o problema (76) caso contrário. Em seguida, através da solução (w^*, b^*) encontrada definimos a função de decisão $F : \mathbb{R}^n \to \mathbb{R}$ dada por

$$F(x) = \text{sgn}((w^*)^T x + b^*),$$

em que $\text{sgn}(a)$ é 1 se a for não-negativo e -1 se a for negativo, e através da qual podemos classificar um novo dado $x \in \mathbb{R}^n$. Assim, se $F(x) > 0$, o ponto x será classificado como da classe $\mathcal{X}^+$, e se $F(x) < 0$, o ponto x será classificado como da classe $\mathcal{X}^-$.

No próximo capítulo utilizaremos o referencial teórico apresentado até o momento, assim como os problemas formulados em (70) e (76), para testar numericamente através de uma implementação computacional a técnica SVM para classificação.

5 EXPERIMENTOS NUMÉRICOS

Este capítulo é dedicado a apresentar alguns experimentos numéricos que desenvolvemos e que têm como objetivo visualizar na prática a implementação da técnica Máquinas de Vetores Suporte (SVM), possibilitando assim analisar suas particularidades. Para tanto, o capítulo é dividido em duas seções. Na primeira, discutimos a aplicação de SVM para classificar o conjunto de dados de flor de Íris entre suas espécies. A segunda seção aborda a aplicação da técnica CSVM para detectar num conjunto de dados sobre células de câncer de mama quais apresentam tumor maligno ou benigno.

Para desenvolvimento desses experimentos utilizamos o *software* de programação Julia em sua versão 1.4.0, além dos pacotes `Plots` para gerar as imagens, `JuMP` [10] e `Ipopt` [26] para resolução dos problemas de otimização, `RDatasets` para ter acesso ao conjunto de dados Íris e o pacote `DataFrame` para gerar as tabelas de dados. Ademais, nos baseamos nos vídeos de Siqueira [24, 25] para desenvolvimento dos códigos na plataforma Jupyter Notebook. Assim, os principais códigos serão apresentados no decorrer das seções seguintes associados às explicações da implementação computacional.

Como abordado no capítulo anterior, Máquinas de Vetores Suporte é uma técnica de Aprendizagem de Máquina muito utilizada para classificação e regressão, e nosso objetivo será sua aplicação em problemas que envolvam a classificação binária de dados. Primeiramente, vamos relembrar que em problemas de classificação estamos interessados, assim como o nome já antecipa, em classificar da melhor maneira possível um determinado conjunto de dados. No caso em que os dados são linearmente separáveis, isto é, existe um hiperplano que os separa corretamente, aplica-se SVM de margem rígida e o problema costuma ter uma resolução mais simples. Entretanto, os problemas de classificação que envolvem situações reais costumam ser mais elaborados, pois neste caso os

dados geralmente não são linearmente separáveis. Nessas situações é necessário utilizar SVM com margem flexível (CSVM), se os dados forem, a grosso modo, mais "comportados", ou a SVM não-linear. Esse último caso exige um desenvolvimento teórico matemático mais avançado e que foge do escopo deste trabalho e portanto não será abordado, se constituindo numa proposta de estudos a ser desenvolvida em projetos futuros.

5.1 IMPLEMENTAÇÃO DE SVM PARA CLASSIFICAÇÃO DO CONJUNTO DE DADOS ÍRIS

Neste primeiro momento nosso propósito será implementar a técnica SVM num exemplo prático: o conjunto de dados flor de Íris [11]. Tal conjunto de dados consiste em 150 amostras de três espécies da planta Íris, sendo 50 amostras da Íris setosa, 50 da Íris virginica e 50 da Íris versicolor. Cada dado amostral contém as medidas de quatro variáveis morfológicas: comprimento e largura das sépalas e das pétalas, medidas em centímetros. É com base nas diferenciações e semelhanças dessas características que é possível distinguir uma espécie da outra.

Figura 16 – As três espécies da flor de Íris: versicolor, setosa e virginica.

Fonte – https://www.datacamp.com/community/tutorials/machine-learning-in-r

Como nossos estudos se concentram na classificação binária, nosso objetivo inicial será aprender a separar os dados em setosa e não setosa e posteriormente em virginica e não virginica. Assim, o conjunto de dados Íris é representado na Figura 17, em que os pontos em azul pertencem à espécie setosa, os pontos em vermelho à espécie versicolor e os pontos em verde à espécie virginica. Tais gráficos foram construídos considerando-se duas características de cada vez.

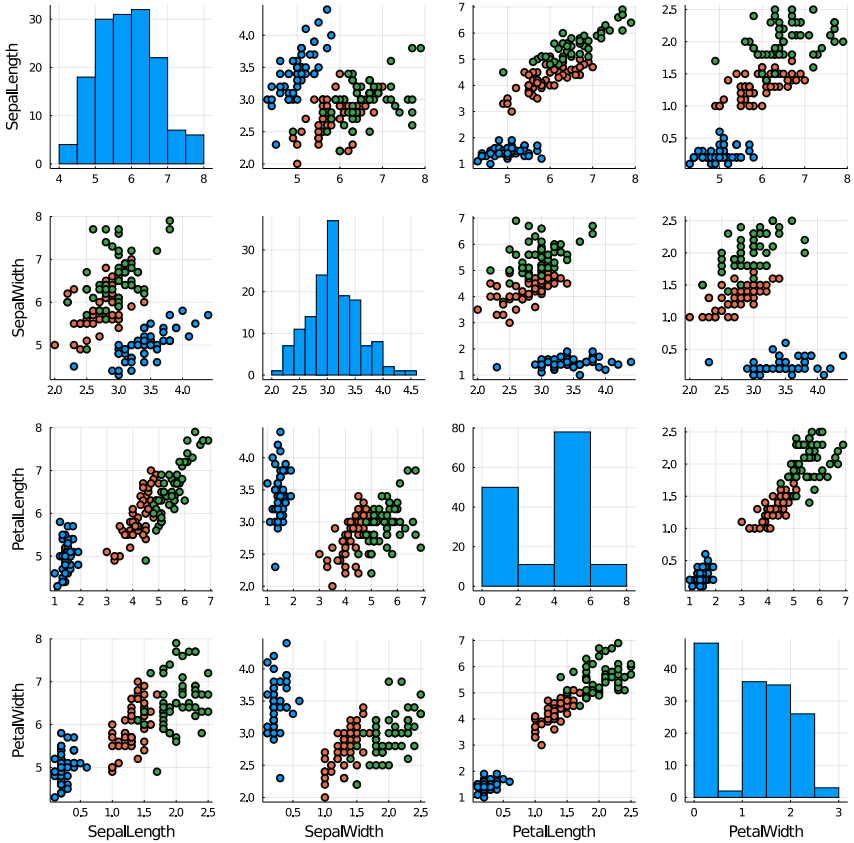


Figura 17 – Conjunto de dados Flor de Íris.

5.1.1 Classificação com duas características

De acordo com a Capítulo 4, a técnica SVM com margem rígida pode ser formulada pelo problema (70). Assim, utilizando esta formulação, apresentamos a seguir uma aplicação na classificação do conjunto de dados Íris em setosa e não setosa com base em duas características apenas.

Inicialmente, é importante lembrar que na modelagem do problema cada dado é representado por um vetor no espaço n-dimensional, em que n corresponde ao número de características do problema em questão. Neste exemplo, como o problema compreende apenas duas características os dados de entrada x^i pertencem ao espaço $\mathbb{R}^2$. Logo, podemos representar estes dados através de uma matriz $X_{150\times 2}$, em que cada linha corresponde a um vetor x^i e cada coluna às suas características, que neste caso serão comprimento e largura das sépalas, medidos em centímetros.

[4]:
```
X = convert(Array,iris[:, 1:2])
p, n = size(X)
```

[4]:
```
(150, 2)
```

Por conseguinte, a SVM é uma técnica de aprendizagem supervisionada e portanto, a obtenção do classificador é feita com base num conjunto de dados de entrada para os quais há o prévio conhecimento da classe y_i a qual cada amostra x^i pertence. Assim, amostras da espécie setosa serão classificadas como 1, enquanto que amostras das espécies versicolor ou virginica serão classificadas em -1, ou seja, $y_i \in \{-1, 1\}$.

[5]:
```
iris_df = DataFrame(X);
iris_df.Y = [species == "setosa" ? 1.0 : -1.0 for species in iris[!,:
↪Species]];
iris_df.Especie = iris.Species
rename!(iris_df, Dict(:x1 => :Comprimento_sepala))
```

```
rename!(iris_df, Dict(:x2 => :Largura_sepala))
first(iris_df, 10);
```

Portanto, nosso objetivo é classificar o conjunto de dados Íris em setosa e não setosa com base nas características "comprimento de sépala" e "largura de sépala". O gráfico a seguir (Figura 18) representa tais dados em relação às duas características citadas. Através dele podemos ter uma noção acerca da separabilidade do conjunto Íris nesse contexto.

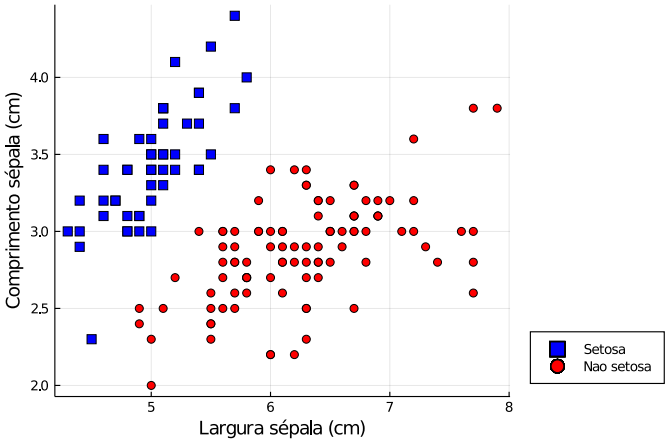


Figura 18 – Conjunto de dados Íris de acordo com comprimento e largura das sépalas.

Assim, no gráfico acima estão representados os 150 dados do conjunto Íris, em que os pontos em azul pertencem à espécie setosa, isto é, os vetores x^i tais que $y_i = 1$, e os pontos em vermelho às espécies versicolor e virginica, que correspondem aos vetores x^i tais que $y_i = -1$. Como os dados aparentam ser linearmente separáveis, aplicaremos primeiramente SVM com margem rígida.

A aprendizagem supervisionada é composta por dois momentos, a fase de treino e a fase de testes. Em vista disso, inicialmente dividimos, de maneira aleatória, o conjunto Íris em dois subconjuntos, o conjunto de treinamento (`train_set`), com 50 dados, e o conjunto de teste (`test_set`), com os 100 dados restantes. É através do conjunto de treinamento que a técnica SVM irá "aprender" a classificar os dados detectando padrões entre suas características e a espécie a qual pertencem. Já o conjunto de teste será utilizado para analisar a eficácia do classificador encontrado (hiperplano ótimo), averiguando se ele gera as saídas corretas para tais dados.

[7]:
```
Random.seed!(0)
trainsize = 50
train_set = sample(1:p,trainsize,replace=false,ordered=true)
test_set = setdiff(1:p,train_set)
Xtrain = X[train_set,:]
Ytrain = iris_df.Y[train_set]
Xtest = X[test_set,:]
Ytest = iris_df.Y[test_set]
ptrain = length(Ytrain)
iris_df.conjunto = fill("treino", p)
iris_df.conjunto[test_set] .= "teste";
```

Como abordado na Capítulo 4, a modelagem do problema de classificação utilizando a técnica SVM consiste em determinar o hiperplano que melhor separa os dados, classificando-os assim em duas classes. Ademais, o hiperplano ótimo $\mathcal{H}(w, b)$ é aquele que maximiza a margem que não contenha nenhum dado, ou seja, desejamos que os pontos x^i satisfaçam a seguinte restrição

$$y_i(w^T x^i + b) \geq 1, \quad i = 1, \ldots, 50.$$

Logo, o problema de encontrar o hiperplano ótimo $\mathcal{H}(w, b)$ é formulado

da seguinte forma

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}\|w\|^2 \\ \text{s.a.} \quad & y_i(w^T x^i + b) \geq 1, \quad i = 1, \ldots, 50, \end{aligned}$$

em que, neste exemplo, $w \in \mathbb{R}^2$ e $b \in \mathbb{R}$. É importante salientar que como neste caso os dados pertencem ao $\mathbb{R}^2$ o hiperplano ótimo será uma reta.

Assim, utilizando o modelo matemático formulado no capítulo anterior, criamos a função `SVM_rigida`, a qual adapta o problema de otimização formulado em (70) para os dados do problema de classificação que desejamos resolver e, com o auxílio do pacote `Ipopt` [26] para resolver este problema, determina o hiperplano ótimo.

[8]:
```
function SVM_rigida(n, ptrain, Xtrain, Ytrain)
    model = Model(optimizer_with_attributes(Ipopt.Optimizer,␣
↪"print_level"=>0))

    @variable(model, w[1:n]) # Aqui declaramos as variáveis.
    @variable(model, b)

    @objective(model, Min, dot(w, w) / 2) # Esta é a função objetivo.

    @constraint(model, [i=1:ptrain], Ytrain[i] * (dot(w, Xtrain[i,:]) + b) ␣
↪1) # Esta é a restrição.

    optimize!(model)

    w, b = value.(w), value.(b) # Com este comando queremos que os valores␣
↪ótimos sejam apresentados.
    return w, b
end
w, b = SVM_rigida(n, ptrain, Xtrain, Ytrain)
```

```
[8]: ([-2.857142829807275, 3.3333333011819755], 4.99999995276859)
```

Assim, com o auxílio da função `SVM_rigida`, encontramos os valores ótimos para w e b, os quais definem o hiperplano separador $\mathcal{H}(w, b)$, que é dado por

$$(-2.857142829807275, 3.333333301181975)^T x + 4.99999995276859 = 0,$$

com $x \in \mathbb{R}^2$.

Para melhor visualizar a classificação dos dados de treinamento vamos representá-los graficamente a seguir junto ao hiperplano ótimo.

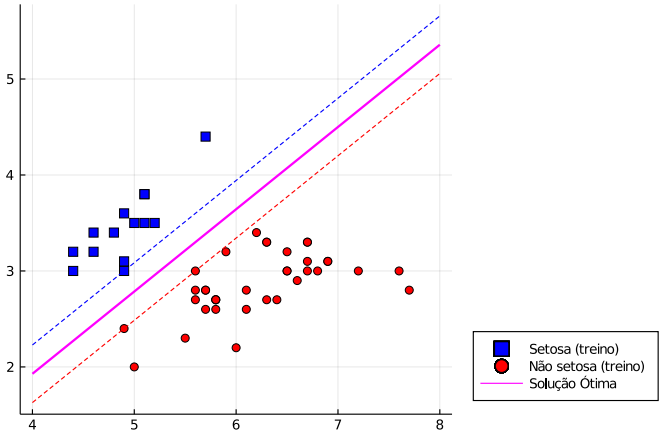


Figura 19 – Separação dos dados de treinamento pelo hiperplano ótimo.

Portanto, no gráfico da Figura 19 a reta rosa corresponde ao hiperplano ótimo e as retas tracejadas são os hiperplanos que delimitam a máxima margem possível. Observe que o conjunto de dados de treinamento é linearmente separável, pois o hiperplano ótimo os está separando corretamente. Além disso, de

acordo com a Definição 4.9 da Seção 4.2, temos que os vetores que estão sobre os hiperplanos da margem são os *vetores suporte*. Observe que tais vetores dão suporte ao hiperplano ótimo, de modo que todos os demais vetores poderiam ser descartados sem alterá-lo.

Agora, tendo determinado o classificador (hiperplano ótimo), vamos para a fase de testes, na qual estamos interessados em analisar se o classificador encontrado é eficaz. Para tanto, vamos acrescentar ao gráfico anterior os dados do conjunto de teste e observar se o hiperplano encontrado também os separa corretamente.

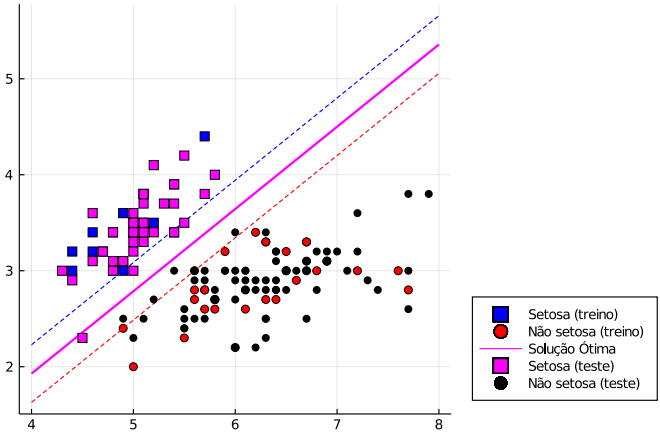


Figura 20 – Classificação do conjunto de teste pelo hiperplano ótimo.

Analisando o gráfico na Figura 20 percebe-se que há um ponto da espécie setosa do conjunto de teste que está localizado ligeiramente do lado direito do hiperplano ótimo, fazendo com que seja classificado incorretamente como não setosa. Assim, como determinar se o classificador encontrado é o melhor?

Para responder a esta pergunta é preciso medir a eficácia do modelo, comparando a real saída dos dados de teste com a respectiva classificação obtida pelo modelo. De modo geral, quanto mais classificações corretas o classificador predizer para o conjunto de teste, mais eficiente ele é.

Para analisarmos o desempenho do classificador utilizaremos a metodologia empregada por Krulikovski [15], que usa os conceitos de matriz de confusão e acurácia. A Matriz de Confusão, também denominada Matriz de Erro, consiste em uma medida de desempenho muito utilizada para fazer avaliações de modelos de classificação da aprendizagem de máquina supervisionada, como a SVM por exemplo. Ela é uma tabela que apresenta quatro combinações entre a classificação real e a prevista, o que nos permite analisar se a previsão sugerida pelo classificador encontrado ao implementar a SVM é condizente com a verdadeira classificação dos dados. Em síntese, a matriz de confusão apresenta as seguintes frequências: Verdadeiro Positivo (VP), Verdadeiro Negativo (VN), Falso Positivo (FP) e Falso Negativo (FN).

Vamos compreender essas terminologias com base na classificação dos dados Íris. Para o problema em que desejamos classificar amostras em setosa ($y_i = 1$) e não setosa ($y_i = -1$), temos que

- Verdadeiro Positivo (VP): quantidade de dados que são setosa ($y_i = 1$) e foram classificados como tal ($y_i = 1$);

- Verdadeiro Negativo (VN): se refere ao número de dados que não são setosa ($y_i = -1$) e foram classificados corretamente como não setosa ($y_i = -1$));

- Falso Positivo (FP): quantidade de dados não setosa ($y_i = -1$) classificados como setosa ($y_i = 1$);

- Falso Negativo (FN): quantidade de dados setosa ($y_i = 1$) que receberam classificação não setosa ($y_i = -1$).

Assim, a matriz de confusão permite observar a relação entre resultados falsos/verdadeiros e negativos/positivos, fornecendo na diagonal principal o número de acertos da classificação predita em relação a real classificação (VP e VN), enquanto os demais elementos correspondem aos erros na classificação (FP e FN). Vale observar que um classificador ideal apresentaria uma matriz de confusão com os elementos não pertencetes a diagonal principal iguais a zero, pois isso significaria que tal classificador não comete erros.

Ademais, usando os valores fornecidos pela matriz de confusão podemos calcular a acurácia do modelo, que fornece a porcentagem de dados positivos e negativos classificados corretamente. Ela é dada pela seguinte fórmula

$$\text{Acurácia} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{FP} + \text{VN} + \text{FN}}.$$

Note que, quanto mais próxima de 1 for a acurácia, maior é a quantidade de dados classificados corretamente.

Portanto, para avaliarmos o classificador encontrado, vamos determinar a matriz de confusão para este caso e calcular sua acurácia.

[11]:
```
Setosa = findall(Xtest*w .+ b .>= 0)
Nonsetosa = findall(Xtest*w .+ b .< 0)

gdf = filter(:conjunto => x -> x == "teste", iris_df)
gdf.Ypredito = fill(1., p - ptrain)
gdf[Nonsetosa, :Ypredito] .= -1

function MatrizConfusao(gdf)
    Verdadeiro_positivo = nrow(filter([:Y, :Ypredito] => (x,y) -> x == 1. &&␣
↪y == 1., gdf))
```

```
    Verdadeiro_negativo = nrow(filter([:Y, :Ypredito] => (x,y) -> x == -1. &&␣
↪y == -1., gdf))
    Falso_positivo = nrow(filter([:Y, :Ypredito] => (x,y) -> x == -1. && y ==␣
↪1., gdf))
    Falso_negativo = nrow(filter([:Y, :Ypredito] => (x,y) -> x == 1. && y ==␣
↪-1., gdf))

    VP, VN, FP, FN = Verdadeiro_positivo, Verdadeiro_negativo,␣
↪Falso_positivo, Falso_negativo
    acuracia = (VP + VN)/(VP + FP + VN + FN)
    matrizconfusao = DataFrame(Classe = ["Real Positiva", "Real Negativa"],
                            Predita_Positiva = [VP , FP],
                            Predita_Negativa = [FN, VN]
                            )
        return  acuracia, matrizconfusao
end
Acuracia_Iris, MatrizConfusao_Iris = MatrizConfusao(gdf);
@show Acuracia_Iris
MatrizConfusao_Iris
```

```
Acuracia_Iris = 0.99
```

[11]:

Classe	Predita_Positiva	Predita_Negativa
Real Positiva	35	1
Real Negativa	0	64

A tabela acima corresponde à matriz de confusão para o hiperplano ótimo encontrado, em que dos 100 dados do conjunto de teste, o número de VP = 35, VN = 64, FN = 1 e FP = 0. Ou seja, um dado da espécie setosa foi classificado incorretamente pelo hiperplano ótimo como pertencente às espécies versicolor ou virginica, implicando numa acurácia de 99%.

É possível visualizar este dado no gráfico apresentado anteriormente e como ele é único analisamos se tal classificação incorreta se repetia nos casos em que a quantidade de dados do conjunto de treinamento fosse maior, averiguando assim se uma maior quantidade de dados para treinar o modelo resultaria numa solução mais eficiente e que classificasse todos os dados de maneira correta. Contudo, o que se observou é que mesmo para um conjunto de treino com 100 dados, por exemplo, ainda ocorria uma classificação incorreta, pois os dados são distribuídos de maneira aleatória entre os conjuntos de treino e teste, de modo que tal controle não seja possível. Ademais, é importante que o conjunto de teste não seja tão pequeno em relação ao conjunto de treino para que não ocorram equívocos na avaliação do classificador.

Em vista disso, aplicamos então a técnica SVM com margem flexível (CSVM) para classificar esse mesmo conjunto de dados Íris e analisar se dessa forma seria possível obter um classificador com uma acurácia de 100% para o conjunto de teste.

Neste caso, de acordo com a teoria desenvolvida na Seção 4.3, o problema de classificação com margem flexível tem o seguinte formato

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{50}\xi_i \\ \text{s.a.} \quad & y_i(w^T x^i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, 50, \\ & \xi_i \geq 0, \qquad\qquad\qquad\quad i = 1, \ldots, 50, \end{aligned}$$

em que $w \in \mathbb{R}^2$, $b \in \mathbb{R}$, $\xi \in \mathbb{R}^{50}$ e $C > 0$.

De maneira análoga à função `SVM_rigida`, definimos então a função `SVM_flexivel` com base no problema acima para aplicar aos problemas que exigem a técnica SVM com margem flexível.

[12]:
```
function SVM_flexivel(n, ptrain, Xtrain, Ytrain, C = 1.0)
    model = Model(optimizer_with_attributes(Ipopt.Optimizer,␣
↪"print_level"=>0))

    @variable(model, w[1:n]) # Aqui declaramos as variáveis.
    @variable(model, b)
    @variable(model, [1:ptrain]  0)

    @objective(model, Min, dot(w, w) / 2 + C * sum()) # Esta é a função␣
↪objetivo.

    @constraint(model, [i=1:ptrain], Ytrain[i] * (dot(w, Xtrain[i,:]) + b)  ␣
↪1- [i]) # Esta é a restrição.

    # print(model)
    optimize!(model)

    w, b,  = value.(w), value.(b), value.() #aqui queremos desenhar os valores␣
↪ótimos.
        return w, b,
end
w, b,  = SVM_flexivel(n, ptrain, Xtrain, Ytrain, 1.0)
```

[12]:
```
([-1.8054054084206983, 1.567567554887955], 4.721081137153697, [0.0,
0.42270269035295177, 0.0, 0.0, 0.0, 0.0, 0.1805405354351823, 0.
↪2659459348641559,
0.0, 0.0    0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0])
```

Portanto, adaptando nosso modelo obtemos uma nova solução que é apresentada acima. Os valores ótimos encontrados determinaram um novo hiperplano classificador, o qual está representado na cor rosa no gráfico da Figura 21.

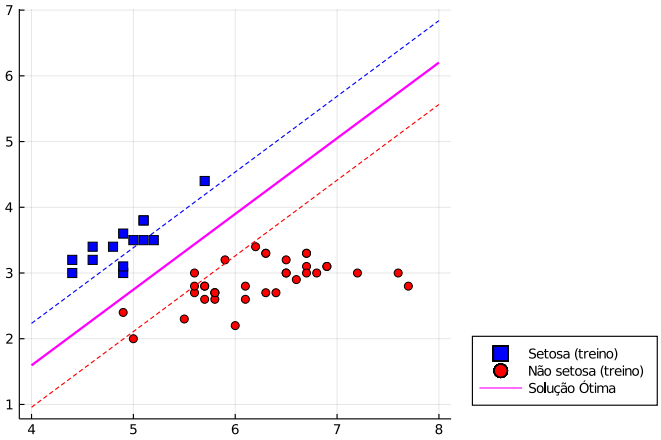


Figura 21 – Hiperplano ótimo determinado pela técnica SVM com margem flexível.

Fazendo uma comparação entre este hiperplano ótimo e aquele determinado no problema em que aplicamos SVM de margem rígida, percebemos que no caso atual os vetores que apresentam $\xi_i > 0$ possuem uma maior "liberdade", de modo que possam estar localizados na região entre as margens e o hiperplano separador. Tal "liberdade" é fruto do relaxamento promovido nas restrições através das variáveis de folga, relaxamento que deve ser regulado. É neste contexto que entra o parâmetro C. De acordo com Krulikovski [15, p. 76], tal parâmetro nos fornece um "(...) equilíbrio entre a maximização da margem e a minimização do erro de classificação". Ou seja, caso o valor atribuído ao parâmetro de penalização C seja pequeno, uma maior quantidade de vetores recebe folga, inclusive alguns para os quais não seria necessário. Caso contrário, se valores muito altos são atribuídos ao parâmetro C, o número de vetores que recebe folga diminui. Contudo, nesse último caso o programa tende a se con-

centrar em minimizar a penalização em vez de maximizar a margem na função objetivo. Em decorrência disso, é de suma importância escolher o valor correto para o parâmetro C.

Assim, ao definir a função `SVM_flexivel` determinamos como padrão $C = 1$, pois como veremos na fase de testes tal valor possibilitou uma boa classificação para o problema atual. No entanto, na próxima seção analisaremos os classificadores obtidos para diferentes valores de C.

Novamente, após determinado o classificador, é preciso avaliar sua eficiência. Portanto, utilizando o conjunto de teste construímos a matriz de confusão e calculamos a acurácia do novo hiperplano encontrado.

[14]:
```
Setosa = findall(Xtest*w .+ b .>= 0)
Nonsetosa = findall(Xtest*w .+ b .< 0)
gdf = filter(:conjunto => x -> x == "teste", iris_df) #filter é uma função.␣
↪Neste caso, ela foi utilizada para "filtrar"/selecionar em um dataframe que␣
↪denominamos por "gdf" todos os dados x que possuiam "teste" na coluna␣
↪"conjunto". Assim, conseguimos analisar a predição do conjunto de teste
gdf.Ypredito = fill(1., p - ptrain) #fill é uma função. Neste caso ela␣
↪preencheu a coluna Ypredito com 1.
gdf[Nonsetosa, :Ypredito] .= -1
Acuracia_Iris, MatrizConfusao_Iris = MatrizConfusao(gdf);
@show Acuracia_Iris
MatrizConfusao_Iris
```

```
Acuracia_Iris = 1.0
```

[14]:

Classe	Predita_Positiva	Predita_Negativa
Real Positiva	36	0
Real Negativa	0	64

Analisando a matriz de confusão temos que os valores de FN e FP são nulos, o que nos permite concluir que o hiperplano ótimo encontrado pela téc-

nica SVM com margem flexível atua como um bom classificador, apresentando uma acurácia igual a 1 e, portanto, 100% de acerto na classificação dos dados de teste em setosa e não setosa. De fato, no gráfico a seguir podemos visualizar que o hiperplano ótimo separa todos os dados dos conjuntos de treino e teste corretamente.

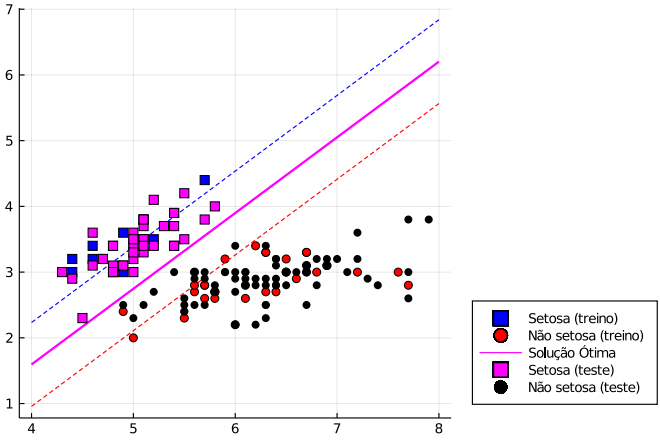


Figura 22 – Classificação do conjunto de teste pelo hiperplano ótimo.

Portanto, utilizando a técnica CSVM, com $C = 1$, foi possível determinar um hiperplano separador com uma acurácia de 100% para o conjunto de teste.

5.1.2 Classificação com quatro características

Nesta seção vamos apresentar a implementação da técnica SVM com margem rígida para classificação do conjunto de dados Íris em setosa e não setosa levando em consideração quatro características: comprimento e largura das sépalas e pétalas.

De modo análogo ao exemplo anterior, os dados pertencentes à espécie setosa serão classificados como 1, e os demais, pertencentes às espécies versicolor ou virginica, classificados como -1. A principal diferença agora é que os dados do conjunto de entrada pertencem ao $\mathbb{R}^4$.

Analisando os gráficos da Figura 17 temos que os dados setosa (em azul) e não setosa (em verde e vermelho) aparentam ser linearmente separáveis. Guiados por essa perspectiva, vamos aplicar a técnica SVM com margem rígida para determinar o classificador.

Inicialmente, assim como no problema anterior, é necessário separar os dados Íris em conjunto de treinamento (`train_set`), com 50 dados, e conjunto de teste (`test_set`), com 100 dados.

[18]:
```
Random.seed!(0)
trainsize = 50
train_set = sample(1:p,trainsize,replace=false,ordered=true)
test_set = setdiff(1:p,train_set)
Xtrain = X[train_set,:]
Ytrain = iris_df.Y[train_set]
Xtest = X[test_set,:]
Ytest = iris_df.Y[test_set]
ptrain = length(Ytrain)
iris_df.conjunto = fill("treino", p)
iris_df.conjunto[test_set] .= "teste"
iris_df;
```

Tendo definido os conjuntos de treino e teste, vamos aplicar a técnica SVM com margem rígida sobre o conjunto de treino para obter o hiperplano separador. Para tanto, utilizaremos a função `SVM_rigida` definida no exemplo anterior.

[19]:
```
w, b = SVM_rigida(n, ptrain, Xtrain, Ytrain)
```

[19]:
```
([-1.9124232444455314e-7, 0.3203662687382003, -0.8237985755072654,
-0.3661327030904415], 1.315789899955701)
```

Para avaliar a eficiência do classificador encontrado, isto é, se ele separa os dados corretamente, tomamos o conjunto de teste, para o qual já conhecemos sua real classificação, e comparamos com a classificação predita pelo hiperplano separador. Os resultados desta comparação podem ser visualizados na matriz de confusão dada a seguir e a taxa de acerto é dada pela acurácia.

[20]:
```
Setosa = findall(Xtest*w .+ b .>= 0)
Nonsetosa = findall(Xtest*w .+ b .< 0)
gdf = filter(:conjunto => x -> x == "teste", iris_df)
gdf.Ypredito = fill(1., p - ptrain)
gdf[Nonsetosa, :Ypredito] .= -1
Acuracia_Iris, MatrizConfusao_Iris = MatrizConfusao(gdf);
@show Acuracia_Iris
MatrizConfusao_Iris
```

```
Acuracia_Iris = 1.0
```

[20]:

Classe	Predita_Positiva	Predita_Negativa
Real Positiva	36	0
Real Negativa	0	64

Assim, como os valores da diagonal secundária são todos nulos temos que o hiperplano ótimo separou todos os dados do conjunto de teste corretamente, apresentando uma acurácia de 100%.

Portanto, para o problema que considera as quatro características dos dados amostrais, temos que a técnica SVM com margem rígida nos fornece um hiperplano que atua como um bom classificador, separando corretamente todos os dados, tanto de treinamento quanto de teste.

5.1.3 Classificação em espécie virginica e não virginica utilizando quatro características

Para finalizar os experimentos numéricos com o conjunto de dados Íris, vamos implementar neste momento a técnica SVM com margem flexível para classificar tal conjunto de dados em virginica e não virginica.

Assim como no exemplo anterior, serão consideradas as características comprimento e largura das sépalas e pétalas, de modo que cada vetor $x^i \in \mathbb{R}^4$. Mas agora, dados da espécie virginica serão classificados com $y_i = 1$ e dados das espécies que não são virginica com $y_i = -1$.

Observando novamente os gráficos apresentados na Figura 17, é possível perceber que alguns dados pertencentes às duas classes distintas (virginica na cor verde e não virginica nas cores azul e vermelha) acabam se sobrepondo. Logo, intuímos que eles não são linearmente separáveis. Em vista disso, aplicaremos a técnica SVM com margem flexível para obter o classificador.

De modo análogo ao que foi desenvolvido anteriormente, é necessário primeiramente dividir aleatoriamente o conjunto total de dados em dois subconjuntos: conjunto de treinamento e conjunto de testes. Neste caso, o conjunto de treinamento será composto por 75 dados e o conjunto de teste pelos 75 dados restantes.

[23]:
```
Random.seed!(0)
trainsize = 75
train_set = sample(1:p,trainsize,replace=false,ordered=true)
test_set = setdiff(1:p,train_set)
Xtrain = X[train_set,:]
Ytrain = iris_df.Y[train_set]
Xtest = X[test_set,:]
Ytest = iris_df.Y[test_set]
ptrain = length(Ytrain)
iris_df.conjunto = fill("treino", p)
```

```
iris_df.conjunto[test_set] .= "teste"
iris_df;
```

Agora, para obter o hiperplano ótimo aplicamos a função `SVM_flexivel` com parâmetro $C = 1$.

[24]:
```
w, b,  = SVM_flexivel(n, ptrain, Xtrain, Ytrain)
```

[24]:
```
([-0.3456754552444701, -0.8049690813571279, 1.7254845918550477,
1.6080094600267831], -6.676855229902285, [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
0.0, 0.0, 0.0    0.0, 0.0, 0.0, 0.8965384097436062, 0.0, 4.754013257179834e-9,
0.0, 0.0, 0.0, 0.0])
```

Assim, a solução acima nos fornece o hiperplano ótimo. Observe também que, de acordo com a solução ótima encontrada, algumas variáveis de folga ξ_i assumiram valores não-nulos.

Para verificarmos se tal hiperplano atua como um bom classificador recorremos à fase de testes. Em vista disso, apresentamos a seguir a matriz de confusão e a acurácia do hiperplano ótimo encontrado ao classificar o conjunto de teste.

[25]:
```
Virginica = findall(Xtest*w .+ b .>= 0)
Nonvirginica = findall(Xtest*w .+ b .< 0)
gdf = filter(:conjunto => x -> x == "teste", iris_df)
gdf.Ypredito = fill(1., p - ptrain)
gdf[Nonvirginica, :Ypredito] .= -1
Acuracia_Iris, MatrizConfusao_Iris = MatrizConfusao(gdf);
@show Acuracia_Iris
MatrizConfusao_Iris
```

```
Acuracia_Iris = 1.0
```

[25]:

Classe	Predita_Positiva	Predita_Negativa
Real Positiva	23	0
Real Negativa	0	52

Felizmente, como a diagonal secundária apresenta somente valores nulos, temos que não ocorreram Falsos Positivos e Falsos Negativos para a classificação do conjunto de teste, resultando uma acurácia de 100%. Logo, o hiperplano ótimo encontrado classifica corretamente os dados do conjunto de treino e de teste, se configurando num ótimo classificador.

5.2 IMPLEMENTAÇÃO DE SVM PARA CLASSIFICAÇÃO DE DADOS DE CÂNCER DE MAMA

De acordo com o Instituto Nacional de Câncer (INCA, [6]),

> Câncer é o nome dado a um conjunto de mais de 100 doenças que têm em comum o crescimento desordenado de células, que invadem tecidos e órgãos. Dividindo-se rapidamente, estas células tendem a ser muito agressivas e incontroláveis, determinando a formação de tumores, que podem espalhar-se para outras regiões do corpo.

O câncer de mama por sua vez ocorre quando há formação de tumor na mama, sendo o tipo de câncer mais incidente entre as mulheres, tanto no Brasil quanto no mundo. Segundo estimativas [6, 13, 20], no mundo, somente em 2018, foram registrados cerca de 2,1 milhões de novos casos, número que equivale a 11,6% de todos os cânceres estimados. No Brasil, em 2017, ocorreram 16.724 óbitos por câncer de mama e, segundo estimativas do INCA [6], são estimados cerca de 66.280 novos casos de câncer de mama no Brasil para cada ano do triênio 2020-2022, o que corresponde a uma taxa de incidência de 61,61 novos casos a cada 100 mil mulheres.

O diagnóstico precoce é um dos principais fatores que contribuem para reduzir a mortalidade por câncer, possibilitando cerca de 95% de chances de cura [23]. Neste contexto,

> A mamografia é uma das melhores técnicas para o rastreamento do câncer de mama disponível atualmente, capaz de registrar imagens da mama com a finalidade de diagnosticar a presença ou ausência de estruturas que possam indicar a doença. Com esse tipo de exame pode-se detectar o tumor antes que ele se torne palpável. (Silva et al. [23, p. 229])

Tendo em vista as altas taxas de incidência e mortes causadas pelo câncer de mama, muitas pesquisas científicas vem sendo desenvolvidas nos últimos anos com o intuito de auxiliar no diagnóstico de doenças, tornando as técnicas de aprendizagem de máquina cada vez mais presentes na área médica. Assim, visto que este tema é de grande relevância e interesse, o próximo experimento numérico será realizado com um conjunto de dados sobre células de câncer de mama retirados de *UC Irvine Machine Learning Repository* [9]. Nosso objetivo será utilizar a técnica SVM com margem flexível para classificar estes dados em tumores malignos ou benignos.

Este conjunto é composto por 569 dados, em que cada um possui um número de identidade (ID), sua classificação em tumor maligno (M) ou benigno (B) e 30 características acerca de núcleos celulares presentes em imagens digitalizadas de um aspirado por agulha fina (do inglês *fine needle aspirate*, FNA), de uma massa mamária [1, 27]. As características que compõem os dados são:

- Raio (média das distâncias do centro aos pontos do perímetro);
- Textura (desvio padrão dos valores da escala cinza);
- Perímetro;
- Área;

- Suavidade (variação local nos comprimentos do raio);
- Compactação $\left(\frac{\text{perímetro}^2}{\text{área} - 1.0}\right)$;
- Concavidade (severidade das porções côncavas do contorno);
- Pontos côncavos (número de partes côncavas do contorno);
- Simetria;
- Dimensão fractal ("aproximação da costa" -1).

Para cada imagem os criadores deste conjunto de dados calcularam a média, o desvio padrão e a média dos três maiores números dos valores atribuídos a cada um dos atributos acima mencionados, resultando em 30 características, algumas das quais são apresentadas na tabela a seguir.

[26]:
```
col_headers = ["id" ,"diagnosis" ,"radius_mean" ,"texture_mean"␣
↪,"perimeter_mean" ,"area_mean" ,"smoothness_mean" ,"compactness_mean"␣
↪,"concavity_mean" ,"concave points_mean" ,"symmetry_mean"␣
↪,"fractal_dimension_mean" ,"radius_se" ,"texture_se" ,"perimeter_se"␣
↪,"area_se" ,"smoothness_se" ,"compactness_se" ,"concavity_se" ,"concave␣
↪points_se" ,"symmetry_se" ,"fractal_dimension_se" ,"radius_worst"␣
↪,"texture_worst" ,"perimeter_worst" ,"area_worst" ,"smoothness_worst"␣
↪,"compactness_worst" ,"concavity_worst" ,"concave points_worst"␣
↪,"symmetry_worst" ,"fractal_dimension_worst"]
cancer_df = CSV.read("cancer_data.csv", header = col_headers,)
cancer_df.Y = [diagnosis == "M" ? 1.0 : -1.0 for diagnosis in cancer_df[!,:
↪diagnosis]];
first(cancer_df,10)
```

[26]:

Tabela 1 – Dados de câncer.

id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean
842302	M	17,99	10,38	122,8	1001,0	0,1184
842517	M	20,57	17,77	132,9	1326,0	0,08474
843 00903	M	19,69	21,25	130,0	1203,0	0,1096
84348301	M	11,42	20,38	77,58	386,1	0,1425
84358402	M	20,29	14,34	135,1	1297,0	0,1003
843786	M	12,45	15,7	82,57	477,1	0,1278
844359	M	18,25	19,98	119,6	1040,0	0,09463
84458202	M	13,71	20,83	90,2	577,9	0,1189
844981	M	13,0	21,82	87,5	519,8	0,1273
84501001	M	12,46	24,04	83,97	475,9	0,1186

Na Figura 23, temos a representação gráfica destes dados considerando-se algumas das características mencionadas. Nesses gráficos, os pontos em azul correspondem aos dados que pertencem à classe benigna e os dados em vermelho à classe maligna.

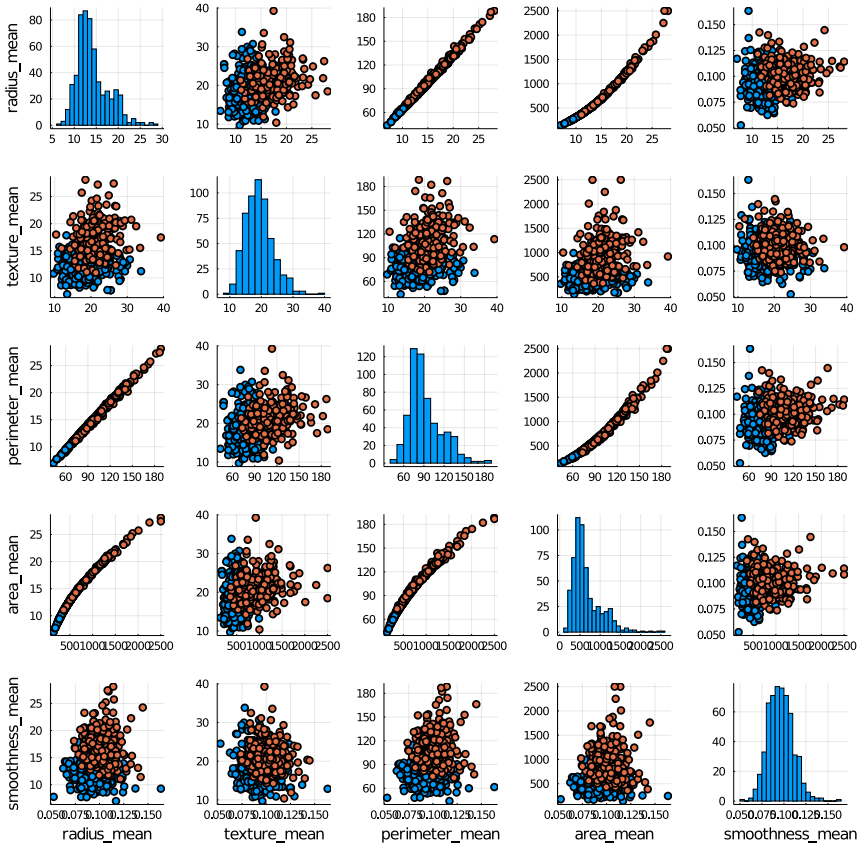


Figura 23 – Conjunto de dados de células de câncer de mama.

Neste exemplo, cada dado x^i do conjunto de entrada pertence ao $\mathbb{R}^{30}$, pois são 30 as características. Ademais, os dados diagnosticados como malignos (M) serão classificados como $y_i = 1$, e os dados com diagnóstico benigno serão classificados com $y_i = -1$.

Com base nos gráficos apresentados na Figura 23, em que apenas duas características foram consideradas em cada vez, é possível intuir que os dados não são linearmente separáveis. Em vista disso, aplicamos a técnica SVM com margem flexível para classificar os dados deste conjunto em câncer Maligno ou Benigno e então discutir os resultados encontrados.

Dos 569 dados do conjunto de entrada, 150 dados foram direcionados ao conjunto de treinamento e os 419 restantes ao conjunto de teste.

[29]:
```
Random.seed!(0)
trainsize = 150
train_set = sample(1:p,trainsize,replace=false,ordered=true)
test_set = setdiff(1:p,train_set)
Xtrain = X[train_set,:]
Ytrain = cancer_df.Y[train_set]
Xtest = X[test_set,:]
Ytest = cancer_df.Y[test_set]
ptrain = length(Ytrain)
cancer_df.conjunto = fill("treino", p)
cancer_df.conjunto[test_set] .= "teste"
cancer_df;
```

Utilizando a função `SVM_flexivel` criada com base na modelagem matemática desenvolvida na Seção 4.3, determinamos o hiperplano ótimo para os seguintes valores para o parâmetro C: $10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3$ e 10^4. Para cada solução encontrada, apresentamos a seguir a matriz de confusão e a acurácia da classificação do conjunto de teste.

[30]:
```
w, b,  = SVM_flexivel(n, ptrain, Xtrain, Ytrain, 10e-4)
```

[31]:
```
Maligno = findall(Xtest*w .+ b .>= 0)
Benigno = findall(Xtest*w .+ b .< 0)
gdf = filter(:conjunto => x -> x == "teste", cancer_df)
gdf.Ypredito = fill(1., p - ptrain)
gdf[Benigno, :Ypredito] .= -1
Acuracia_Cancer1, MatrizConfusao_Cancer = MatrizConfusao(gdf);
@show Acuracia_Cancer1
MatrizConfusao_Cancer
```

```
Acuracia_Cancer1 = 0.9284009546539379
```

[31]:

Classe	Predita_Positiva	Predita_Negativa
Real Positiva	134	26
Real Negativa	4	255

[32]:
```
w, b,  = SVM_flexivel(n, ptrain, Xtrain, Ytrain, 10e-3)
```

[33]:
```
Maligno = findall(Xtest*w .+ b .>= 0)
Benigno = findall(Xtest*w .+ b .< 0)
gdf = filter(:conjunto => x -> x == "teste", cancer_df)
gdf.Ypredito = fill(1., p - ptrain)
gdf[Benigno, :Ypredito] .= -1
Acuracia_Cancer2, MatrizConfusao_Cancer = MatrizConfusao(gdf);
@show Acuracia_Cancer2
MatrizConfusao_Cancer
```

```
Acuracia_Cancer2 = 0.9451073985680191
```

[33]:

Classe	Predita_Positiva	Predita_Negativa
Real Positiva	143	17
Real Negativa	6	253

[34]:
```
w, b,  = SVM_flexivel(n, ptrain, Xtrain, Ytrain, 10e-2)
```

[35]:
```
Maligno = findall(Xtest*w .+ b .>= 0)
Benigno = findall(Xtest*w .+ b .< 0)
gdf = filter(:conjunto => x -> x == "teste", cancer_df)
gdf.Ypredito = fill(1., p - ptrain)
gdf[Benigno, :Ypredito] .= -1
Acuracia_Cancer3, MatrizConfusao_Cancer = MatrizConfusao(gdf);
@show Acuracia_Cancer3
MatrizConfusao_Cancer
```

```
Acuracia_Cancer3 = 0.9474940334128878
```

[35]:

Classe	Predita_Positiva	Predita_Negativa
Real Positiva	144	16
Real Negativa	6	253

[36]:
```
w, b,  = SVM_flexivel(n, ptrain, Xtrain, Ytrain, 1.0)
```

[37]:
```
Maligno = findall(Xtest*w .+ b .>= 0)
Benigno = findall(Xtest*w .+ b .< 0)
gdf = filter(:conjunto => x -> x == "teste", cancer_df)
gdf.Ypredito = fill(1., p - ptrain)
gdf[Benigno, :Ypredito] .= -1
Acuracia_Cancer4, MatrizConfusao_Cancer = MatrizConfusao(gdf);
@show Acuracia_Cancer4
MatrizConfusao_Cancer
```

```
Acuracia_Cancer4 = 0.954653937947494
```

[37]:

Classe	Predita_Positiva	Predita_Negativa
Real Positiva	151	9
Real Negativa	10	249

[38]:
```
w, b,  = SVM_flexivel(n, ptrain, Xtrain, Ytrain, 10)
```

[39]:
```
Maligno = findall(Xtest*w .+ b .>= 0)
Benigno = findall(Xtest*w .+ b .< 0)
gdf = filter(:conjunto => x -> x == "teste", cancer_df)
gdf.Ypredito = fill(1., p - ptrain)
gdf[Benigno, :Ypredito] .= -1
Acuracia_Cancer5, MatrizConfusao_Cancer = MatrizConfusao(gdf);
@show Acuracia_Cancer5
MatrizConfusao_Cancer
```

```
Acuracia_Cancer5 = 0.9618138424821002
```

[39]:

Classe	Predita_Positiva	Predita_Negativa
Real Positiva	153	7
Real Negativa	9	250

[40]:
```
w, b,  = SVM_flexivel(n, ptrain, Xtrain, Ytrain, 10e1)
```

[42]:
```
Maligno = findall(Xtest*w .+ b .>= 0)
Benigno = findall(Xtest*w .+ b .< 0)
gdf = filter(:conjunto => x -> x == "teste", cancer_df)
gdf.Ypredito = fill(1., p - ptrain)
gdf[Benigno, :Ypredito] .= -1
Acuracia_Cancer6, MatrizConfusao_Cancer = MatrizConfusao(gdf);
@show Acuracia_Cancer6
```

```
MatrizConfusao_Cancer
```

```
Acuracia_Cancer6 = 0.954653937947494
```

[42]:

Classe	Predita_Positiva	Predita_Negativa
Real Positiva	150	10
Real Negativa	9	250

[43]:
```
w, b,  = SVM_flexivel(n, ptrain, Xtrain, Ytrain, 10e2)
```

[45]:
```
Maligno = findall(Xtest*w .+ b .>= 0)
Benigno = findall(Xtest*w .+ b .< 0)
gdf = filter(:conjunto => x -> x == "teste", cancer_df)
gdf.Ypredito = fill(1., p - ptrain)
gdf[Benigno, :Ypredito] .= -1
Acuracia_Cancer7, MatrizConfusao_Cancer = MatrizConfusao(gdf);
@show Acuracia_Cancer7
MatrizConfusao_Cancer
```

```
Acuracia_Cancer7 = 0.9522673031026253
```

[45]:

Classe	Predita_Positiva	Predita_Negativa
Real Positiva	149	11
Real Negativa	9	250

[46]:
```
w, b,  = SVM_flexivel(n, ptrain, Xtrain, Ytrain, 10e3)
```

[48]:
```
Maligno = findall(Xtest*w .+ b .>= 0)
Benigno = findall(Xtest*w .+ b .< 0)
gdf = filter(:conjunto => x -> x == "teste", cancer_df)
gdf.Ypredito = fill(1., p - ptrain)
```

```
gdf[Benigno, :Ypredito] .= -1
Acuracia_Cancer8, MatrizConfusao_Cancer = MatrizConfusao(gdf);
@show Acuracia_Cancer8
MatrizConfusao_Cancer
```

```
Acuracia_Cancer8 = 0.9522673031026253
```

[48]:

Classe	Predita_Positiva	Predita_Negativa
Real Positiva	149	11
Real Negativa	9	250

[49]:

```
Resultados = DataFrame(C = [10e-4 , 10e-3, 10e-2 , 1, 10, 10e1, 10e2, 10e3],
                       Acurácia = [Acuracia_Cancer1, Acuracia_Cancer2,␣
↪Acuracia_Cancer3,
                                   Acuracia_Cancer4, Acuracia_Cancer5,␣
↪Acuracia_Cancer6,
                                   Acuracia_Cancer7, Acuracia_Cancer8]
                       )
```

[49]:

Tabela 2 – Parâmetro C e respectiva acurácia.

C	Acurácia
0,001	0,928401
0,01	0,945107
0,1	0,947494
1,0	0,954654
10,0	0,961814
100,0	0,954654
1000,0	0,952267
10000,0	0,952267

Para facilitar a comparação entre as acurácias calculadas para cada classificador encontrado ao atribuir diferentes valores ao parâmetro C, construímos a Tabela 2 acima, que apresenta na primeira coluna os diferentes valores atribuídos ao parâmetro C e, na segunda coluna, a respectiva acurácia na classificação dos dados de teste pelo hiperplano ótimo encontrado em cada caso.

Observe que, conforme diferentes valores são atribuídos ao parâmetro C a solução ótima também varia, porém nenhuma delas apresenta 100% de acerto na classificação dos dados de teste. Em vista disso, para escolhermos o melhor hiperplano classificador dentre os hiperplanos calculados, é necessário analisar qual é o mais eficiente. Para realizar essa escolha é preciso compreender que no contexto do nosso problema o classificador que desejamos obter fornecerá o diagnóstico de câncer de mama em maligno ou benigno. Desse modo, além de possuir uma alta acurácia, é de suma importância que o classificador apresente uma baixa quantidade de falsos negativos, pois neste caso uma pessoa com câncer maligno receberia um diagnóstico benigno, comprometendo seu tratamento médico.

Primeiramente, analisando as soluções ótimas encontradas, notamos que conforme o parâmetro C aumenta, a quantidade de variáveis de folga não-nulas diminui. Isso também implica que quanto maior o parâmetro C, maior é a penalização sobre as variáveis de folga e, portanto, menor a margem de separação. Tanto é que a partir de C próximo a 100 as variáveis de folga passam a ser todas nulas, pois neste caso obtemos $\|\xi\| = 0$.

Por conseguinte, com base na Tabela 2, os parâmetros $C = 10^3$ e $C = 10^4$ apresentam a mesma acurácia de 95,22%, sugerindo que as soluções encontradas são muito próximas. Com efeito, calculando a norma da diferença entre a solução w encontrada para cada um desses parâmetros obtemos um valor muito pequeno.

Por fim, temos que o parâmetro $C = 10$ determina a solução com maior

acurácia, $96,18\%$, e menor quantidade de falsos negativos, que totalizam 7. Enquanto que para as demais soluções a acurácia varia entre $92,84\%$ e 95,46% e a quantidade de falsos negativos fica entre 9 e 26. Portanto, tomando $C = 10$ temos, dentre as escolhas de C, o classificador que fornece o melhor diagnóstico para os dados de câncer.

Analisando os experimentos numéricos realizados neste capítulo concluímos que a técnica SVM, tanto com margem rígida quanto com margem flexível, apresenta resultados satisfatórios na classificação de dados. A implementação da técnica para classificação do conjunto de dados Íris foi realizado com fins mais didáticos, através da qual foi possível compreender a atuação prática da técnica SVM e suas particularidades, além de visualizar as etapas (treinamento e teste) que são características das técnicas de aprendizagem supervisionada. Posteriormente, a aplicação da técnica SVM com margem flexível na classificação de dados de células de câncer de mama em tumor maligno e benigno consiste num problema prático. Nesse experimento, percebemos que a eficiência do classificador encontrado depende da escolha do parâmetro de penalização C, escolha esta que está relacionada à natureza do problema.

6 CONSIDERAÇÕES FINAIS

Este trabalho apresentou uma análise teórica matemática, do ponto de vista da otimização, da técnica Máquinas de Vetores Suporte (SVMs) aplicada à classificação binária de dados [8, 15].

Para tanto, vimos que a formulação matemática da técnica SVM para classificação se concentra em obter um hiperplano que melhor separa os dados em duas classes, de modo a possibilitar a máxima margem de separação. A partir disso, derivamos o problema de otimização com restrições lineares cuja solução fornece o hiperplano separador que atuará como classificador para novos dados. Essa abordagem é um dos fatores que torna a técnica SVM tão interessante, pois permite pensar a aprendizagem de máquina supervisionada baseada em argumentos geométricos, e não somente estatísticos como em outras técnicas.

Ademais, verificamos no Capítulo 4 que o problema empregado pela técnica SVM na fase de treinamento para determinar o hiperplano separador é um problema de programação quadrática convexa com restrições lineares. Tendo isso em vista, foi de suma importância as discussões do Capítulo 2 em relação a conceitos e resultados da teoria de otimização, apresentando o desenvolvimento das condições de otimalidade para problemas de otimização irrestrita e com restrições lineares, dentre as quais as condições de Karush-Kuhn-Tucker, por exemplo, bem como do Capítulo 3, em que tratamos de alguns elementos da teoria de otimização convexa, a qual fornece resultados importantes aos problemas de otimização, como a garantia de que minimizadores locais são globais.

Finalmente, no Capítulo 5, implementamos a técnica SVM para classificação binária do conjunto de dados de flores Íris [11] e de um conjunto de dados de células de câncer de mama [1, 9, 27]. Através destes experimentos numéricos concluímos que a técnica SVM apresenta, de fato, bons resultados na determinação de um classificador, de modo que foi possível atingir uma acurácia de

100% nos testes realizados com o conjunto Íris, e uma acurácia de 96,18% na classificação dos dados de câncer em tumor maligno ou benigno. Além disso, na implementação da técnica SVM para classificar os dados de câncer, que não são linearmente separáveis, constatamos que a eficiência do classificador está relacionada com a escolha adequada do parâmetro de penalização.

É importante pontuar que neste trabalho nos limitamos a formular somente o problema primal da técnica SVM, tanto de margem rígida quanto flexível. Assim, algumas sugestões para serem desenvolvidas em trabalhos futuros são realizar um estudo da teoria de dualidade para obter o dual dos problemas (70) e (76), assim como, desenvolver a formulação matemática da técnica SVM não-linear, que utiliza *Kernels* não-lineares para classificar conjuntos de dados não-linearmente separáveis nos casos em que a técnica com margem flexível não fornece um bom classificador.

REFERÊNCIAS

[1] Kristin P. Bennett e O. L. Mangasarian. “Robust Linear Programming Discrimination of Two Linearly Inseparable Sets”. Em: *Optimization Methods and Software* 1.1 (jan. de 1992), pp. 23–34. DOI: 10.1080/10556789208805504.

[2] D.P. Bertsekas. *Nonlinear Programming*. Athena scientific optimization and computation series. Athena Scientific, 2016. ISBN: 9781886529052. URL: https://books.google.com.br/books?id=TwOujgEACAAJ.

[3] Jeff Bezanson, Alan Edelman, Stefan Karpinski e Viral B Shah. “Julia: A Fresh Approach to Numerical Computing”. Em: *SIAM Rev.* 59.1 (fev. de 2017), pp. 65–98.

[4] C.M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer New York, 2016. ISBN: 9781493938438. URL: https://books.google.com.br/books?id=kOXDtAEACAAJ.

[5] Bernhard E. Boser, Isabelle M. Guyon e Vladimir N. Vapnik. “A Training Algorithm for Optimal Margin Classifiers”. Em: *Proceedings of the fifth annual workshop on Computational learning theory*. ACM, 1992, pp. 144–152.

[6] Instituto Nacional de Câncer (INCA). *Câncer de mama*. 2020. URL: https://www.inca.gov.br/tipos-de-cancer/cancer-de-mama.

[7] Corina Cortes e Vladimir Vapnik. “Support-Vector Networks”. Em: *Machine Learning. Springer* 20.3 (1995), pp. 273–297.

[8] Peter Deisenroth, A. Aldo Faisal e Cheng Soon Ong. *Mathematics for Machine Learning*. Boston: Cambridge University Press, 2019.

[9] Dheeru Dua e Casey Graff. *UCI Machine Learning Repository*. 2017. URL: `http://archive.ics.uci.edu/ml`.

[10] Iain Dunning, Joey Huchette e Miles Lubin. “JuMP: A Modeling Language for Mathematical Optimization”. Em: *SIAM Review* 59.2 (2017), pp. 295–320. DOI: `10.1137/15M1020575`.

[11] R. A. Fisher. “The Use of Multiple Measurements in Taxonomic Problems”. Em: *Annals of Eugenics* 7.2 (set. de 1936), pp. 179–188. DOI: `10.1111/j.1469-1809.1936.tb02137.x`.

[12] Ana Friedlander. *Elementos de Programação Não-Linear*. Unicamp, 1994.

[13] Federação Brasileira de Instituições Filantrópicas de Apoio à Saúde da Mama (FEMAMA). *O câncer de mama em números*. 2019. URL: `https://www.femama.org.br/site/br/noticia/o-cancer-de-mama-em-numeros`.

[14] Alexey Izmailov e Mikhail Solodov. *Otimização. Condições de Otimalidade. Elementos de Análise Convexa e de Dualidade*. 3ª ed. Vol. I. Rio de Janeiro: IMPA, 2014.

[15] Evelin Heringer Manoel Krulikovski. "Análise Teórica de Máquinas de Vetores Suporte e Aplicação a Classificação de Caracteres". Dissertação de Mestrado em Matemática. Universidade Federal do Paraná, 2017.

[16] Elon Lages Lima. *Álgebra Linear*. 8ª ed. Coleção Matemática Universitária. Rio de Janeiro: IMPA, 2014. ISBN: 9788524400896.

[17] Elon Lages Lima. *Curso de análise*. 15ª ed. Vol. 1. Coleção Projeto Euclides. Rio de Janeiro: IMPA, 2019. ISBN: 9788524404689.

[18] Ana Carolina Lorena e André C. P. L. F. de Carvalho. "Uma Introdução às Support Vector Machines". Em: *Revista de Informática Teórica e Aplicada* 14.2 (2007), pp. 43–67.

[19] D.G. Luenberger e Y. Ye. *Linear and Nonlinear Programming*. International Series in Operations Research & Management Science. Springer US, 2008. ISBN: 9780387745039. URL: https://books.google.com.br/books?id=EJTrgq79QWUC.

[20] Sociedade Brasileira de Mastologia. *INCA lança estimativa da incidência de câncer de mama no Brasil*. 2020. URL: https://www.sbmastologia.com.br/noticias/inca-lanca-estimativa-da-incidencia-de-cancer-de-mama-no-brasil/.

[21] C.D. Meyer. *Matrix Analysis and Applied Linear Algebra*. Other Titles in Applied Mathematics. Society for Industrial e Applied Mathematics, 2000. ISBN: 9780898714548. URL: https://books.google.com.br/books?id=Zg4M0iFlbGcC.

[22] Ademir A. Ribeiro e Elizabeth W. Karas. *Otimização Contínua: Aspectos teóricos e computacionais*. Cengage Learning, 2013.

[23] R. M. Silva, M. R. R. Leal e F. M. Lima. "Predição do Câncer de Mama com Aplicação de Modelos de Inteligência Computacional". Em: *Tendências em Matemática Aplicada e Computacional* 20.2 (2019), pp. 229–240.

[24] Abel Siqueira. *Implementando métodos de Machine Learning do zero em Julia - MeLOne.jl: SVM - parte 2*. 2020. URL: `https://www.youtube.com/watch?v=hr0kvvv1KZY&pbjreload=101`.

[25] Abel Siqueira. *Implementando métodos de Machine Learning do zero em Julia - SVM parte 1*. 2020. URL: `https://www.youtube.com/watch?v=GBTHeEJlqB8&t=3687s`.

[26] Andreas Wächter e Lorenz T. Biegler. "On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming". Em: *Mathematical Programming* 106 (2006), pp. 25–57. DOI: `10.1007/s10107-004-0559-y`.

[27] W. H. Wolberg, M. A. Tanner e W. Y. Loh. "Diagnostic Schemes for Fine Needle Aspirates of Breast Masses". Em: *Analytical and Quantitative Cytology and Histology* 10.3 (jun. de 1988), pp. 225–228.