



**Simpósio de Métodos
Numéricos em Engenharia**

25 a 27 de outubro, 2017

Análise teórica de Máquinas de Vetores Suporte

Evelin Heringer Manoel Krulikowski¹
Orientadores: Profa. Dra. Mael Sachine
Prof. Dr. Ademir Alves Ribeiro

Programa de Pós-Graduação em Matemática
Universidade Federal do Paraná
Curitiba, Brasil
evelin.hmk@gmail.com¹

Resumo—O objetivo geral deste trabalho foi realizar um estudo teórico sobre Máquinas de Vetores Suporte, que inclui relatar justificativas para o uso de tal técnica e exibir sua interpretação geométrica e perspectiva analítica. Para aplicar a técnica em problemas de classificação, buscamos fundamentar matematicamente sua utilização, visto que envolve um problema de programação quadrática, convexa e com restrições. Para a análise da técnica, utilizamos a teoria de dualidade Lagrangiana, que notamos facilitar os cálculos e a análise das soluções. Estabelecemos algumas implicações e exibimos alguns contraexemplos, para mostrar que certos resultados decorrentes da técnica encontrados na literatura não são precisos.

Palavras-chave—Máquinas de Vetores Suporte; Programação não linear; Otimização com restrições; Dualidade Lagrangiana;

I. INTRODUÇÃO

Vários campos da ciência fazem uso da Otimização para ajudar na tomada de decisões. Em particular, isso é observado na Aprendizagem de Máquina, que tem apresentado a intenção de criar um algoritmo tão eficiente em decisão quanto o cérebro humano. Neste trabalho, temos como objetivo realizar um estudo teórico de uma técnica supervisionada de Aprendizagem de Máquina: as Máquinas de Vetores Suporte (SVM, do Inglês *Support Vector Machine*).

A técnica SVM foi desenvolvida por Vladimir Vapnik, Bernhard Boser, Isabelle Guyon e Corrina Cortes, com seus fundamentos provenientes da Teoria de Aprendizagem Estatística [4]–[6], que

vem despertando interesse nos últimos anos. Os livros de Nello Cristianini e John Shawe-Taylor [2], [3] são referências muito utilizadas, na primeira tem uma introdução para tal técnica. Na sequência apresentamos as principais ideias sobre SVM.

Tal técnica concentra a atenção em um problema de programação quadrática, convexa e com restrições do tipo

$$\begin{aligned} \min_{w,b} \quad & f(w, b) \quad \text{com } w \in \mathbb{R}^n \text{ e } b \in \mathbb{R} \\ \text{s.a.} \quad & g(w, b) \leq 0, \end{aligned} \quad (1)$$

onde as funções $f : \mathbb{R}^{n+1} \rightarrow \mathbb{R}$ e $g : \mathbb{R}^{n+1} \rightarrow \mathbb{R}$ são continuamente diferenciáveis. Neste trabalho, usamos SVM para classificação. O objetivo é obter um classificador linear $\text{sign}(w^T x + b)$, a partir da solução (w, b) . A Fig. 1 representa um problema de classificação que pode ser modelado pelo problema (1), o qual estamos interessados em resolver. A partir da solução deste problema construímos um hiperplano definido pela equação $w^T x + b = 0$, que será usado para classificação de novos dados.

Trataremos da modelagem SVM com margem rígida [5] e suas generalizações, que são margem flexível [4] e não linear [6], conforme Fig. 1, Fig. 2 e Fig. 3, respectivamente. Por questão de espaço, omitimos algumas demonstrações.

A. Principais contribuições

- Abordamos um resultado que não é tão clássico: Teorema 2.3, que trata da existência de solução global para programação quadrática.

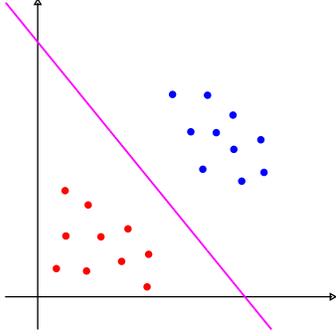


Figura 1: Margem rígida.

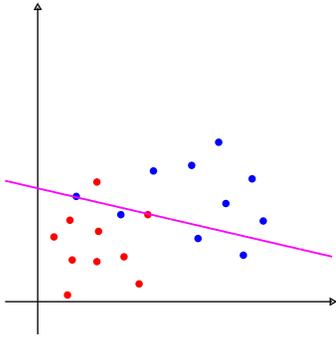


Figura 2: Margem flexível.

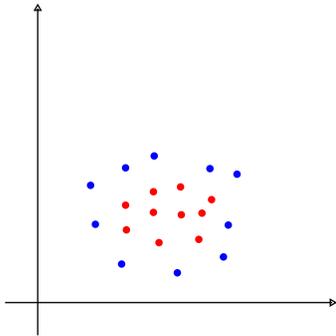


Figura 3: Não linear.

- Fundamentamos teoricamente a existência e a unicidade das soluções dos problemas primal e dual, obtidos em SVM.
- Discutimos a respeito das definições de vetores suporte, presentes na literatura [1], [3], tentando deixar clara a diferença entre elas.
- Provamos que ao retirar pontos que não são vetores suporte o classificador não muda.
- Apresentamos modificações nos resultados sobre a unicidade da solução primal, para SVM com margem flexível, encontrados em [8].
- Enunciamos resultados sobre as funções *Kernel*.

II. CONCEITOS IMPORTANTES

Os próximos resultados serão importantes para estabelecer uma propriedade muito útil das funções *Kernel*, que definiremos adiante.

Teorema 2.1: Sejam $\mathcal{X} \subset \mathbb{R}^n$ um conjunto compacto e $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ uma função contínua e simétrica tal que

$$\int_{\mathcal{X} \times \mathcal{X}} K(x, z) f(x) f(z) dx dz \geq 0$$

para toda $f \in L_2(\mathcal{X})$. Então, dados $x^1, \dots, x^m \in \mathcal{X}$ e $c_1, \dots, c_m \in \mathbb{R}$, temos

$$\sum_{i,j=1}^m K(x^i, x^j) c_i c_j \geq 0.$$

Vejamos que a recíproca do corolário também é verdadeira.

Teorema 2.2: Sejam $\mathcal{X} \subset \mathbb{R}^n$ um conjunto compacto e $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ uma função contínua e simétrica. Suponha que

$$\sum_{i,j=1}^m K(x^i, x^j) c_i c_j \geq 0$$

para todos $x^1, \dots, x^m \in \mathcal{X}$, $m \in \mathbb{N}$ e $c_1, \dots, c_m \in \mathbb{R}$. Então,

$$\int_{\mathcal{X} \times \mathcal{X}} K(x, z) f(x) f(z) dx dz \geq 0$$

para toda $f \in L_2(\mathcal{X})$.

No que segue, trataremos de condições que garantem a existência de minimizador global para um problema quadrático. Consideremos um problema de Otimização no qual a função objetivo é quadrática e o conjunto viável é um poliedro. Isto é,

$$\begin{aligned} \min_x \quad & f(x) = x^T H x + c^T x \\ \text{s.a} \quad & A x + b \leq 0, \end{aligned} \quad (2)$$

onde $H \in \mathbb{R}^{n \times n}$ é simétrica, $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$ e $b \in \mathbb{R}^m$. Este problema pode não ter solução. No entanto, se f for limitada inferiormente, então é garantida a existência de um minimizador. Este resultado não é muito clássico e se deve a Frank e Wolfe [9]. Uma demonstração alternativa deste é apresentada por Blum e Oettli [7].

Teorema 2.3: Considere o problema quadrático (2), suponha que o seu conjunto viável seja não vazio e que a função objetivo seja limitada inferiormente neste conjunto. Então o problema tem um minimizador global.

A seguir, veremos o desenvolvimento de SVM com dados linearmente separáveis e nas seções seguintes suas generalizações. Consideramos em todos os casos o conjunto de treinamento

$$X = \{(x^1, y_1), \dots, (x^m, y_m), x^i \in \mathbb{R}^n, y_i \in \{-1, 1\}\}, \text{ com}$$

$$X_1 = \{x^i \in X \mid y_i = 1\} \quad \text{da classe positiva e}$$

$$X_2 = \{x^i \in X \mid y_i = -1\} \quad \text{da classe negativa.}$$

III. MÁQUINAS DE VETORES SUPORTE - MARGEM RÍGIDA

Primeiramente, trataremos sobre o problema com margem rígida, ou seja, quando os dados são linearmente separáveis. Para tanto, vamos definir alguns conceitos sobre hiperplano e formular o problema com o objetivo de separar os dados da melhor maneira.

Considere a Fig. 4, na qual os pontos em azul representam dados de treinamento da classe positiva e os em vermelho dados de treinamento da classe negativa.

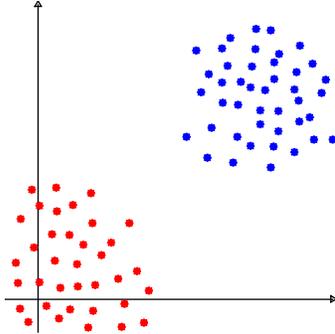


Figura 4: Dados de treinamento.

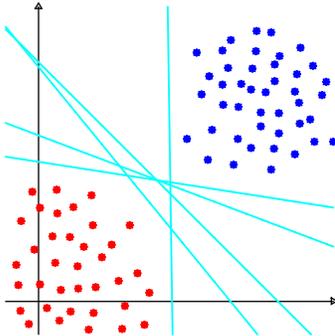


Figura 5: Hiperplanos separadores.

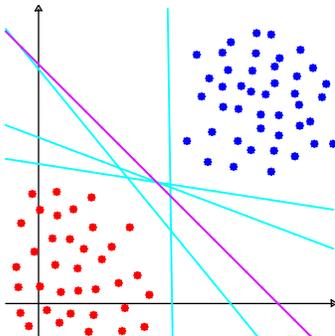


Figura 6: Hiperplano ótimo.

Note que todos os hiperplanos representados na Fig. 5 separam os dados. No entanto, dentre todos os hiperplanos separadores, aquele

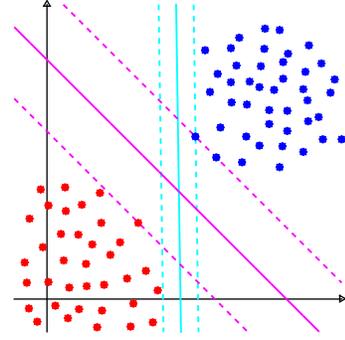


Figura 7: Máxima margem.

que está destacado na cor magenta, na Fig. 6, é o que melhor separa os dados pois possibilita a maior faixa que não contém nenhum atributo, conforme Fig. 7. Vamos agora formalizar estas ideias com o objetivo de modelar matematicamente o problema de classificação.

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

$$\begin{aligned} \min_{w,b} \quad & (1/2)\|w\|^2 \\ \text{s.a} \quad & y_i(w^T x^i + b) \geq 1, \quad i = 1, \dots, m, \end{aligned} \quad (3)$$

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

Como a função objetivo é limitada inferiormente e o conjunto viável $\Omega = \{(w, b) \in \mathbb{R}^{n+1} \mid g_i(w, b) \leq 0, i = 1, \dots, m\}$ é um poliedro não vazio (pois os conjuntos X_1 e X_2 são linearmente separáveis), o Teorema 2.3 garante a existência de um minimizador global para o problema (3). Também podemos garantir a unicidade da solução. Para mostrar este resultado, precisamos do seguinte lema.

Lema 3.1: Se (w^*, b^*) é ótimo para o problema (3) então existem $\bar{x} \in X_1$ e $\tilde{x} \in X_2$ tais que $(w^*)^T \bar{x} + b^* = 1$ e $(w^*)^T \tilde{x} + b^* = -1$.

Teorema 3.2: O minimizador global para o problema (3) é único.

Demonstração. Suponha que (\bar{w}, \bar{b}) e (\tilde{w}, \tilde{b}) são soluções ótimas para o problema (3). Então $(1/2)\|\bar{w}\|^2 = (1/2)\|\tilde{w}\|^2$ e consequentemente $\|\bar{w}\| = \|\tilde{w}\|$. Como Ω é um conjunto convexo, temos que $(\hat{w}, \hat{b}) = (1/2) \left((\bar{w}, \bar{b}) + (\tilde{w}, \tilde{b}) \right)$ pertence a Ω . Além disso, temos que $\|\hat{w}\| < \|\bar{w}\|$ o que implica que $f(\hat{w}, \hat{b}) < f(\bar{w}, \bar{b})$, contradizendo o fato de (\bar{w}, \bar{b}) ser uma solução ótima. Portanto, temos a unicidade de w .

Agora, mostraremos a unicidade de b . Suponha que (w, \bar{b}) e (w, \tilde{b}) são soluções ótimas para o problema (3) com, por exemplo, $\bar{b} < \tilde{b}$. Pelo Lema 3.1, existe $x \in X_1$ tal que $w^T x + \bar{b} = 1$. Por outro lado,

$$1 \leq w^T x + \tilde{b} < w^T x + \bar{b} = 1,$$

o que é uma contradição. \square

Assim, depois de resolvido o problema primal (3) e encontrados w^* e b^* , podemos classificar um novo dado $x \in \mathbb{R}^n$ usando a função

de decisão $F : \mathbb{R}^n \rightarrow \mathbb{R}$ dada por

$$F(x) = \text{sinal}((w^*)^T x + b^*). \quad (4)$$

Caso $F(x) > 0$, o ponto x será classificado como da classe positiva e se $F(x) < 0$, o ponto x será classificado como da classe negativa.

Consideremos a situação ilustrada na Fig. 8. Ao resolver o problema primal associado, encontramos a solução (w^*, b^*) . Mas, como será provado no Teorema 3.3, apenas três pontos definem completamente o hiperplano separador. A formulação primal só detecta os pontos desnecessários após resolver o problema primal e avaliar todas as restrições na solução, enquanto a formulação dual consegue detectar isso apenas com a positividade da sua solução, como mostraremos na sequência do texto.

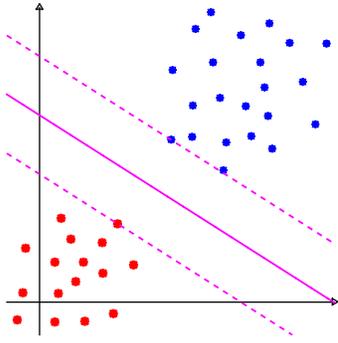


Figura 8: Dados sobre os hiperplanos tracejados correspondem as restrições ativas.

Teorema 3.3: Considere (w^*, b^*) a solução do problema (3) e $I^* = I(w^*, b^*)$ o conjunto dos índices das restrições ativas na solução. Então a solução do problema

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2} \|w\|^2 \\ \text{s.a.} \quad & y_i(w^T x^i + b) \geq 1, \quad i \in I^* \end{aligned} \quad (5)$$

é (w^*, b^*) .

Demonstração. Seja (\bar{w}, \bar{b}) a solução do problema (5). Como o conjunto viável do problema (3) está contido no conjunto viável do problema (5), temos que $f(\bar{w}, \bar{b}) \leq f(w^*, b^*)$. Considere agora, para cada $t \in (0, 1)$,

$$(w_t, b_t) = (1-t)(w^*, b^*) + t(\bar{w}, \bar{b}).$$

Para $i \in I^*$, a convexidade da restrição g_i e a viabilidade das soluções garantem que

$$g_i(w_t, b_t) \leq (1-t)g_i(w^*, b^*) + tg_i(\bar{w}, \bar{b}) \leq 0.$$

Por outro lado, para $i \notin I^*$, temos $g_i(w^*, b^*) < 0$ e portanto, pela continuidade de g_i , vale $g_i(w_t, b_t) < 0$ para t suficientemente pequeno. Portanto, (w_t, b_t) é viável para o problema (3). Afirmamos que $f(\bar{w}, \bar{b}) = f(w^*, b^*)$. De fato, se $f(\bar{w}, \bar{b}) < f(w^*, b^*)$, então

$$\|w_t\| = \|(1-t)w^* + t\bar{w}\| \leq (1-t)\|w^*\| + t\|\bar{w}\| < \|w^*\|,$$

implicando $f(w_t, b_t) < f(w^*, b^*)$, o que é uma contradição. Concluímos então que (\bar{w}, \bar{b}) e (w^*, b^*) são viáveis para o problema (5) com o mesmo valor (ótimo) da função objetivo, ou seja, ambos são minimizadores globais do problema (5). Pelo Teorema 3.2, segue que $(\bar{w}, \bar{b}) = (w^*, b^*)$. \square

Vamos agora apresentar o dual do problema (3),

$$\begin{aligned} \max_{\alpha} \quad & -\frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j (x^i)^T x^j + \sum_{i=1}^m \alpha_i \\ \text{s.a.} \quad & \sum_{i=1}^m \alpha_i y_i = 0, \\ & \alpha_i \geq 0 \quad i = 1, \dots, m. \end{aligned} \quad (6)$$

Este problema tem solução e o *gap* de dualidade é zero. Assim, depois de encontrado α^* , podemos facilmente obter w^* por

$$w^* = \sum_{i=1}^m \alpha_i^* y_i x^i = \sum_{i \in I} \alpha_i^* y_i x^i. \quad (7)$$

Além disso, pela complementariedade, temos

$$b^* = \frac{1 - y_i (w^*)^T x^i}{y_i} = y_i - \sum_{j \in I} \alpha_j^* y_j (x^j)^T x^i, \quad (8)$$

para dado $i \in I = \{i \mid \alpha_i^* > 0\}$.

Note que uma solução do dual consegue detectar de modo simples quais vetores do conjunto de dados são desnecessários para a obtenção da solução primal (w^*, b^*) : todos os pontos x^i tais que $i \notin I$.

Veremos agora uma outra forma de obter a solução primal, tanto teoricamente quanto computacionalmente mais simples, conforme o que será estabelecido no Teorema 3.4. Para tanto, se $I = \{i_1, \dots, i_p\}$, vamos denotar $\alpha_I = (\alpha_{i_1}, \dots, \alpha_{i_p})^T \in \mathbb{R}^p$.

Teorema 3.4: Considere (w^*, b^*) a solução do problema (3), $\alpha^* \in \mathbb{R}_+^m$ uma solução global arbitrária do problema dual (6) e $I = \{i_1, \dots, i_p\} = \{i \mid \alpha_i^* > 0\}$. Então (w^*, b^*) é a solução do problema primal reduzido

$$\begin{aligned} \min_{w,b} \quad & (1/2) \|w\|^2 \\ \text{s.a.} \quad & y_i(w^T x^i + b) \geq 1, \quad i \in I \end{aligned} \quad (9)$$

e α_I^* é uma solução global do problema dual reduzido de (9),

$$\begin{aligned} \max_{\beta \in \mathbb{R}^p} \quad & -\frac{1}{2} \sum_{k,l=1}^p \beta_k \beta_l y_{i_k} y_{i_l} (x^{i_k})^T x^{i_l} + \sum_{k=1}^p \beta_k \\ \text{s.a.} \quad & \sum_{k=1}^p \beta_k y_{i_k} = 0, \\ & \beta_k \geq 0, \quad k = 1, \dots, p. \end{aligned} \quad (10)$$

Demonstração. Considere $\theta : \mathbb{R}^m \rightarrow \mathbb{R}$ a função objetivo do problema dual (6) e $\theta_r : \mathbb{R}^p \rightarrow \mathbb{R}$ a função objetivo do problema dual reduzido (10). Dado $\beta \in \mathbb{R}^p$ satisfazendo as restrições deste problema, defina

o vetor $\alpha \in \mathbb{R}^m$ por

$$\alpha_i = \begin{cases} \beta_k, & \text{se } i = i_k \\ 0, & \text{se } i \notin I. \end{cases}$$

Desta forma temos $\alpha \geq 0$ e

$$\sum_{i=1}^m \alpha_i y_i = \sum_{k=1}^p \beta_k y_{i_k} = 0,$$

o que significa que α é viável para o problema (6). Portanto, $\theta(\alpha^*) \geq \theta(\alpha)$. Por outro lado, temos que α_I^* é viável para o problema (10) (pois α^* é viável para o problema (6)) e

$$\theta_r(\alpha_I^*) = \theta(\alpha^*) \geq \theta(\alpha) = \theta_r(\beta),$$

onde segue a segunda afirmação do teorema. Para provar a outra afirmação, note que a solução de (9), que vamos denotar (\bar{w}, \bar{b}) , pode ser obtida a partir de uma solução do seu dual (10), ou seja,

$$\bar{w} = \sum_{i \in I} \alpha_i^* y_i x^i = w^* \quad \text{e} \quad \bar{b} = y_i - \bar{w}^T x^i = b^*.$$

□

Mostramos que a solução do problema primal (3) é única. Entretanto, não podemos garantir a unicidade da solução do problema dual (6), conforme o Exemplo 3.5.

Exemplo 3.5: Considere o seguinte conjunto de treinamento,

$$TR = \{(0, 1), (1, 1), (2, 1), (0, -1), (1, -1), (2, -1)\},$$

com saídas $\{1, 1, 1, -1, -1, -1\}$, respectivamente, representado pela Fig. 9. Os pontos em vermelho e azul representam as classes negativa e positiva, respectivamente.

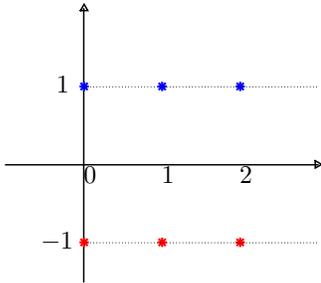


Figura 9: Conjunto linearmente separável.

Resolvendo o problema dual associado, podemos obter soluções distintas, dentre as quais, destacamos

$$\alpha^* = \frac{1}{6} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad \hat{\alpha} = \begin{bmatrix} 0 \\ 0.5 \\ 0 \\ 0.25 \\ 0 \\ 0.25 \end{bmatrix} \quad \text{e} \quad \tilde{\alpha} = \begin{bmatrix} 0.5 \\ 0 \\ 0 \\ 0.5 \\ 0 \\ 0 \end{bmatrix}.$$

À luz deste exemplo vamos apresentar agora duas definições de vetores suporte, bastante presentes na literatura [1], [3], e fazer uma discussão a respeito delas com o objetivo de propor algumas melhorias neste conceito.

Definição 3.6: Considere um conjunto X de vetores linearmente separáveis e (w^*, b^*) a solução do problema primal (3). Os vetores suporte são os vetores $x^i \in X$ tais que $y_i((w^*)^T x^i + b^*) = 1$.

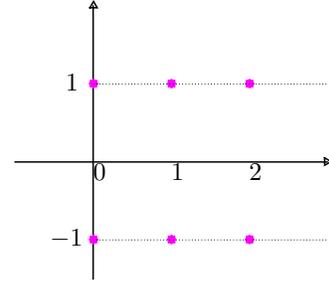


Figura 10: Dados em magenta são vetores suporte, segundo Definição 3.6.

Esta definição, apesar de ser intrínseca aos dados do problema, contempla vetores que não trazem contribuição para a obtenção do hiperplano separador. De fato, conforme o Exemplo 3.5, podemos ter vetores satisfazendo $y_i((w^*)^T x^i + b^*) = 1$ com multiplicador correspondente igual a zero. Isto motiva uma segunda definição de vetores suporte.

Definição 3.7: Considere um conjunto X de vetores linearmente separáveis e α^* uma solução do problema dual (6). Os vetores suporte são os vetores $x^i \in X$ tais que $\alpha_i^* > 0$.

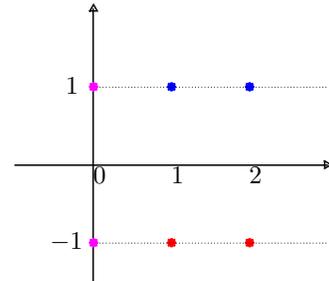


Figura 11: Dados em magenta são vetores suporte, segundo Definição 3.7.

Nesta seção, discutimos SVM apenas para dados linearmente separáveis, obtendo a máxima margem entre eles. Entretanto, são raros os problemas reais onde os dados sejam linearmente separáveis. Nas próximas seções, trataremos os problemas onde isso não ocorre.

IV. MÁQUINAS DE VETORES SUPORTE - MARGEM FLEXÍVEL

Consideramos agora dados não linearmente separáveis, como por exemplo, o da Fig. 2. Para este problema relaxamos as restrições adicionando variáveis de folga e, adicionamos na função objetivo do problema (3) um termo para controlar o valor destas variáveis. Isto pode ser feito considerando o problema

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

onde $C > 0$ é um parâmetro de regularização que serve para controlar a importância das variáveis de folga.

Conseguimos garantir a existência para o problema (11). Mas, somente sob certas condições podemos garantir a unicidade da solução. A componente w^* de qualquer solução é única. Este resultado pode ser encontrado em [8], mas por conveniência apresentamos aqui com algumas modificações no enunciado e na demonstração.

Teorema 4.1: Sejam (w^*, b^*, ξ^*) e $(\bar{w}, \bar{b}, \bar{\xi})$ soluções arbitrárias do problema (11). Então $w^* = \bar{w}$.

Demonstração. Como o problema (11) é convexo, temos

$$f(w^*, b^*, \xi^*) = f(\bar{w}, \bar{b}, \bar{\xi}) = f^* \stackrel{\text{def}}{=} \inf_{(w,b,\xi) \in \Omega} f(w, b, \xi).$$

Mais ainda, definindo $(w_t, b_t, \xi_t) = (1-t)(w^*, b^*, \xi^*) + t(\bar{w}, \bar{b}, \bar{\xi})$, temos que

$$\frac{1}{2}\|w_t\|^2 + C \sum_{i=1}^m (\xi_t)_i = f(w_t, b_t, \xi_t) = f^*$$

para todo $t \in [0, 1]$. Portanto,

$$\frac{1}{2} \sum_{i=1}^n ((1-t)w_i^* + t\bar{w}_i)^2 + C \sum_{i=1}^m ((1-t)\xi_i^* + t\bar{\xi}_i) = f^*$$

para todo $t \in [0, 1]$. Derivando em relação a t , obtemos

$$\sum_{i=1}^n ((1-t)w_i^* + t\bar{w}_i)(\bar{w}_i - w_i^*) + C \sum_{i=1}^m (\bar{\xi}_i - \xi_i^*) = 0.$$

Derivando novamente em relação a t , obtemos

$$\sum_{i=1}^n (\bar{w}_i - w_i^*)^2 = 0,$$

donde segue que $w^* = \bar{w}$. \square

Teorema 4.2: Temos $y_i((w^*)^T x^i + b^*) \geq 1$ se, e somente se, $\xi_i^* = 0$. Além disso, se $\xi_i^* > 0$, então $\xi_i^* = 1 - y_i((w^*)^T x^i + b^*)$.

Pelos Teoremas 4.1 e 4.2, o vetor ξ^* fica completamente determinado por w^* e b^* , uma vez que

$$\xi_i^* = \max\{0, 1 - y_i((w^*)^T x^i + b^*)\}.$$

Vamos agora apresentar os resultados tirados de [8] a respeito da unicidade da solução para o problema (11). Novamente por conveniência e também para esclarecer as ideias, reestruturamos os enunciados.

Teorema 4.3: Considere um conjunto de dados $X \subset \mathbb{R}^n$, com saída $y \in \mathbb{R}^n$, uma constante $C > 0$ e o problema primal (11) associado. Considere também uma solução (w^*, b^*, ξ^*) deste problema e os conjuntos de índices

$$\begin{aligned} N_1 &= \{i \mid y_i = 1 \text{ e } (w^*)^T x^i + b^* < 1\}, \\ N_2 &= \{i \mid y_i = -1 \text{ e } (w^*)^T x^i + b^* > -1\}, \\ N_3 &= \{i \mid y_i = 1 \text{ e } (w^*)^T x^i + b^* = 1\}, \\ N_4 &= \{i \mid y_i = -1 \text{ e } (w^*)^T x^i + b^* = -1\}, \\ N_5 &= \{i \mid y_i = 1 \text{ e } (w^*)^T x^i + b^* > 1\} \text{ e} \\ N_6 &= \{i \mid y_i = -1 \text{ e } (w^*)^T x^i + b^* < -1\}. \end{aligned}$$

Então a solução é única se, e somente se,

$$|N_2 \cup N_4| \neq |N_1| \quad \text{e} \quad |N_1 \cup N_3| \neq |N_2|. \quad (12)$$

Existe solução para o problema dual de (11). A partir da solução dual obtemos a solução primal,

$$w^* = \sum_{i=1}^m \alpha_i^* y_i x^i = \sum_{i \in I} \alpha_i^* y_i x^i, \quad (13)$$

onde $I = \{i \mid \alpha_i^* > 0\}$.

Para obtermos b^* e ξ^* , separamos em dois casos, conforme os resultados a seguir.

Teorema 4.4: Suponha que existe $j \in \{1, \dots, m\}$ tal que $0 < \alpha_j^* < C$. Então, o problema primal (11) possui uma única solução (w^*, b^*, ξ^*) , onde w^* é definido por (13),

$$b^* = y_j - (w^*)^T x^j \quad \text{e} \quad \xi_i^* = \begin{cases} 1 - y_i((w^*)^T x^i + b^*), & \text{se } i \in I \\ 0, & \text{se } i \notin I. \end{cases}$$

Teorema 4.5: Suponha que toda solução do problema dual de (11) satisfaz $\alpha_i = 0$ ou $\alpha_i = C$ para todo $i = 1, \dots, m$. Então este problema dual tem uma única solução α^* . Além disso, definindo $\beta^* = Ce - \alpha^*$ e $I = \{i \mid \alpha_i^* = C\}$, a(s) solução(ões) do problema primal (11) é(são) o(s) vetor(es) (w^*, b^*, ξ^*) , onde w^* é dado por (13) e b^* e ξ^* satisfazem

$$\begin{cases} y_i((w^*)^T x^i + b^*) + \xi_i^* = 1, & \text{se } i \in I \\ y_i((w^*)^T x^i + b^*) \geq 1, & \text{se } i \notin I \\ \xi_i^* = 0, & i \notin I \\ \xi_i^* \geq 0. \end{cases} \quad (14)$$

Veremos na próxima seção como lidar com casos nos quais mesmo introduzindo variáveis de folga não conseguimos obter um bom classificador.

V. MÁQUINAS DE VETORES SUPORTE - NÃO LINEAR

Para problemas em que o conjunto de dados não é linearmente separável, isto é, quando não é possível obter um hiperplano separador, nem mesmo usando regularização, Boser, Guyon e Vapnik [6], em 1992, sugeriram uma maneira de criar classificadores não lineares, aplicando o Truque do *Kernel*. Esta ferramenta permite mapear os dados do espaço original para um espaço de dimensão mais elevada (chamado *espaço de características*), na esperança de que neste novo espaço seja possível realizar uma classificação linear da mesma maneira que tratamos nas seções anteriores. Considere então uma função $\phi : \mathbb{R}^n \rightarrow \mathbb{R}^N$, com $N > n$ que mapeia os atributos $x^i \in \mathbb{R}^n$ do espaço de entrada para o espaço de características \mathbb{R}^N .

Exemplo 5.1: Na Fig. 12, temos um conjunto de treinamento $X = \{(-2, -2), (-2, 2), (-1, -1), (-1, 1), (0, 0), (1, -1), (1, 1), (2, -2), (2, 2)\}$ com saídas $\{-1, -1, 1, 1, 1, 1, 1, -1, -1\}$ respectivamente, que não é linearmente separável.

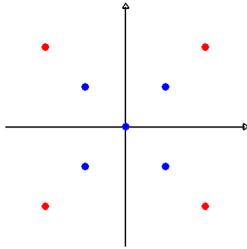


Figura 12: Dados não linearmente separáveis.

Usando o mapeamento $\phi(x) = (x_1, x_2, x_1^2 + x_2^2)^T$ obtemos um conjunto linearmente separável em \mathbb{R}^3 , conforme vemos na Fig. 13.

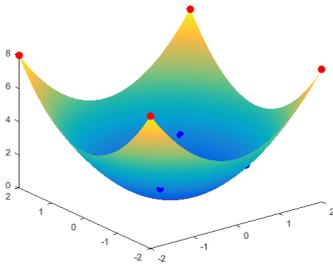


Figura 13: Dados mapeados em \mathbb{R}^3 .

Assim, podemos resolver o problema (3) trocando x^i por $\phi(x^i)$, obtendo um hiperplano ótimo em \mathbb{R}^3 , ilustrado na Fig. 14. Realizando a interseção do hiperplano ótimo em \mathbb{R}^3 com a superfície descrita por ϕ , como ilustrado na Fig. 15(a), obtemos uma superfície separando os dados no espaço de entrada, veja Fig. 15(b).

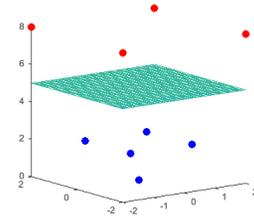
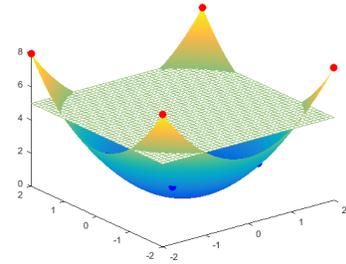
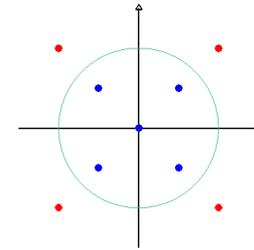


Figura 14: Hiperplano no espaço de características.



(a) Interseção Hiperplano e Superfície.



(b) Superfície separadora em \mathbb{R}^2 .

O problema primal com regularização associado aos dados mapeados no espaço de características pode ser escrito como

$$\begin{aligned} \min_{w,b,\xi} \quad & (1/2)\|w\|^2 + C \sum_{i=1}^m \xi_i \quad w \in \mathbb{R}^N, b \in \mathbb{R}, \xi \in \mathbb{R}^m \\ \text{s.a.} \quad & y_i(w^T \phi(x^i) + b) \geq 1 - \xi_i, \quad i = 1, \dots, m, \\ & \xi_i \geq 0, \quad i = 1, \dots, m. \end{aligned} \quad (15)$$

Note que se N for muito grande, o problema acima pode ficar difícil de trabalhar computacionalmente. Assim, usaremos a formulação dual do problema (15), dada por

$$\begin{aligned} \max_{\alpha} \quad & -(1/2) \sum_{i=1}^m \alpha_i \alpha_j y_i y_j \phi(x^i)^T \phi(x^j) + \sum_{i=1}^m \alpha_i \\ \text{s.a.} \quad & \sum_{i=1}^m \alpha_i y_i = 0, \\ & 0 \leq \alpha_i \leq C, i = 1, \dots, m. \end{aligned} \quad (16)$$

Notamos que todos os resultados estabelecidos nas Seções III e IV se aplicam aqui também, bastando trocar x^i por $\phi(x^i)$. No caso do problema dual (16), temos o produto interno $\phi(x^i)^T \phi(x^j)$. Assim, precisamos saber como calcular produtos internos no espaço de características e, para isso, usamos as chamadas funções *Kernel*.

Definição 5.2: Considere um conjunto $\mathcal{X} \subset \mathbb{R}^n$. Uma função $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ é chamada *Kernel* se existe $\phi : \mathcal{X} \rightarrow \mathbb{R}^N$ tal que

$$K(x, z) = \phi(x)^T \phi(z)$$

para todos $x, z \in \mathcal{X}$. Neste contexto, \mathbb{R}^N é chamado espaço de características e ϕ um mapeamento de \mathcal{X} em \mathbb{R}^N .

O Truque do *Kernel* consiste em substituir o produto interno $\phi(x^i)^T \phi(x^j)$ que aparece no problema (16) pelo número real $K(x^i, x^j)$. A vantagem é que o espaço de características não precisa ser construído explicitamente. Logo, a função de decisão é dada por

$$F(x) = \text{signal} \left(\sum_{i \in I} \alpha_i^* y_i K(x^i, x) + b^* \right).$$

Uma questão natural que surge agora é saber quais são as condições para que uma dada função $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ seja *Kernel*, sendo assim admissível para uso em SVM. Os dois resultados a seguir mostram algumas condições necessárias, clássicas na literatura [3].

Teorema 5.3: Considere um conjunto $\mathcal{X} \subset \mathbb{R}^n$ e suponha que $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ seja *Kernel*. Então,

- (a) K é simétrica, no sentido de que $K(x, z) = K(z, x)$;
- (b) Vale a desigualdade de Cauchy–Schwarz generalizada

$$K(x, z)^2 \leq K(x, x)K(z, z).$$

Teorema 5.4: Se $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ é *Kernel*, então a matriz de Gram associada a qualquer conjunto finito $X = \{x^1, \dots, x^m\} \subset \mathcal{X}$, definida por $G \in \mathbb{R}^{m \times m}$, $G(i, j) = K(x^i, x^j)$, é simétrica semidefinida positiva.

O interessante agora é que se o conjunto \mathcal{X} for compacto e K for contínua, então vale a recíproca do Teorema 5.4.

Teorema 5.5: Sejam $\mathcal{X} \subset \mathbb{R}^n$ um conjunto compacto e $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ uma função contínua. Suponha que para todo conjunto finito $X = \{x^1, \dots, x^m\} \subset \mathcal{X}$ a matriz de Gram associada seja simétrica semidefinida positiva. Então K é *Kernel*, mais precisamente, existe uma função $\phi : \mathcal{X} \rightarrow \ell_2$ tal que

$$K(x, z) = \langle \phi(x), \phi(z) \rangle = \sum_{k=1}^{\infty} \phi_k(x) \phi_k(z).$$

Outra forma de estabelecer condições suficientes é por meio do clássico Teorema de Mercer, apresentado a seguir. Mais detalhes podem ser encontrados em [3].

Teorema 5.6 (Mercer): Sejam $\mathcal{X} \subset \mathbb{R}^n$ um conjunto compacto e $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ uma função contínua e simétrica tal que o operador

integral $T_K : L_2(\mathcal{X}) \rightarrow L_2(\mathcal{X})$ dado por

$$T_K(f)(\cdot) = \int_{\mathcal{X}} K(\cdot, z) f(z) dz$$

é semidefinido positivo, ou seja,

$$\int_{\mathcal{X} \times \mathcal{X}} K(x, z) f(x) f(z) dx dz \geq 0, \quad (17)$$

para todo $f \in L_2(\mathcal{X})$. Então, existe uma função $\phi : \mathcal{X} \rightarrow \ell_2$ tal que

$$K(x, z) = \langle \phi(x), \phi(z) \rangle = \sum_{k=1}^{\infty} \phi_k(x) \phi_k(z).$$

Quando o conjunto \mathcal{X} não é finito podemos ainda estabelecer relações entre a positividade dada em (17) e a positividade das matrizes de Gram.

Teorema 5.7: Sejam $\mathcal{X} \subset \mathbb{R}^n$ um conjunto compacto e $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ uma função contínua e simétrica. Então,

$$\int_{\mathcal{X} \times \mathcal{X}} K(x, z) f(x) f(z) dx dz \geq 0,$$

para todo $f \in L_2(\mathcal{X})$ se, e somente se, a matriz de Gram associada a qualquer conjunto finito $X = \{x^1, \dots, x^m\} \subset \mathcal{X}$ é simétrica semidefinida positiva.

Demonstração. Segue direto do Teorema 2.1 e do Teorema 2.2, baseadas em integrais múltiplas, como somas de Riemann. \square

CONSIDERAÇÕES FINAIS

O foco deste trabalho foi realizar uma análise teórica matemática sobre a SVM, apresentando fundamentos de existência e unicidade para obtenção do hiperplano ótimo.

AGRADECIMENTOS

Agradeço a Profa. Mael Sachine e Prof. Ademir Alves Ribeiro. Também ao professor Celso Carnieri por ter apresentado a SVM.

REFERÊNCIAS

- [1] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.
- [2] N. Cristianini e J. Shawe-Taylor, *Kernel methods for pattern analysis*. Cambridge University Press, 2004.
- [3] N. Cristianini e J. Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press, 2000.
- [4] C. Cortes e V. N. Vapnik, *Support-vector networks*, Machine learning. Springer. v. 20. n. 3, pp. 273-297, 1995.
- [5] V. N. Vapnik, *Pattern recognition using generalized portrait method*. Automation and remote control. v. 24. pp. 774-780, 1963
- [6] I. M. Guyon e V. N. Vapnik, *A training algorithm for optimal margin classifiers*. ACM. pp. 144-152, 1992.
- [7] E. Blum e W. Oettli, *Technical Note – Direct proof of the Existence Theorem for Quadratic Programming*. Operations Research. v. 20. n. 1, pp. 165-167, 1972.
- [8] C. J. C. Burges et al, *Uniqueness of the SVM Solution*. NIPS. v. 99, pp. 223-229, 1999.
- [9] M. Frank e P. Wolfe, *An Algorithm for Quadratic Programming*. Naval Research Logistics Quarterly. v. 3. pp. 95-110, 1956.