Máquinas de Vetores de Suporte
Support 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

UFPE

Centro de Informática UFPE

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.

U F P E

Centro de Informática
UFPE

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.

UFPE

Centro de Informática UFPE

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

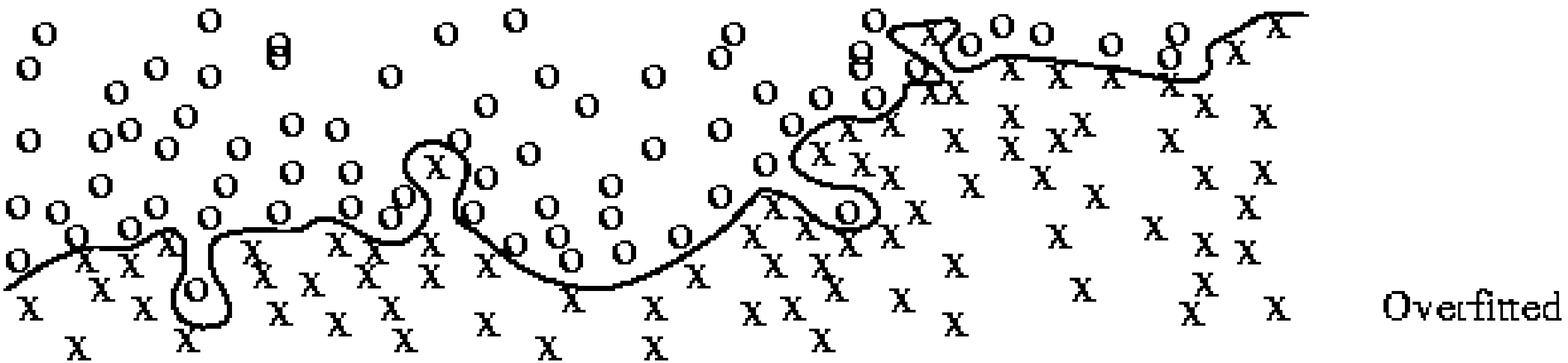
- A tarefa a ser realizada:
  - Um conjunto de dados finito {(**x**, *y*)} onde **x** representa uma entrada e *y* uma das duas classes à qual ela pode pertencer.

{0,1}, {-1,+1}, {o,x}, {♦,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.

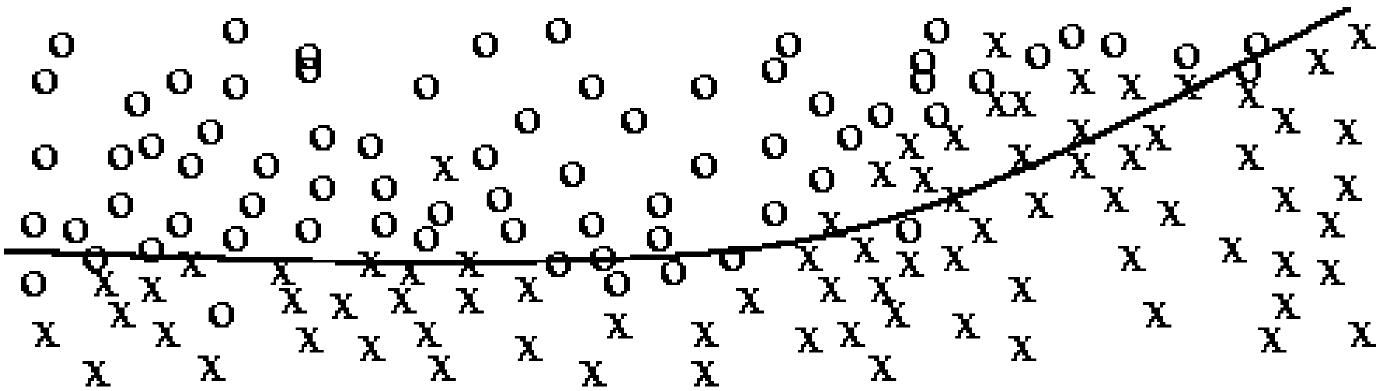
Classificadores Binários
Função de Separação

- 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 sobretreinamento (*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*.



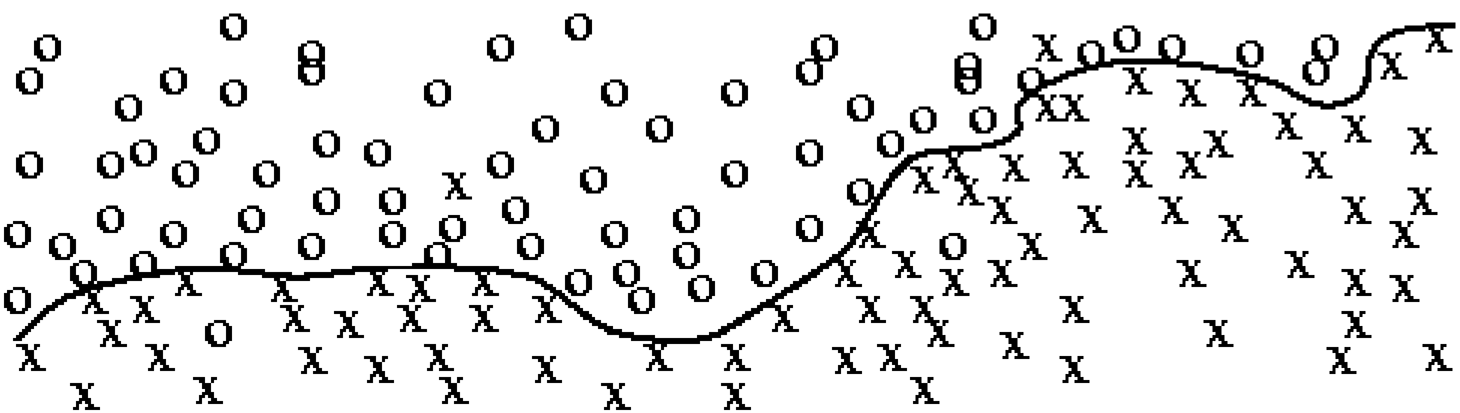
Aprendizagem Estatística
Minimização do Risco Estrutural

- 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.

$$\text{risco esperado} \leq \text{risco empírico} + \text{termo de complexidade}$$



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.

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

UFPE

Centro de Informática UFPE

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.

U F P E

Centro de Informática UFPE

Aprendizagem Estatística
Dimensão-VC

- A dimensão Vapnik-Chervonenkis (VC) é entendida como uma medida da capacidade ou complexidade de um conjunto de funções que podem ser aprendidas por um algoritmo estatístico de classificação binária, originalmente definida por Vladimir Vapnik e Alexey Chervonenkis;
- É determinada pela cardinalidade do maior conjunto de pontos que um algoritmo pode separar corretamente:
  - Um conjunto de *n* pontos que podem ser corretamente separados em classes dicotômicas por um classificador, atribuição correta de todos 2^n rótulos, e que não seja possível encontrar nenhum conjunto de n+1 pontos que permita particionamento válido;
  - Então a dimensão VC é igual a *n*.

Aprendizagem Estatística
Dimensão-VC

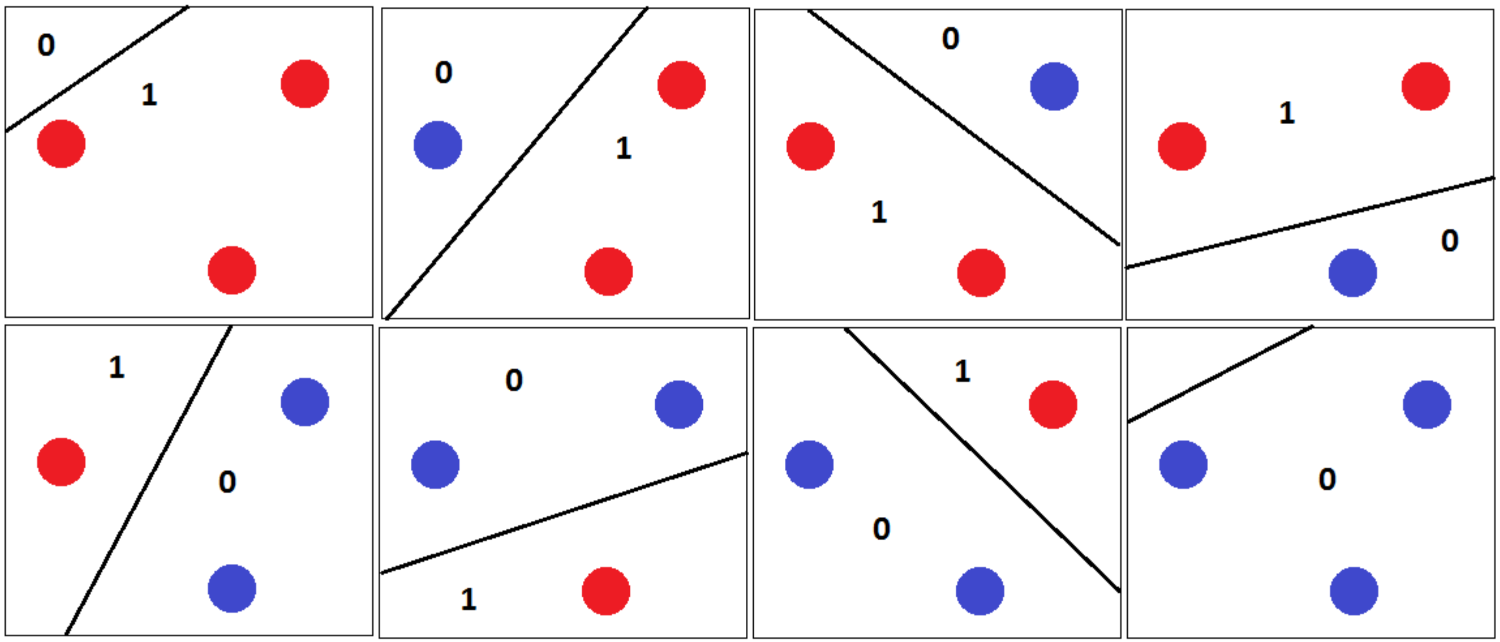
- Definição de Dimensão-VC:
  - Definição 1 (Particionamento de conjunto): Um subconjunto S de instâncias de um conjunto X é particionado por uma coleção de funções F se $\forall$ S' $\subseteq$ S, existe uma função $f \in F$ tal que:

$$f(x) = \begin{cases} 1, & x \in S' \\ 0, & x \in S - S' \end{cases}$$

  - Definição 2 (Dimensão-VC): A dimensão VC de um conjunto de funções F é definida como a cardinalidade do maior conjunto de dados que pode ser particionado por F.

Aprendizagem Estatística
Dimensão-VC

- Seja *h* = {conjunto de classificadores lineares 2D} de forma que quaisquer 3 pontos podem ser classificados por *h* corretamente com a separação do hiperplano,



- A dimensão VC é *h*= 3 pois para quaisquer 4 pontos no plano 2D, um classificador linear não separa todas as combinações dos pontos.

Aprendizagem Estatística
Minimização do Risco Estrutural

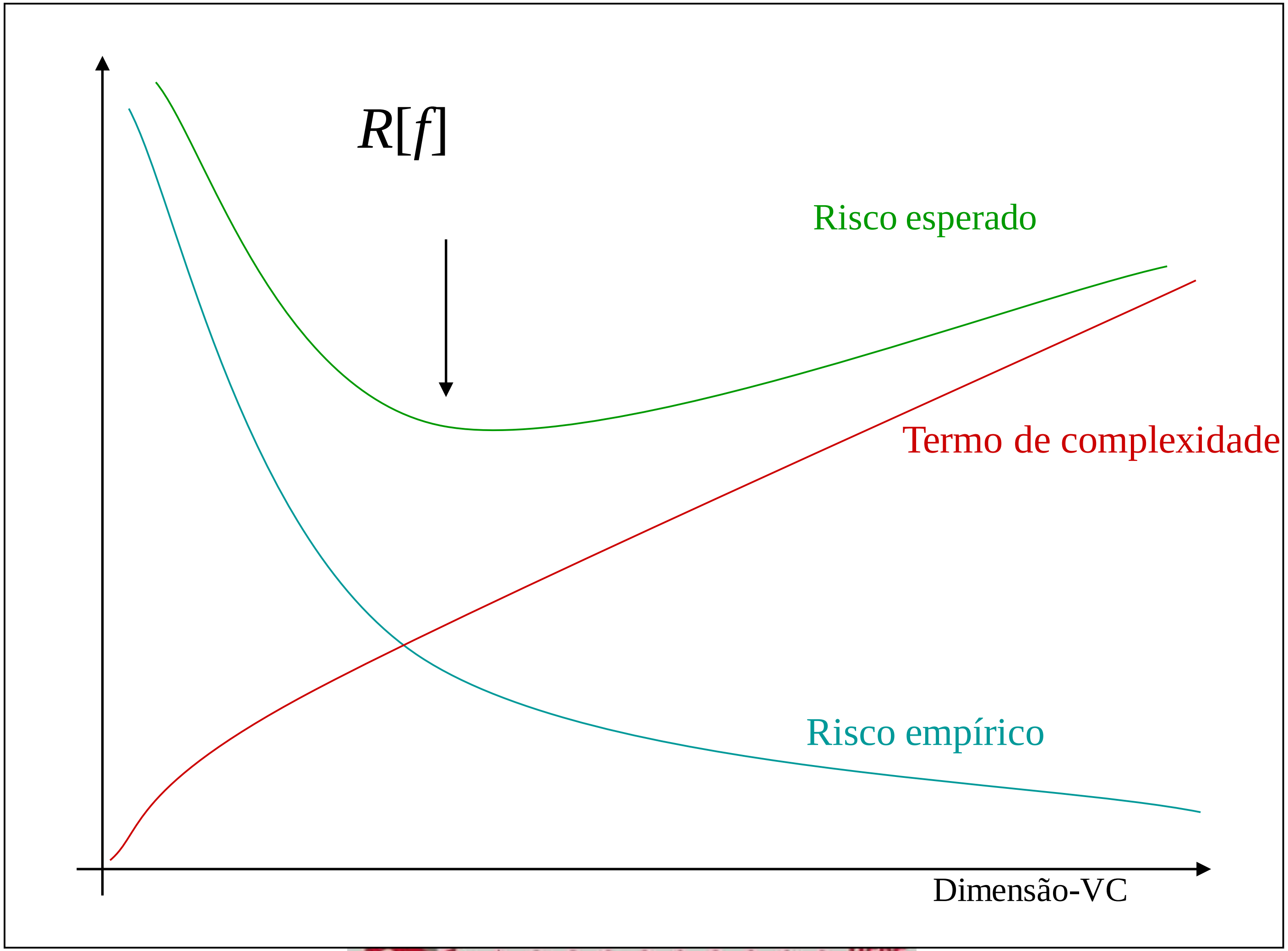
- A equação de delimitação pode ser re-escrita empregando a dimensão-VC, isto é, usando *h*.
  - Probabilidade da equação abaixo ser verdadeira: 1-δ.
  - O número de exemplos de treinamento é *n*.

$$\text{risco esperado} \leq \text{risco empírico} + \text{termo de complexidade}$$

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

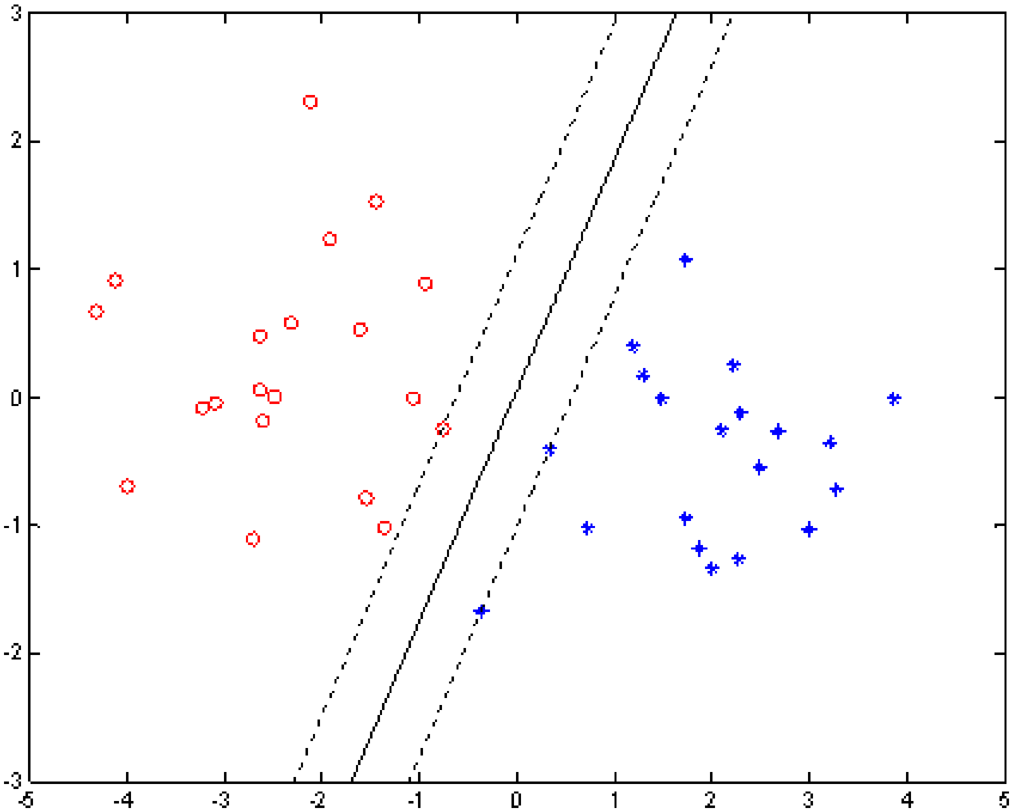
  - O crescimento de δ acarreta o aumento do risco esperado.

Aprendizagem Estatística
Minimização do Risco Estrutural



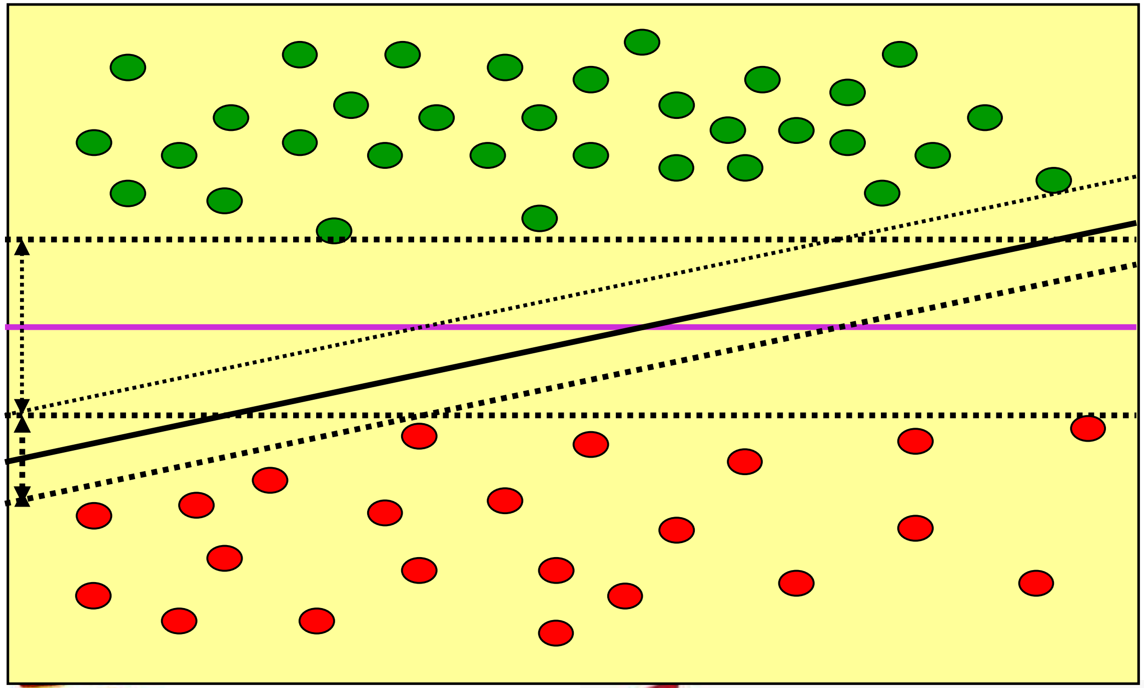
Aprendizagem Estatística
Margem de Separação

- 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.



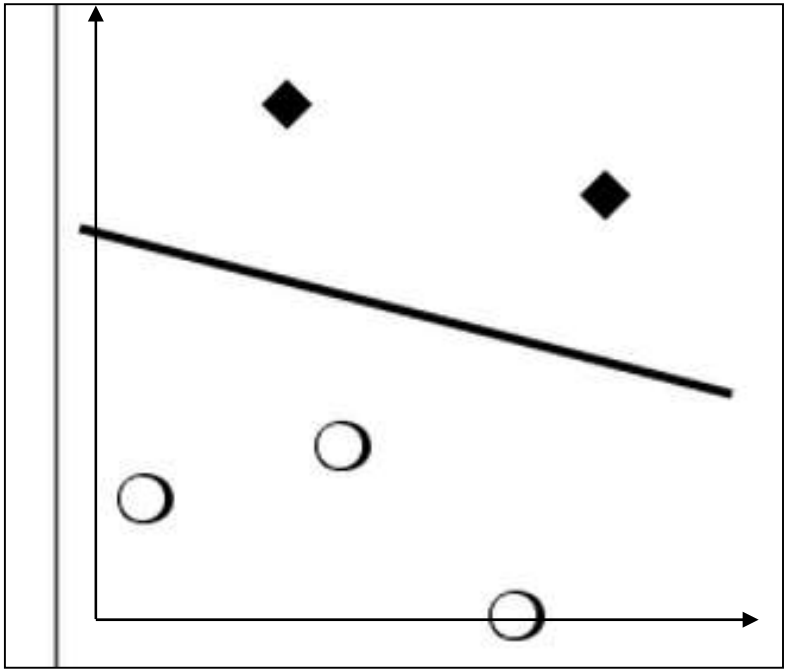
•Quanto maior a margem de um classificador menor será sua dimensão VC (prova está em teorema).

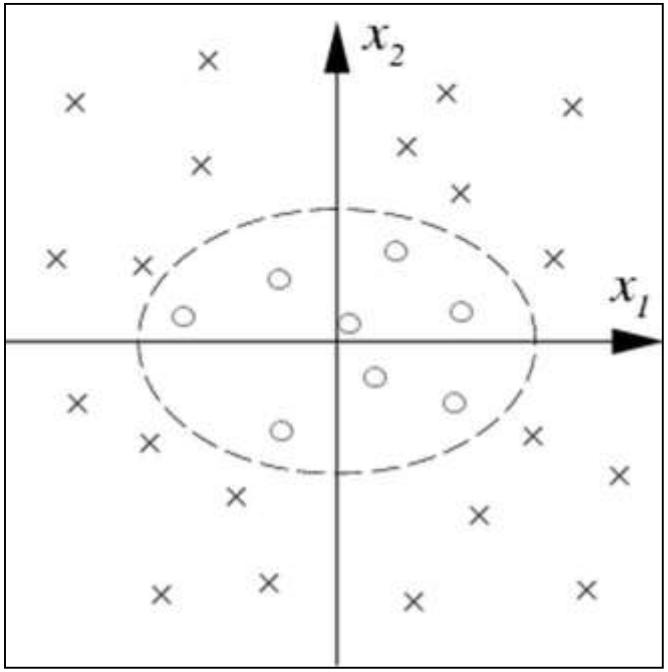
•Hiperplano com margem alta e que minimize os erros de treinamento e teste é chamado de hiperplano ótimo.

Centro de Informática UFPE

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 separa-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 (**w**) e o vetor de entrada. O par (**w**,*b*) é determinado durante o treinamento.

SVM com Margens Rígidas
Hiperplano de Separação

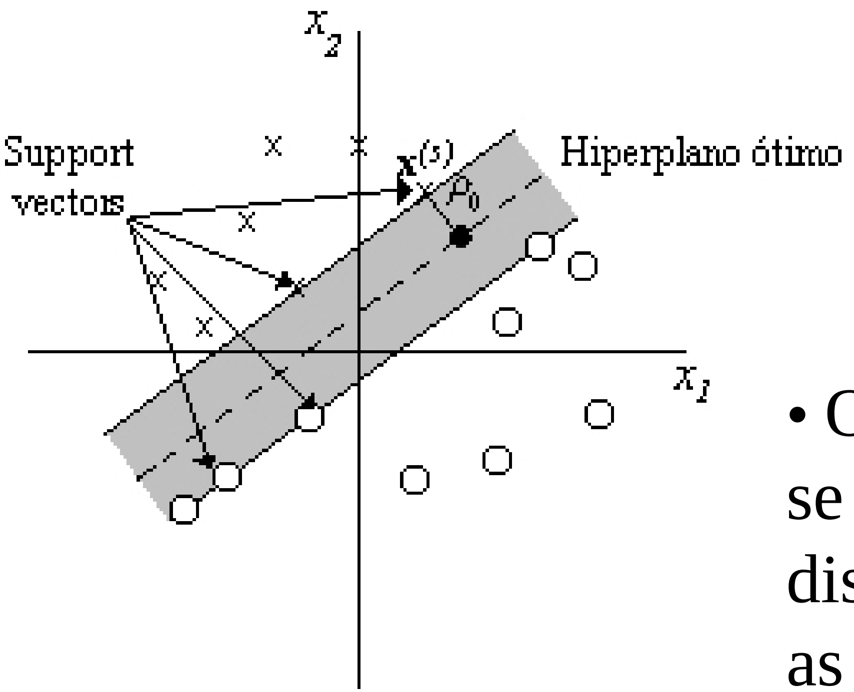
• A equação do hiperplano divide o espaço de entrada em duas regiões que produzem dois tipos de saídas através da uma função sinal:

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

• 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.



$\mathbf{w_o}^T\mathbf{x} + b_\mathbf{o} = 0$, eq. Hiperplano ótimo

$\mathbf{w}_o$, vetor de pesos ótimo

b_o, 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 ($\mathbf{w_o}$) e do bias (b_o) da seguinte forma: $\mathbf{w_o}^T\mathbf{x}+ b_o = 0$.

- A função discriminante $g(\mathbf{x}) = \mathbf{w_o}^T\mathbf{x}+ b_o$ dá uma medida algébrica da distância de $\mathbf{x}$ para o hiperplano ótimo. Neste caso, pode-se escrever:

$$\mathbf{x} = \mathbf{x}_p + r\frac{\mathbf{w}_0}{\|\mathbf{w}_0\|}$$

onde $\mathbf{x_p}$ é a projeção de $\mathbf{x}$ no hiperplano ótimo.

Para encontrar a distância r faz-se:

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

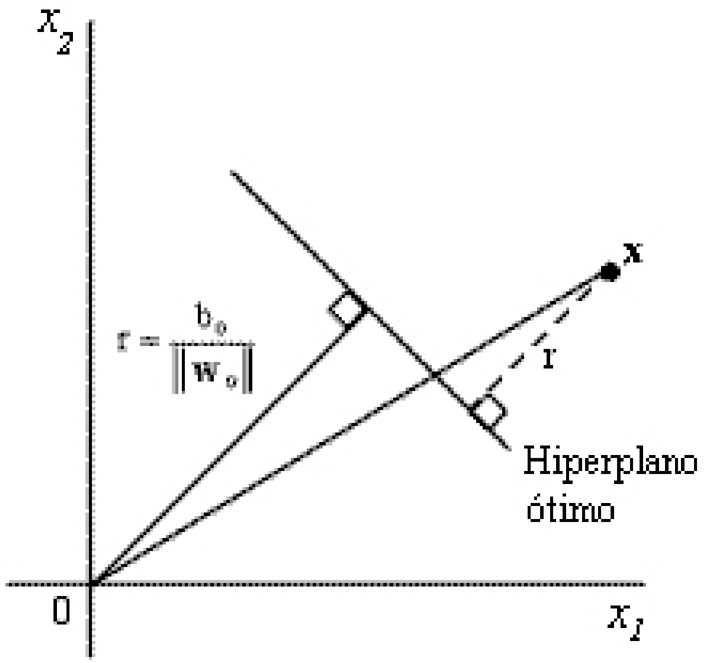
UFPE

Centro de Informática UFPE

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



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 que **w** e *b* para que os pontos mais próximos do hiperplano separador que satisfaçam $|\mathbf{w}^T.\mathbf{x} + b| = 1$. 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

• o valor de r dos vetores suporte para o hiperplano ótimo é calculada:

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

• 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 $\mathbf{w}_0{}^{T}\mathbf{x}^{(s)} \pm b_0 = \pm 1$:

$$\rho_0 = 2r = \frac{2}{\|\mathbf{w}_0\|}$$

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

UFPE

Centro de Informática
UFPE

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 $\mathbf{w}_o$ 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 **w**.

• O problema de otimização com restrições a ser resolvido é:

- Dado o conjunto de treinamento ($\mathbf{x}_i$, d_i), $i=1, ..., N$; Encontre o vetor de pesos **w** e o *bias* b ótimos que satisfaçam às restrições: $d_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1$, e **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:

$$
\begin{aligned}
&\max\ (\min)\ \ f(\mathbf{x}) \\
&\text{s.a.}\quad g_i(\mathbf{x}) = 0, \qquad i = 1, \ldots, N
\end{aligned}
$$

onde f e g_i (i=1,..,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:

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

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

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

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

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

Centro de Informática
UFPE

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:

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

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

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

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

$$\mathbf{w_0} = \sum_{i=1}^{N} \alpha_{0,\,i} d_i \mathbf{x}_i \qquad b_0 = 1 - \mathbf{w_0^T x^{(s)}}, \quad \text{para } d^{(s)} = 1$$

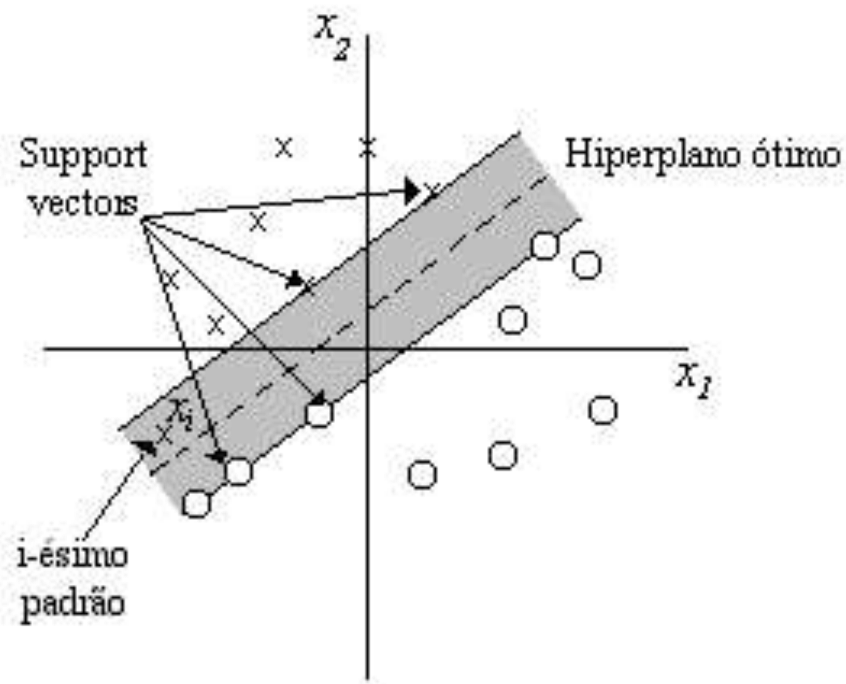
UFPE

Centro de Informática UFPE

SVM com Margens Rígidas
Padrões Não-linearmente Separáveis

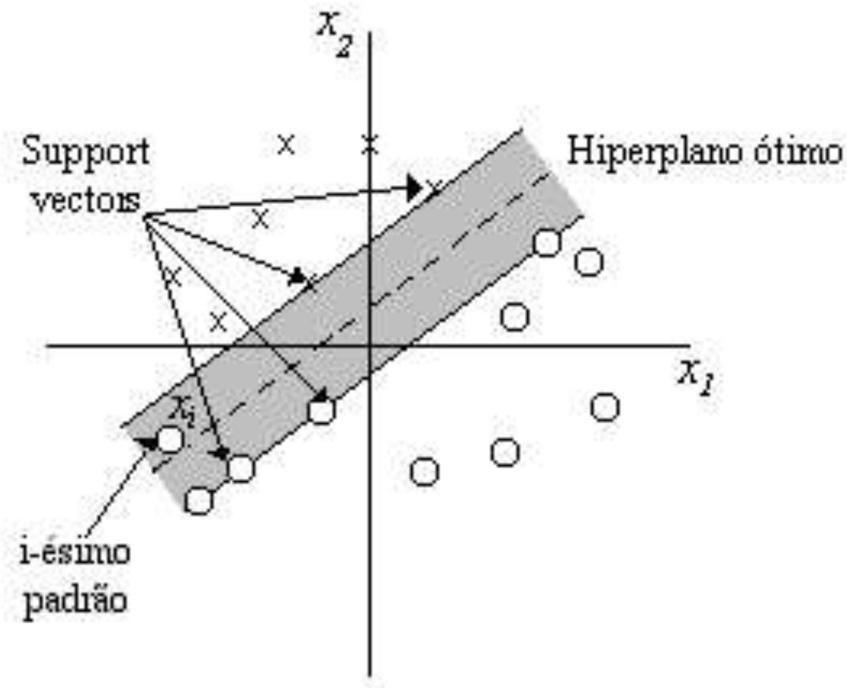
A condição $d_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1$, para $i = 1, 2, \ldots N$

pode ser violada em duas situações:



- 1^a 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



- 2^a 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 $\{\xi_i\}_{i=1}^{N}$.

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

$$0 \leq \xi_i \leq 1\text{:}\ \ 1^a \text{ situação}$$

$$\xi_i > 1\text{:}\ \ 2^a \text{ situação}$$

- 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:

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

$$\sum_{i=1}^{N} \alpha_i d_i = 0 \quad \text{e} \quad 0 \le \alpha_i \le C, \text{ para i = 1, 2, ... N}$$

Logo, $\mathbf{w}_0$ é obtido por: $$\mathbf{w}_0 = \sum_{i=1}^{N} \alpha_{0,1} d_i \mathbf{x}_i$$

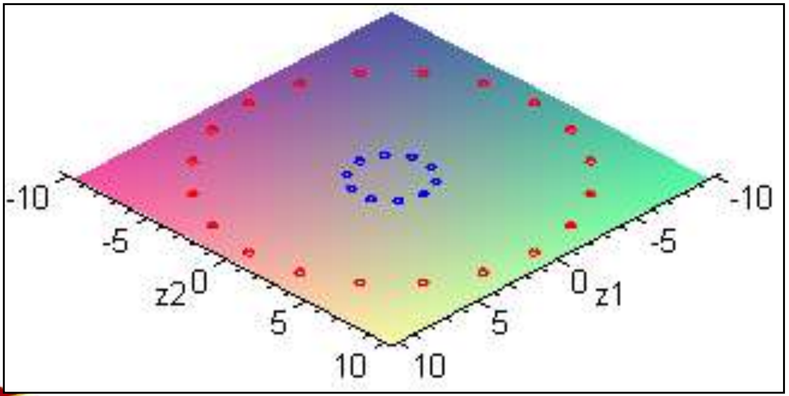
e b_0 através de: $$\alpha_i[y_i\,(\mathbf{w}_0{}^T\mathbf{x}_i + b_0) - 1 + \xi_i] = 0$$

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.

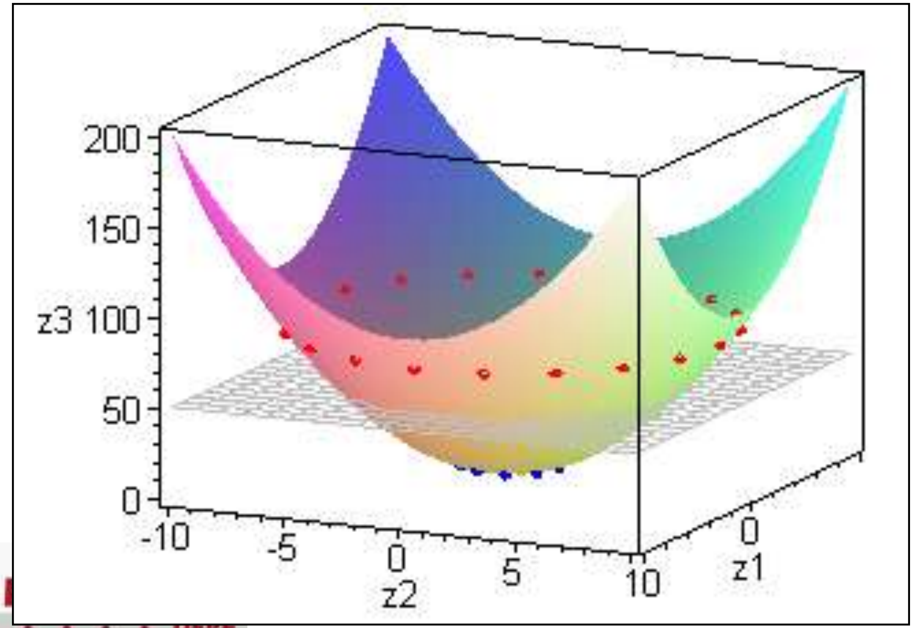
• Dados não-linearmente separáveis podem se tornar linearmente separáveis, em um espaço transformado através de um mapeamento Φ. Este novo espaço é chamado de espaço de características (*feature space*).

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



$\xrightarrow{\Phi}$

Feature Space



SVM Separando Padrões Não-linearmente Separáveis - Mapeamento Φ

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

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

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

• 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 $\mathbf{\Phi}(\mathbf{x}_i)\mathbf{\Phi}(\mathbf{x}_j)$ sem mesmo conhecer o mapeamento Φ.

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

•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:

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

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

Centro de Informática UFPE

U F P E

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:

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

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

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

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

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

Centro de Informática UFPE

U F P E

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:

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

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

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

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

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

SVM e a Função Kernel
Definição e Papel

- Teorema de Mercer: Toda função semi-positiva definida simétrica é um kernel;
- Exemplos de funções de kernel:
- Kernel linear: $K(\mathbf{x}_i,\mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$
- Kernel polinomial: $K(\mathbf{x}_i,\mathbf{x}_j) = (1+\mathbf{x}_i^T \mathbf{x}_j)^p$
- Kernel gaussiano: $K(\mathbf{x}_i,\mathbf{x}_j) = \exp(-\frac{\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2}{2\sigma^2})$
- Kernel sigmoidal: $K(\mathbf{x}_i,\mathbf{x}_j) = \tanh(\beta_0 \mathbf{x}_i^T \mathbf{x}_j + \beta_1)$

UFPE

Centro de Informática UFPE

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

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

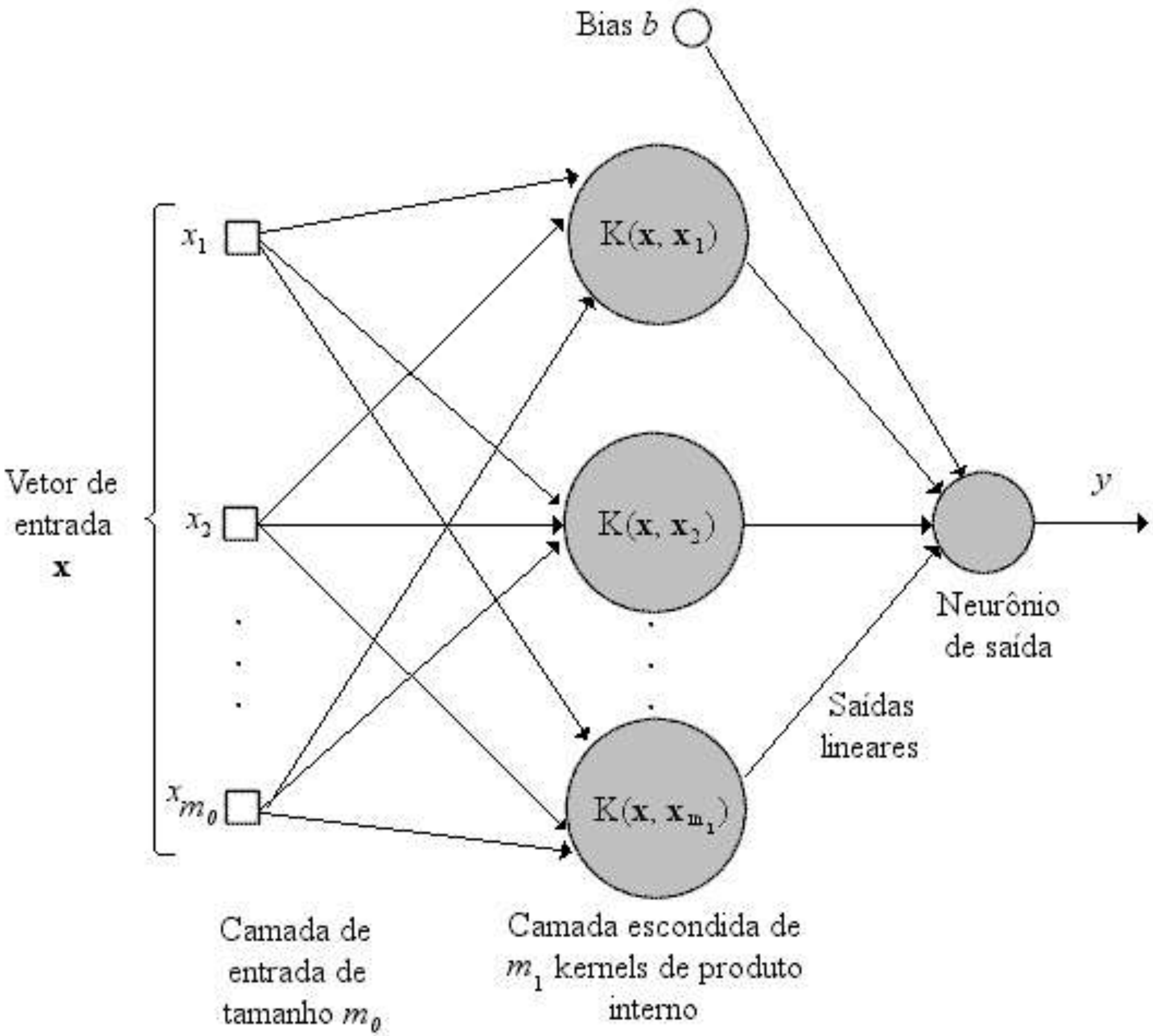
s.a. $\sum_{i=1}^{N} \alpha_i d_i = 0$

$0 \le \alpha_i \le C,$ para $i = 1,2,\ldots,N$ e C é determinado pelo usuário.

SVM e a Função Kernel
Definição e Papel

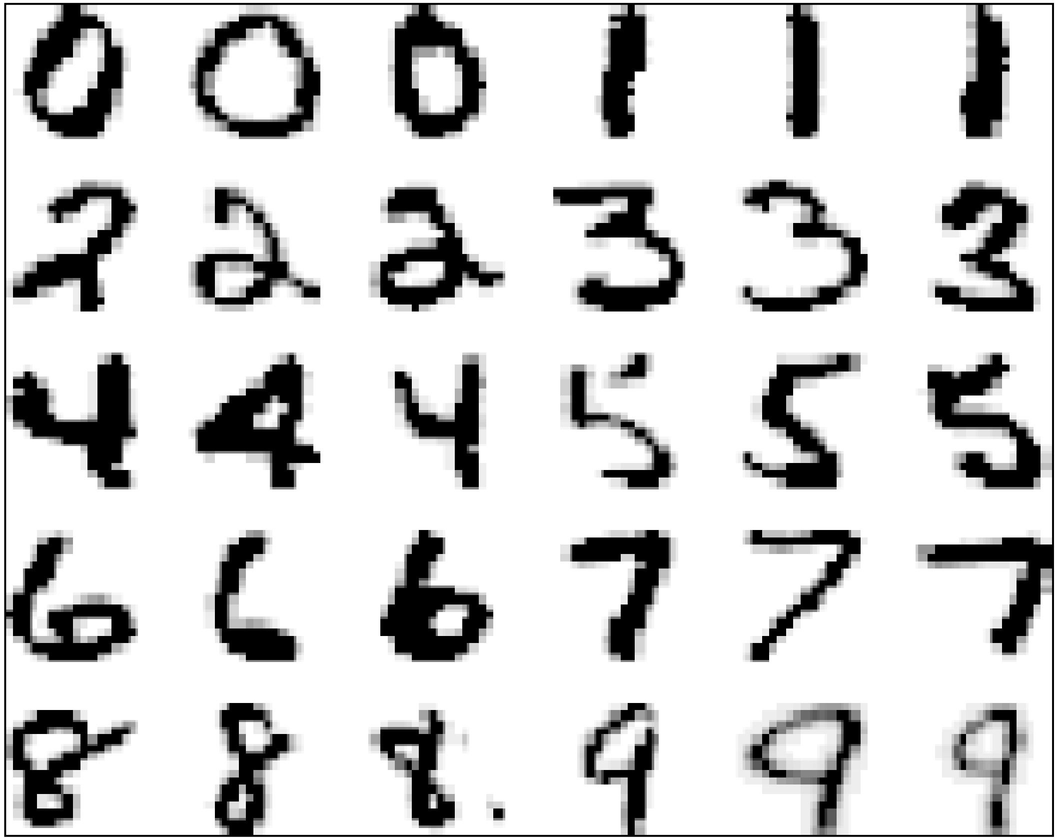
- Três idéias fundamentais:
  - Definição de um hiperplano ótimo de modo que ele possa ser identificado em maneira computacional eficiente: Maximize a margem.
  - Extensão da definição acima para problemas linearmente não-separáveis: Considere uma penalidade para termos equivocadamente classificados.
  - Mapeamento dos dados para um espaço de dimensão mais alta no qual é viável 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

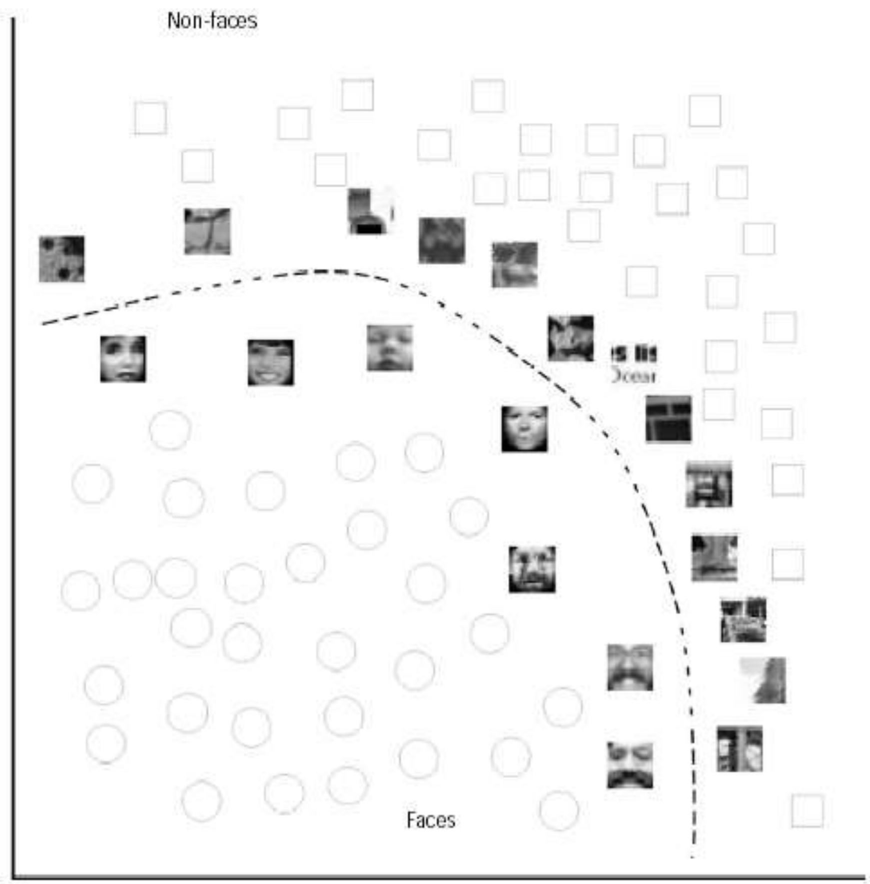
linear PCA & linear SVM (Schölkopf et. al. [11])	8.7%
k-Nearest Neighbor	5.7%
LeNet1 (LeCun et. al. [132], [133], [134])	4.2%
Regularized RBF Networks (Rätsch [128])	4.1%
Kernel-PCA & linear SVM (Schölkopf et. al. [11])	4.0%
SVM (Schölkopf et. al. [120])	4.0%
Virtual SVM (Schölkopf [4])	3.0%
Invariant 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

- Detecção de seres humanos em pé nas imagens
  - Classificação binária da detecção de objetos na imagem;
  - Usa-se o HOG (histogramas de gradientes orientados) para extração de características;
- Exemplo positivos: Exemplo negativo

Discussão

- Os parâmetros têm grande influência no treinamento;
- SVM usa programação matemática para aprender;
- SVM emprega a transformação pelo *kernel* para mapear saídas para espaços de dimensões mais altas;
- SVM tem se caracterizado por bom desempenho, robustez, eficiência e versatilidade ao mesmo tempo que há teoria fundamentando sua capacidade de generalização;
- SVM perde desempenho com dados ruidosos;
- SVM lida com duas classes, portanto, para *n* classes:
  - Constrói-se SVMs que determinar o pertencimento a uma classe contra não pertencer a ela;

U F P E

Centro de Informática UFPE

Sítios

- Alguns sítios interessantes de SVM:
  - https://www.csie.ntu.edu.tw/~cjlin/libsvm/
  - https://www.csie.ntu.edu.tw/~cjlin/libsvm/

Referências

- Du, K.-L. & Swamy M. N. S. (2019). *Neural Networks and Statistical Learning*. Springer, 2nd edition.
- Haykin, S. (2009). *Neural Networks and Learning Machines*. Third Edition. Pearson.
- 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.