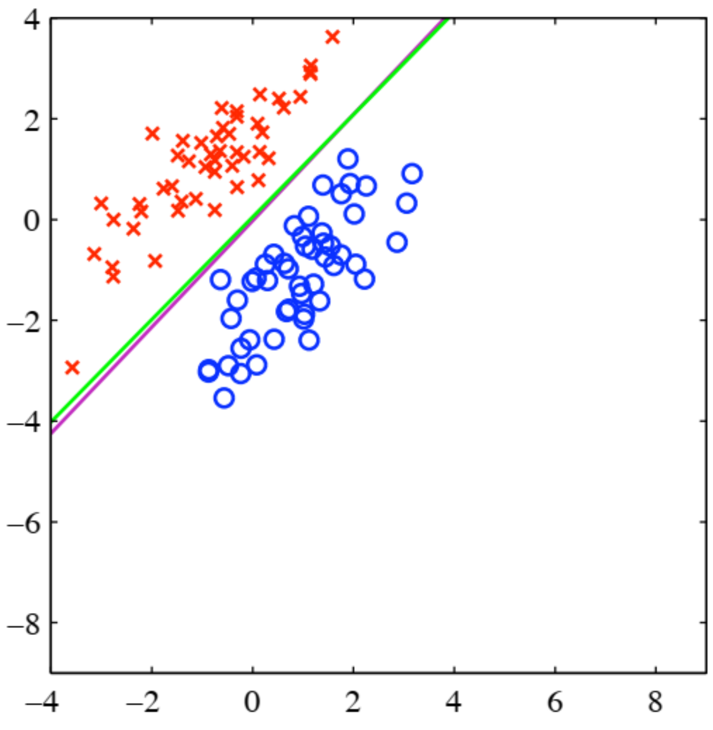
Sistemas Inteligentes

CLASSIFICADOR DE MÍNIMOS QUADRADOS, LDA E QDA

Modelos Lineares para Classificação

Modelos lineares sãocapazes de separar o espaço de características por meio de fronteiras de decisão lineares



Por exemplo, considere que os padrões a serem classificados correspondem a vetores com duas components, $\mathbf{x} = [x_1, x_2]^T$.

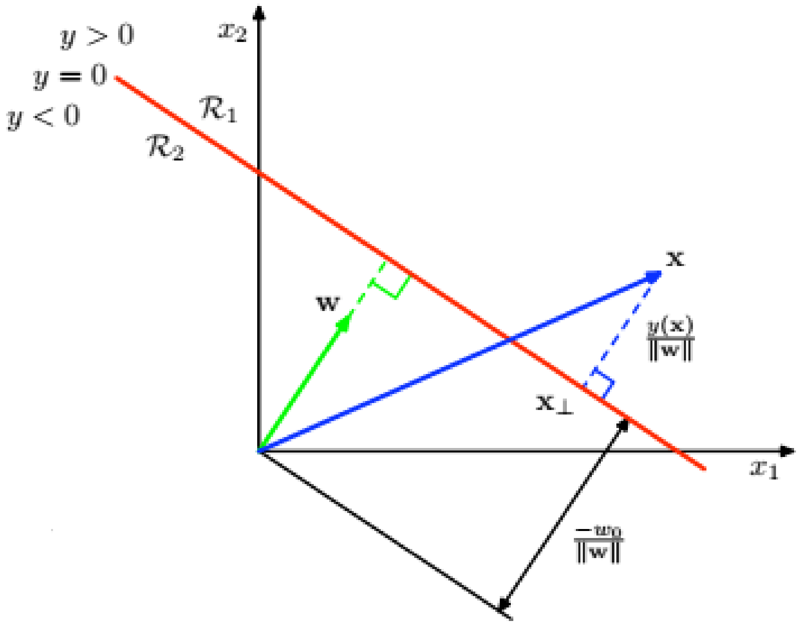
Um função discriminante linear é dada matematicamente por

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

Discriminante de duas classes

Considere que $f(x) = x$, i.e.,

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



Assim,

- $y(\mathbf{x}) \geq 0 \rightarrow \mathbf{x}$ pertence à classe C_1
- $y(\mathbf{x}) < 0 \rightarrow \mathbf{x}$ pertence à classe C_2

Ajustando os parâmetros por mínimos quadrados

Para obter os parâmetros $\mathbf{w}$ e w_0 que realizam a classificação da melhor maneira possível podemos empregar o **método dos mínimos quadrados**.

Para isso, podemos escrever a equação da função discriminante como

$$y(\mathbf{x}) = [w_0 \quad \mathbf{w}^T] \begin{bmatrix} 1 \\ \mathbf{x} \end{bmatrix} = \widetilde{\mathbf{w}}^{\boldsymbol{T}} \tilde{\mathbf{x}}$$

Supondo que os dados de treinamento para um classificador binário correspondem aos pares de vetores e rótulos dados por $(\mathbf{x}_n, t_n)$, para $n = 1, \ldots, N$ (rótulos binários) o método de mínimos quadrados busca ajustar o vetor de parâmetros $\widetilde{\mathbf{w}}$ de maneira a minimizar o erro quadrático da classificação, i.e.,

$$\min \textstyle\sum_n e_n^2 = \min \textstyle\sum_n \left(\widetilde{\mathbf{w}}^{\boldsymbol{T}} \tilde{\mathbf{x}}_n - t_n\right)^2$$

Ajustando os parâmetros por mínimos quadrados

Se os dados de treinamento $\mathbf{x}_n$ e t_n forem organizados de acordo com

$$\widetilde{\mathbf{X}} = \begin{bmatrix} 1 & \mathbf{x}_0^T \\ \vdots & \vdots \\ 1 & \mathbf{x}_N^T \end{bmatrix} \text{ e } \mathbf{T} = \begin{bmatrix} t_0 \\ \vdots \\ t_n \end{bmatrix}$$

A solução para o conjunto de parâmetros ótimos é dada por

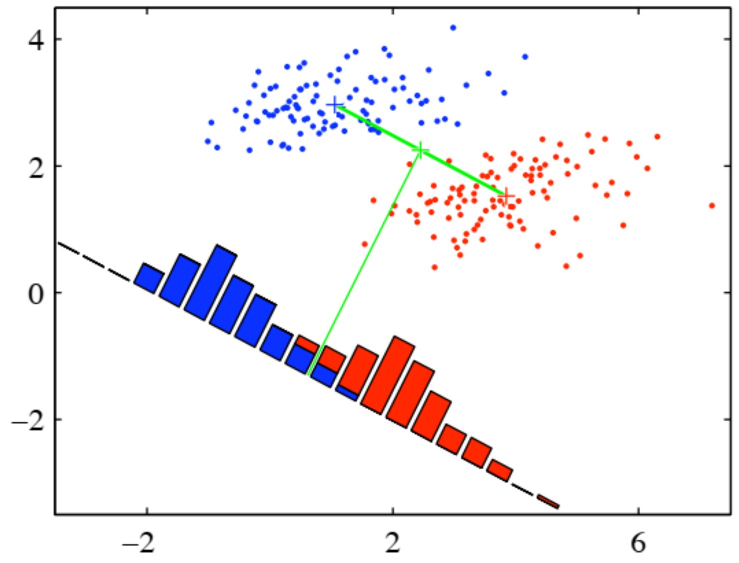
$$\widetilde{\mathbf{w}} = \begin{bmatrix} w_0 \\ \mathbf{w} \end{bmatrix} = \left(\widetilde{\mathbf{X}}^T \widetilde{X}\right)^{-1} \widetilde{\mathbf{X}}^T \mathbf{T}$$

Outra possibilidade – Discriminante de Fischer

Outra forma de interpretar os discriminantes lineares: encontrar o subespaço 1D que maximiza a separação entre as duas classes

Em outras palavras, procuramos um vetor $\mathbf{w}$ (que define uma direção no espaço dos dados) sobre o qual faremos a projeção dos dados, e gostaríamos que a separação entre os dados projetados da classe 1 e da classe 2 seja a maior possível.

A projeção do dado $\mathbf{x}_i$ na direção definida por $\mathbf{w}$ é dada por $\mathbf{w}^T\mathbf{x}_i$



Discriminante de Fischer

Como gostaríamos que a separação dos dados projetados seja a maior possível, é necessário definir uma medida que quantifique essa separação. Considerando

$$\tilde{\mu}_1 = \frac{1}{N_1}\sum_{n\in C_1} \mathbf{w}^T\mathbf{x}_n \qquad\qquad \tilde{\mu}_2 = \frac{1}{N_2}\sum_{n\in C_2} \mathbf{w}^T\mathbf{x}_n$$

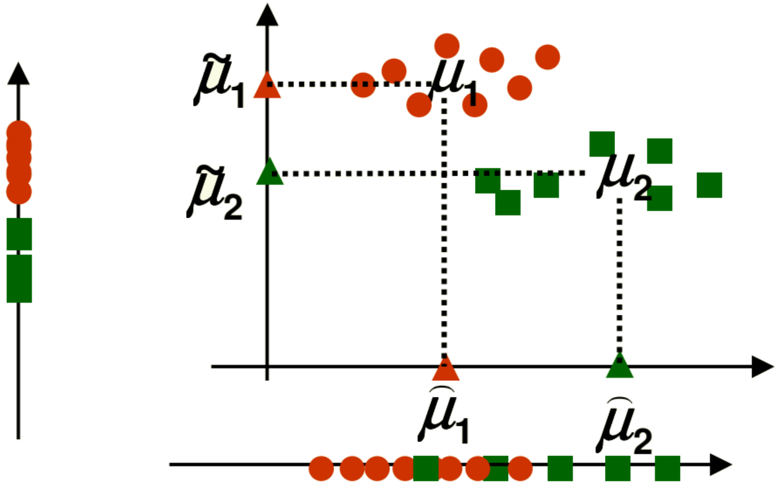
as médias dos dados projetados da Classe 1 e Classe 2, respectivamente, uma possibilidade de medida de separação entre os dados projetados seria

$$|\tilde{\mu}_1 - \tilde{\mu}_2| = \left|\frac{1}{N_1}\sum_{n\in C_1} \mathbf{w}^T\mathbf{x}_n - \frac{1}{N_2}\sum_{n\in C_2} \mathbf{w}^T\mathbf{x}_n\right|$$

$$= \left|\mathbf{w}^T \overbrace{\left(\frac{1}{N_1}\sum_{n\in C_1} \mathbf{x}_n\right)}^{\mathbf{m}_1} - \mathbf{w}^T \overbrace{\left(\frac{1}{N_2}\sum_{n\in C_2} \mathbf{x}_n\right)}^{\mathbf{m}_2}\right| = |\mathbf{w}^T(\mathbf{m}_1 - \mathbf{m}_2)|$$

onde $\mathbf{m}_1$ e $\mathbf{m}_2$ correspondem à media dos dados da Classe 1 e 2, respectivamente.

Discriminante de Fischer



A figura ilustra a diferença entre as médias dos dados projetados em duas direções distintas.

Embora $|\hat{\mu}_1 - \hat{\mu}_2|$ seja maior do que $|\tilde{\mu}_1 - \tilde{\mu}_2|$, nota-se claramente que a separação entre os dados projetados é maior quando consideramos a projeção no eixo vertical

O problema é que o critério adotado não considera a variância dos dados de uma mesma classe.

Como incorporar a variância dos dados da classe?

Considere uma medida de dispersão dos dados projetados de uma mesma classe (intra-classe), dada por

$$\tilde{s}_i^2 = \sum_{n \in C_i} (\mathbf{w}^T \mathbf{x}_n - \tilde{\mu}_i)^2$$

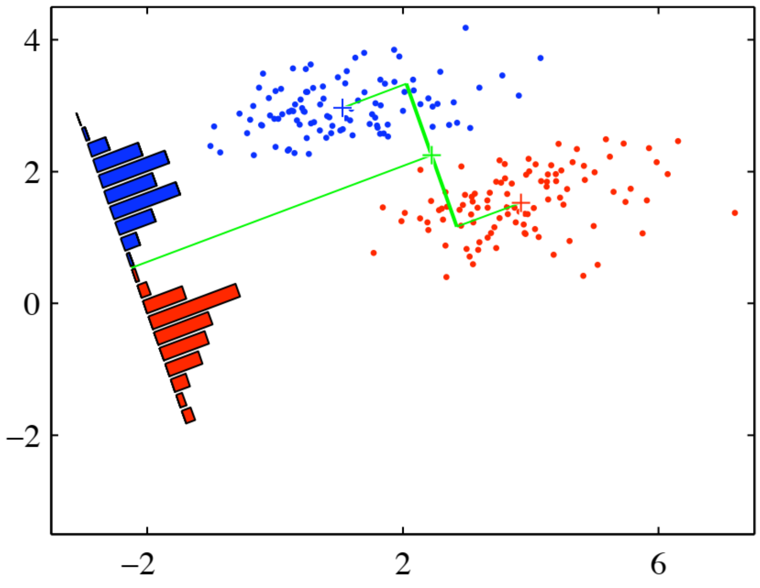
i.e., uma estimativa da variância dos dados projetados da classe C_i multiplicada pelo número de dados dessa classe.

Com isso, podemos normalizar a medida de separação $|\tilde{\mu}_1 - \tilde{\mu}_2|$ pela dispersão dos dados projetados, i.e., obter $\mathbf{w}$ de maneira a maximizar

Médias projetadas afastadas

$$J(\mathbf{w}) = \frac{(\tilde{\mu}_1 - \tilde{\mu}_2)^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$$

Dispersões tão pequenas quanto possível



Solução para o discriminante de Fischer

Definindo-se as matrizes de dispersão dos dados de cada classe

$$\mathbf{S}_1 = \sum_{n \in C_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^T$$

$$\mathbf{S}_2 = \sum_{n \in C_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^T$$

é possível definir a matriz de dispersão intra-classe como

$$\mathbf{S}_w = \mathbf{S}_1 + \mathbf{S}_2$$

A solução para o discriminante de Fischer é dada por

$$\mathbf{w} = \mathbf{S}_w^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

Linear Discriminant Analysis (LDA)

Na literatura, o discriminante de Fischer e a LDA são utilizados quase como sinônimos. A abordagem da LDA, entretanto, é desenvolvida tendo como base algumas hipóteses que não foram diretamente exploradas por Fischer:

- Os dados das classes possuem distribuição Gaussiana
- As matrizes de covariância dos dados de cada classe são iguais (Homoscedasticidade)

Embora matematicamente a LDA e o discriminante de Fischer sejam semelhantes, a derivação do LDA parte de um modelo estatístico para os dados, considerando uma distribuição condicional para os dados

$$p(\mathbf{x}|\text{Classe} = k)$$

Veremos mais adiante no curso que essa distribuição condicional está relacionada a um método importante de estimação de parâmetros, denominado de método de máxima verossimilhança.

Linear Discriminant Analysis (LDA)

Utilizando a regra de Bayes é possível escrever

$$p(\text{Classe} = k|\mathbf{x}) = \frac{p(\mathbf{x}|\text{Classe} = k)p(\text{Classe} = k)}{p(\mathbf{x})}$$

ou seja, uma vez que tenhamos observado o vetor $\mathbf{x}$ existe uma probabilidade dele pertencer à Classe k dada por $p(\text{Classe} = k|\mathbf{x})$*, a distribuição a posteriori da Classe.*

A hipótese central do LDA consiste em considerar $p(\mathbf{x}|\text{Classe} = k)$ como uma distribuição Gaussiana multivariada, i.e.,

$$p(\mathbf{x}|\text{Classe} = k) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k)\right)$$

onde $\boldsymbol{\mu}_k$ e Σ_k denotam, respectivamente, o vetor de média e a matriz de covariância da distribuição Gaussiana, e d é a dimensão de $\mathbf{x}$ (número de features).

Linear Discriminant Analysis (LDA)

Assim, a decisão sobre a classe a qual pertence um vetor de dados $\mathbf{x}$ pode ser feita comparando-se o valor de $p(\text{Classe} = k|\mathbf{x})$ para diferentes k.

A comparação pode ser feita por meio da razão

$$\log\left(\frac{p(\text{Classe} = k|\mathbf{x})}{p(\text{Classe} = l|\mathbf{x})}\right) = \log\left(\frac{p(\mathbf{x}|\text{Classe} = k)p(\text{Classe} = k)}{p(\mathbf{x}|\text{Classe} = l)p(\text{Classe} = l)}\right)$$

se o valor for positivo, o vetor pertence à classe k, e se for negativo pertence à classe l.

Linear Discriminant Analysis (LDA)

O limiar, portanto, pode ser obtido avaliando-se quando a equação se iguala a zero. Como na LDA considera-se que todas as matrizes de covariância das classes são iguais, i.e., $\Sigma_k = \Sigma$, o limiar é dado por

$$\log\left(\frac{p(\mathbf{x}|\text{Classe} = k)p(\text{Classe}=k)}{p(\mathbf{x}|\text{Classe} = l)p(\text{Classe}=l)}\right) = 0$$

$$\log\left(\frac{p(\mathbf{x}|\text{Classe} = k)}{p(\mathbf{x}|\text{Classe} = l)}\right) + \log\left(\frac{p(\text{Classe}=k)}{p(\text{Classe}=l)}\right)$$

$$= (\boldsymbol{\mu}_k - \boldsymbol{\mu}_l)^T\Sigma^{-1}\mathbf{x} - \frac{1}{2}\left(\boldsymbol{\mu}_k^T\Sigma^{-1}\boldsymbol{\mu}_k - \boldsymbol{\mu}_l^T\Sigma^{-1}\boldsymbol{\mu}_l\right) + \log\left(\frac{p(\text{Classe}=k)}{p(\text{Classe}=l)}\right) = 0$$

ou seja,

$$\underbrace{(\boldsymbol{\mu}_k - \boldsymbol{\mu}_l)^T\Sigma^{-1}\mathbf{x}}_{\mathbf{w}^T\mathbf{x}} = \boxed{\frac{1}{2}\left(\boldsymbol{\mu}_k^T\Sigma^{-1}\boldsymbol{\mu}_k - \boldsymbol{\mu}_l^T\Sigma^{-1}\boldsymbol{\mu}_l\right) - \log\left(\frac{p(\text{Classe} = k)}{p(\text{Classe} = l)}\right)}$$

$\mathbf{w}^T\mathbf{x}$

Fronteira de decisão linear

Valor escalar

Quadratic Discriminant Analysis (QDA)

Na QDA, entretanto, não assume-se que as matrizes de covariância são iguais. O procedimento para obter a fronteira de decisão é o mesmo que na LDA, i.e.

$$\log\left(\frac{p(\mathbf{x}|\text{Classe} = k)p(\text{Classe}=k)}{p(\mathbf{x}|\text{Classe} = l)p(\text{Classe}=l)}\right) = 0$$

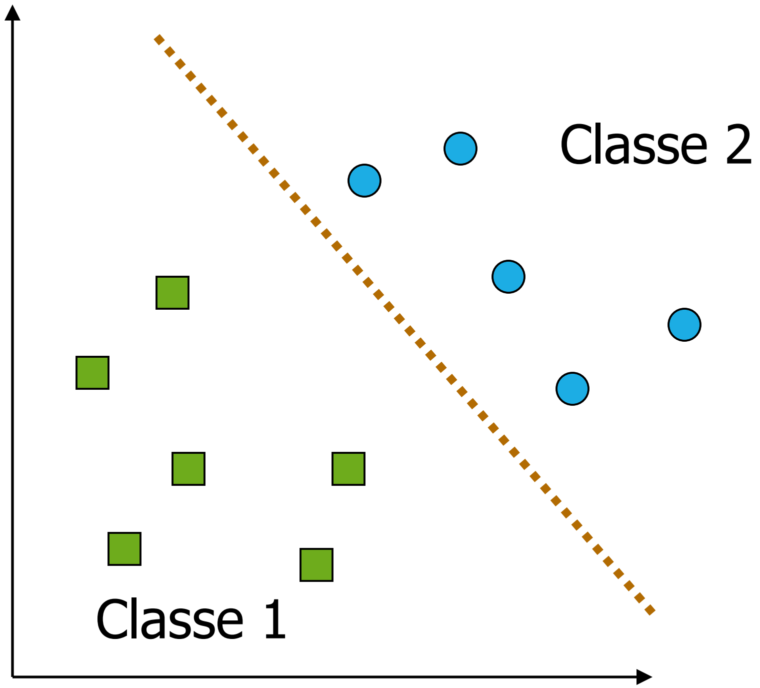
$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T\Sigma_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_l)^T\Sigma_l^{-1}(\mathbf{x}-\boldsymbol{\mu}_l) + \log\left(\frac{p(\text{Classe}=k)}{p(\text{Classe}=l)}\right) = 0$$

Nesse caso, entretanto, não é possível realizar todas as simplificações que no LDA, e assim a superfície de decisão ótima deixa de ser linear (torna-se quadrática, conforme indica o nome da ferramenta)

Sistemas Inteligentes

SUPPORT VECTOR MACHINES

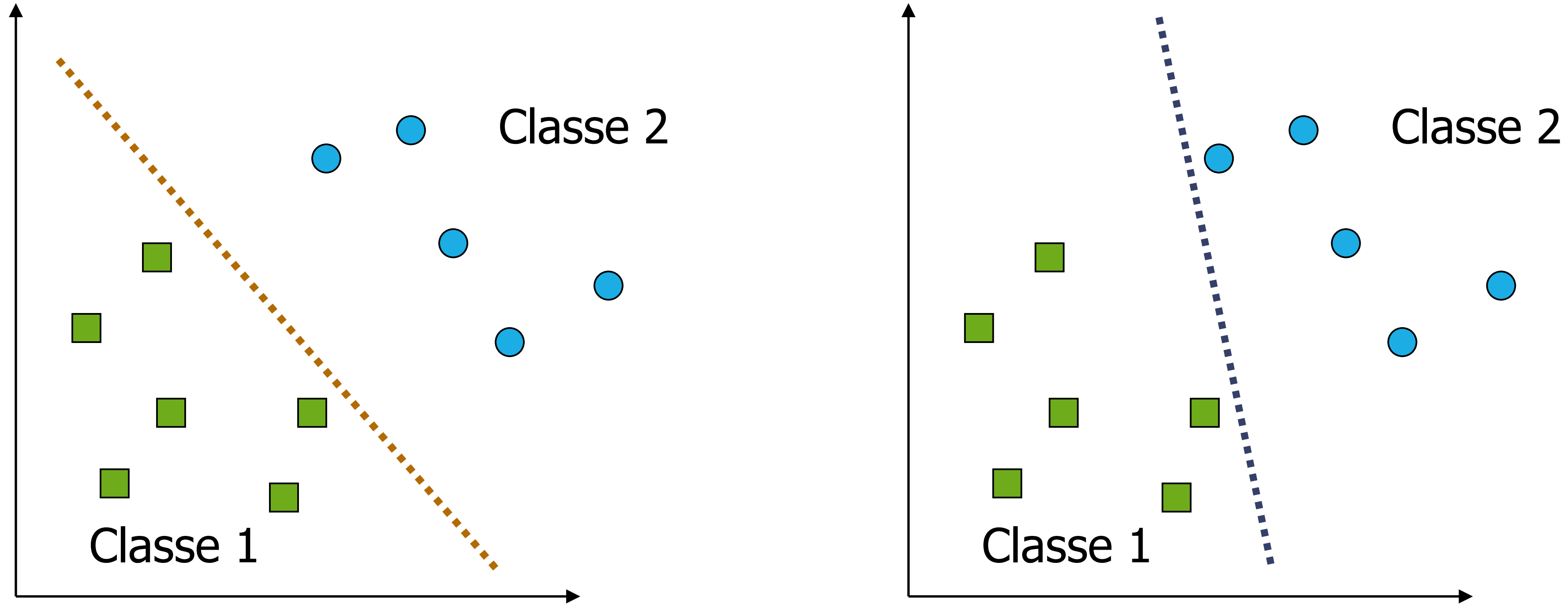
Problema de Classificação de Padrões – Classes linearmente separáveis



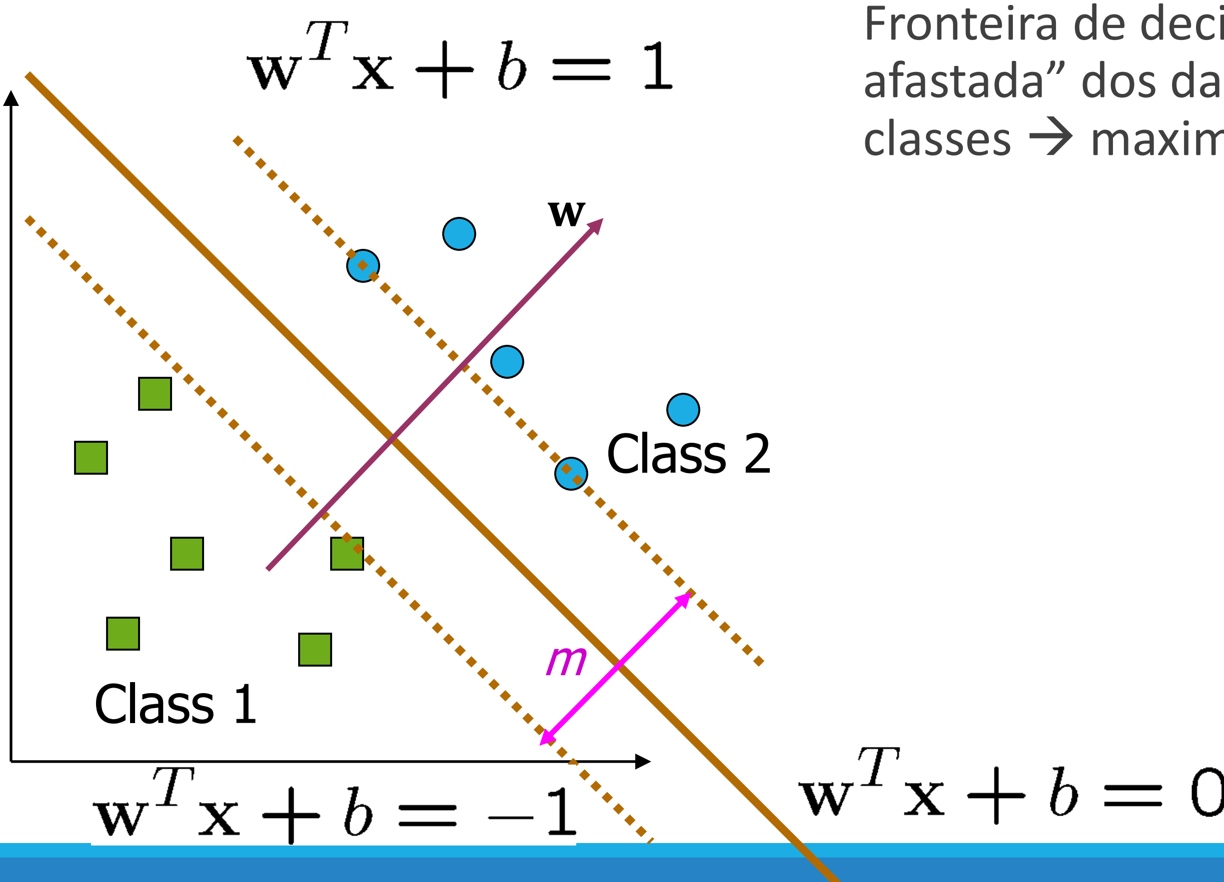
Várias fronteiras de decisão podem separar as classes

Qual delas escolher?

Examplos de possíveis fronteiras de decisão



Critério de escolha: maximização da margem



Fronteira de decisão deve ser "a mais afastada" dos dados de ambas as classes → maximização da margem *m*

Definição do problema

Conjunto de *n* padrões de treinamento $(\mathbf{x}_i, di)$ onde $\mathbf{x}_i$ denota um vetor de características e d_i é a saída desejada. Seja $d_i = +1$ para exemplos positivos e $d_i = -1$ para exemplos negativos.

Considere que os padrões são linearmente separáveis → fronteira de separação dada por um hiperplano

Definindo o hiperplano

Hiperplano de separação é descrito por:

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

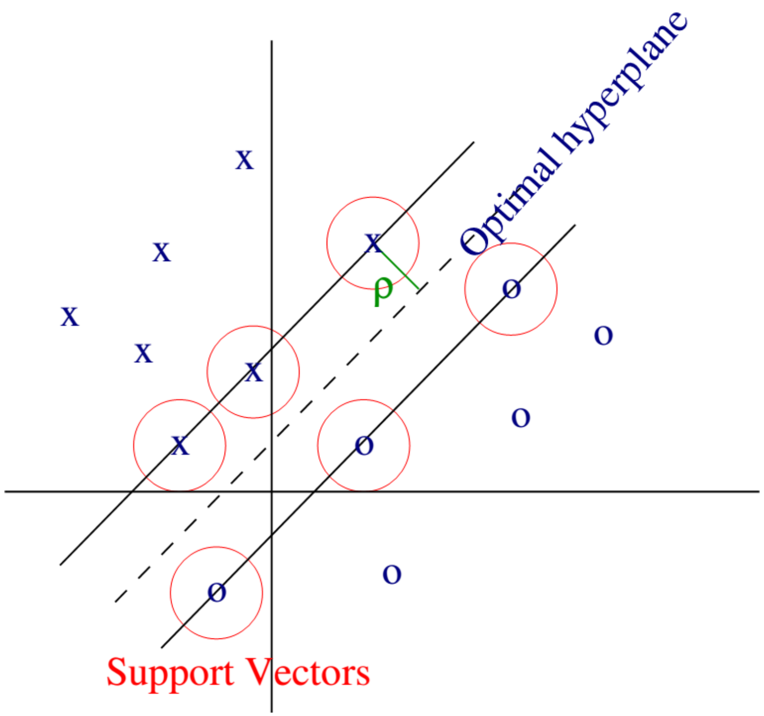
- **w** é o vetor de parâmetros
- **x** é o vetor de entrada
- *b* é o *bias*

Assim, temos que

$$\mathbf{w}^T\mathbf{x} + b > 0 \text{ for } d_i = +1$$
$$\mathbf{w}^T\mathbf{x} + b < 0 \text{ for } d_i = -1$$

$\mathbf{w}_0 \; e \; b_0$: parâmetros ótimos

Hiperplano Ótimo e Vetores Suporte



Vetores Suporte: Vetores de entrada mais próximos do hiperplano de separação

Margem de Separação ρ: distância entre o hiperplano de separação e o vetor suporte

Pode-se mostrar que maximizar a margem entre duas classes é equivalente a minimizar a norma Euclidiana do vetor de pesos $\mathbf{w}_0$!

Solução Ótima obtida por Otimização com restrições

Portanto, o conjunto de parâmetros ótimos é obtido por meio de um problema de minimização com restrições, i.e.,

$$\min \left\{ \Phi(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} \right\}$$

Sujeito à

$$d_i(\mathbf{w}_o^T\mathbf{x}_i + b_0) \geq 1$$

para $i = 1,2, \ldots, N$

A solução pode ser obtida com o método de multiplicadores de Lagrange

Otimização com restrições – Multiplicadores de Lagrange

No caso da SVM, a função custo é convexa, e os pontos que satisfazem as restrições forma um conjunto convexo. Dessa forma, o problema possui apena um mínimo global.

O método de multiplicadores de Lagrange baseia-se na construção de uma função Lagrangiana, que incorpora as restrições (ponderadas pelos multiplicadores),

$$L(\mathbf{w}, b, \boldsymbol{\lambda}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_i \lambda_i (d_i(\mathbf{w}^T\mathbf{x}_i + b) - 1\,)$$

Para obter a solução, temos de obter

$$\frac{\partial L}{\partial b} = 0 \;\text{ e } \frac{\partial}{\partial \mathbf{w}} = 0$$

Multiplicadores de Lagrange – forma dual

$$\frac{\partial L}{\partial b} = -\sum_i \lambda_i d_i = 0 \qquad \rightarrow \qquad \sum_i \lambda_i d_i = 0$$

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_i \lambda_i d_i \mathbf{x}_i \qquad \rightarrow \qquad \mathbf{w} = \sum_i \lambda_i d_i \mathbf{x}_i$$

Substituindo os resultados na função Lagrangiana, obtemos a forma dual do problema de otimização

$$L(\mathbf{w}, b, \boldsymbol{\lambda}) = \frac{1}{2}\|\mathbf{w}\|^2 + \sum_i \lambda_i - \sum_i \lambda_i d_i \mathbf{w}^T \mathbf{x}_i - \sum_i \lambda_i d_i b$$

$$= \sum_i \lambda_i - \frac{1}{2}\sum_i \sum_j \lambda_i \lambda_j d_i d_j (\mathbf{x}_i^T \mathbf{x}_j)$$

Produto interno entre os padrões

com as restrições

$$\begin{cases} \lambda_i \geq 0, \qquad \forall_i = 1, \dots, n \\ \displaystyle\sum_i \lambda_i d_i = 0 \end{cases}$$

Solução do problema de otimização

Encontrar a solução para o problema de otimização (forma primal ou dual) requer o uso de ferramentas de otimização - algumas delas especificamente desenvolvidas para SVMs.

De qualquer forma, pela teoria de otimização, a solução ótima do problema posto deve respeitar a condição

$$\lambda_i(d_i(\mathbf{w}^T\mathbf{x}_i + b) - 1\ ) = 0, \qquad \forall_i$$

e, para isso, há duas possibilidades:

- $\lambda_i = 0$, o que significa que $d_i(\mathbf{w}^T\mathbf{x}_i + b) - 1$ pode assumir qualquer valor
- $d_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 = 0$, que significa que $(\mathbf{w}^T\mathbf{x}_i + b) = \pm 1$, ou seja, $\mathbf{x}_i$ é um vetor suporte.

São justamente os vetores suporte que participam na determinação da fronteira de decisão.

Classificação com Margem Suave

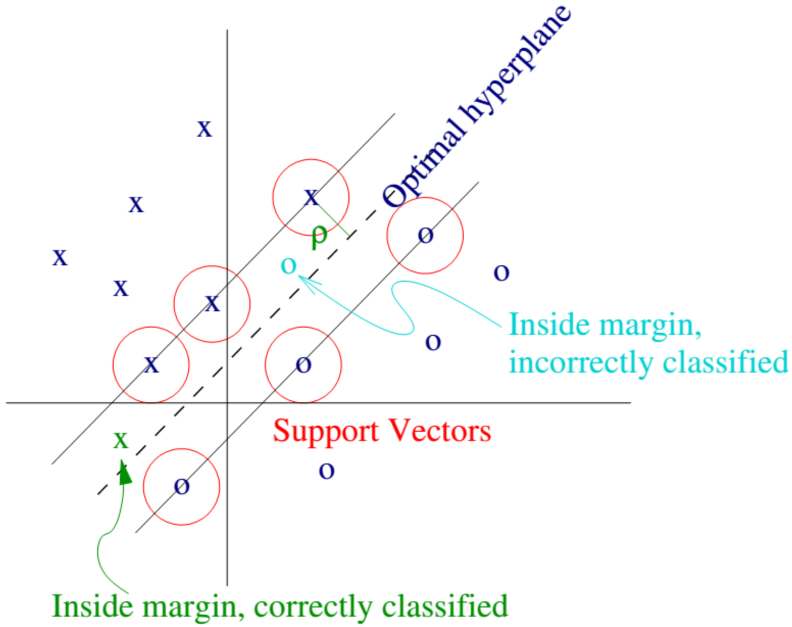
Alguns problemas podem violar a condição

$$d_i(\mathbf{w}_o^T\mathbf{x}_i + b_0) \geq 1$$

Nesse caso, podemos introduzir um novo conjunto de variáveis $\{\xi_i\}_{i=1}^{N}$:

$$d_i(\mathbf{w}_o^T\mathbf{x}_i + b_0) \geq 1 - \xi_i$$

onde ξ_i é chamada de variável de folga (*slack variable*)



Classificação com Margem Suave

Objetivo: encontrar o hiperplano que minimize

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

Solução

$$\mathbf{w}_o = \sum_{i=1}^{N_s} \alpha_{o,i} d_i \mathbf{x}_i$$

$$b_o = d_i - \mathbf{w}_o^T \mathbf{x}^{(s)}$$

C representa um parâmetro a ser definido pelo usuário

- Valores altos: alta confiança nos dados de treinamento
- Valores baixos: baixa confiança nos dados (ruidosos)