

-
-
-
-
-

Máquinas de Vetores de Suporte

Supprot Vector Machine

Aluizio Fausto Ribeiro Araújo
Universidade Federal de Pernambuco
Centro de Informática



-
-
-
-
-

Conteúdo

1. Introdução
2. Classificadores Binários
3. Aprendizagem Estatística
4. SVM com Margens Rígidas
5. SVM com Margens Rígidas: Hiperplano Ótimo
6. SVM com Margens Rígidas: Método de Multiplicadores de Lagrange.
7. SVM com Margens Rígidas: Padrões Não-linearmente Separáveis
8. SVM Separando Padrões Não-linearmente Separáveis
9. SVM e a Função Kernel
10. Aplicações
11. Discussão



Introdução

- As Máquinas de Vetores Suporte (*Support Vector Machines - SVMs*) são baseadas na Teoria de Aprendizagem Estatística (TAE) proposta por Vapnik e Chernovemkis nas décadas de 1960 e 1970 (Vapnik, 1995).
- A Teoria de Aprendizagem Estatística visa encontrar condições matemáticas para escolha de uma função que separe dados a serem aprendidos em problemas de categorização. Esta separação deve considerar o menor erro de treinamento ao mesmo tempo que deve maximizar a capacidade de generalização de um classificador (para aprendizagem supervisionada).

Introdução

- Método para escolha de função de separação de dados em categorias: Minimizar o erro de treinamento e a complexidade da função selecionada.
 - O nível da complexidade está associado com a capacidade de generalização.
- O conceito dimensão Vapnik-Chervonenkis (VC) é útil para obter as condições mencionadas acima. Ela mede a complexidade das hipóteses (funções) consideradas por um algoritmo de busca por soluções.

Introdução

-
-
- Características favoráveis ao uso de SVMs:
 - i. Capacidade de generalização alta, evitando sobretreinamento (*overfitting*).
 - ii. Robustez para categorização de dados com dimensões altas, que tendem a ser sobretreinados em outros classificadores pois muitas micro-características são pouco discriminantes.
 - iii. Convexidade da função objetivo pois esta é uma função quadrática com apenas um ótimo global.
 - iv. Teoria bem estabelecida nas áreas de matemática e estatística.

Introdução

- Treinamento: Supervisionado ou Não-supervisionado que não tem conhecimento prévio sobre o domínio do problema.
- Classes de problemas em que são comumente usadas SVM:
 - i. Classificação de padrões;
 - ii. Regressão;
 - iii. Reconhecimento de padrões;
 - iv. Agrupamento.
- Exemplos de áreas de aplicação (dimensão alta dos dados):
 - Detecção de faces em imagens; Categorização de textos; Regressão linear; Bioinformática.

Classificadores Binários

Função de Separação

- A tarefa a ser realizada:
 - Um conjunto de dados finito $\{(\mathbf{x}, y)\}$ onde \mathbf{x} representa uma entrada e y uma das duas classes à qual ela pode pertencer.

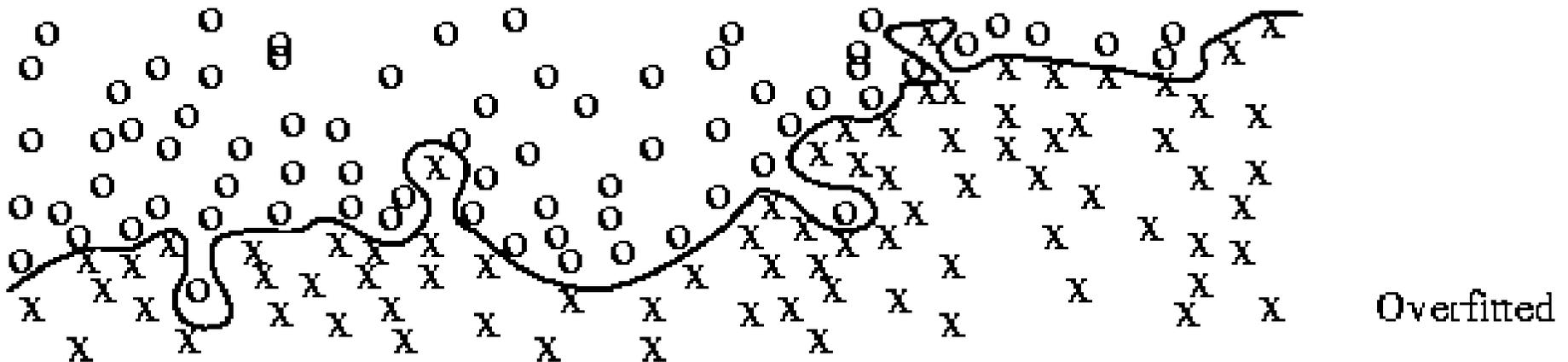
$\{0,1\}, \{-1,+1\}, \{o,x\}, \{\blacklozenge,o\}...$

- A solução:
 - Aprender uma função que baseada em um grupo de padrões de treinamento (que pode ser muito pequeno), possa associar dados não vistos anteriormente à classe correta.

Classificadores Binários

Função de Separação

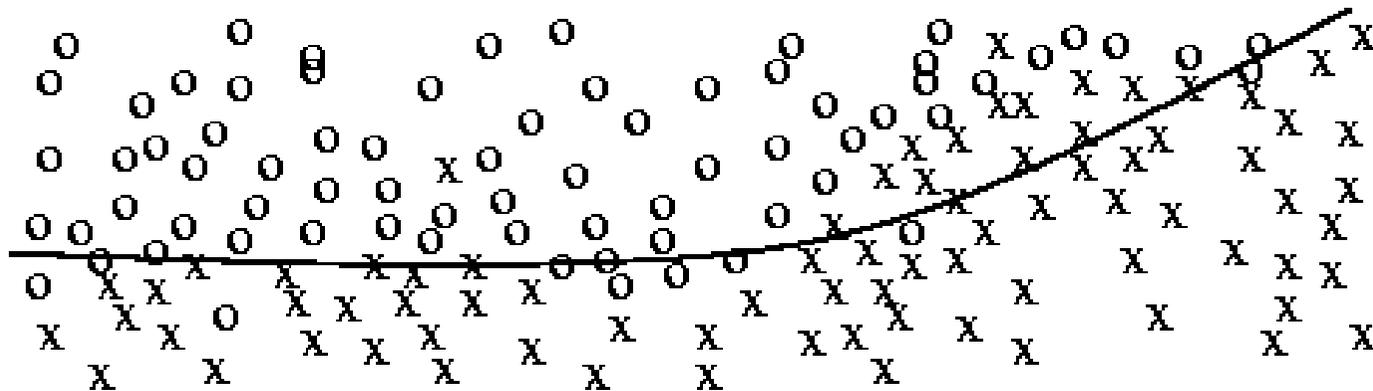
- A abordagem clássica é tomar uma função, como um polinômio, e ajustar seus parâmetros para separar os dados de treinamento colocando-os em uma das duas classes.
- No treinamento, aumentando o grau do polinômio é possível reduzir o erro nos dados de treinamento.
 - Esta estratégia pode levar ao sobre-treinamento (*overfitting*) implicando em baixa capacidade de generalização.



Classificadores Binários

Função de Separação

- Procedimento alternativo:
 - Redução significativa do grau do polinômio.
 - Esta opção pode levar ao aumento do erro de classificação para os dados de treinamento, o *underfitting*.



Aprendizagem Estatística

Minimização do Risco Estrutural

- A teoria de Aprendizagem Estatística visa determinar condições matemáticas para escolha de um classificador com desempenho desejado para conjuntos de treinamento e teste.
- É sempre possível encontrar um polinômio de alto grau que separe duas classes quaisquer.
 - Logo o risco empírico pode sempre ser minimizado para zero ao custo de uma função de decisão muito complexa.
 - A distribuição dos dados de treinamento pode não ser tão complexa mas, fatores como ruído podem fazer a distribuição parecer mais complexa para a máquina de aprendizagem.
- A teoria da Minimização do Risco Estrutural (MRE) formaliza o conceito de controle de complexidade e minimização de risco empírico.

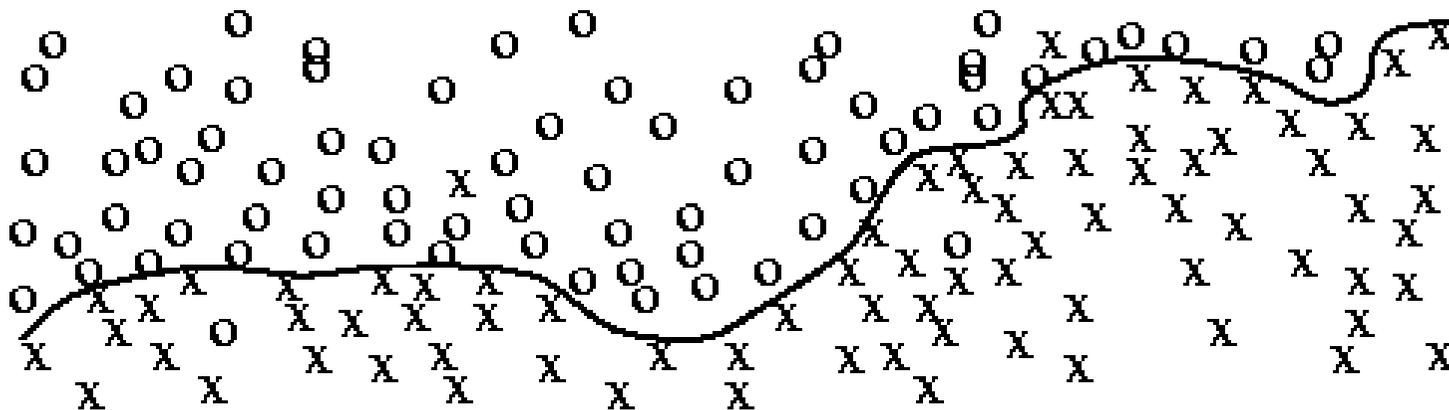


Aprendizagem Estatística

Minimização do Risco Estrutural

- Se uma máquina de aprendizagem, como rede neural ou máquina de vetor suporte, pretende minimizar o risco esperado, ela deve minimizar tanto o risco empírico quanto o termo de complexidade.

risco esperado \leq risco empírico + termo de complexidade



Well-Trained

Aprendizagem Estatística

Minimização do Risco Empírico (treinamento)

- Critérios considerados para escolha de um classificador (f):
 - Minimização do risco empírico, relativo a erro durante o treinamento, no qual se considera:
 - O número de pares entrada-saída.
 - A função de custo que relacione a previsão de saída com a saída desejada.

$$R_{emp}(f) = \frac{1}{n} \sum_{i=1}^n \frac{1}{2} c(f(\mathbf{x}_i), y_i)$$

Aprendizagem Estatística

Minimização do Risco Funcional (generalização)

- Critérios considerados para escolha de um classificador (f):
 - Minimização do risco funcional, relativo a erro durante a validação (generalização), no qual se considera:
 - Função de custo relacionando a previsão de saída com a saída desejada.
 - Distribuição de probabilidade dos pares.

$$R(f) = \int \frac{1}{2} c(f(\mathbf{x}), y) dP(\mathbf{x}, y)$$

Aprendizagem Estatística

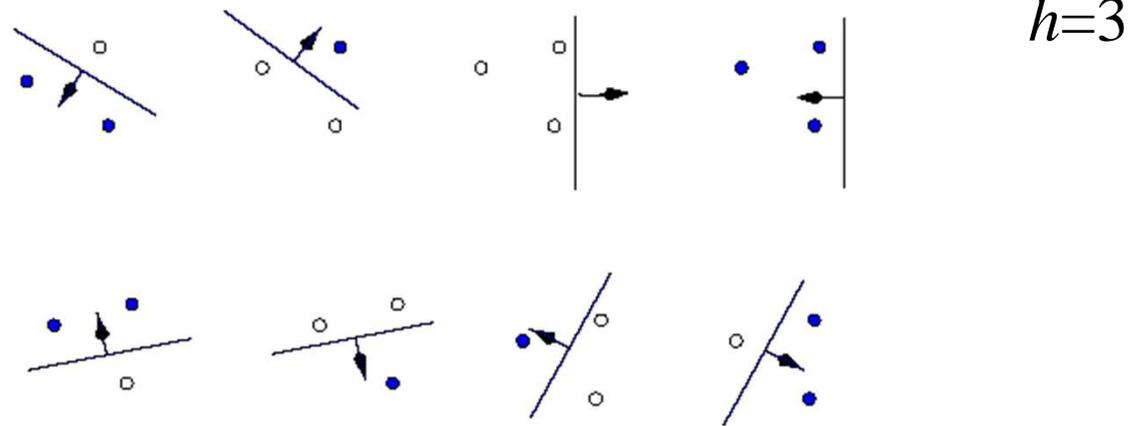
Minimização do Risco Funcional (generalização)

- Limites do risco funcional determinam a escolha do classificador:
 - Os limites do risco funcional para funções sinal (classe de funções aqui considerada) relacionam o número de exemplos de treinamento, o risco empírico para este conjunto e a complexidade do espaço de hipóteses.
 - O risco funcional de uma função classificadora é minimizado se o número de observações do conjunto de treinamento for suficientemente grande.
 - A complexidade do espaço de hipóteses é medida através da dimensão Vapnik-Chervonenkis (VC).
 - O risco médio de uma função classificadora é minimizado se a dimensão VC do conjunto destas funções for suficientemente pequena.

Aprendizagem Estatística

Dimensão-VC

- A complexidade de um grupo de funções de decisão pode ser medida por um valor h , chamado Dimensão-VC, que, avalia a quantidade máxima de pontos que podem ser separados por este grupo de funções se todas as permutações de rótulos ocorrerem.
- Aqui trabalha-se com dicotomias: funções sinais que dividem o espaço de entradas em dois subconjuntos disjuntos.
- Valor alto de dimensão VC implica em grande complexidade das funções de decisão.



Aprendizagem Estatística

Minimização do Risco Estrutural

- A equação de delimitação pode ser re-escrita empregando a dimensão-VC, isto é, usando h .
 - Probabilidade da equação abaixo ser verdadeira: $1-\delta$.
 - O número de exemplos de treinamento é n .

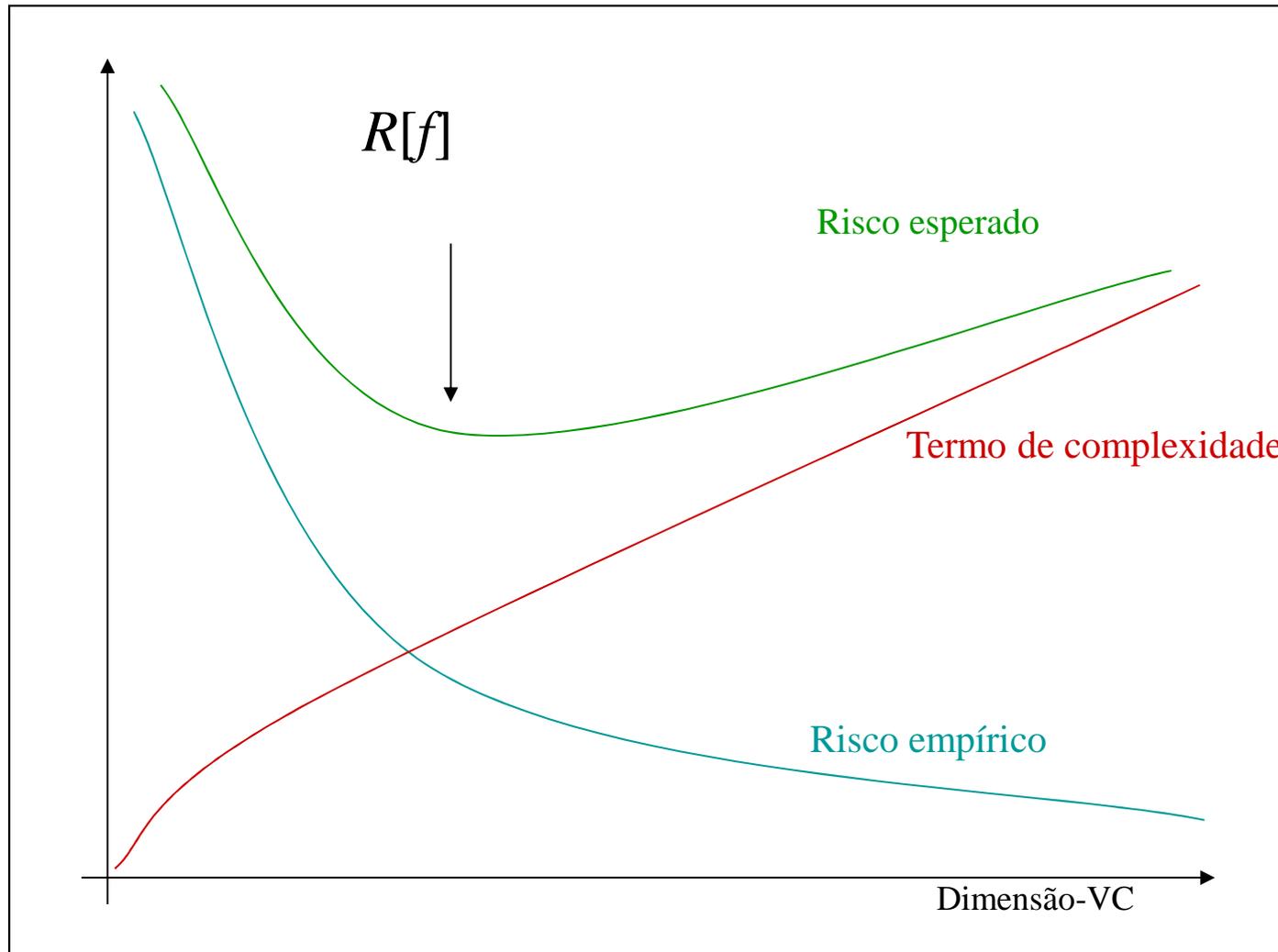
risco esperado \leq risco empírico + termo de complexidade

$$R[f] \leq R_{emp}[f] + \sqrt{\frac{h \left(\ln \frac{2n}{h} + 1 \right) - \ln \frac{\delta}{4}}{n}}$$

- O crescimento de δ acarreta o aumento do risco esperado.

Aprendizagem Estatística

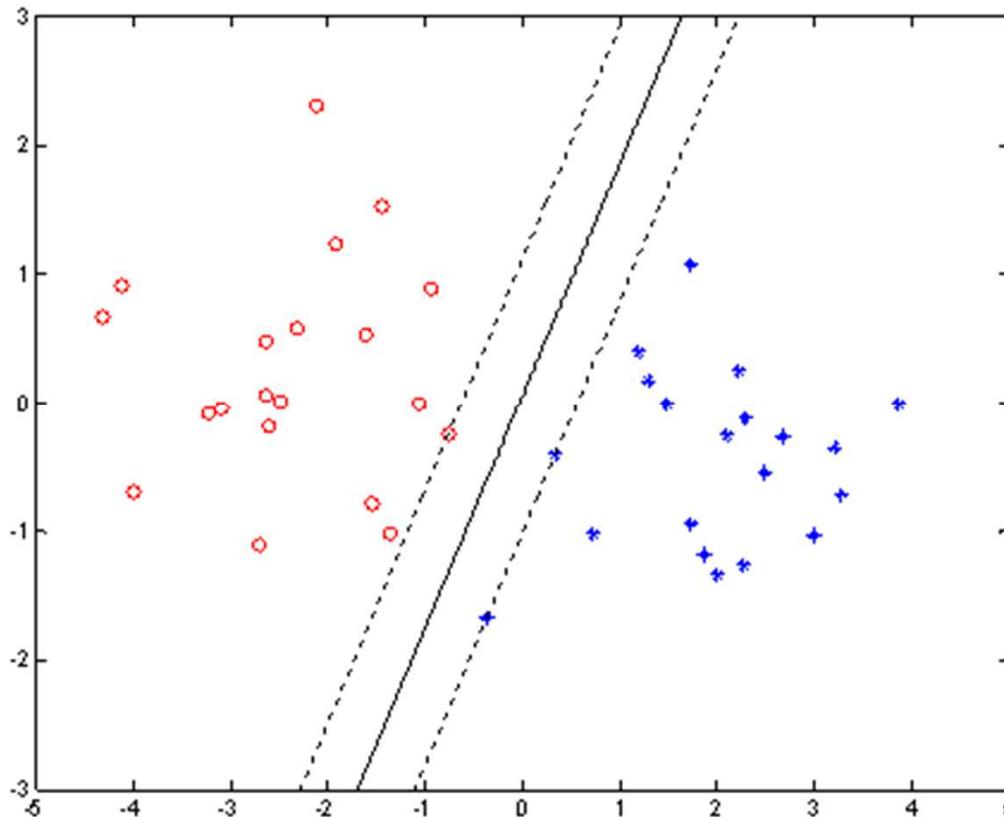
Minimização do Risco Estrutural



Aprendizagem Estatística

Margem de Separação

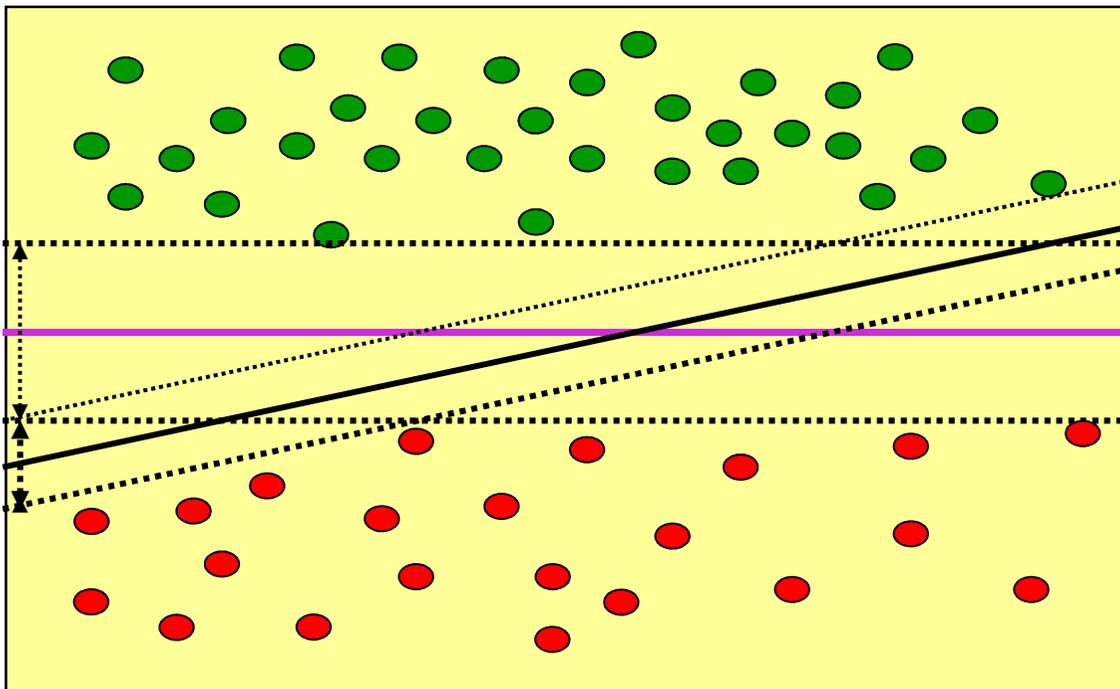
- A margem de separação de um classificador é definida como a menor distância entre exemplos do conjunto de treinamento e o hiperplano utilizado na separação destes dados em classes.



Aprendizagem Estatística

Margem de Separação

- Podem existir vários hiperplanos separando os dados corretamente, contudo existe ao menos um melhor que os demais.
 - Pode-se notar que o hiperplano com maior margem de separação tem melhor capacidade de generalização pois diminui a possibilidade de erro.

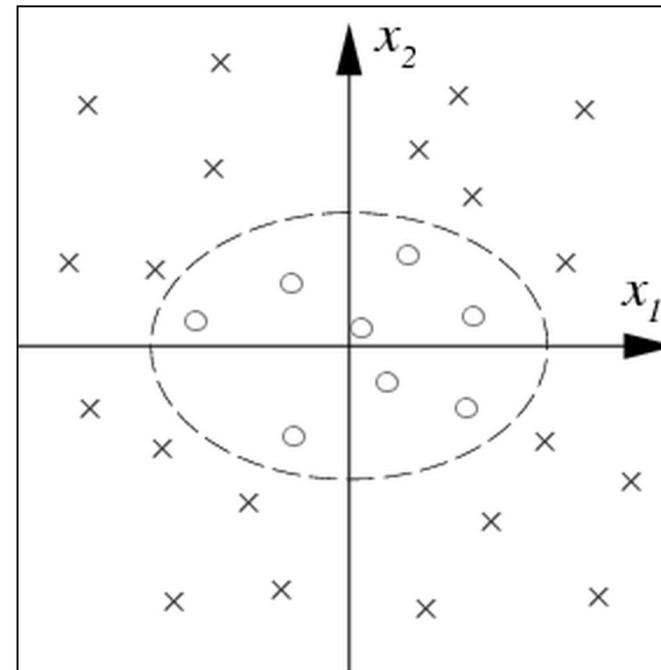
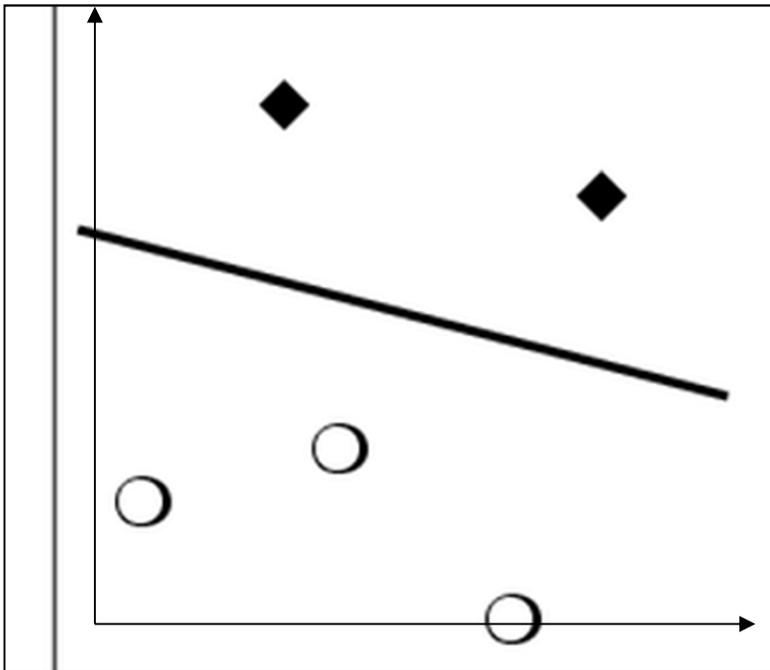


- Quanto maior a margem de um classificador menor será sua dimensão VC (prova está em teorema).
- Hiperplano com margem alta e que minimize os erros de treinamento e teste é chamado de hiperplano ótimo.

SVM com Margens Rígidas

Separabilidade Linear

- Um conjunto de pontos de treinamento é chamado linearmente separável se existe ao menos um hiperplano que é capaz de separá-los corretamente.



SVM com Margens Rígidas

Hiperplano de Separação

- As SVMs foram originalmente projetadas para classificação de dados em duas classes, gerando dicotomias.
 - Problema de classificação considerado: Classificar objetos m -dimensionais (vetores) nas classes $+1$ e -1 .
 - Conjunto de treinamento: formado por n observações dos vetores de entradas com suas respectivas classificações binárias.
- Um conjunto de dados é linearmente separável se for possível dividir seus elementos em duas classes através de ao menos um hiperplano. Estes classificadores lineares podem ser definidos por:

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

- O produto escalar envolve um vetor normal ao hiperplano (\mathbf{w}) e o vetor de entrada. O par (\mathbf{w}, b) é determinado durante o treinamento.

SVM com Margens Rígidas

Hiperplano de Separação

- A equação do hiperplano divide o espaço de entrada em duas regiões que produzem dois tipos de saídas através da uma função sinal:

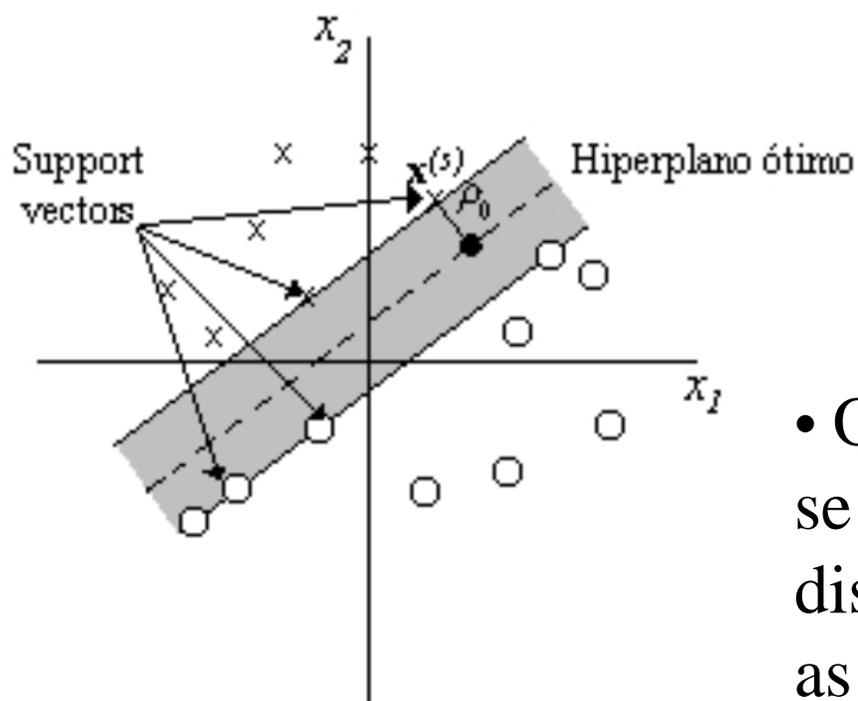
$$y_i = \begin{cases} +1, & \text{se } \mathbf{w}^T \cdot \mathbf{x}_i + b > 0 \\ -1, & \text{se } \mathbf{w}^T \cdot \mathbf{x}_i + b < 0 \end{cases}$$

- Logo, um conjunto de treinamento será linearmente separável se for possível determinar ao menos um par (\mathbf{w}, b) que faça a função sinal classificar corretamente os exemplos de tal conjunto.

SVM com Margens Rígidas

Hiperplano Ótimo

- Deseja-se determinar o hiperplano ótimo para padrões linearmente separáveis. O hiperplano ótimo é aquele cuja margem de separação (ρ_0) é máxima.



$\mathbf{w}_0^T \mathbf{x} + b_0 = 0$, eq. Hiperplano ótimo

\mathbf{w}_0 , vetor de pesos ótimo

b_0 , bias ótimo

- Os vetores suporte são aqueles que se situam sobre os hiperplanos que distam ρ_0 do hiperplano que separa as classes.

SVM com Margens Rígidas

Hiperplano Ótimo

- O hiperplano ótimo é definido pelos valores ótimos do vetor de pesos (\mathbf{w}_0) e do bias (b_0) da seguinte forma: $\mathbf{w}_0^T \mathbf{x} + b_0 = 0$.
- A função discriminante $g(\mathbf{x}) = \mathbf{w}_0^T \mathbf{x} + b_0$ dá uma medida algébrica da distância de \mathbf{x} para o hiperplano ótimo. Neste caso, pode-se escrever:

$$\mathbf{x} = \mathbf{x}_p + r \frac{\mathbf{w}_0}{\|\mathbf{w}_0\|} \quad \text{onde } \mathbf{x}_p \text{ é a projeção de } \mathbf{x} \text{ no hiperplano ótimo.}$$

Para encontrar a distância r faz-se:

$$g(\mathbf{x}) = \mathbf{w}_0^T \mathbf{x} + b_0 = \mathbf{w}_0^T \left(\mathbf{x}_p + r \frac{\mathbf{w}_0}{\|\mathbf{w}_0\|} \right) + b_0 = \mathbf{w}_0^T \mathbf{x}_p + r \frac{\mathbf{w}_0^T \mathbf{w}_0}{\|\mathbf{w}_0\|} + b_0$$

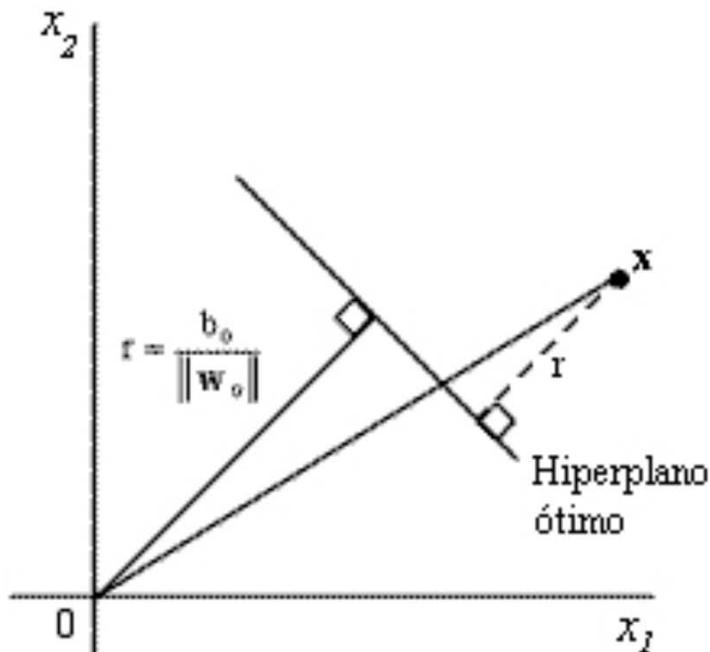


SVM com Margens Rígidas

Hiperplano Ótimo

$$\therefore g(\mathbf{x}) = (\mathbf{w}_0^T \mathbf{x}_p + b_0) + r \frac{\|\mathbf{w}_0\|^2}{\|\mathbf{w}_0\|} \therefore g(\mathbf{x}) = g(\mathbf{x}_p) + r \|\mathbf{w}_0\| \therefore r = \frac{g(\mathbf{x})}{\|\mathbf{w}_0\|}$$

Se \mathbf{x} estiver na origem então $r = \frac{b_0}{\|\mathbf{w}_0\|}$



Se $b_0 > 0$, a origem está no lado positivo do hiperplano ótimo;

Se $b_0 < 0$, a origem está no negativo do hiperplano ótimo;

Se $b_0 = 0$, o hiperplano ótimo passa pela origem.

SVM com Margens Rígidas

Vetores de Suporte

- Para um conjunto de treinamento linearmente separável, pode-se re-escalonar \mathbf{w} e b para que os pontos mais próximos do hiperplano separador que satisfaçam $|\mathbf{w}^T \cdot \mathbf{x} + b| = 1$. Isto permite a obtenção da representação canônica do hiperplano que facilita futuras considerações na determinação do hiperplano ótimo.
- Um vetor suporte é definido como: $g(\mathbf{x}^{(s)}) = \mathbf{w}_0^T \mathbf{x}^{(s)} \pm b_0 = \pm 1$,
para $d^{(s)} = \pm 1$.
- Os vetores suporte são os mais difíceis para classificar por estarem mais próximos da superfície de decisão.

SVM com Margens Rígidas

Vetores de Suporte

- A distância dos vetores suporte para o hiperplano ótimo é calculada:

$$r = \frac{g(\mathbf{x}^{(s)})}{\|\mathbf{w}_0\|} = \begin{cases} \frac{1}{\|\mathbf{w}_0\|} & \text{se } d^{(s)} = +1 \\ -\frac{1}{\|\mathbf{w}_0\|} & \text{se } d^{(s)} = -1 \end{cases}$$

- Tem-se que ρ_0 é o valor ótimo da margem de separação entre as duas classes que formam o conjunto de treinamento. Assim tem-se que a expressão a seguir mede a distância entre os hiperplanos

$$\mathbf{w}_0^T \mathbf{x}^{(s)} \pm b_0 = \pm 1: \quad \rho_0 = 2r = \frac{2}{\|\mathbf{w}_0\|}$$

- Conclui-se da expressão acima que a maximização da margem de separação é obtida pela minimização da norma Euclidiana de \mathbf{w}_0 .



SVM com Margens Rígidas

Determinação dos Pesos Ótimos

- O hiperplano ótimo definido por $\mathbf{w}_0^T \mathbf{x} + b_0 = 0$ é único pois o vetor de pesos ótimo \mathbf{w}_0 dá a separação máxima possível de exemplos positivos e os negativos. A condição ótima é atendida pela minimização da norma euclidiana do vetor de pesos \mathbf{w} .
- O problema de otimização com restrições a ser resolvido é:
 - Dado o conjunto de treinamento (\mathbf{x}_i, d_i) , $i=1, \dots, N$; Encontre os vetor de pesos \mathbf{w} e do *bias* b ótimos que satisfaçam as restrições: $d_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1$, e \mathbf{w} minimize a função de custo: $\Phi(\mathbf{w}) = (1/2) \mathbf{w}^T \mathbf{w}$
 - O fator de escala $(1/2)$ é incluído por conveniência, a função de custo é convexa, as restrições são lineares.
 - Este problema pode ser resolvido através do Método de Multiplicadores de Lagrange.

SVM com Margens Rígidas

Pesos Ótimos por Multiplicadores de Lagrange

- Método dos Multiplicadores de Lagrange: Empregado para resolver problemas de extremos sujeitos a restrições de igualdade.
- Seja o problema a seguir:

$$\max (\min) f(\mathbf{x})$$

$$\text{s.a. } g_i(\mathbf{x}) = 0, \quad i = 1, \dots, N$$

onde f e g_i ($i=1, \dots, N$) são funções reais de n ($n > N$) variáveis e duas vezes diferenciáveis num determinado conjunto D .

- Chama-se função de Lagrange ou lagrangiano à função:

$$L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \sum_{i=1}^N \lambda_i g_i(\mathbf{x})$$

SVM com Margens Rígidas

Pesos Ótimos por Multiplicadores de Lagrange

• Função Lagrangiana:

$$J(\mathbf{w}, b, \alpha) = \frac{1}{2} \mathbf{w}^T \mathbf{w} - \sum_{i=1}^N \alpha_i [d_i (\mathbf{w}^T \mathbf{x}_i + b) - 1]$$

• O problema consiste em encontrar um ponto de sela que minimize $J(\cdot)$ em relação a \mathbf{w} e b e maximize-a com respeito aos multiplicadores de Lagrange (α).

- Minimizando $J(\mathbf{w}, b, \alpha)$ em relação a \mathbf{w} e b .

Condição 1:
$$\frac{\partial J(\mathbf{w}, b, \alpha)}{\partial \mathbf{w}} = 0 \therefore \mathbf{w} = \sum_{i=1}^N \alpha_i d_i \mathbf{x}_i$$

Condição 2:
$$\frac{\partial J(\mathbf{w}, b, \alpha)}{\partial b} = 0 \therefore \sum_{i=1}^N \alpha_i d_i = 0$$

SVM com Margens Rígidas

Pesos Ótimos por Multiplicadores de Lagrange

- Expandindo a Função Lagrangiana tem-se:

$$J(\mathbf{w}, b, \alpha) = \frac{1}{2} \mathbf{w}^T \mathbf{w} - \sum_{i=1}^N \alpha_i [d_i (\mathbf{w}^T \mathbf{x}_i + b) - 1] \therefore$$

$$J(\mathbf{w}, b, \alpha) = \frac{1}{2} \mathbf{w}^T \mathbf{w} - \sum_{i=1}^N \alpha_i d_i \mathbf{w}^T \mathbf{x}_i - b \sum_{i=1}^N \alpha_i d_i + \sum_{i=1}^N \alpha_i$$

- Para a expressão acima, tem-se que

$$\sum_{i=1}^N \alpha_i d_i = 0;$$

$$\mathbf{w} = \sum_{i=1}^N \alpha_i d_i \mathbf{x}_i;$$

$$\mathbf{w}^T \mathbf{w} = \sum_{i=1}^N \alpha_i d_i \mathbf{w}^T \mathbf{x}_i = \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j d_i d_j \mathbf{x}_i^T \mathbf{x}_j$$

-As expressões à esquerda geram o problema dual em função de α

- Os vetores \mathbf{x}_i e \mathbf{x}_j são o vetor de entrada e o padrão de entrada pertencente ao j -ésimo exemplo,



SVM com Margens Rígidas

Pesos Ótimos por Multiplicadores de Lagrange

- Deve-se encontrar os multiplicadores de Lagrange que maximize a Função Objetivo:

$$\text{Max } J(\mathbf{w}, b, \alpha) = Q(\alpha) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j d_i d_j \mathbf{x}_i^T \mathbf{x}_j$$

$$\text{s.a. } \sum_{i=1}^N \alpha_i d_i = 0$$

$$\alpha_i \geq 0, \quad \text{para } i = 1, 2, \dots, N$$

- Após determinar os multiplicadores ótimos $(\alpha_{0,i})$, \mathbf{w}_0 e b_0 são obtidos:

$$\mathbf{w}_0 = \sum_{i=1}^N \alpha_{0,i} d_i \mathbf{x}_i \quad b_0 = 1 - \mathbf{w}_0^T \mathbf{x}^{(s)}, \quad \text{para } d^{(s)} = 1$$

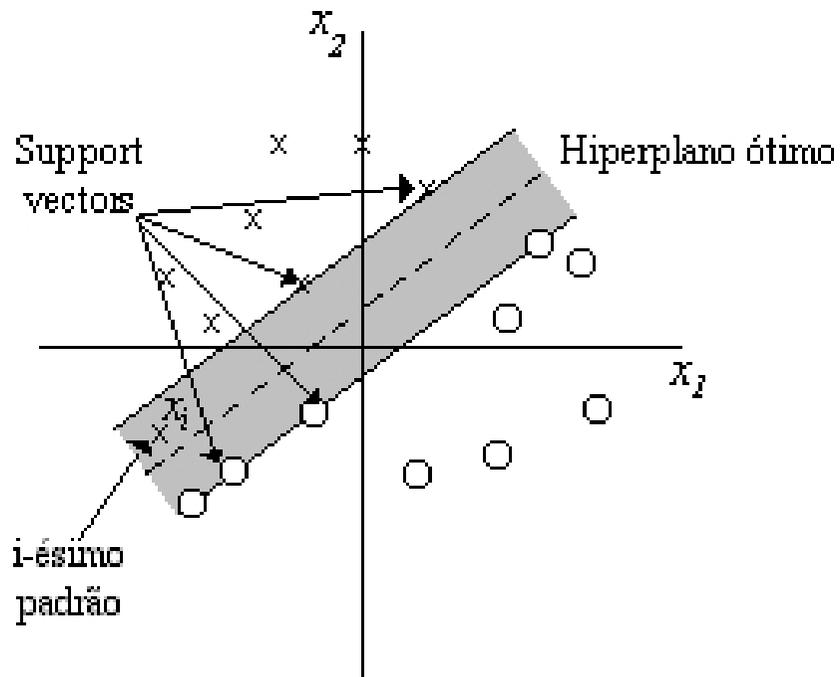


SVM com Margens Rígidas

Padrões Não-linearmente Separáveis

A condição $d_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1$, para $i = 1, 2, \dots, N$

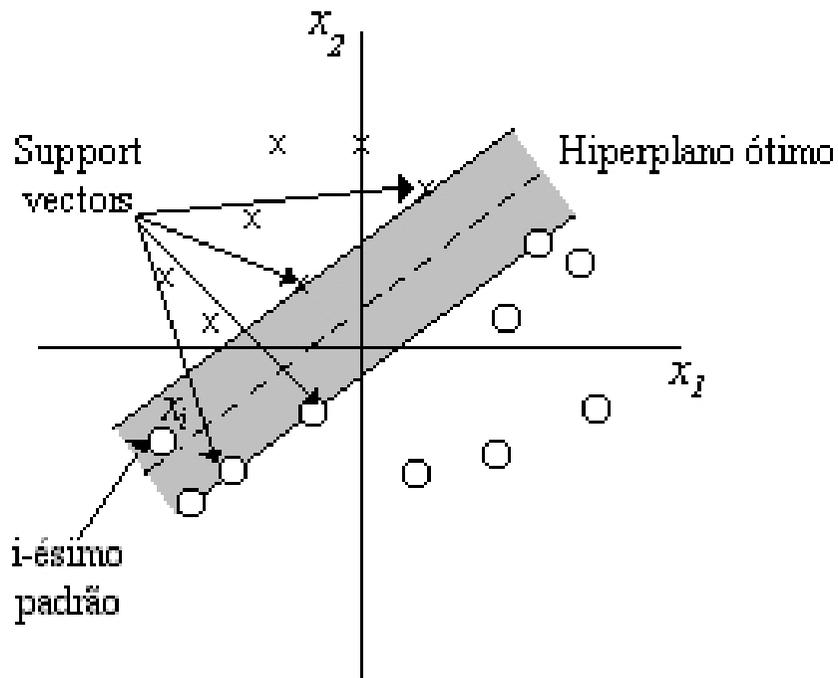
pode ser violada em duas situações:



- 1ª situação de violação:
 - Ponto (\mathbf{x}_i, d_i) está na região de separação, mas do lado correto da superfície de decisão.

SVM com Margens Rígidas

Padrões Não-linearmente Separáveis



- 2ª situação de violação:
 - Ponto (\mathbf{x}_i, d_i) está no lado incorreto da superfície de decisão.

SVM com Margens Rígidas

Padrões Não-linearmente Separáveis

- A equação anterior pode ser re-escrita, com a introdução de um conjunto de variáveis escalares não negativas $\{\xi_i\}_{i=1}^N$.

$$d_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i, \text{ para } i = 1, 2, \dots, N \quad (21)$$

$0 \leq \xi_i \leq 1$: 1ª situação

$\xi_i > 1$: 2ª situação

- O conjunto $\{\xi_i\}_{i=1}^N$ é adicionado à função de custo:

$$\Phi(\mathbf{w}, \xi) = \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^N \xi_i$$

- que deve ser minimizada, sujeita às restrições: Eq. (21) e $\xi_i \geq 0$.

SVM com Margens Rígidas

Padrões Não-linearmente Separáveis

- A maximização de $Q(\alpha)$ é realizada com alteração em uma de suas restrições:

$$J(\mathbf{w}, b, \alpha) = Q(\alpha) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j d_i d_j \mathbf{x}_i^T \mathbf{x}_j$$

$$\sum_{i=1}^N \alpha_i d_i = 0$$

$$\text{e } 0 \leq \alpha_i \leq C, \text{ para } i = 1, 2, \dots, N$$

Logo, \mathbf{w}_0 é obtido por:

$$\mathbf{w}_0 = \sum_{i=1}^N \alpha_{0,1} d_i \mathbf{x}_i$$

e b_0 através de:

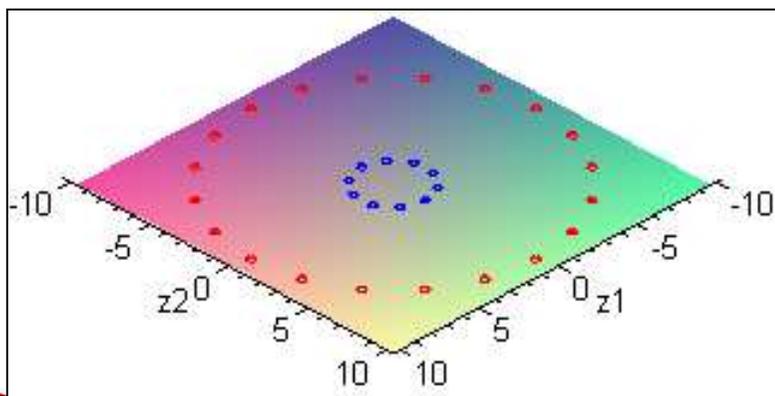
$$\alpha_i [y_i (\mathbf{w}_0^T \mathbf{x}_i + b_0) - 1 + \xi_i] = 0$$

SVM Separando Padrões Não-linearmente Separáveis - Mapeamento Φ

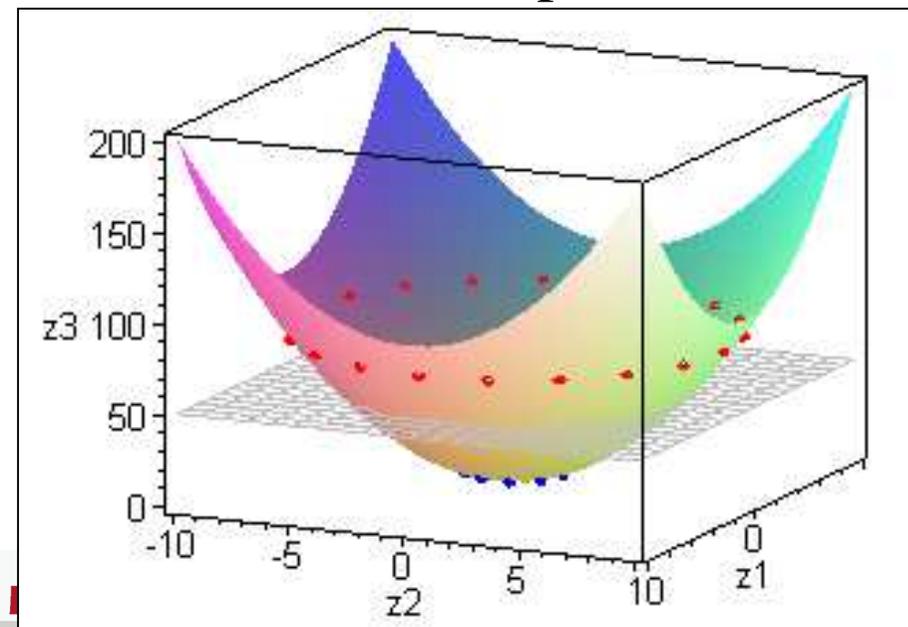
- Classificadores lineares são limitados, veja a porta XOR. Contudo, eles possuem boas propriedades como função de decisão fácil.
- Dados não-linearmente separáveis podem se tornar linearmente separáveis, em um espaço transformado através de um mapeamento Φ . Este novo espaço é chamado de espaço de características (*feature space*).

Feature Space

$$\phi := (x_1, x_2) \rightarrow (z_1, z_2, z_3) = (x_1, x_2, x_1^2 + x_2^2)$$



Φ



SVM Separando Padrões Não-linearmente Separáveis - Mapeamento Φ

- Deve-se substituir cada produto escalar no espaço de entrada por pontos transformados.

$$f(\mathbf{x}_j) = \text{sgn}\left(\sum_{i=1}^N d_i \alpha_i (\mathbf{x}_j^T \cdot \mathbf{x}_i) + b_j\right) \Rightarrow$$

$$f(\mathbf{x}_j) = \text{sgn}\left(\sum_{i=1}^N d_i \alpha_i (\Phi(\mathbf{x}_j^T) \cdot \Phi(\mathbf{x}_i)) + b_j\right)$$

- Possível problema:
 - O espaço transformado pode ter número muito alto, até infinito, de dimensões, impossibilitando o cálculo do produto interno.
 - É difícil também encontrar a função Φ que resolva o problema.

SVM e a Função Kernel

Definição e Papel

- Com uma função especial, chamada função kernel é possível calcular o produto escalar $\Phi(\mathbf{x}_i)\Phi(\mathbf{x}_j)$ sem mesmo conhecer o mapeamento Φ .

$$f(\mathbf{x}_j) = \text{sgn}\left(\sum_{i=1}^N d_i \alpha_i (\mathbf{x}_j^T \cdot \mathbf{x}_i) + b_j\right) = \text{sgn}\left(\sum_{i=1}^N d_i \alpha_i K(\mathbf{x}_j, \mathbf{x}_i) + b_j\right)$$

• Definição do kernel do produto interno

- O produto interno de dois vetores induzidos no espaço de características por \mathbf{x}_i e \mathbf{x}_j compõem a definição do referido kernel:

$$K(\mathbf{x}_j, \mathbf{x}_i) = \Phi^T(\mathbf{x}_j) \cdot \Phi(\mathbf{x}_i) = \sum_{l=1}^N \Phi(\mathbf{x}_l) \Phi(\mathbf{x}_l)$$

- O kernel do produto interno é comutativo com respeito a seus argumentos.

SVM e a Função Kernel

Definição e Papel

- A definição para $K(\mathbf{x}_i, \mathbf{x}_j)$ é um caso particular do teorema de Mercer no âmbito de análise funcional:

- Seja $K(\mathbf{x}, \mathbf{x}')$ um kernel contínuo e simétrico que é definido no intervalo fechado $\mathbf{a} \leq \mathbf{x} \leq \mathbf{b}$ e da mesma forma para \mathbf{x}' . O kernel pode ser expandido pela série:

$$K(\mathbf{x}, \mathbf{x}') = \sum_{l=1}^{\infty} \lambda_l \Phi_l(\mathbf{x}) \Phi_l(\mathbf{x}'), \quad \forall \lambda_l > 0$$

- Expansão válida e convergente, absoluta e uniformemente, se e só se:

$$\int_b^a \int_b^a K(\mathbf{x}, \mathbf{x}') \Psi(\mathbf{x}) \Psi(\mathbf{x}') d\mathbf{x} d\mathbf{x}'$$

vale para quando $\int_b^a \Psi^2(\mathbf{x}) d\mathbf{x} < \infty$

- As funções Φ_l são chamadas autofunções e os números λ_l são denominados autovalores.



SVM e a Função Kernel

Definição e Papel

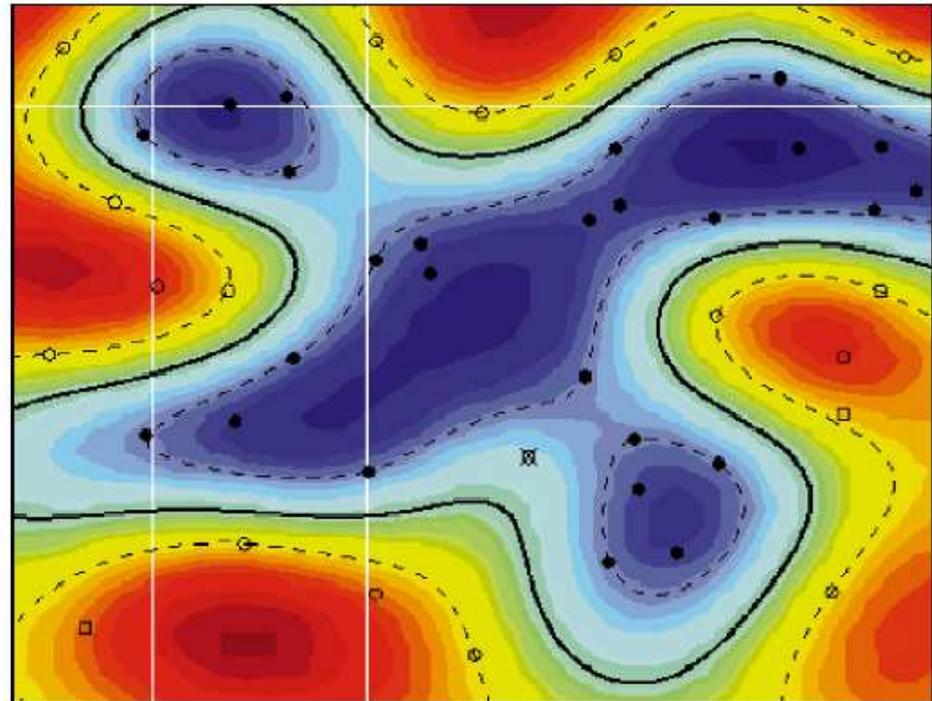
Exemplos de função kernel:

RBF Guassiana : $K(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|^2 / c)$; Polinomial : $((\mathbf{x} \cdot \mathbf{y}) + \theta)^d$

Sigmoidal : $\tanh(K(\mathbf{x} \cdot \mathbf{y}) + \theta)$; Multiquadrática Inv.: $\frac{1}{\sqrt{\|\mathbf{x} - \mathbf{y}\|^2 + c^2}}$

Exemplo com o Kernel RBF

- Nestes exemplos de funções kernel, geralmente, possuem parâmetros escolhidos pelo usuário e faixa de validade destes parâmetros para o Teorema de Mercer.



SVM e a Função Kernel

Definição e Papel

- A expansão de $K(\mathbf{x}_j, \mathbf{x}_i)$ permite a construção de superfície de decisão não-linear no espaço de entrada, com imagem linear no espaço de características. Tal expansão viabiliza o enunciado da forma dual da otimização com restrições de uma SVM:

Dado um conjunto de treinamento $\{(x_i, d_i)\}_{i=1}^N$, encontre os multiplicadores de Lagrange que maximizam a função objetivo

$$Q(\alpha) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j d_i d_j K(\mathbf{x}_i, \mathbf{x}_j)$$

$$\text{s.a.} \quad \sum_{i=1}^N \alpha_i d_i = 0$$

$$0 \leq \alpha_i \leq C,$$

para $i = 1, 2, \dots, N$ e C é determinado pelo usuário.

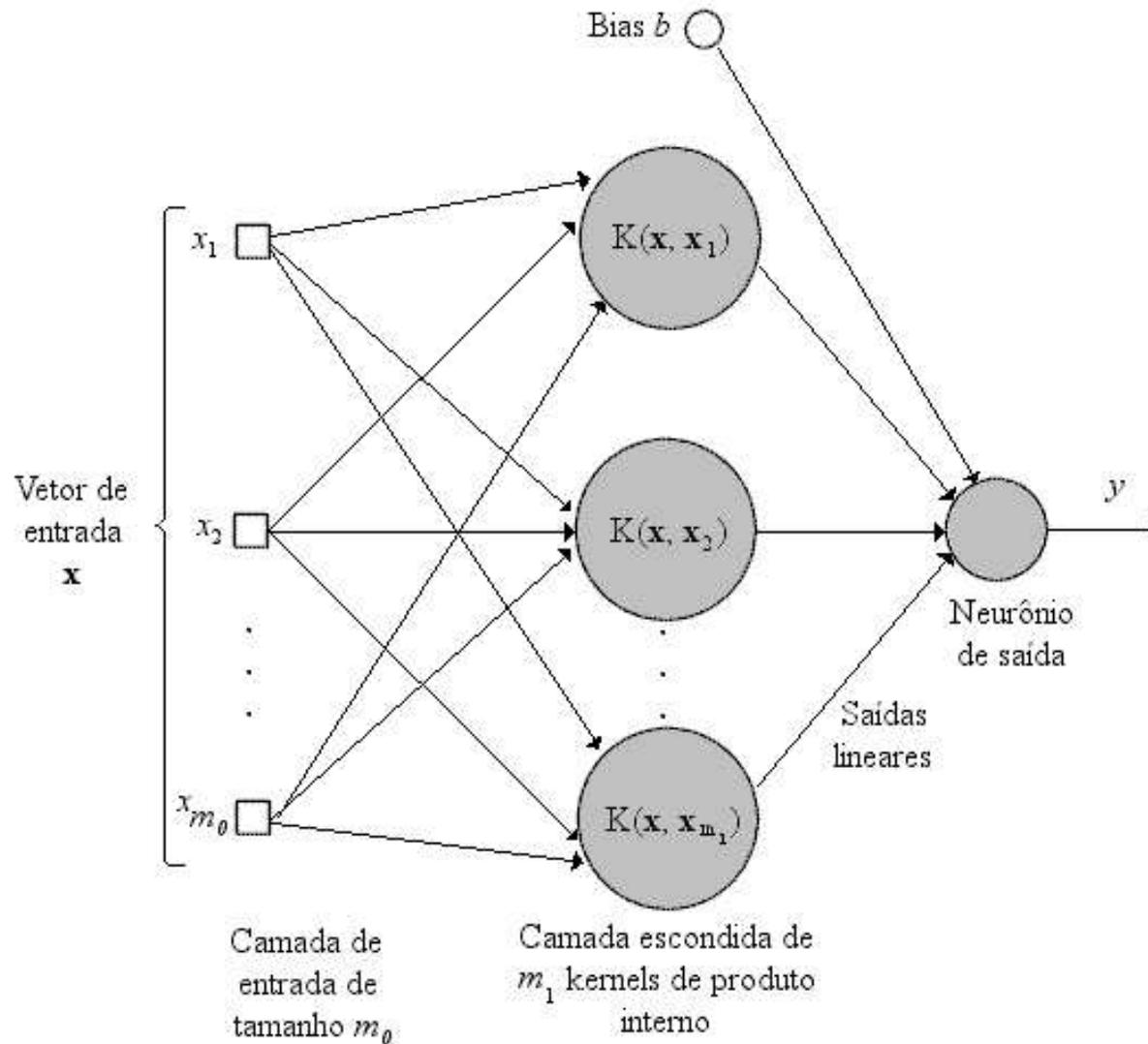


SVM e a Função Kernel

Definição e Papel

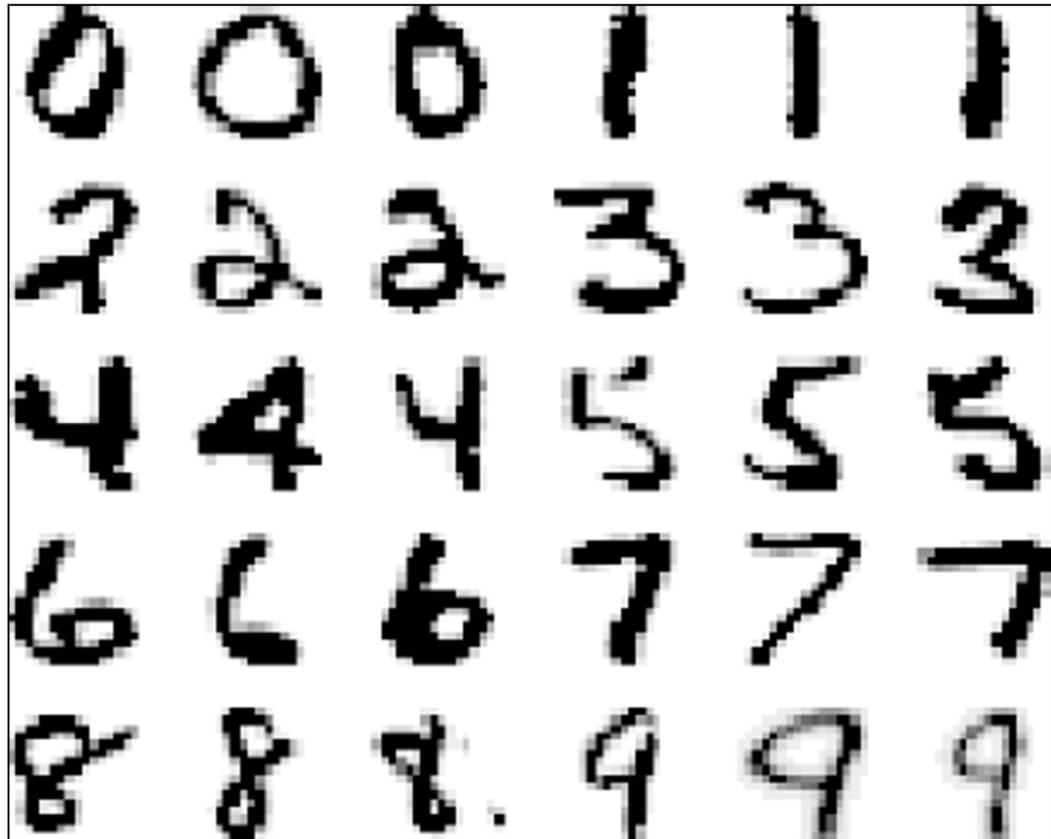
- Três idéias fundamentais:
 - Definição de um hiperplano ótimo de modo que ele possa ser identificado em maneira computacional eficiente: Maximize a margem.
 - Extensão da definição acima para problemas linearmente não-separáveis: Considere uma penalidade para termos equivocadamente classificados.
 - Mapeamento dos dados para um espaço de dimensão mais alta no qual é mais fácil realizar classificação com superfícies lineares de decisão: reformula o problema tal que os dados são mapeados implicitamente para este espaço.

SVM e a Função Kernel Arquitetura



SVM: Aplicações

- Reconhecimento de caracteres manuscritos:
 - Exemplos de caracteres:



SVM: Aplicações

- Reconhecimento de caracteres manuscritos:

- Desempenho de máquinas de aprendizagem distintas:

CLASSIFICATION ERROR IN % FOR OFF-LINE HANDWRITTEN CHARACTER RECOGNITION ON THE USPS WITH 7291 PATTERNS. INVARIANT SVMs ARE ONLY SLIGHTLY BELOW THE BEST EXISTING RESULTS (PARTS OF THE TABLE ARE FROM [136]). THIS IS EVEN MORE REMARKABLE SINCE IN [135]–[137], A LARGER TRAINING SET WAS USED, CONTAINING SOME ADDITIONAL MACHINE-PRINTED DIGITS WHICH HAVE BEEN FOUND TO IMPROVE THE ACCURACY

linear PCA & linear SVM (Schölkopf et. al. [11])	8.7%
k-Nearest Neighbor	5.7%
→ LeNet1 (LeCun et. al. [132], [133], [134])	4.2%
Regularized RBF Networks (Rätsch [128])	4.1%
Kernel-PCA & linear SVM (Schölkopf et. al. [11])	4.0%
SVM (Schölkopf et. al. [120])	4.0%
Virtual SVM (Schölkopf [4])	3.0%
→ Invariant SVM (Schölkopf et. al. [131])	3.0%
→ Boosting (Drucker et. al. [137])	2.6%
→ Tangent Distance (Simard et. al. [135], [136])	2.5%
→ Human error rate	2.5%

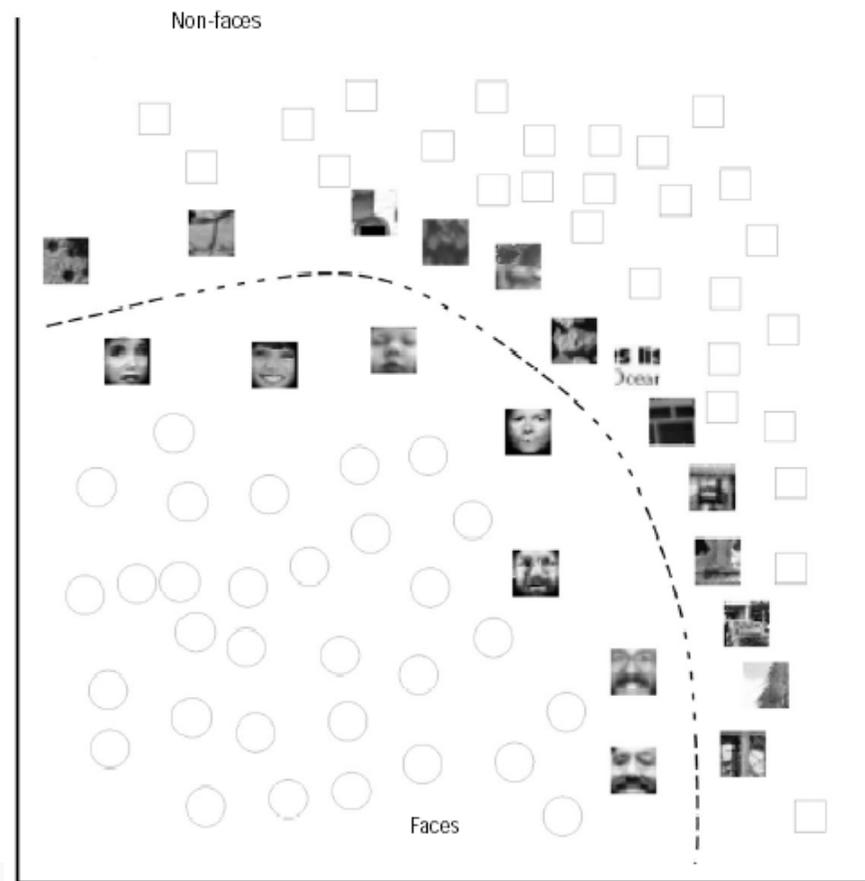
SVM: Aplicações

- Detecção de faces (definição): Dada uma imagem digital arbitrária determine se existe faces humanas nesta imagem.
 - Se existirem, retorne uma codificação de sua localização.
 - Codificação significa acomodar cada face em uma caixa de fronteiras definida pelas coordenadas das esquinas na imagem.
 - Pode ser extendida para reconhecimento de faces, HCI, sistemas de vigilância, etc.



SVM: Aplicações

- Detecção de faces (processo):
 - SVM treinada para padrões com tamanho fixo de face e não face.
 - Teste de candidatos de localização de imagens para padrões locais com procedimento de classificação que determina se padrão de imagem local é uma face.
 - Este problema de classificação, tem duas classes dicotômicas.



SVM: Aplicações

- Resultados experimentais em imagens estáticas:
 - Conjunto A: 313 com alta qualidade, mesmo número de faces.
 - Conjunto B: 23 com qualidade misturada, total de 155 faces.

	TEST SET A		TEST SET B	
	DETECT RATE (%)	FALSE ALARMS	DETECT RATE (%)	FALSE ALARMS
SVM	97.1	4	74.2	20
Sung	94.6	2	74.2	11



SVM: Aplicações

- Visão Computacional:
 - Detecção de pele.



Discussão

- Os parâmetros têm grande influência no treinamento.
- Interface de treinamento geral.
- Tempo de treinamento depende da CPU.
- Necessidade de interface para aplicações.
- SVM expressa aprendizagem como um programa matemático empregando a teoria de otimização.
- SVM emprega a transformação pelo kernel para mapear indiretamente para espaços de dimensões mais altas.
- SVM tem se caracterizado por bom desempenho, robustez, eficiência e versatilidade ao mesmo tempo que existem indicações teóricas dos motivos de sua capacidade de generalização.

Referências

- Haykin, S. (1999). *Neural Networks – A Comprehensive Survey*. Prentice-Hall, New Jersey, second edition.
- Smola, A. J., Barlett, P., Schölkopf, B., & Schuurmans, D. (1999). *Advances in Large Margin Classifiers*. The MIT Press (<http://www.kernel-machines.org/nips98/lmc-book.pdf>).
- Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag.