Maximização de Margem

Victor Ruela

Programa de Pós-Graduação em Engenharia Elétrica
Universidade Federal de Minas Gerais

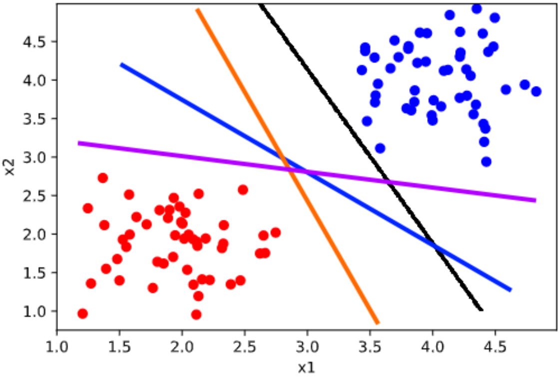
victorspruela@ufmg.br

9 de fevereiro de 2021

Agenda

Definição do Problema

- Dados de treinamento: $\mathcal{T} = \{(\mathbf{x}_1, d_1), \ldots, (\mathbf{x}_N, d_N)\}$
- Problema de classificação binário: $d_i = \{-1, 1\}$
- Linearmente separável
- Existem infinitos hiperplanos separadores!



Definição do Problema

- O uso do Perceptron, por exemplo, encontrará um hiperplano que separa estes padrões:

$$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b = 0 \tag{1}$$

- Entretanto, considerar a distância de todos os pontos para treinamento não é uma garantia de otimalidade
- E se fossem usados somente aqueles mais difíceis de serem separados e próximos da superfície de separação?

Maximização de Margem

Abordagem introduzida por Vapnik [Vapnik, 1992] para encontrar uma superfície de separação ótima com boa generalização, aplicável a diferentes modelos lineares

O Hiperplano Ótimo

- Quando a escolha de $\mathbf{w}$ e b maximizam a margem de separação (ρ), o hiperplano é dito **ótimo** [Haykin, 2007]:

$$\mathbf{w}_o^T \mathbf{x} + b_o = 0 \tag{2}$$

- Para otimalidade, o par $(\mathbf{w}_o, b_o)$ deve satisfazer:

$$\begin{cases} \mathbf{w}_o^T \mathbf{x}_i \geq 1, & d_i = +1 \\ \mathbf{w}_o^T \mathbf{x}_i \leq -1, & d_i = -1 \end{cases} \tag{3}$$

- Os pontos $(\mathbf{x}_i, d_i)$ que satisfazem com igualdade estas restrições são chamados de **vetores de suporte**
- Eles são os pontos mais próximos do hiperplano e consequentemente mais difíceis de classificar [Haykin, 2007].

O Hiperplano Ótimo

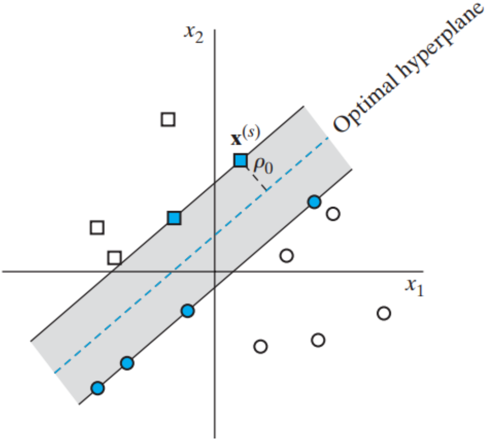


Figura: Representação do hiperplano ótimo e seus elementos. Extraído de [Haykin, 2007]

O Hiperplano Ótimo

- Pela definição do vetor de suporte $\mathbf{x}^{(s)}$:

$$\mathbf{w}_o^T \mathbf{x}^{(s)} + b_o = \mp 1 \tag{4}$$

- Logo, sua distância ao hiperplano separador é dada por:

$$r = \frac{g(\mathbf{x}^{(s)})}{\|\mathbf{w}_o\|} \tag{5}$$

$$= \begin{cases} \frac{1}{\|\mathbf{w}_o\|} & \text{se } d^{(s)} = +1 \\ \frac{-1}{\|\mathbf{w}_o\|} & \text{se } d^{(s)} = -1 \end{cases} \tag{6}$$

- Finalmente, a margem ótima ρ é:

$$\rho = \frac{2}{\|\mathbf{w}_o\|} \tag{7}$$

Conclusão

Maximizar a margem de separação entre classes binárias é equivalente a minimizar a norma Euclidiana do vetor de pesos $\mathbf{w}$

O Hiperplano Ótimo

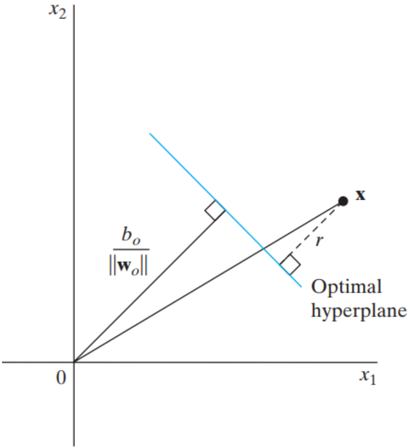


Figura: Distância de um ponto ao hiperplano ótimo. Extraído de [Haykin, 2007]

Formulação do Problema

- Combinando as Equações (4) e (7), podemos formular o problema de encontrar a margem ótima como:

$$\begin{aligned} \underset{\mathbf{w},b}{\text{minimizar}} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} \\ \text{sujeito a} \quad & d_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, \quad i = 1, 2, \ldots, N \end{aligned} \tag{8}$$

- Problema com objetivo quadrático e restrições lineares.
- Por ser convexo, possuirá solução única.
- É reformulado e resolvido através da técnica de multiplicadores de Lagrange [Haykin, 2007]
- Mais fácil de ser reolvido e fornece os vetores de suporte

Resultado

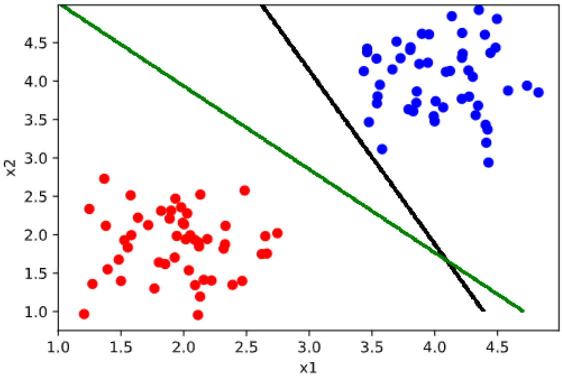


Figura: Superfícies de separação: MSE (preta) e margem máxima (verde)

Padrões Não Separáveis

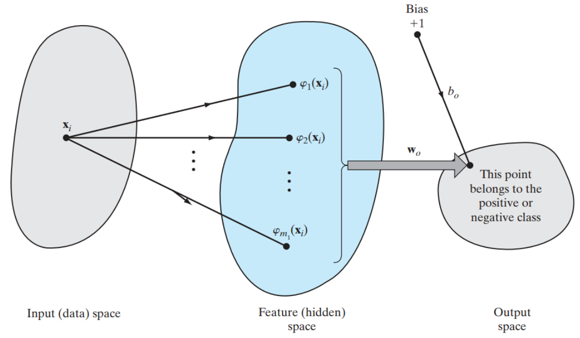
- Em aplicações práticas, não podemos garantir que os dados sejam perfeitamente separáveis
- Logo, é adicionada uma variável de folga ξ para representar estas discrepâncias:

$$\begin{aligned} \underset{\mathbf{w},b}{\text{minimizar}} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\xi_i \\ \text{sujeito a} \quad & d_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i \quad i = 1, 2, \ldots, N \\ & \xi_i \geq 0 \end{aligned} \tag{9}$$

- É reformulado e resolvido de forma similar ao Problema (9)
- A constante C controla a generalização do modelo
- Deve ser definida usando validação cruzada, por exemplo

Problemas não-linearmente separáveis

- Problemas reais nem sempre podem ser separados linearmente
- Entretanto, podemos mapeá-lo para um novo espaço de alta dimensão onde ele é mais provável de ser linearmente separável: **Teorema de Cover** [Cover, 1965]
- Este mapeamento é feito por funções não-lineares conhecidas como *Kernels*



Truque do Kernel

- No espaço original (primal):

$$g(\mathbf{x})^p = \mathbf{w}^T.\varphi(\mathbf{x}) + b \qquad (10)$$

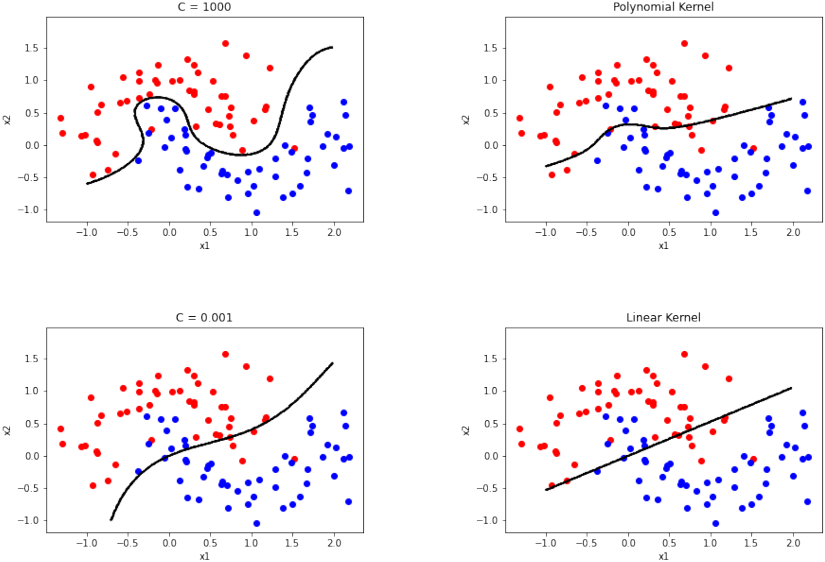
- Após a transformação para o espaço dual:

$$g(\mathbf{x})^d = \sum_{k=1}^{N} \alpha_k K(\mathbf{x}_k, \mathbf{x}) + b \qquad (11)$$

- Escolhendo um Kernel apropriado [Mercer, 1909], as Equações (10) e (11) são representações duais da mesma superfície de decisão, portanto:

$$w_i = \sum_{k=1}^{N} \alpha_k \varphi_i(\mathbf{x}_k) \qquad (12)$$

Exemplo - SVM


C = 1000
Polynomial Kernel
C = 0.001
Linear Kernel
x1
x2
−1.0 −0.5 0.0 0.5 1.0 1.5 2.0

Referências

- Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992, July). A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual workshop on Computational learning theory (pp. 144-152).
- Haykin, S. (2007). Neural Networks: A Comprehensive Foundation (3rd Edition).
- Cover, T. M. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. IEEE transactions on electronic computers, (3), 326-334.
- Mercer, J. (1909). Xvi. functions of positive and negative type, and their connection the theory of integral equations. Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character, 209(441-458), 415-446.

Obrigado!