Carlos A. Silva

PATTERN RECOGNITION

Introdução: Motivação

Classificação: Abordagens – Model driven X Data driven

- Abordagens à extração de modelos empíricos a partir dos dados:
  - ***Statistical model estimation*:**
    - O objectivo é estimar o modelo "verdadeiro" usando metodologias estatísticas de estimação de modelos.

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})}$$

    - Quais são os desafios nesta abordagem ?
  - ***Preditive learning*:**
    - O objectivo é a estimação de modelos com boa capacidade de generalização em vez de procurar estimar o modelo "verdadeiro".

PI, Carlos A. Silva

Classificação: Definições

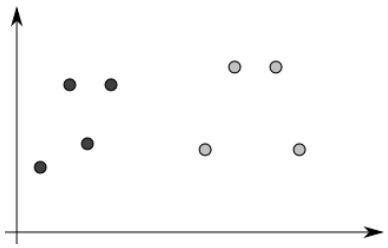
- Definamos:
  - Vectores de treino: x_i, i= 1, ..., L.
  - Vectores de características.
    - Por exemplo,
      - Um paciente= [Altura, peso, ...]
  - Considere o caso em que temos apenas duas classes:
    - Defina um vector indicador y, tal que

$$y_i = \begin{cases} 1, & \text{se } x_i \in \text{classe } 1 \\ -1, & \text{se } x_i \in \text{classe } 2 \end{cases}$$

  - Um hiperplano que separe todos os dados

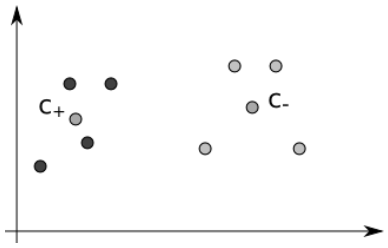
PI, Carlos A. Silva

Classificação: Support Vector Machine



PI, Carlos A. Silva

Classificação: Support Vector Machine

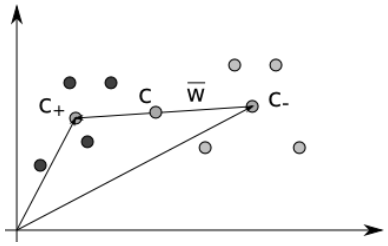


$$\mathbf{c}_+ = \frac{1}{m_+} \sum_{\{i|y_i=+1\}} \mathbf{x}_i,$$

$$\mathbf{c}_- = \frac{1}{m_-} \sum_{\{i|y_i=-1\}} \mathbf{x}_i,$$

PI, Carlos A. Silva

Classificação: Support Vector Machine



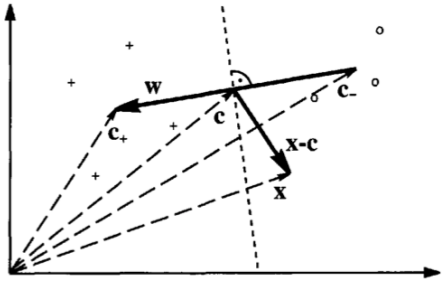
$$\mathbf{c} := (\mathbf{c}_+ + \mathbf{c}_-)/2.$$

$$\mathbf{w} := \mathbf{c}_+ - \mathbf{c}_-$$

Como podemos classificar uma nova amostra com base nos elementos definidos ?

PI, Carlos A. Silva

Classificação: Support Vector Machine



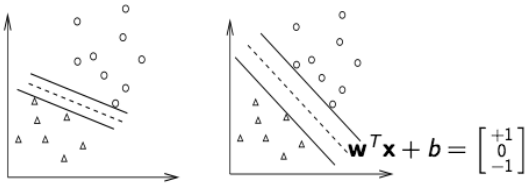
$$y = \operatorname{sgn} \langle (\mathbf{x} - \mathbf{c}), \mathbf{w} \rangle$$
$$= \operatorname{sgn} \langle (\mathbf{x} - (\mathbf{c}_+ + \mathbf{c}_-)/2), (\mathbf{c}_+ - \mathbf{c}_-) \rangle$$
$$= \operatorname{sgn} (\langle \mathbf{x}, \mathbf{c}_+ \rangle - \langle \mathbf{x}, \mathbf{c}_- \rangle + b).$$

$$b := \frac{1}{2}(\|\mathbf{c}_-\|^2 - \|\mathbf{c}_+\|^2),$$

PI, Carlos A. Silva

Classificação: Margem Máxima

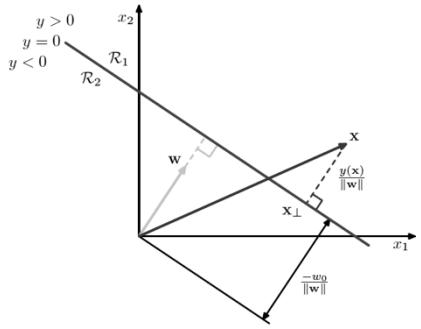
- Pressupondo que os dados são separáveis linearmente.



- Qual é a relação entre o vector de parâmetros **w** e o hiperplano de decisão ?

PI, Carlos A. Silva

Classificação: Margem Máxima



$$y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0$$

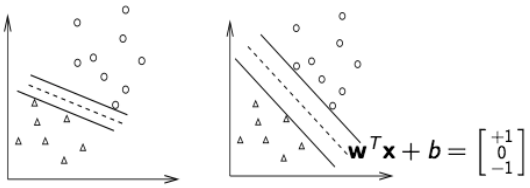
Dois pontos x_A e x_B que estão sobre o hiperplano de decisão, por definição, satisfazem a equação abaixo:

$$y(\mathbf{x}_{\mathrm{A}}) = y(\mathbf{x}_{\mathrm{B}}) = 0$$

$$\mathbf{w}^{\mathrm{T}}(\mathbf{x}_{\mathrm{A}} - \mathbf{x}_{\mathrm{B}}) = 0$$

PI, Carlos A. Silva

Classificação: Margem Máxima



- Equação do hiperplano separador:

$$\mathbf{w}^T\mathbf{x}_i + \mathbf{b} > 0, \quad \text{se } y_i = 1 \qquad \mathbf{w}^T\mathbf{x} + \mathbf{b} = 0$$

$$\mathbf{w}^T\mathbf{x}_i + \mathbf{b} < 0, \quad \text{se } y_i = -1$$

- A função de decisão é dada por $f(\mathbf{x}) = \mathrm{sgn}\left(\mathbf{w}^T\mathbf{x} + \mathbf{b}\right)$, onde x é um vector de teste.
  - Existirão muitas escolhas possíveis para **w** e **b**.

PI, Carlos A. Silva

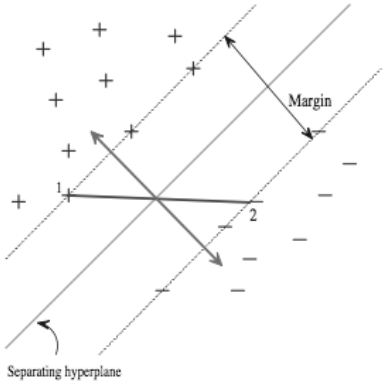
SVM: Determinação da Margem

- Problema em programação quadrática

$$\min_{w,b}\left\{\frac{1}{2}\mathbf{w}^T\mathbf{w}\right\}$$

Satisfazendo a

$$y_i\left(\mathbf{w}^T\mathbf{x}+\mathbf{b}\right)\geq 1,$$
$$i=1,\ldots,L$$



PI, Carlos A. Silva

SVM: Determinação da Margem

- O sistema de equações anteriores pode ser rescrito na seguinte equação de optimização:

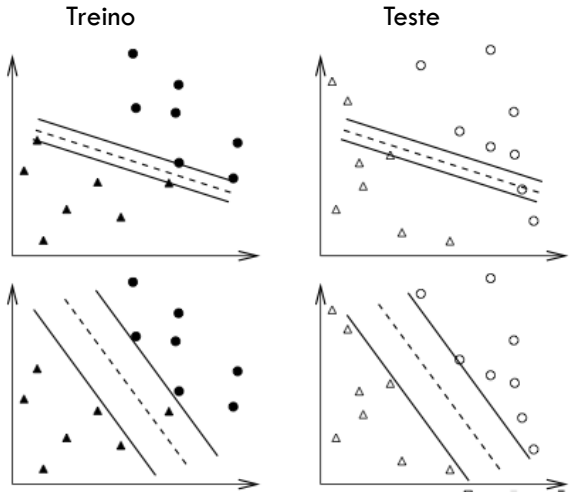
$$L(\mathbf{w},b)=\frac{1}{2}(\mathbf{w}\cdot\mathbf{w})-\sum_{i=1}^{m}\alpha_i\left[y_i((\mathbf{w}\cdot\mathbf{x}_i)+b)-1\right]$$

- Que resulta na seguinte equação de decisão:

$$D(\mathbf{z})=sign\left[\sum_{j=1}^{m}\alpha_j y_j\left(\mathbf{x}_j\cdot\mathbf{z}\right)+b\right]$$

PI, Carlos A. Silva

SVM: Outliers e ruído



PI, Carlos A. Silva

SVM: Reformulação para outliers

- Support Vector Classifier:

$$\underset{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n}{\text{maximize}} \quad M$$

$$\text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 = 1,$$

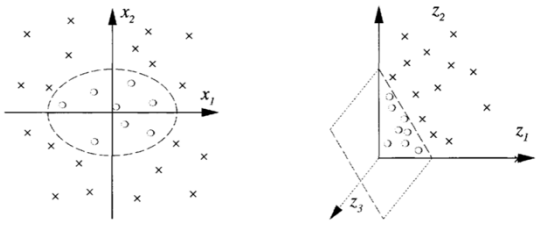
$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i),$$

$$\epsilon_i \geq 0, \quad \sum_{i=1}^{n} \epsilon_i \leq C,$$

PI, Carlos A. Silva

SVM: Espaços não linearmente separáveis

- Exemplo:



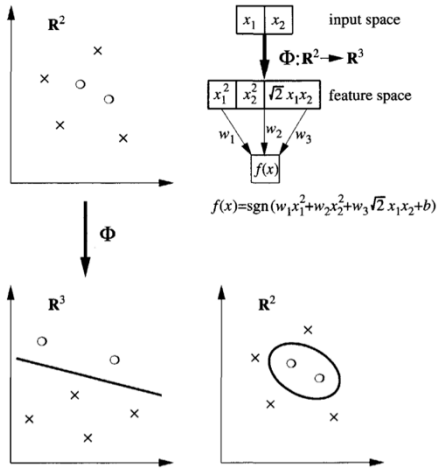
- Podemos permitir erros durante o treino.
- Transformar o espaço de características noutro espaço de ordem superior (possivelmente infinito).

$$\Phi(\mathbf{x}) = (\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \ldots) \qquad \mathbf{x} \in \Re^3,\ \Phi(\mathbf{x}) \in \Re^{10}$$

  - Exemplo:

$$\Phi(\mathbf{x}) = \left(1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_3, x_1^2, x_2^2, x_3^2, \sqrt{2}x_1x_2, \sqrt{2}x_1x_3, \sqrt{2}x_2x_3\right)$$

SVM: Espaços não linearmente separáveis



P201

SVM: Determinação da função de decisão

- O vector **w** pode ter dimensão infinita.
- Neste caso resolvemos o problema de optimização através da optimização da equação dual.

$$\min_{\alpha}\left\{\frac{1}{2}\alpha^T Q\alpha + e^T\alpha\right\}$$

Satisfazendo a $0 \le \alpha_i \le C,\ i = 1,\ldots,L$

$$\mathbf{y}^T\alpha = 0$$

onde $Q_{ij} = y_i y_j \phi(x_i)^T \phi(x_j)$ e $e = [1,\ldots,1]^T$

SVM: Determinação da função de decisão

- Após a terminação dos multiplicadores de Lagrange (a_i) teremos:

$$\mathbf{w} = \sum_{i=1}^{L}\alpha_i y_i \phi(x_i)$$

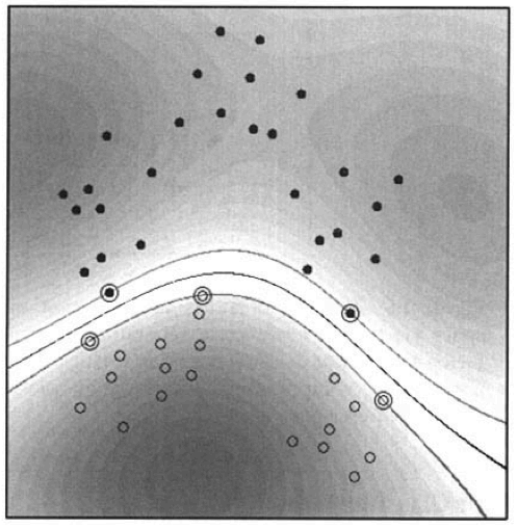
- Com base nos vectores óptimos, a classificação/resposta do sistema a futuras entrada é obtida por

$$f(\mathbf{x}) = \operatorname{sgn}\left(\mathbf{w}^T\phi(\mathbf{x}) + \mathbf{b}\right)$$

$$= \left[\sum_{i=1}^{L}\alpha_i y_i \phi(x_i)\right]^T \phi(\mathbf{x}) + b = \sum_{i=1}^{L}\alpha_i y_i K(x_i, \mathbf{x}) + b$$

- Usamos apenas os $\phi(x_i)$ para $\alpha_i > 0$ (*Support vectors*).

SVM: Determinação da função de decisão



PI, Carlos A. Silva

SVM: Determinação da função de decisão

- No caso finito: # variáveis= # dados de treino.
- No equação dual temos

$$Q_{ij} = y_i y_j \phi(x_i)^T \phi(x_j)$$

ao produto $\phi(x_i)^T \phi(x_j)$ chamamos de *kernel*, tal que

$K(\mathbf{x},\mathbf{y}) = \phi(\mathbf{x})^T \phi(\mathbf{y})$

PI, Carlos A. Silva

SVM: Kernels

- Os *kernels* mais comuns são:

$$K(\mathbf{x},\mathbf{y}) = \phi(\mathbf{x})^T \phi(\mathbf{y})$$

  - Radial Basis Function (RBF),

$$K(x_i, x_j) = \mathrm{e}^{-\gamma\|x_i - x_j\|^2}$$

  - Kernel polinomial,

$$K(\mathbf{x}_i, \mathbf{x}_j) = \left(\frac{1}{a}\mathbf{x}_i^T\mathbf{x}_j + b\right)^d$$

P207

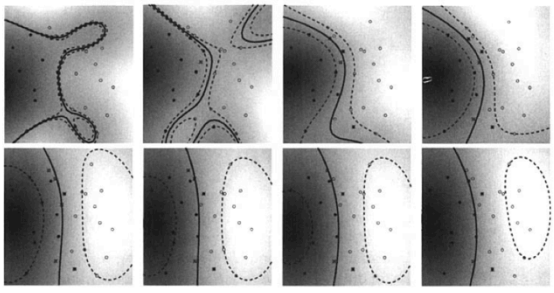


Figure 7.9 Toy problem (task: separate circles from disks) solved using ν-SV classification, with parameter values ranging from $\nu = 0.1$ (top left) to $\nu = 0.8$ (bottom right). The larger we make ν, the more points are allowed to lie inside the margin (depicted by dotted lines). Results are shown for a Gaussian kernel, $k(x, x') = \exp(-\|x - x'\|^2)$.

ν	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
fraction of errors	0.00	0.07	0.25	0.32	0.39	0.50	0.61	0.71
fraction of SVs	0.29	0.36	0.43	0.46	0.57	0.68	0.79	0.86
margin $\rho/\|\mathbf{w}\|$	0.005	0.018	0.115	0.156	0.364	0.419	0.461	0.546