2. Redes Neurais Artificiais

Prof. Renato Tinós

Depto. de Computação e Matemática (FFCLRP/USP)

2.5. Support Vector Machines

2.5. Support Vector Machines (SVM)

2.5.1. Hiperplano Ótimo

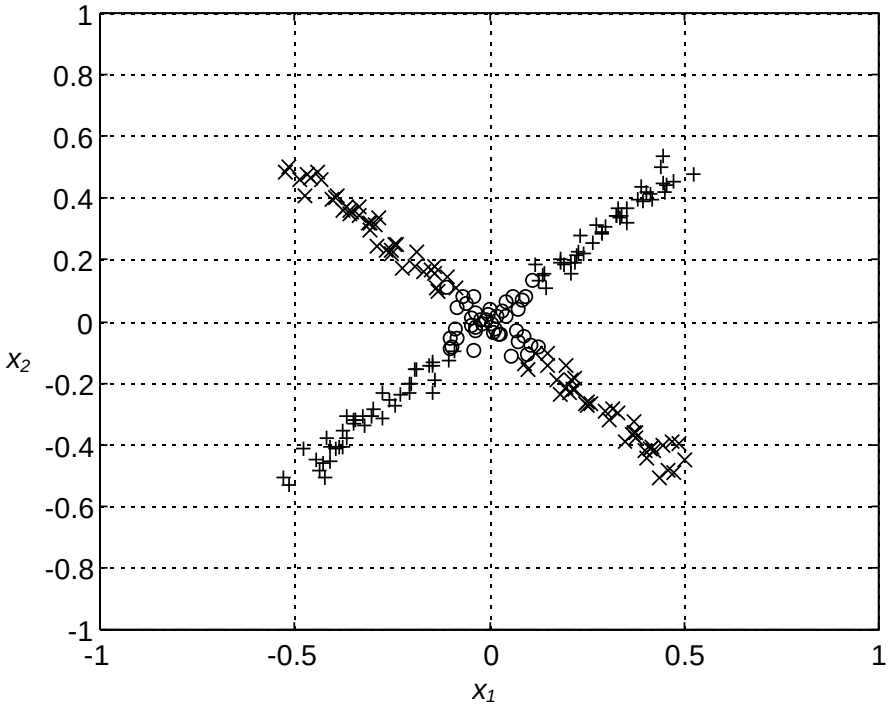
2.5.2. Kernels

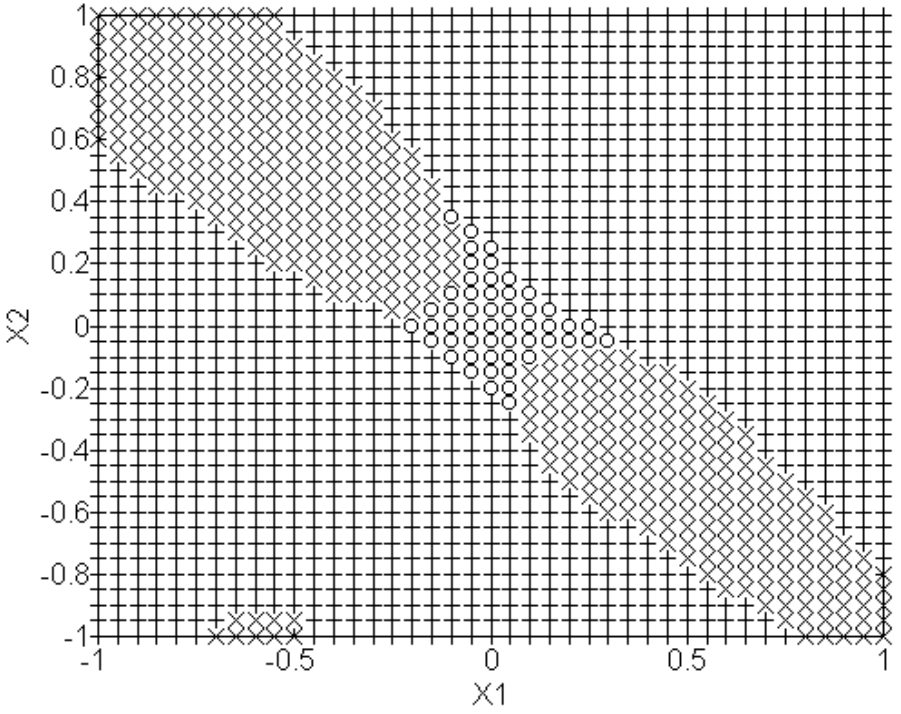
2.5.1. Support Vector Machines

- **Support Vector Machines** (**SVM**) ou **Máquinas de Vetor de Suporte**
  - São redes feedforward
    - » Propostas por Vapnik (1992 -1995)
    - » Bastante popular, especialmente na primeira década nos anos 2000
  - Como outros modelos vistos aqui, são aproximadores universais
    - » Podem portanto ser usadas em classificação e regressão

2.5.1. Hiperplano Ótimo

- **Como definir hiperplanos para o problema de classificação?**
  - Diversos modelos vistos aqui: os hiperplanos são otimizados de acordo com o erro médio quadrático sobre o conjunto de treinamento
    - » Exemplo: espaço de classificação gerado pelo MLP

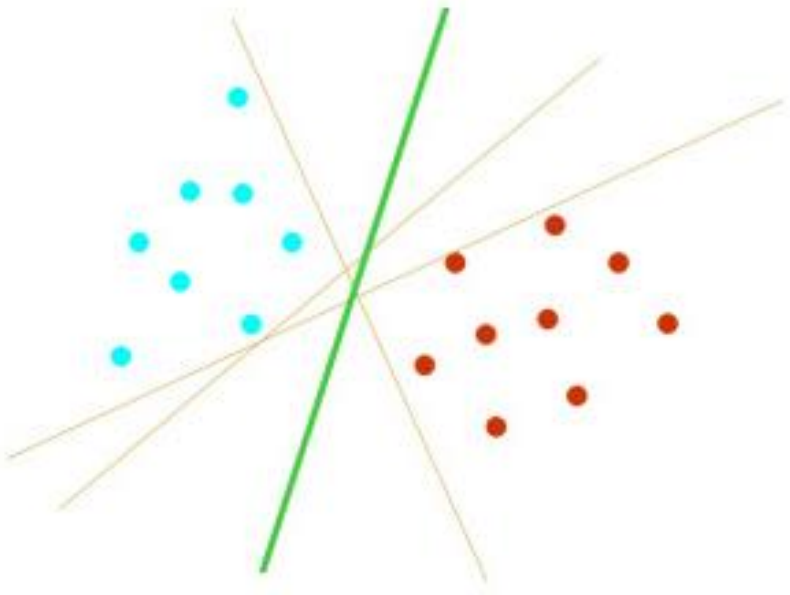




2.5.1. Hiperplano Ótimo

- **Problema de classificação linearmente separável**
  - Qual é o melhor hiperplano?

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



2.5.1. Hiperplano Ótimo

- **Ideia principal nas SVMs**

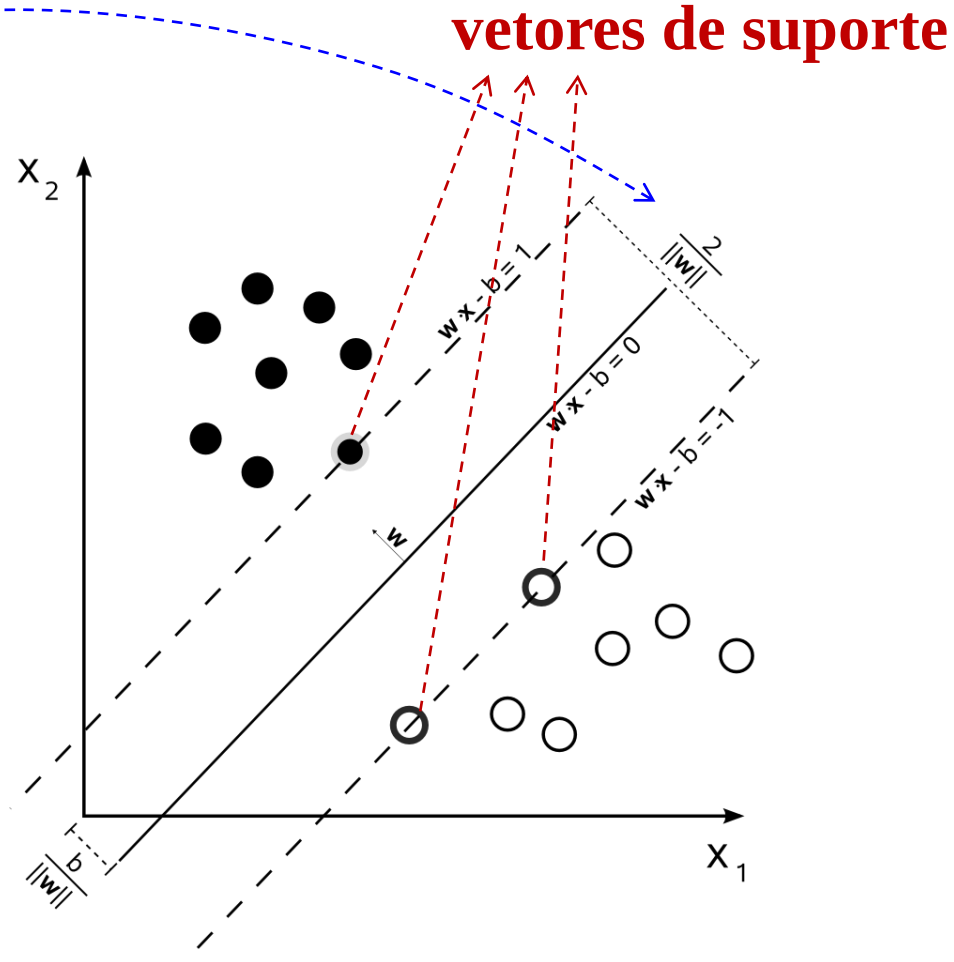
Construir um hiperplano de tal forma que a margem de separação entre exemplos positivos e negativos seja máxima

» É baseada no **método de minimização estrutural de riscos**

- ❑Fundamentada na teoria de aprendizagem estatística
- ❑Este princípio indutivo é baseado no fato de que a taxa de erro de uma máquina de aprendizagem sobre dados de teste (generalização) é limitada pela soma da taxa de erro de treinamento e por um termo que depende da dimensão de Vapnik-Chervonendis (V-C)

2.5.1. Hiperplano Ótimo

- **Margem de separação** (ρ)
  - Separação entre o hiperplano e o dado mais próximo
    - » A equação é dada na figura (ver desenvolvimento em [Haykin, 2001])



Fonte: Graphic showing the maximum separating hyperplane and the margin. https://commons.wikimedia.org/wiki/File:Svm_max_sep_hyperplane_with_margin.png

2.5.1. Hiperplano Ótimo

- **Como encontrar o hiperplano ótimo?**
  - Em SVM, a ideia é encontrar o vetor de pesos **w** que fornece a máxima separação possível entre exemplos positivos e negativos
    - » Maximizar ρ significa minimizar a norma Euclidiana de **w**
  - Fixando ρ=1, temos o seguinte problema de otimização para um conjunto de treinamento ($\mathbf{x}_i$,y_i) para i=1,....,n

Minimizar	$\|\| \mathbf{w} \|\|$
Sujeito às restrições:	$y_i (\mathbf{w}^T\mathbf{x}_i + b) \geq 1,$
	para i=1,....,n

2.5.1. Hiperplano Ótimo

- **Este é um problema de otimização restrito**
  - com função de custo convexa
  - com restrições lineares em relação à **w**

- **Portanto, podemos utilizar o método dos multiplicadores de Lagrange para encontrar os pontos ótimos**
  - Este é um grande atrativo de SVMs

2.5.1. Hiperplano Ótimo

- **Método de otimização**
  - Função Lagrangiana a ser otimizada

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i (y_i(\langle \mathbf{w} \cdot \mathbf{x_i} \rangle + b) - 1)$$

sendo que α_i são os multiplicadores de Lagrange

  - Soluções ótimas: pontos de sela da função Lagrangiana

$$\frac{\partial L}{\partial b} = 0$$

$$\frac{\partial L}{\partial \mathbf{w}} = 0$$

2.5.1. Hiperplano Ótimo

- **Método de otimização**
  - Resolvendo as equações (ver [Haykin, 2001]), obtemos o seguinte problemas

$$\text{Maximizar: } \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x_i} \cdot \mathbf{x_j}$$

$$\text{Sujeito a: } \begin{cases} \alpha_i \geq 0, \; i = 1, \; \ldots, \; n \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \end{cases}$$

  - Sendo que o hiperplano ótimo é definido por

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x_i}$$

$$b^* = -\frac{1}{2} \left[\max_{\{i | y_i = -1\}} (\mathbf{w}^* \cdot \mathbf{x_i}) + \min_{\{i | y_i = +1\}} (\mathbf{w}^* \cdot \mathbf{x_i}) \right]$$

$$= -\frac{1}{2} \left[\max_{\{i | y_i = -1\}} \left(\sum_{j=1}^{n} y_j \alpha_j^* \mathbf{x_i} \cdot \mathbf{x_j} \right) + \min_{\{i | y_i = +1\}} \left(\sum_{j=1}^{n} y_j \alpha_j^* \mathbf{x_i} \cdot \mathbf{x_j} \right) \right]$$

2.5.1. Hiperplano Ótimo

Algoritmo 3.1 Determinação do hiperplano ótimo para conjuntos linearmente separáveis (Vert, 2001).

1: Para cada conjunto de treinamento linearmente separável $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$

2: Seja $\alpha^* = (\alpha_1^*, \ldots, \alpha_n^*)$ a solução do seguinte problema de otimização com restrições:

3: Maximizar: $\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j \mathbf{x_i} \cdot \mathbf{x_j}$

4: Sob as restrições: $\begin{cases} \sum_{i=1}^{n} y_i \alpha_i = 0 \\ \alpha_i \geq 0, \, i = 1, \ldots, n \end{cases}$

5: O par $(\mathbf{w}^*, b^*)$ apresentado a seguir define o hiperplano ótimo.

6: $\mathbf{w}^* = \sum_{i=1}^{n} \alpha_i{}^* y_i \mathbf{x_i}$

7: $b^* = -1/2 \left[\max_{\{i|y_i=-1\}} (\mathbf{w}^* \cdot \mathbf{x_i}) + \min_{\{i|y_i=+1\}} (\mathbf{w}^* \cdot \mathbf{x_i}) \right]$

2.5.1. Hiperplano Ótimo

- **E quando os padrões não são linearmente separáveis?**

1. Construir margens de separação suaves

$$\textbf{Minimizar: } \|\mathbf{w}\|^2 + C \sum_{i=1}^{n} \varepsilon_i$$

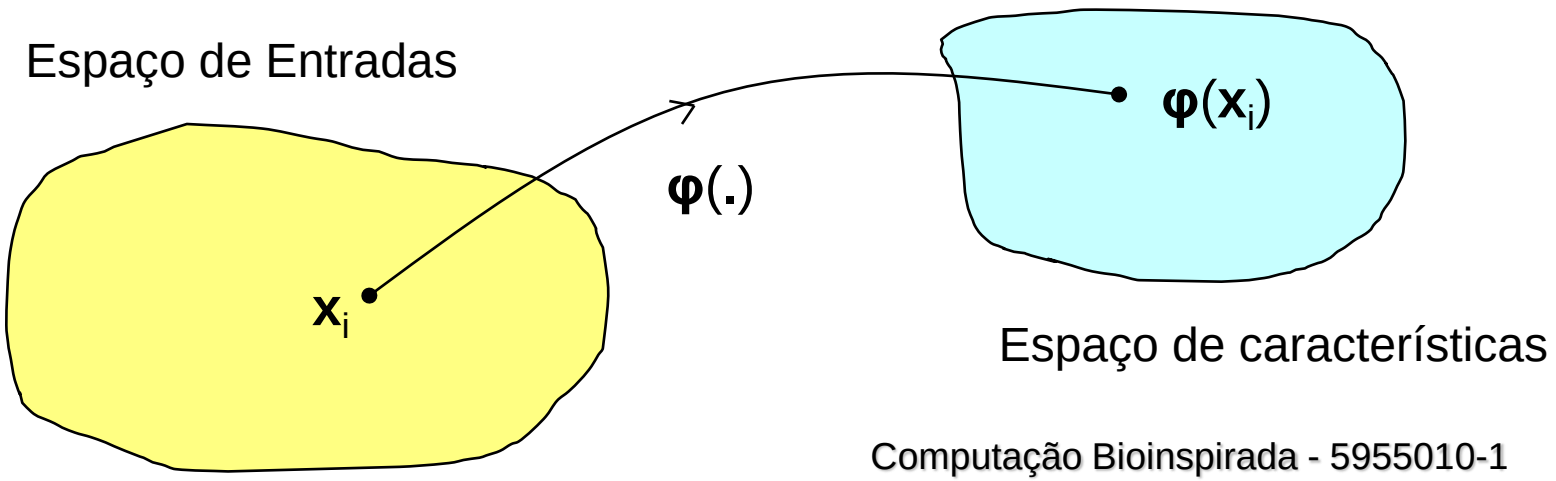
$$\textbf{Sob as restrições: } \begin{cases} \varepsilon_i \geq 0 \\ y_i \left(\mathbf{w} \cdot \mathbf{x_i} + b\right) \geq 1 - \varepsilon_i \end{cases}$$

2.5.1. Hiperplano Ótimo

- **E quando os padrões não são linearmente separáveis?**

2. Mapear não-linearmente o vetor de entrada para um espaço de característica de mais alta dimensionalidade

» Oculto dos espaços de entrada e saída



2.5.2. *Kernels*

- **Teorema de Cover sobre a separabilidade de padrões**

  "*Um problema complexo de classificação de padrões disposto não-linearmente em um espaço de alta dimensão tem maior probabilidade de ser linearmente separável do que em um espaço de baixa dimensionalidade*"

  (ver Seção 5.2 de [Haykin, 2001])

2.5.2. *Kernels*

- **Utilizando o vetor de características, ao invés do vetor de entradas, teríamos**
  - Hiperplano

$$\mathbf{w}^{\mathrm{T}} \cdot \boldsymbol{\varphi}(\mathbf{x}) = 0$$

  - Vetor para o hiperplano ótimo

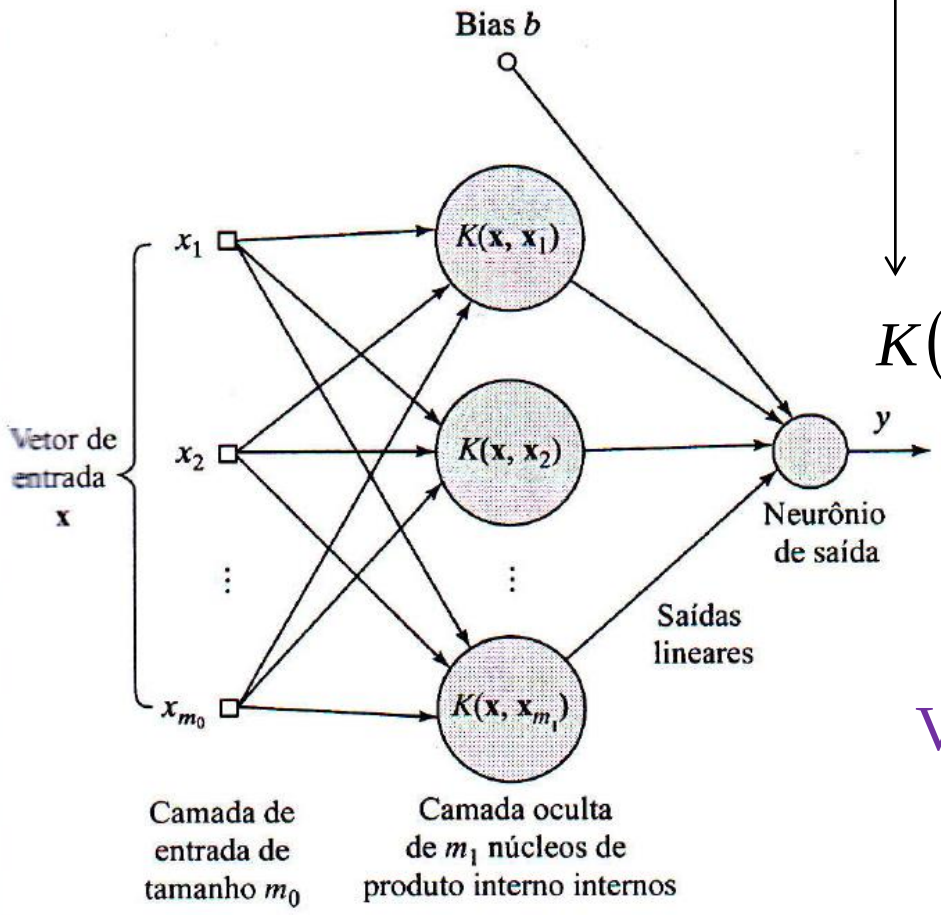
$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{\varphi}(\mathbf{x}_i)$$

    - » Substituindo na primeira equação

$$\sum_{i=1}^{N} \alpha_i y_i \boldsymbol{\varphi}(\mathbf{x}_i)^T \boldsymbol{\varphi}(\mathbf{x}) = 0$$

2.5.2. *Kernels*

*Núcleo do produto interno (**Kernel**)*



Dimensão do espaço de características

$$K(\mathbf{x},\mathbf{x}_i) = \boldsymbol{\varphi}^{\mathrm{T}}(\mathbf{x}) \cdot \boldsymbol{\varphi}(\mathbf{x}_i) = \sum_{j=0}^{m_1} \varphi_j(\mathbf{x}) \cdot \varphi_j(\mathbf{x}_i)$$

i-ésimo exemplo de treinamento

Vetor de características

FIGURA 6.5 Arquitetura da máquina de vetor de suporte

Fonte: Haykin, 2001

2.5.2. *Kernels*

- **Novo hiperplano**

$$\sum_{i=1}^{N} \alpha_i y_i K(\mathbf{x},\mathbf{x}_i) = 0$$

- **Kernels usuais**

TABELA 6.1 Resumo dos Núcleos de Produto Interno

Tipo de máquina de vetor de suporte	Núcleo de produto interno $K(\mathbf{x}, \mathbf{x}_i)$, $i = 1, 2, \ldots, N$	Comentários
Máquina de aprendizagem polinomial	$(\mathbf{x}^T\mathbf{x}_i + 1)^p$	A potência p é especificada *a priori* pelo usuário
Rede de função de base radial	$\exp\left(-\frac{1}{2\sigma^2}\|\mathbf{x} - \mathbf{x}_i\|^2\right)$	A largura σ^2, comum a todos os núcleos, é especificada *a priori* pelo usuário
Perceptron de duas camadas	$\tanh(\beta_0\mathbf{x}^T\mathbf{x}_i + \beta_1)$	O teorema de Mercer é satisfeito apenas para alguns valores de β_0 e β_1

2.5.2. *Kernels*

- **E quando existem várias classes?**
  - Solução usual:
    - » Decomposição um-contra-todos
      - ❑Gera um classificador para cada classe

Comentários

- **Referências**

  - **Haykin, S. S.. *Redes neurais: princípios e prática.* 2ª ed., Bookman, 2001.**
    - » Capítulo 6