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 [2], 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.

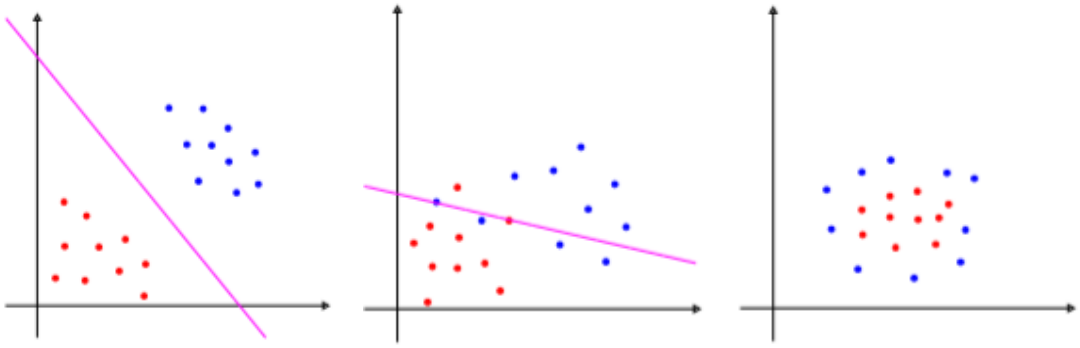
A aprendizagem supervisionada é composta por uma etapa denominada fase de treinamento, na qual é dado um conjunto de treinamento formado por vários dados de entrada e saída que funcionam como exemplos, a partir dos quais a máquina detecta padrões e cria um modelo para deduzir a saída de novos dados. 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 [2] são exemplos de técnicas para a aprendizagem não supervisionada.

As Máquinas de Vetores Suporte (SVM, do inglês *Support Vector Machine*), conforme mencionado por Krulikovski [2], 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 [2], 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 [2]

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

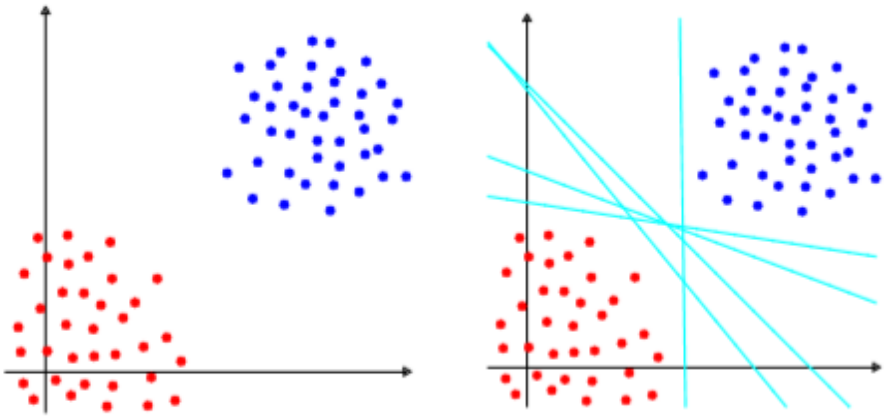
O hiperplano $\mathcal{H}(w, b)$ divide o espaço $\mathbb{R}^n$ em dois semiespaços, dados por

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

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.

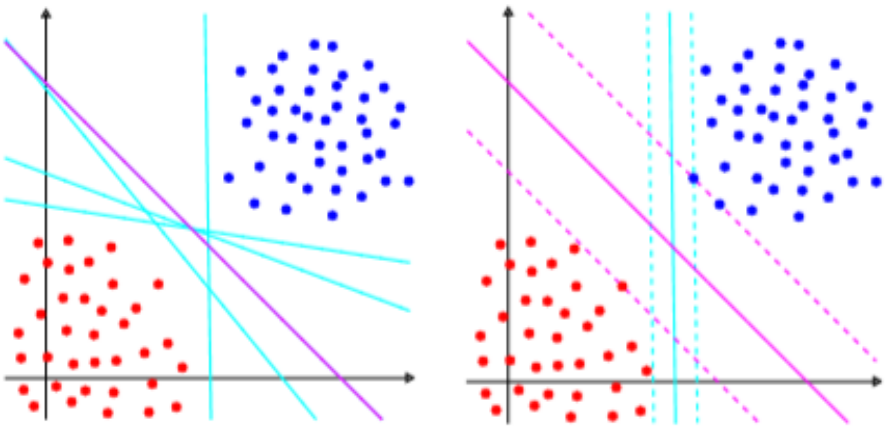
Definição 2. Os conjuntos $\mathcal{X}^+, \mathcal{X}^- \subset \mathbb{R}^n$ são ditos linearmente separáveis quando existem $w \in \mathbb{R}^n$ e $b \in \mathbb{R}$ tais que $w^T x + b > 0$ para todo $x \in \mathcal{X}^+$ e $w^T x + b < 0$ para todo $x \in \mathcal{X}^-$. O hiperplano $\mathcal{H}(w,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



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

Figura 3: Hiperplano Ótimo.
Fonte: Krulikovski [2]

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 $b \in \mathbb{R}$.

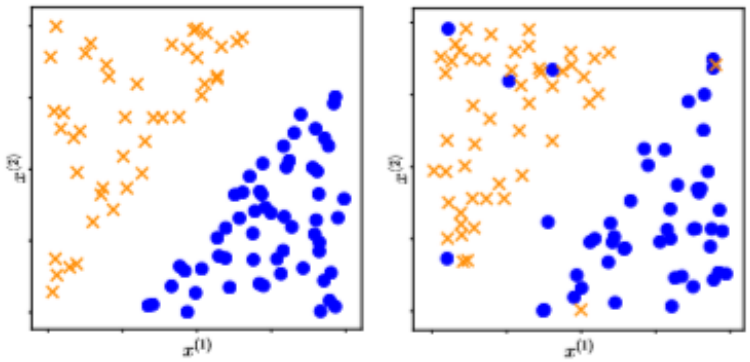
2 Máquinas de Vetores Suporte - CSVM

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

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

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



(a) Dados linearmente separáveis.

(b) Dados não linearmente separáveis.

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

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.

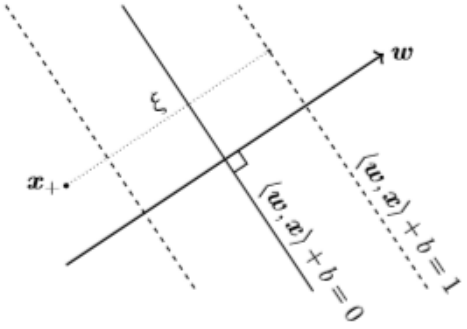


Figura 5: Variáveis de folga
Fonte: Deisenroth, Faisal e Ong [1]

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 [2], a **??** 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).

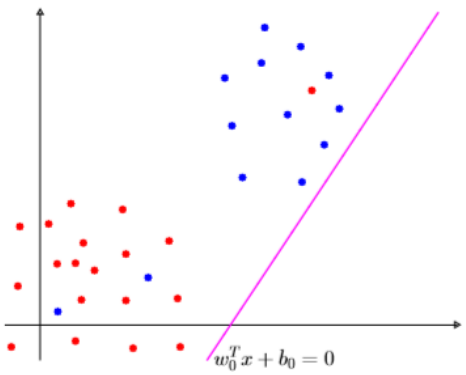


Figura 6: Pensar.
Fonte: Krulikovski [2]

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

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

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

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

Referências

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

[2] 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.