Métodos de Otimização e Máquinas de Vetores Suporte

Qualificação de Trabalho de Conclusão de Curso

Paula Cristina Rohr Ertel*

Orientador: Luiz Rafael dos Santos

Universidade Federal de Santa Catarina - Campus Blumenau

18 de Novembro de 2019

1 Introdução às Máquinas de Vetores Suporte

A Aprendizagem de Máquina (do inglês *Machine Learning*) é o estudo do uso de técnicas computacionais para automaticamente detectar padrões em dados e usá-los para fazer predições e tomar decisões. De acordo com Krulikovski [4], existem dois tipos de Aprendizagem de Máquina, a aprendizagem supervisionada, em que a partir de um conjunto de dados de entrada e saída a máquina constrói um modelo que deduz a saída para novas entradas, e a não supervisionada, na qual a máquina cria sua própria solução.

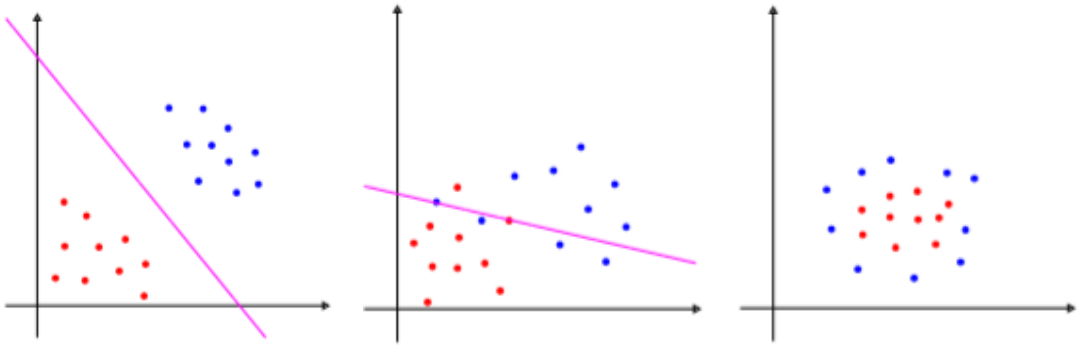
A aprendizagem supervisionada é composta por uma etapa denominada fase de treinamento, na qual é dado um conjunto de treinamento formado por vários dados de entrada e saída que funcionam como exemplos, a partir dos quais a máquina detecta padrões e cria um modelo para deduzir a saída de novos dados. Após essa fase novas entradas são testadas, denominadas conjunto de teste, no intuito de analisar se a máquina está gerando as saídas corretas. Algumas técnicas para aprendizagem de máquina supervisionada são as Máquinas de Vetores Suporte, Regressão Linear, Regressão Logística e Redes Neurais. Enquanto que a *Singular Value Decomposition* (SVD), Clusterização e

*Acadêmica do curso de Licenciatura em Matemática/UFSC-Blumenau

Análise de Componentes Principais [4] são exemplos de técnicas para a aprendizagem não supervisionada.

As Máquinas de Vetores Suporte (SVM, do inglês *Support Vector Machine*), conforme mencionado por Krulikovski [4], são indicadas nos casos em que ocorrem dados de dimensões elevadas e com altos níveis de ruídos, além de apresentar uma boa capacidade de generalização. Esta técnica pode ser aplicada tanto para problemas de regressão como de classificação. Segundo Krulikovski [4], essa técnica foi desenvolvida por Vladimir Vapnik, Bernhard Boser, Isabelle Guyon e Corrina Cortes, com base na Teoria de Aprendizagem Estatística. Algumas aplicações de SVM em problemas práticos são o reconhecimento facial, leitura de placas automotivas e detecção de spam.

Agora, vamos formular matematicamente o problema de classificação utilizando as Máquinas de Vetores Suporte. Para tanto, considere um conjunto de dados, pertencentes a duas classes distintas, conforme Figura 1.



(a) Linear. (b) Flexível. (c) Não Linear.

Figura 1: Dados lineares, com margem flexível e não lineares.
Fonte: Krulikovski [4]

Observe que na Figura 1a os dados podem ser classificados corretamente através de uma reta. Já na Figura 1b é possível encontrar uma reta que separa alguns poucos dados, porém incorretamente. E na Figura 1c não é possível classificar os dados como nos casos anteriores. Nestes exemplos temos representados os três casos de SVM: o linear com margem rígida, o linear com margem flexível e o não linear, respectivamente.

A modelagem do problema de classificação, utilizando a técnica de SVM, consiste em encontrar um hiperplano ótimo que melhor separe os dados de entrada x^i em duas saídas y_i através de uma função de decisão. Matematicamente, mostraremos que trata-se um problema de programação quadrática convexa com restrições lineares, que pode ser

formulado como

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

com $w \in \mathbb{R}^n$ e $b \in \mathbb{R}$, em que $f : \mathbb{R}^n \to \mathbb{R}$ é uma função quadrática e $g : \mathbb{R}^{n+1} \to \mathbb{R}^m$ é linear. Note também que f e g são continuamente diferenciáveis.

Para formular matematicamente o problema de classificação, considere os conjuntos de entrada $\mathcal{X} = \{x^1, \ldots, x^m\} \subset \mathbb{R}^n$ e de treinamento $\mathcal{Y} = \{(x^1, y_1), \ldots, (x^m, y_m) \mid x^i \in \mathcal{X} \, e \, y_i \in \{-1, 1\}\}$, com a partição

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

dos conjuntos formados pelos atributos pertencentes às classes positiva e negativa, respectivamente.

Definição 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 $\mathcal{H}(w, b) = \{x \in \mathbb{R}^n \mid w^T x + b = 0\}$.

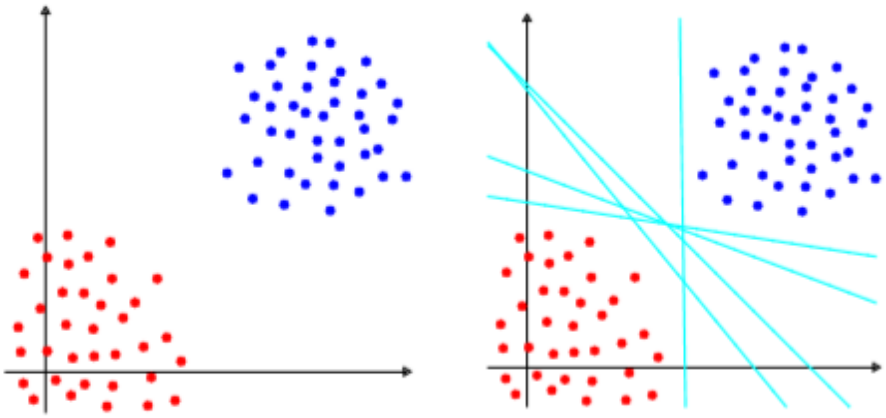
O hiperplano $\mathcal{H}(w, b)$ divide o espaço $\mathbb{R}^n$ em dois semiespaços, dados por

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

Considere dois conjuntos de dados de treinamento representados no $\mathbb{R}^2$ como na Figura 2a, em que os pontos em azul representam a classe positiva, e os pontos em vermelho a classe negativa. Perceba na Figura 2b que todos os hiperplanos representados separam corretamente os dados, porém nosso objetivo será encontrar o hiperplano que melhor separa esses dados, o qual está representado na Figura 3a pela cor violeta. Logo, desejamos encontrar o hiperplano que possibilita a maior faixa que não contém nenhum dado, pois caso a faixa seja muito estreita pequenas perturbações no hiperplano ou no conjunto de dados podem resultar uma classificação incorreta.

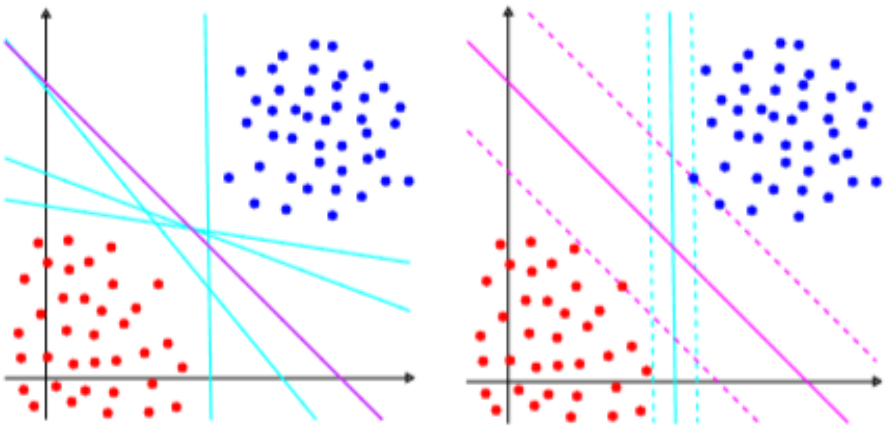
Definição 2. Os conjuntos $\mathcal{X}^+, \mathcal{X}^- \subset \mathbb{R}^n$ são ditos linearmente separáveis quando existem $w \in \mathbb{R}^n$ e $b \in \mathbb{R}$ tais que $w^T x + b > 0$ para todo $x \in \mathcal{X}^+$ e $w^T x + b < 0$ para todo $x \in \mathcal{X}^-$. O hiperplano $\mathcal{H}(w, b)$ é chamado hiperplano separador dos conjuntos $\mathcal{X}^+$ e $\mathcal{X}^-$.

Lema 1. *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 $\mathcal{H}(w, b)$*



(a) Dados de treinamento. (b) Hiperplanos separadores.

Figura 2: Conjunto de Dados e Hiperplanos.
Fonte: Krulikovski [4]



(a) Hiperplano ótimo. (b) Máxima margem.

Figura 3: Hiperplano Ótimo.
Fonte: Krulikovski [4]

pode ser descrito por

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

satisfazendo

$$\bar{w}^T x + \bar{b} \geq 1, \text{ para todo } x \in \mathcal{X}^+, \tag{1}$$
$$\bar{w}^T x + \bar{b} \leq -1, \text{ para todo } x \in \mathcal{X}^-. \tag{2}$$

Demonstração. Pela Definição 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 $\bar{w} := \dfrac{w}{\gamma}$ e $\bar{b} := \dfrac{b}{\gamma}$, obtemos as desigualdades (1) e (2).

□

A partir do Lema 1 temos que $\mathcal{H}^+ := \{x \in \mathbb{R}^n \mid w^T x + b = 1\}$ e $\mathcal{H}^- := \{x \in \mathbb{R}^n \mid w^T x + b = -1\}$ são os hiperplanos que definem a faixa que separa os conjuntos $\mathcal{X}^+$ e $\mathcal{X}^-$.

Proposição 1. *A projeção ortogonal de um vetor $\bar{x} \in \mathbb{R}^n$ sobre um hiperplano afim $\mathcal{H}(w, b)$, é dada por*

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

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

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^* a projeção ortogonal de $\bar{x}$ sobre $\mathcal{H}(w,b)$. Assim, temos que

$$w^T(x^* - \bar{z}) = 0 \tag{3}$$

e

$$\bar{x} - x^* = \lambda w \Longrightarrow x^* = \bar{x} - \lambda w. \tag{4}$$

Substituindo (4) em (3), 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 $\text{proj}_{\mathcal{H}}(\bar{x})$ 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

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

Assim, temos

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

ou seja,

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

□

Utilizando a Proposição 1 podemos demonstrar o Lema 2, o qual estabelece a largura da faixa entre os hiperplanos separadores $\mathcal{H}^+$ e $\mathcal{H}^-$.

Lema 2. *A distância entre os hiperplanos $\mathcal{H}^+$ e $\mathcal{H}^-$ é dada por* $\text{dist}(\mathcal{H}^+, \mathcal{H}^-) = \dfrac{2}{\|w\|}$.

Demonstração. Considere um ponto arbitrário $\bar{x} \in \mathcal{H}^+$ e seja $x^* \in \mathcal{H}^-$ a projeção ortogonal de $\bar{x}$ sobre $\mathcal{H}^-$. Usando a Proposição 1, temos

$$x^* = \text{proj}_{\mathcal{H}^-}(\bar{x}) = \bar{x} - \frac{w^T\bar{x} + b + 1}{\|w\|^2} w. \tag{5}$$

Além disso, a distância entre dois conjuntos é definida por

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

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

$$\text{dist}(\mathcal{H}^+, \mathcal{H}^-) = \|\bar{x} - x^*\|. \tag{6}$$

Substituindo (5) em (6) obtemos

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

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

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

concluindo que

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

□

1.1 Formulação Matemática do Problema de Classificação - Margem Rígida

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

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

implica

$$\|w\| \geq \|w^*\|. \tag{7}$$

Logo, $w^* = \arg\min \|w\|$. Além disso, como $\|\cdot\|$ é não negativa, elevando ao quadrado ambos os lados da desigualdade (7) 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

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

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

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}\|w\|^2 \\ \text{s.a.} \quad & y_i(w^T x^i + b) \geq 1, \quad i = 1, \ldots, m, \end{aligned} \tag{8}$$

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

O problema (8) possui função objetivo

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

convexa, 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

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

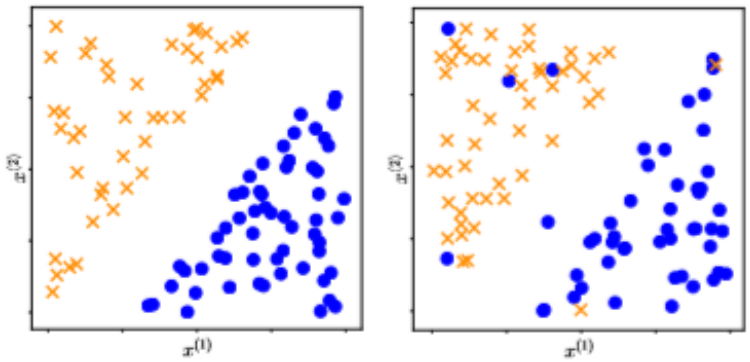
com e sendo o vetor cujas m componentes são todas iguais a 1, $Y = \text{diag}(y_i)$, $X = \text{diag}(x^i)$, $y^T = [y_1 \ \ldots \ y_m]$, $w \in \mathbb{R}^n$ e $b \in \mathbb{R}$.

1.2 Formulação Matemática do Problema de Classificação - Margem Flexível (CSVM)

Situações reias dificilmente envolvem problemas cujos dados são linearmente separáveis. Em vista disso, faz-se necessário estender os conceitos e resultados estudados nas SVMs lineares de margem rígida para o caso de SVM com margem flexível, quando os dados não são linearmente separáveis. Para tanto, considere um conjunto de dados não linearmente sepavável como da Figura 4b, isto é, não existe um hiperplano separador.

Neste caso, temos que o conjunto viável

$$\{(w, b) \in \mathbb{R}^{n+1} \mid 1 - y_i(w^T x^i + b) \leq 0, \quad i = 1, \ldots, m\}$$



(a) Dados linearmente separáveis.

(b) Dados não linearmente separáveis.

Figura 4: Fonte: Deisenroth, Faisal e Ong [1]

é vazio e, portanto, a formulação dada pelo problema (8) não fornece um classificador.

Assim, no intuito de contornar esse problema utilizamos regularização para suavizar as margens, acrescentando variáveis de folga $\xi_i \geq 0$ associadas aos dados de treinamento x_i, com $i = 1, \ldots, m$, e permitindo, assim, uma flexibilização do problema de estimar as variáveis w e b. Em outras palavras, a restrição $1 - y_i(w^T x^i + b) \leq 0$ é relaxada e substituída por $1 - y_i(w^T x^i + b) \leq \xi_i$, com $\xi_i \geq 0$. Cada variável de folga ξ_i mensura a distância que determinado dado x_i está do seu respectivo hiperplano separador, caso este dado esteja do lado errado.

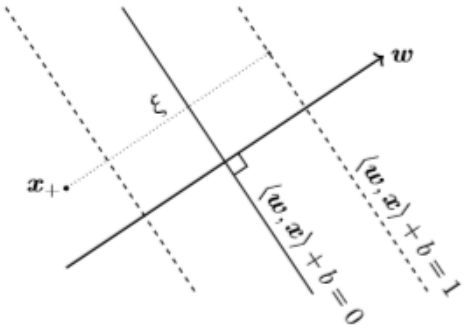


Figura 5: Variáveis de folga
Fonte: Deisenroth, Faisal e Ong [1]

Tal procedimento permite que pontos da classe positiva permaneçam fora do semiespaço $\mathcal{S}^+ = \{x \in \mathbb{R}^n \mid w^T x + b \geq 1\}$ e/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 é denominado hiperplano de margem flexível

e as restrições dos hiperplanos separadores são reformuladas da seguinte maneira

$$w^T x^i + b \geq 1 - \xi_i, \text{ para todo } x^i \in \mathcal{X}^+, \tag{9}$$
$$w^T x^i + b \leq -1 + \xi_i, \text{ para todo } x^i \in \mathcal{X}^-. \tag{10}$$

Agora nosso objetivo é encontrar w e b ótimos de modo a obter um bom classificador. Primeiramente, observe que dados w e b arbitrários, podemos escolher $\xi_i \geq 0$ de modo que as restrições (9) e (10) sejam satisfeitas. Para tanto, podemos definir

$$\xi_i = \begin{cases} \max\{0, 1 - w^T x^i - b\}, & \text{se} \quad x^i \in \mathcal{X}^+, \\ \max\{0, 1 + w^T x^i + b\}, & \text{se} \quad x^i \in \mathcal{X}^-. \end{cases}$$

Desse modo, para obter um bom classificador não basta apenas maximizar a margem definida pelos hiperplanos $\mathcal{H}^+$ e $\mathcal{H}^-$ e introduzir as variáveis de folga nas restrições, mantendo a mesma função objetivo, pois, como exemplificado por Krulikovski [4], a Figura 6 ilustra o hiperplano dado por $w_0^T x + b = 0$, que não poder ser usado para classificar os dados, mas satisfaz as restrições (9) e (10).

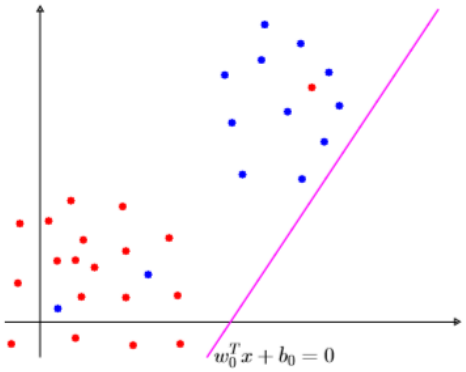


Figura 6: Pensar.
Fonte: Krulikovski [4]

Portanto, para reformular o problema original de maximizar a margem é preciso também controlar o valor dessas variáveis de modo a estimular uma classificação correta, pois quanto maior o valor delas mais será permitido violar as restrições. Em vista disso, é acrescentada na função objetivo uma parcela que corresponde à penalização das vio-

lações e o problema (8) é reformulado da seguinte forma

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i \\ \text{s.a.} \quad & y_i(w^T x^i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, m, \\ & \xi_i \geq 0, \quad i = 1, \ldots, m, \end{aligned} \tag{11}$$

em que $C > 0$ é um parâmetro de regularização que tem o objetivo de controlar a importância das variáveis de folga. O valor do parâmetro C que fornece uma boa classificação dos dados é escolhido de maneira heurística na fase de treinamento, geralmente a partir da natureza do problema.É devido a utilização desse parâmetro esta modelagem de SVM também é conhecida como C-SVM.

O termo $C\sum_{i=1}^{m}\xi_i$ na função objetivo do problema (11) pode ser pensado como uma medida de erro de classificação, pois minimiza o valor das variáveis de folga e reduz desse modo o número de pontos classificados incorretamente. De fato, aumentando o valor do parâmetro C aumenta-se a penalização sobre a violação da restrição original do problema SVM. Por outro lado, diminuindo o valor de C o modelo se torna mais flexível a esse tipo de violação.

O problema de margem flexível (11), assim como o problema (8), também possui restrições lineares

$$\begin{aligned} g_i(w,b,\xi) =& 1 - \xi_i - y_i(w^T x^i + b) \leq 0 \quad \text{e} \\ h_i(w,b,\xi) =& -\xi_i \leq 0, \quad i = 1, \ldots, m, \end{aligned}$$

e assim, o conjunto viável

$$\Omega = \{(w,b,\xi) \in \mathbb{R}^{n+1+m} \mid g_i(w,b,\xi) \leq 0,\ h_i(w,b,\xi) \leq 0,\ i = 1, \ldots, m\}$$

é um poliedro não vazio.

Ademais, a função objetivo f é quadrática e limitada inferiormente em Ω, pois

$$f(w,b,\xi) = \frac{1}{2}\|w\|^2 + \underbrace{C}_{>0}\sum_{i=1}^{m}\underbrace{\xi_i}_{\geq 0} \geq 0.$$

2 Conceitos de Otimização

Tendo em vista que o problema de clasificação trata-se de um problema de Otimização, neste capítulo pretendemos discutir alguns conceitos e resultados de otimização para o problema

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{s.a} \quad & x \in \Omega, \end{aligned} \tag{12}$$

em que $f : \mathbb{R}^n \to \mathbb{R}$ é chamada *função objetivo*, $\Omega \subset \mathbb{R}^n$ é chamado *conjunto factível* do problema (12) e os pontos de Ω são chamados de *pontos factíveis*.

Inicialmente, faz-se necessário caracterizar os pontos que são solução do problema (12).

Definição 3. Dizemos que um ponto $x^* \in \Omega$ é

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

(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 3 forem estritas para $x \neq x^*$, diremos que x^* é minimizador estrito.

Pela Definição 3, todo minimizador global é também minimizador local, porém a recíproca não é verdadeira. É interessante salientar que ná prática, do ponto de vista teórico e computacional, em muitas circunstâncias iremos nos contentar com um ponto de mínimo local. Haja vista que em geral, condições globais e soluções globais só podem ser encontradas se o problema possuir certas propriedades de convexidade que garantem, essencialmente, que qualquer mínimo local é mínimo global.

Definição 4. Dizemos que $\bar{v} \in [-\infty, +\infty)$ definido por

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

é o *valor ótimo* do problema (12).

Observação 1. 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.

Quando o conjunto factível $\Omega = \mathbb{R}^n$ dizemos o problema (12) é irrestrito. No caso em que Ω é definido por um sistema de igualdades e/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 desiguladade, haja vista que o problema de classificação, o qual estamos interessados em resolver, possui este formato.

No estudo do problema de otimização irrestrita uma das principais questões que surge diz respeito a existência da solução. Observe que se na Definição 4 temos $\bar{v} = -\infty$, o problema (12) não admite solução global, pois neste caso f é ilimitada inferiormente no conjunto factível. Outro caso em que também não minimizador global ocorre quando $\bar{v}$ não é atingido em nenhum ponto factível.

Exemplo 2.0.1. Seja $f : \mathbb{R} \to \mathbb{R}$, $f(x) = e^x$, $\Omega = \mathbb{R}$. 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 $f(0) = 0$. 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 1. ***(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. □

Definição 5. Dado $\bar{x} \in \Omega$, dizemos que uma restrição de desigualdade g_i é ativa em $\bar{x}$ quando $g_i(\bar{x}) = 0$. Caso $g_i(\bar{x}) < 0$, dizemos que g_i é inativa em $\bar{x}$.

O conjunto dos índices das restrições de desigualdade ativas é denotado por

$$I(\bar{x}) = \{i \mid g_i(\bar{x}) = 0\}.$$

Definição 6. Um ponto $x^* \in \mathbb{R}^n$ é dito *estacionário* para o problema (12) quando existirem $\mu^* \in \mathbb{R}^m$, $\lambda^* \in \mathbb{R}^p$ (*multiplicadores de Lagrange*) tais que

$$\begin{aligned} \nabla f(x^*) + \sum_{i=1}^{m} \mu_i^* \nabla g_i(x^*) + \sum_{i=1}^{p} \lambda_i^* \nabla h_i(x^*) = 0, \\ g(x^*) \leq 0, h(x^*) = 0, \\ \mu_i^* \geq 0, \quad i = 1, \ldots, m, \\ \lambda_i^* g_i(x^*) = 0, \quad i = 1, \ldots, m. \end{aligned} \tag{13}$$

As condições (13) são conhecidas como condições de Karush-Kuhn-Tucker (KKT). Assim, em outras palavras, x é dito estacionário quando satisfaz as condições KKT.

Para que as condições KKT possam ser consideradas uma condição de otimalidade, isto é, que sejam satisfeitas por um ponto que é minimizador, é necessário assumir alguma hipótese adicional às restrições do problema, a qual será chamada de condição de qualificação.

Definição 7. Dizemos que as restrições $g(x) \leq 0$ e $h(x) = 0$ cumprem uma condição de qualificação em $x^* \in \Omega$ quando, dada qualquer função diferenciável f que tenha mínimo em x^*, relativamente a Ω, sejam satisfeitas as condições de otimalidade de KKT.

Logo, um ponto $x \in \mathbb{R}^n$ é dito qualificado quando atende uma condição de qualificação. Vejamos primeiramente a noção de *ponto regular*.

Definição 8. Seja x^* um ponto que satisfaz as restrições

$$h(x^*) = 0, \quad g(x^*) \leq 0, \tag{14}$$

e seja $I(x^*)$ o conjunto dos índices i tais que $g_i(x^*) = 0$. Então x^* é chamado *ponto regular* das restrições (14) se os gradientes das restrições de igualdade e desilgualdade ativas, isto é, $\nabla h_j(x^*)$ e $\nabla g_i(x^*)$, com $1 \leq j \leq m$ e $i \in I(x^*)$, são linearmente independentes.

Esta definição é chamada de condição de regularidade, também conhecida como condição de qualificação de independência linear dos gradietes das restrições ativas (LICQ, do inglês *Linear Independence Constraint Qualification*) e enunciada da seguinte forma:

Condição de qualificação de independência linear: Dizemos que a condição de qualificação de independência linear (LICQ) é satisfeita em $\bar{x}$ quando o conjunto formado pelos gradientes das restrições de igualdade e das restrições de desigualdade ativas é linearmente independente, isto é,

$$\{\nabla g_i(\bar{x}) \mid i \in I(\bar{x})\} \cup \{\nabla h_i(\bar{x}), i = 1, \ldots, p\} \quad \text{é LI.}$$

Outra condição interessante aos nossos estudos por mencionar a convexidade é a condição de qualificação de Slater.

Condição de qualificação de Slater: Dizemos que a condição de qualificação de Slater é satisfeita quando h é linear, cada componente g_i, $i = 1, \ldots, m$, é convexa e existe $\tilde{x} \in \Omega$ tal que $h(\tilde{x}) = 0$ e $g(\tilde{x}) < 0$.

Teorema 2. *(*KKT*) Seja $x^* \in \mathbb{R}^n$ um minimizador local do problema* (12) *e suponha que seja satisfeita uma condição de qualificação. Então existem vetores $\mu^* \in \mathbb{R}^m$ e $\lambda^* \in \mathbb{R}^p$ tais que (x^*, μ^*, λ^*) cumpre* (13)*.*

É importante observar que se não for verificada nenhuma condição de qualificação podemos ter minimizadores que não cumprem KKT, o que dificulta a caracterização de tais pontos. Vejamos a seguir um exemplo que justifica essa afirmação.

Exemplo 2.0.2. Considere o problema

$$\begin{aligned} \min_x \quad & f(x) = x_1 \\ \text{s.a} \quad & g_1(x) = -x_1^3 + x_2 \leq 0, \\ & g_2(x) = -x_2 \leq 0. \end{aligned}$$

O ponto $x^* = 0$ é o minimizador deste problema, mas não cumpre as condições KKT. De fato, observe que $0 \leq x_2 \leq x_1^3$, o que implica em $f(x) = x_1 \geq 0 = f(x^*)$, para todo ponto viável x.

No entanto,

$$\nabla f(x^*) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \nabla g_1(x^*) = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad \text{e} \quad \nabla g_2(x^*) = \begin{pmatrix} 0 \\ -1 \end{pmatrix},$$

ou seja, os gradientes das restrições não são LI e, portanto, x^* não é um ponto regular. Assim, as condições KKT não são satisfeitas, pois não existem $\mu_1^*, \mu_2^* \in \mathbb{R}^+$ tais que

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix} + \mu_1^* \begin{pmatrix} 0 \\ 1 \end{pmatrix} + \mu_2^* \begin{pmatrix} 0 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Tendo em vista que em nosso problema o conjunto factível é poliedral, abordaremos nas seções seguintes como resolver o problema de minimização com restrições nesse contexto. Para tanto, começaremos analisando o problema de minimização com restrições lineares de igualdade.

2.1 Minimização com Restrições Lineares de Igualdade

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

$$\begin{aligned} \min_x \quad & f(x) \\ \text{s.a} \quad & Ax = b, \end{aligned} \tag{15}$$

em que $A \in \mathbb{R}^{m \times n}$, $1 \leq m < n$ e posto $A = m$.

O conjunto

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

é chamado *conjunto de factibilidade* de (15). Este conjunto é a variedade afim de soluções do sistema linear

$$Ax = b, \tag{16}$$

e de modo geral, S 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 $n > 3$ e $m = 1$ caracterizamos S como um hiperplano.

Para definir as condições de otimalidade para o problema (15), será necessário primeiramente determinar o conjunto de direções factíveis em S. Para tanto, associado a S temos o conjunto de soluções do sistema homogêneo $Ax = b$, 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 $A = m$, e é interessante observar que $\mathcal{N}(A)$ é um subespaço paralelo à variedade afim S e passa pela origem. Esta noção é representada na Figura 7.

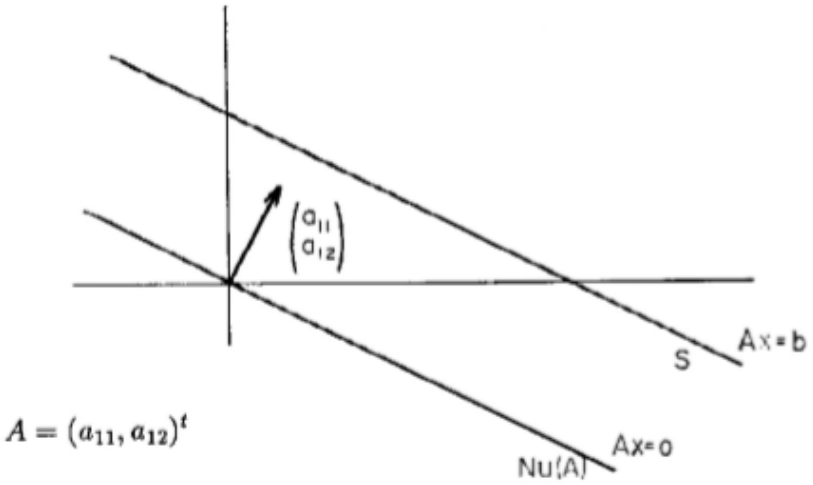


Figura 7: Fonte: Friedlander [2]

Como posto $A = m$, as m linhas de A formam um conjunto de vetores linearmente independentes que geram o subespaço Imagem de A^T de dimensão m, que é denotado por $Im(A^T)$. Dessa forma, como é possível observar na Figura 7 as linhas de A são ortogonais a $\mathcal{N}(A)$, ou, em outras palavras, $\mathcal{N}(A)$ é o complemento ortogonal de $Im(A^T)$. Vamos demonstrar este resultado.

Proposição 2. *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

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

A partir disso, se $d \in \mathcal{N}(A)$ e $\tilde{x}$ é uma solução da restrição (16), então $x := \tilde{x} + \alpha d$ também é solução de (16), 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 no espaço na qual podemos nos deslocar a partir de uma solução factível sem correr o risco de abandonar a região de factibilidade.

A recíproca também é verdadeira, pois se a partir de uma solução factível $\tilde{x}$, andando numa direção $d \in \mathbb{R}^n$ obtemos

$$x = \tilde{x} + \alpha d \quad \text{e} \quad Ax = b,$$

então $A(\tilde{x}+\alpha d) = b$ implica em $Ad = 0$ e, portanto, $d \in \mathcal{N}(A)$. Dessa forma, concluímos que $\mathcal{N}(A)$ é o conjunto de direções factíveis em S.

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 $\gamma \in \mathbb{R}^{n-m}$ tal que $d = Z\gamma$, ou seja, d é escrito como combinação linear dos vetores da base do núcleo de A. Assim, se $\tilde{x}$ é uma solução de (16), podemos caracterizar o conjunto factível da seguinte forma

$$S = \{x \in \mathbb{R}^n \mid x = \tilde{x} + Z\gamma, \gamma \in \mathbb{R}^{n-m}\}. \tag{17}$$

2.1.1 Condições Necessárias de Primeira Ordem

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 iguladade. Para obter esta condição utilizaremos a parametrização do conjunto factível proposta em (17), transferindo as restrições de (15) 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 S dada em (17) permite definir a seguinte função $\varphi : \mathbb{R}^{n-m} \to \mathbb{R}$ dada por

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

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

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

Primeiramente, vejamos que os problemas (15) e (19) são equivalentes.

Proposição 3. *γ^* é um minimizador local (global) de φ em $\mathbb{R}^{n-m}$ se, e somente se,*

$x^* := \tilde{x} + Z\gamma^*$ *é um minimizador local (global) de* (15).

Demonstração. Seja γ^* um minimizador local (global) de φ em $\mathbb{R}^{n-m}$. Então,

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

para todo $\gamma \in \mathbb{R}^{n-m}$. Logo, chamando $x^* := \tilde{x} + Z\gamma^*$, temos

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

para todo $x \in \mathbb{R}^n$, com $x^* \in S$. Portanto, x^* é um minimizador local (global) de (15).

Reciprocamente, suponhamos que $x^* = \tilde{x} + Z\gamma^*$ é um minimizador local (global) de (15). Então,

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

para todo $x \in \mathbb{R}^n$, o que implica em

$$\varphi(\gamma^*) \leq \varphi(\gamma),$$

para todo $\gamma \in \mathbb{R}^{n-m}$. Portanto, γ^* é um minimizador local (global) de φ em $\mathbb{R}^{n-m}$. □

Agora, como (19) é um problema de minimização irrestrito, a condição necessária de primeira ordem para ele é

$$\nabla\varphi(x^*) = 0. \tag{20}$$

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

$$\varphi'(\gamma) = f'(g(\gamma))g'(\gamma) = \nabla f(g(\gamma))Z$$

e, portanto

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

Desse modo, da condição (20), resulta que

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

Portanto, uma condição necessária de primeira ordem para que x^* seja minimizador local de (15) é que

$$Z^T\nabla f(x^*) = 0, \tag{21}$$

ou seja, que $\nabla f(x^*)$ seja ortogonal a $\mathcal{N}(A)$. Logo, pelas considerações feitas anteriormente, 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 $\lambda^* \in \mathbb{R}^m$ tal que

$$\nabla f(x^*) = A^T \lambda^*. \tag{22}$$

Observe que as condições propostas em (21) e (22) são equivalentes. Com efeito, se (21) se verifica, isso implica que $\nabla f(x^*) \in \mathcal{N}(A)^\perp = Im(A^T)$ e portanto (22) é verdadeiro. Reciprocamente, se (22) ocorre, temos que, pela Proposição 2, $Z^T \nabla f(x^*) = Z^T(A^T \lambda^*) = 0$.

Em vista disso, se x^* é um minimizador local de (15), então, por (22), existe $\lambda^* \in \mathbb{R}^m$ tal que (x^*, λ^*) é solução do seguinte sistema de $(n+m)$ equações

$$\begin{aligned} &\nabla f(x^*) = A^T \lambda^* \\ &Ax^* = b. \end{aligned} \tag{23}$$

A solução de (15) é necessariamente solução de (23), porém para a recíproca ser verdadeira é necessária a informação de segunda ordem.

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

Vejamos a seguir um exemplo em que as condições de otimalidade de primeira ordem nos permitem determinar a expressão geral para uma solução do problema de minimizar $f(x) = \|x\|$ sujeita a restrições de igualdade.

Exemplo 2.1.1. Considere o seguinte problema

$$\begin{aligned} \min_x \quad & \|x\| \\ \text{s.a} \quad & Ax = b, \end{aligned} \tag{24}$$

com $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $m < n$ e posto $A = m$. Então a solução $\tilde{x}$ desse problema pode ser escrita como $\tilde{x} = \tilde{A}b$, em que $\tilde{A} \in \mathbb{R}^{n \times m}$ e $A\tilde{A} = I$.

Com efeito, seja $\tilde{x}$ solução de (24). Então, $\tilde{x}$ também minimiza $\dfrac{1}{2}\|x\|$, pois ambos os problemas são equivalentes. Assim, inicialmente $\tilde{x}$ deve satisfazer a restrição de igualdade, isto é,

$$A\tilde{x} = b. \tag{25}$$

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

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

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

$$\tilde{x} = A^T\tilde{\lambda}. \tag{26}$$

Substituindo (26) em (25), temos

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

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

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

o que implica em

$$\tilde{\lambda} = (AA^T)^{-1}b. \tag{27}$$

Agora, substituindo (27) em (26), concluímos que

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

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

Observação 2. A matriz $\tilde{A}$ é também chamada de pseudo-inversa de Moore-Penrose e geralmente denotada por $A^\dagger$.

2.1.2 Condições Necessárias e Suficientes de Segunda Ordem

A condição necessária de segunda ordem para uma solução do problema (19) é

$$\nabla^2\varphi(\gamma^*) \succcurlyeq 0 \quad \text{(semidefinida positiva)}. \tag{28}$$

Aplicando a regra da cadeia em $\nabla\varphi(\gamma) = Z^T\nabla f(\tilde{x} + Z\gamma)$, obtemos

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

Assim, por (28) temos que a condição necessária de segunda ordem para que x^* seja minimizador local de (15) 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

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

Analogamente, a partir das condições suficientes para o problema (19), podemos determinar as seguintes condições suficientes de segunda ordem para (15):

Se $x^* \in \mathbb{R}^n$ verifica $Ax^* = b$ e

(i) $Z^T\nabla f(x^*) = 0$,

(ii) $Z^T\nabla^2 f(x^*)Z \succ 0$ (definida positiva),

então x^* é um minimizador local de (15).

Proposição 4. *Seja $f : \mathbb{R}^n \to \mathbb{R}$, $f \in \mathcal{C}^2$. Seja $\tilde{x} \in \mathbb{R}^n$ tal que $A\tilde{x} = b$ ($A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$) e tal que existe $\lambda \in \mathbb{R}^m$ com $\nabla f(\tilde{x}) = A^T\lambda$ e $\nabla^2 f(\tilde{x})$ definida positiva. Então $\tilde{x}$ é um minimizador local de f sujeita a $Ax = b$.*

Demonstração. Para provar que $\tilde{x}$ é minimizador local de f vamos verificar se ele satisfaz as condições suficientes de segunda ordem. Por hipótese temos que $\tilde{x} \in \mathbb{R}^n$ satisfaz a restrição $A\tilde{x} = b$. Por conseguinte, existe $\lambda \in \mathbb{R}^m$ tal que $\nabla f(\tilde{x}) = A^T\lambda$. Assim, seja Z a matriz cujas colunas são os vetores da base do $\mathcal{N}(A)$, sabemos que as linhas de A são ortogonais a $\mathcal{N}(A)$ e portanto,

$$Z^T\nabla f(\tilde{x}) = Z^T A^T\lambda = 0.$$

Além disso, como $\nabla^2 f(\tilde{x})$ é definida positiva, temos que

$$Z^T\nabla^2 f(\tilde{x})Z \succ 0.$$

Portanto, $\tilde{x}$ é minimizador local de f. □

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 2.1.2. Considere o problema

$$\begin{aligned} \min_x \quad & \frac{1}{2}x^TQx + p^Tx + q \\ \text{s.a} \quad & Ax = b, \end{aligned} \tag{29}$$

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^TQZ é definida positiva. Seja x^0 tal que $Ax^0 = b$. Então a solução $\tilde{x}$ é dada por

$$\tilde{x} = x^0 - Z(Z^TQZ)^{-1}Z^T(Qx^0 + p). \tag{30}$$

Para provar que (30) 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,

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

pois $AZ = 0$ pelo fato de $\mathcal{N}(A)$ e $Im(A^T)$ serem ortogonais.

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

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

(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

$$Z^TQZ \succ 0.$$

Portanto, $\tilde{x}$ é solução do problema (29).

3 Otimização Convexa

Em otimização uma hipótese com ótimas consequências é a convexidade, pois ela garante que pontos estacionários são minimizadores e permite concluir que minimizadores locais são globais. A partir disso, os problemas de classificação podem ser formulados em termos de otimização convexa. Os principais conceitos definidos neste capítulo são conjuntos convexos e funções convexas e a partir deles apresentamos alguns resultados importantes da análise convexa. Para desenvolvimento desse capítulo as principais referências utilizadas foram Izmailov e Solodov [3], Krulikovski [4] e Ribeiro e Karas [5].

3.1 Conjuntos Convexos

Definição 9. Um conjunto $C \subset \mathbb{R}^n$ é dito convexo quando dados $x, y \in C$, o segmento $[x, y] = \{(1-t)x + ty \mid t \in [0, 1]\}$ estiver inteiramente contido em C.

A Figura 8 apresenta a noção de convexidade de conjuntos, ilustrando um conjunto convexo e outro não convexo. Em outras palavras um conjunto convexo se caracteriza por conter todos os segmentos cujos extremos pertencem ao conjunto, o que não ocorre no segundo conjunto da Figura 8 por exemplo.

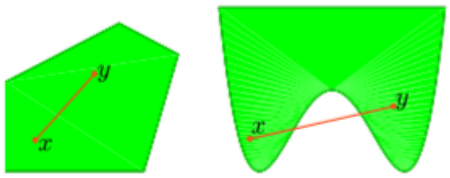


Figura 8: Exemplo de conjunto convexo e não convexo.
Fonte: Krulikovski [4]

Alguns exemplos de conjuntos convexos são o conjunto vazio, o espaço $\mathbb{R}^n$, qualquer hiperplano do $\mathbb{R}^n$ e um conjunto que contém um ponto só.

A seguir, apresentamos alguns resultados importantes da análise convexa.

Lema 3. *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

$$(1-t)^2u^Tu + 2t(1-t)u^Tv + t^2v^Tv = \|(1-t)u+tv\|^2 = r^2. \tag{31}$$

Como $u^Tu = v^Tv = r^2$ e $t \in (0,1)$, substituindo em (31) e desenvolvendo, obtemos

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

Evidenciando r^2, temos

$$(2t-2t^2)r^2 = (2t-2t^2)u^Tv,$$

e, portanto,

$$r^2 = u^Tv. \tag{32}$$

Assim, por (32),

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

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 distânica 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 4. *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

$$\alpha \leq \|z - x^n\| \leq \alpha + \frac{1}{n}. \tag{33}$$

Em particular, $\|z - x^n\| \leq \alpha + 1$, para todo $n \in \mathbb{N}$. Logo, 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 (33), temos que $\|z - x^n\| \longrightarrow \alpha$, e portanto, concluímos que $\|z - \bar{z}\| = \alpha$.

□

Lema 5. *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 4. Para provar a unicidade suponha que existam $\bar{z}, \tilde{z} \in S$, com $\bar{z} \neq \tilde{z}$, tais que

$$\|z - \bar{z}\| \leq \|z - x\| \quad \text{e} \quad \|z - \tilde{z}\| \leq \|z - x\|, \tag{34}$$

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, 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 (34) teríamos

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

Portanto, $\bar{z} = \tilde{z}$.

□

No Lema 5 denotamos $\bar{z} = \text{proj}_S(z)$.

Teorema 3. *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 4 $\bar{z} \in S$ 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

$$2(z - \bar{z})^T(x - \bar{z}) \leq t\|x - \bar{z}\|^2. \tag{35}$$

Passando o limite em (35) quando $t \to 0$, obtemos

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

completando a demonstração. □

O Teorema 3 estabelece uma condição necessária e suficiente para caracterizar a projeção. Este resultado é provado no lema seguinte.

Lema 6. *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

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

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, $\bar{z} = \text{proj}_S(z)$. □

O Lema 6 fornece uma condição necessária de otimalidade ao minimizar uma função em um conjunto convexo fechado.

Teorema 4. ***(Taylor de Primeira Ordem)*** *[5, p.25] Considere $f : \mathbb{R}^n \to \mathbb{R}$ uma função diferenciável e $\bar{x} \in \mathbb{R}^n$. Então podemos escrever*

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

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

Lema 7. *Sejam $f : \mathbb{R}^n \to \mathbb{R}$ uma função diferenciável e $C \subset \mathbb{R}^n$ um conjunto convexo e fechado. Se $x^* \in C$ é minimizador local de f em C, então*

$$\text{proj}_C(x^* - \alpha\nabla f(x^*)) = x^*,$$

para todo $\alpha \geq 0$.

Demonstração. Seja $x^* \in C$ minimizador local de f em C. Fixando $x \in C$, como C é

um conjunto convexo, temos

$$f(x^*) \leq f((1-t)x^* + tx), \tag{36}$$

para todo $t \geq 0$ suficientemente pequeno. Pelo Teorema 4,

$$f(x^* + t(x - x^*)) = f(x^*) + t\nabla f(x^*)^T(x - x^*) + r(t), \tag{37}$$

em que $\lim_{t\to 0} \dfrac{r(t)}{t} = 0$. Dessa forma, por (36) e (37), temos

$$0 \leq f(x^* + t(x - x^*)) - f(x^*) = t\nabla f(x^*)^T(x - x^*) + r(t),$$

e portanto,

$$t\nabla f(x^*)^T(x - x^*) + r(t) \geq 0. \tag{38}$$

Dividindo por t e passando o limite quando $t \to 0$ em (38), obtemos

$$\nabla f(x^*)^T(x - x^*) \geq 0.$$

Assim, dado $\alpha \geq 0$,

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

para todo $x \in C$.

Portanto, pelo Lema 6 concluímos que $\text{proj}_C(x^* - \alpha\nabla f(x^*)) = x^*$, para todo $\alpha \geq 0$. □

3.2 Funções Convexas

Definição 10. 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.

A noção geométrica da Definição 10 é apresentada na Figura 9, em que qualquer arco do gráfico de uma função convexa está sempre abaixo do segmento que liga as extremidades.

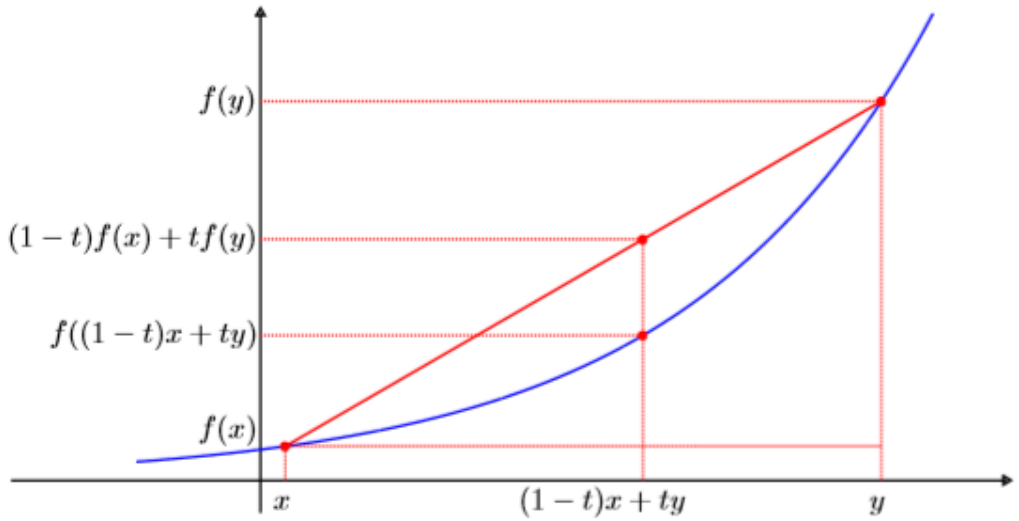


Figura 9: Função Convexa.
Fonte: Krulikovski [4]

Definição 11. Se $C \subset \mathbb{R}^n$ é um conjunto convexo, dizemos que $f : C \to \mathbb{R}$ é uma *função côncava* em C, quando a função $(-f)$ é convexa em C.

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

O teorema seguinte justifica o fato da convexidade ser uma propriedade tão importante em otimização.

Teorema 5. *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

$$f(x^*) \leq f(x) \leq (1 - t)f(x^*) + tf(y) = f(x^*) + t(f(y) - f(x^*)),$$

donde segue que $f(x^*) \leq f(y)$.

Portanto, x^* é minimizador global de f. □

A seguir apresentamos outra forma de caracterizar a convexidade de uma função quando temos hipóteses de diferenciabilidade.

Teorema 6. *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. Seja f uma função convexa. Para $x, y \in C$ e $t \in (0, 1]$ quaisquer, temos $(1 - t)x + ty = x + t(y - x)$. Assim, definindo $d = y - x$, temos que $x + td \in C$ e

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

logo

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

o que 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, considere $z = (1-t)x + ty$ e 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 $(1-t)$ e a segunda por t, obtemos

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

Portanto, a função f é convexa em C. □

A seguir, apresentamos alguns resultados necessários para demonstrar o Teorema 9.

Teorema 7. *(**Taylor de Segunda Ordem**) [5, p.26] Se $f : \mathbb{R}^n \to \mathbb{R}$ é uma função duas vezes diferenciável e $\bar{x} \in \mathbb{R}^n$, então*

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

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

Teorema 8. *(**Taylor com Resto de Lagrange**) [5, p.26] 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*

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

Lema 8. *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. Por demonstrar. □

Teorema 9. *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. Considere $x \in \operatorname{int} C$. Então, dado $d \in \mathbb{R}^n$ temos que $x + td \in C$, para t suficientemente pequeno. Portanto, pela convexidade de f e pelo Teorema 6, temos

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

e assim,

$$f(x + td) - f(x) - t\nabla f(x)^T d \geq 0. \tag{39}$$

Aplicando o Teorema 7 para $f(x + td)$, temos

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

e substituindo em (39), obtemos

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

com $\lim_{t\to 0} \dfrac{r(t^2)}{t^2} = 0$.

Dividindo por t^2 e passando o limite com $t \to 0$, temos

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

Agora, considere $x \in C$ arbitrário. Como existe $y \in \operatorname{int} C$, o Lema 8 garante que todos os pontos do segmento $(x, y] \subset \operatorname{int} C$. 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 $x \in C$.

Reciprocamente, dados $x \in C$ e $d \in \mathbb{R}^n$ tal que $x + d \in C$, pelo Teorema 8,

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

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

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

Logo, pelo Teorema 6, f é convexa. □

Em geral, a função objetivo dos problemas de SVM serão funções quadráticas. Assim, faz-se necessário mostrar que a função quadrática é convexa. Tal resultado é apresentado a seguir.

Teorema 10. *Seja $C \in \mathbb{R}^n$ um conjunto convexo e Q uma matriz quadrada. Seja $f : C \to \mathbb{R}$ tal que $f(x) = x^T Q x$ é uma função quadrática. Então, f é convexa se, e somente se, Q é semidefinida positiva.*

Demonstração. Por demonstrar. □

Referências

[1] Peter Deisenroth, A. Aldo Faisal e Cheng Soon Ong. *Mathematics for Machine Learning*. Boston: Cambridge University Press, 2019.

[2] Ana Friedlander. *Elementos de Programação Não-Linear*. Unicamp, 1994.

[3] 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.

[4] 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.

[5] Ademir A. Ribeiro e Elizabeth W. Karas. *Otimização Contínua: Aspectos teóricos e computacionais*. Cengage Learning, 2013.