

Análise Teórica de Máquinas de Vetores Suporte

Aluna: Paula Cristina Rohr Ertel

Orientador: Luiz Rafael dos Santos

Universidade Federal de Santa Catarina - Campus Blumenau

18 de Novembro de 2019

1 Introdução à SVM

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 previsões e tomar decisões. De acordo com Krulikovski [5], 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 Análise de Componentes Principais [5] são exemplos de técnicas para a aprendizagem não supervisionada.

As Máquinas de Vetores Suporte (SVM, do inglês *Support Vector Machine*), conforme mencionado por Krulikovski [5], são indicadas nos casos em que ocorrem dados de di-

mensões 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 [5], 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.

1.1 Objetivos

Desenvolver um estudo teórico das Máquinas de Vetores Suporte, em particular, estudar a SVM para o problema de classificação, abordando inicialmente o problema de classificação binária, isto é, quando o conjunto de valores possíveis que a classe de saída pode atingir é binário. Para tanto, neste trabalho o conjunto de saída será denotado por $\{1, -1\}$. Entretanto, quaisquer dois valores distintos poderiam ser utilizados, como $\{0, 1\}$, $\{True, False\}$, $\{red, blue\}$.

1.2 Referencial Teórico

Primeiramente, fez-se necessário estudar os aspectos teóricos-matemáticos dos Métodos de Otimização relacionados ao aprendizado de máquina, mais especificamente os associados as Máquinas de Vetores Suporte. Para tanto, assim como proposto por Deisenroth, Faisal e Ong [1], realizou-se uma revisão dos principais conceitos de Álgebra Linear e do Cálculo de Várias Variáveis relacionados ao assunto. Por conseguinte, desenvolveu-se um estudo teórico das condições de otimalidade para problemas de otimização sem restrições baseando-se em Ribeiro e Karas [6]. Para dar continuidade ao desenvolvimento desse projeto será necessário estudar as condições de otimalidade para problemas com restrições, haja vista que um dos problemas que se pretende resolver trata-se de um problema de programação quadrática convexa com restrições lineares. Em vista disso, pretende-se utilizar como referência Friedlander [2], Izmailov e Solodov [3, 4] e Ribeiro e Karas [6] para estudar a teoria de otimização com restrições, programação quadrática e dualidade. O estudo específico acerca dos aspectos teóricos-matemáticos da técnica de Máquinas de Vetores Suporte será desenvolvido a partir de Krulikovski [5].

1.3 Máquinas de Vetores Suporte - Margem Rígida

Considere um conjunto de dados, pertencentes a duas classes distintas, conforme Figura 1.

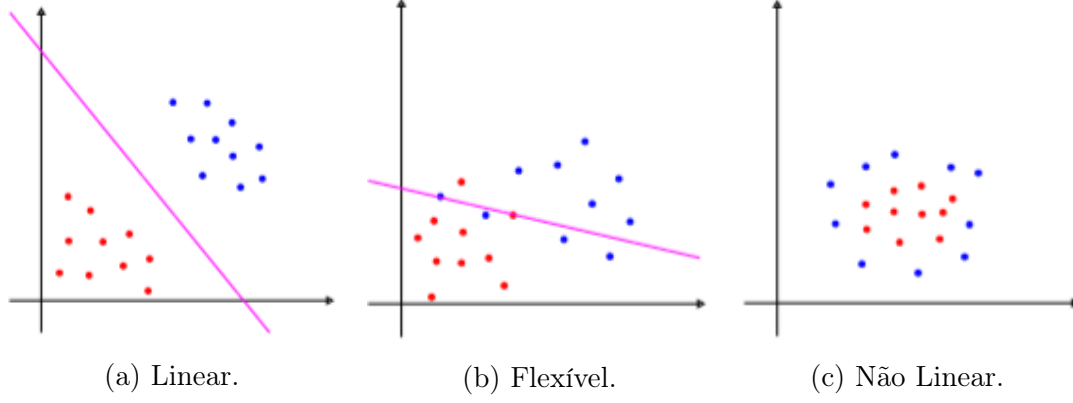


Figura 1: Dados lineares, com margem flexível e não lineares.
Fonte: Krulikowski [5]

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

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

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

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

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

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.

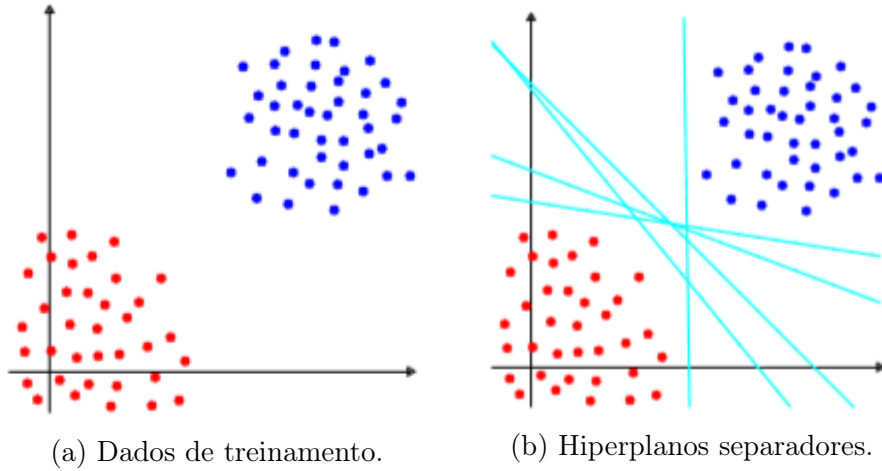


Figura 2: Conjunto de Dados e Hiperplanos.
Fonte: Krulikovski [5]

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}^- .

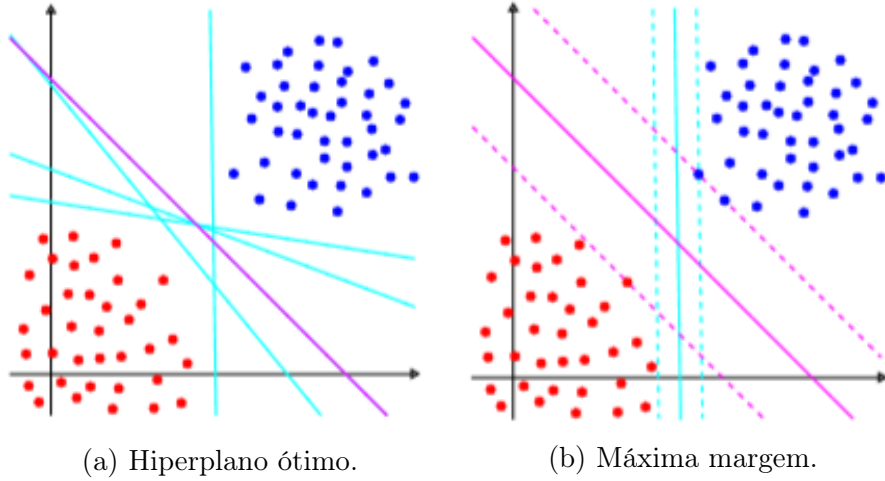


Figura 3: Hiperplano Ótimo.
Fonte: Krulikovski [5]

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 $\bar{w} \in \mathbb{R}^n$ e $\bar{b} \in \mathbb{R}$ tais que $\mathcal{H}(w, b)$ pode ser descrito por*

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

satisfazendo

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

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

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

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 $\bar{w} := \frac{w}{\gamma}$ e $\bar{b} := \frac{b}{\gamma}$, 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

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

Substituindo (4) em (3), obtemos

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

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 $d = \frac{2}{\|w\|}$.*

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

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

Além disso, a distância entre dois conjuntos é definida por

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

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

$$d(\mathcal{H}^+, \mathcal{H}^-) = \|\bar{x} - x^*\|. \quad (6)$$

Substituindo (5) em (6) obtemos

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

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

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

concluindo que

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

□

1.4 Formulação do Problema de Classificação

Encontrar o hiperplano que melhor separa os dados implica maximizar a largura da margem, isto é, maximizar $d = \frac{2}{\|w\|}$. 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

$$\|w\| \geq \|w^*\|. \quad (7)$$

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

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

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

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

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

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

em que a função $g : \mathbb{R}^{n+1} \rightarrow \mathbb{R}^m$ 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 $b \in \mathbb{R}$.

2 Projetos Futuros

O problema de classificação utilizando a técnica de SVM trata-se de um problema de programação quadrática convexa com restrições lineares. Para dar continuidade ao projeto será necessário estudar a teoria de otimização com restrições e a teoria de dualidade, em particular a relacionada ao problema de programação quadrática com restrições lineares. Por fim, pretendemos realizar uma pequena implementação computacional da técnica de Máquinas de Vetores Suporte a um problema de classificação. Para tanto, utilizaremos a linguagem de programação `Julia`, sobre a qual também será preciso estudar e se aperfeiçoar.

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. Vol. I. IMPA, 2014.
- [4] Alexey Izmailov e Mikhail Solodov. *Otimização. Métodos Computacionais*. 3ª. Vol. II. IMPA, 2014.
- [5] 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.
- [6] Ademir A. Ribeiro e Elizabeth W. Karas. *Otimização Contínua: Aspectos teóricos e computacionais*. Cengage Learning, 2013.