Discriminantes Lineares

Tópicos

- Discriminantes Lineares - duas classes
- Método dos Mínimos Quadrados
- Função do Erro Quadrático Médio
- Discriminantes Lineares - multi-classe

Discriminantes Lineares - duas classes

- Modelo linear de classificação:

$$\hat{y} = w_0 + w_1 x_1 + \ldots + w_d x_d = \mathbf{w}^\top \mathbf{x} = \begin{bmatrix} w_0 & w_1 & \ldots & w_d \end{bmatrix} \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$$

$\mathbf{x} \in \varpi_1$ se $\hat{y} < 0$, e $\mathbf{x} \in \varpi_2$ se $\hat{y} \geq 0$

- Pode-se estimar o valor óptimo $\mathbf{w}$ através do método do mínimos quadrados. Esta técnica minimiza a função de custo do erro quadrático médio.

Discriminantes Lineares - duas classes

- O método dos mínimos quadrados é uma técnica matemática para resolver um sistema sobre determinado de equações (com mais equações que incógnitas). O método permite estimar analiticamente a solução minimizando o erro quadrático médio entre a predição do modelo $\hat{y}$ e o seu valor desejado y.
- Função do erro quadrático médio:
  - Conjunto de dados: $\mathcal{X} = \{\mathbf{x}[1], \mathbf{x}[2], \dots, \mathbf{x}[N]\}$
    Cada vector $\mathbf{x}$ pertence a uma de duas classes $\Omega = \{\varpi_1, \varpi_2\}$
  - Cada vector $\mathbf{x}$ tem associado um escalar $y \in \{-1, +1\}$ que representa a classe: $\mathbf{x} \in \varpi_1$ se $y = -1$ e $\mathbf{x} \in \varpi_2$ se $y = +1$.
  - Erro Quadrático Médio (parábola):

$$\mathcal{E}(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N}(y[n] - \hat{y}[n])^2 = \frac{1}{N}\sum_{n=1}^{N}\left(y[n] - \mathbf{w}^\top \mathbf{x}[n]\right)^2$$

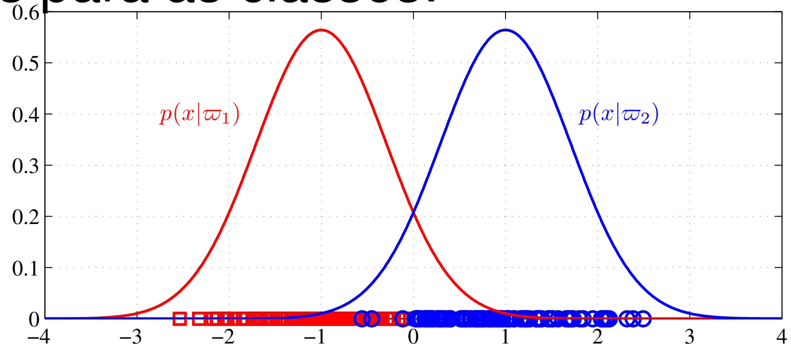
  - Solução: $\mathbf{w}_{\text{opt}} \implies$ resolver $\dfrac{\partial \mathcal{E}(\mathbf{w})}{\partial \mathbf{w}} = 0$

Discriminantes Lineares - duas classes

Exemplo: dados 1D, 2 classes

Considere as seguintes distribuições para as classes:

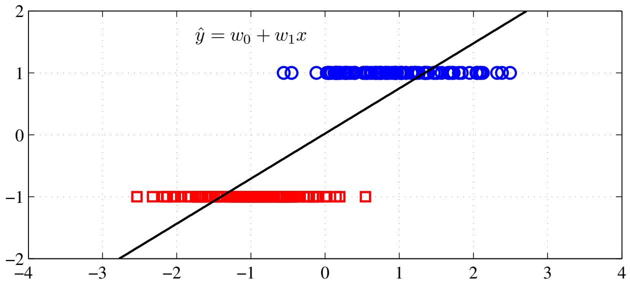
$$
\begin{aligned}
p(x|\varpi_1) &= \mathcal{N}\left(-1, \tfrac{1}{2}\right) \\
p(x|\varpi_2) &= \mathcal{N}\left(+1, \tfrac{1}{2}\right) \\
\text{com} \quad & p(\varpi_1) = p(\varpi_2)
\end{aligned}
$$



- Modelo de classificação: $\hat{y} = w_0 + w_1 x = \begin{bmatrix} w_0 & w_1 \end{bmatrix} \begin{bmatrix} 1 \\ x \end{bmatrix} = \mathbf{w}^\top \mathbf{x}$

  se $\hat{y} < 0,\ x \in \varpi_1,$ se $\hat{y} \geq 0, x \in \varpi_2$

- Saídas desejadas: $y = -1,\ \text{se } x \in \varpi_1,\ y = +1,\ \text{se } x \in \varpi_2$



Discriminantes Lineares - duas classes

Exemplo: dados 1D, 2 classes
Para estimar os parâmetros $\mathbf{w}$ é necessário derivar $\mathcal{E}(\mathbf{w})$, igualar a zero e resolver o sistema de equações resultante.

- $$\frac{\partial \mathcal{E}(\mathbf{w})}{\partial w_0} = \frac{1}{N}\sum_{n=1}^{N}\frac{\partial}{\partial w_0}(y[n] - w_0 - w_1 x[n])^2 = \frac{-2}{N}\sum_{n=1}^{N}(y[n] - w_0 - w_1 x[n]) = 0$$

$$\Longrightarrow \quad \frac{-2}{N}\left(\sum_{n=1}^{N} y[n] - N w_0 - w_1 \sum_{n=1}^{N} x[n]\right) = 0$$

$$\Longrightarrow \quad N w_0 + w_1 \sum_{n=1}^{N} x[n] = \sum_{n=1}^{N} y[n]$$

- $$\frac{\partial \mathcal{E}(\mathbf{w})}{\partial w_1} = \frac{1}{N}\sum_{n=1}^{N}\frac{\partial}{\partial w_0}(y[n] - w_0 - w_1 x[n])^2 = 0$$

$$\Longrightarrow \quad w_0 \sum_{n=1}^{N} x[n] + w_1 \sum_{n=1}^{N} x[n]^2 = \sum_{n=1}^{N} y[n] x[n]$$

Discriminantes Lineares - duas classes

Exemplo: dados 1D, 2 classes
Obtém-se um sistema de duas equações com duas incógnitas:

- $$N w_0 + w_1 \sum_{n=1}^{N} x[n] = \sum_{n=1}^{N} y[n]$$
  $$w_0 \sum_{n=1}^{N} x[n] + w_1 \sum_{n=1}^{N} x[n]^2 = \sum_{n=1}^{N} y[n]x[n]$$
- $$\begin{bmatrix} N & \sum_{n=1}^{N} x[n] \\ \sum_{n=1}^{N} x[n] & \sum_{n=1}^{N} x[n]^2 \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} \sum_{n=1}^{N} y[n] \\ \sum_{n=1}^{N} y[n]x[n] \end{bmatrix}$$
- $$\mathbf{R_x w} = \mathbf{r_{xy}} \quad \Longrightarrow \quad \mathbf{w} = \mathbf{R_x^{-1} r_{xy}}$$

Discriminantes Lineares - duas classes

Generalização: 2 classes, dados a d-dimensões

- Modelo de classificação:

$$\hat{y} = w_0 + w_1 x_1 + \ldots + w_d x_d = \begin{bmatrix} w_0 & w_1 & \ldots & w_d \end{bmatrix} \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} = \mathbf{w}^\top \mathbf{x}$$

  se $\hat{y} < 0,\ \mathbf{x} \in \varpi_1,$ se $\hat{y} \geq 0, \mathbf{x} \in \varpi_2$

- Saídas desejadas: $y = -1,\ \text{se}\ \mathbf{x} \in \varpi_1,\ y = +1,\ \text{se}\ \mathbf{x} \in \varpi_2$

Discriminantes Lineares - duas classes

Generalização: 2 classes, dados a d-dimensões

- Modelo de classificação: $\hat{y} = w_0 + w_1 x_1 + \ldots + w_d x_d = \mathbf{w}^\top \mathbf{x}$
  se $\hat{y} < 0,\ \mathbf{x} \in \varpi_1,$ se $\hat{y} \geq 0, \mathbf{x} \in \varpi_2$

  Para estimar os parâmetros $\mathbf{w}$ é necessário derivar $\mathcal{E}(\mathbf{w})$, igualar a zero e resolver o sistema de equações resultante.
- Função do erro:

$$\mathcal{E}(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N}(y[n] - \mathbf{w}^\top \mathbf{x}[n])^2 = \frac{1}{N}\sum_{n=1}^{N}(y[n] - w_0 - w_1 x_1[n] \ldots - w_d x_d[n])^2$$

- Derivadas:

$$\begin{aligned}
\frac{\partial \mathcal{E}(\mathbf{w})}{\partial w_0} = 0 \quad &\Longrightarrow \quad \tfrac{2}{N}\textstyle\sum_{n=1}^{N}\left(y[n] - \mathbf{w}^\top \mathbf{x}[n]\right)(-1) = 0 \\
\frac{\partial \mathcal{E}(\mathbf{w})}{\partial w_1} = 0 \quad &\Longrightarrow \quad \tfrac{2}{N}\textstyle\sum_{n=1}^{N}\left(y[n] - \mathbf{w}^\top \mathbf{x}[n]\right)(-x_1[n]) = 0 \\
&\vdots \\
\frac{\partial \mathcal{E}(\mathbf{w})}{\partial w_d} = 0 \quad &\Longrightarrow \quad \tfrac{2}{N}\textstyle\sum_{n=1}^{N}\left(y[n] - \mathbf{w}^\top \mathbf{x}[n]\right)(-x_d[n]) = 0
\end{aligned}$$

Discriminantes Lineares - duas classes

Generalização: 2 classes, dados a d-dimensões

- Modelo de classificação: $\hat{y} = w_0 + w_1 x_1 + \ldots + w_d x_d = \mathbf{w}^\top \mathbf{x}$
  se $\hat{y} < 0,\ \mathbf{x} \in \varpi_1,$ se $\hat{y} \geq 0, \mathbf{x} \in \varpi_2$
- Função do erro:

$$\mathcal{E}(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N}(y[n] - \mathbf{w}^\top \mathbf{x}[n])^2 = \frac{1}{N}\sum_{n=1}^{N}(y[n] - w_0 - w_1 x_1[n] \ldots - w_d x_d[n])^2$$

- Derivadas:

$$\frac{\partial \mathcal{E}(\mathbf{w})}{\partial \mathbf{w}} = 0 \quad \Longrightarrow \quad \frac{-2}{N}\sum_{n=1}^{N}\left(y[n] - \mathbf{w}^\top \mathbf{x}[n]\right)\mathbf{x}[n] = 0$$

$$\Longrightarrow \quad \sum_{n=1}^{N} y[n]\mathbf{x}[N] - \left(\sum_{n=1}^{N}\mathbf{x}[n]\mathbf{x}[n]^\top\right)\mathbf{w} = 0$$

Discriminantes Lineares - duas classes

Generalização: 2 classes, dados a d-dimensões

- Modelo de classificação: $\hat{y} = w_0 + w_1 x_1 + \ldots + w_d x_d = \mathbf{w}^\top \mathbf{x}$
  se $\hat{y} < 0,\ \mathbf{x} \in \varpi_1,$ se $\hat{y} \geq 0, \mathbf{x} \in \varpi_2$
- Função do erro:

$$\mathcal{E}(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N}(y[n] - \mathbf{w}^\top \mathbf{x}[n])^2 = \frac{1}{N}\sum_{n=1}^{N}(y[n] - w_0 - w_1 x_1[n] \ldots - w_d x_d[n])^2$$

- Derivadas:

$$\frac{\partial \mathcal{E}(\mathbf{w})}{\partial \mathbf{w}} = 0 \quad \Longrightarrow \quad \underbrace{\sum_{n=1}^{N} y[n]\mathbf{x}[n]}_{(d+1)\times 1} - \underbrace{\sum_{n=1}^{N}\left(\mathbf{x}[n]\mathbf{x}[n]^\top\right)}_{(d+1)\times(d+1)} \mathbf{w} = 0$$

$$\Longrightarrow \quad \mathbf{r}_{\mathbf{x}y} - \mathbf{R}_{\mathbf{x}}\mathbf{w} = 0$$

$$\Longrightarrow \quad \mathbf{w} = \mathbf{R}_{\mathbf{x}}^{-1}\mathbf{r}_{\mathbf{x}y}$$

Discriminantes Lineares - duas classes

Generalização: 2 classes, dados a d-dimensões

- Modelo de classificação: $\hat{y} = w_0 + w_1 x_1 + \ldots + w_d x_d = \mathbf{w}^\top \mathbf{x}$
  se $\hat{y} < 0,\ \mathbf{x} \in \varpi_1,$ se $\hat{y} \geq 0, \mathbf{x} \in \varpi_2$
- Função do erro:

$$\mathcal{E}(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N}(y[n] - \mathbf{w}^\top \mathbf{x}[n])^2 = \frac{1}{N}\sum_{n=1}^{N}(y[n] - w_0 - w_1 x_1[n] \ldots - w_d x_d[n])^2$$

- Solução: $\mathbf{w} = \mathbf{R}_{\mathbf{x}}^{-1} \mathbf{r}_{\mathbf{x}y}$
- Notação matricial

$$\mathbf{X} = \underbrace{\begin{bmatrix} \mathbf{x}[1] & \mathbf{x}[2] & \ldots & \mathbf{x}[N] \end{bmatrix}}_{\text{matriz de } (d+1)\times N} = \begin{bmatrix} 1 & 1 & 1 & \ldots & 1 \\ x_1[1] & x_1[2] & x_1[3] & \ldots & x_1[N] \\ x_2[1] & x_2[3] & x_2[3] & \ldots & x_2[N] \\ \vdots & & & \ddots & \vdots \\ x_d[n] & x_d[2] & x_d[3] & \ldots & x_d[N] \end{bmatrix}$$

$$\mathbf{Y} = \underbrace{\begin{bmatrix} y[1] & y[2] & \ldots & y[N] \end{bmatrix}}_{\text{matriz de } 1\times N \text{ com } \pm 1_{\mathrm{s}}} \quad \mathrm{e}\ \hat{\mathbf{Y}} = \mathbf{w}^\top \mathbf{X}$$

Discriminantes Lineares - duas classes

Generalização: 2 classes, dados a d-dimensões

- Modelo de classificação: $\hat{y} = w_0 + w_1 x_1 + \ldots + w_d x_d = \mathbf{w}^\top \mathbf{x}$
  se $\hat{y} < 0,\ \mathbf{x} \in \varpi_1,$ se $\hat{y} \geq 0, \mathbf{x} \in \varpi_2$
- Função do erro:

$$\mathcal{E}(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N}(y[n] - \mathbf{w}^\top \mathbf{x}[n])^2 = \frac{1}{N}\sum_{n=1}^{N}(y[n] - w_0 - w_1 x_1[n] \ldots - w_d x_d[n])^2$$

- Solução: $\mathbf{w} = \mathbf{R}_{\mathbf{x}}^{-1}\mathbf{r}_{\mathbf{x}y} = (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{Y}^\top$
- Notação matricial

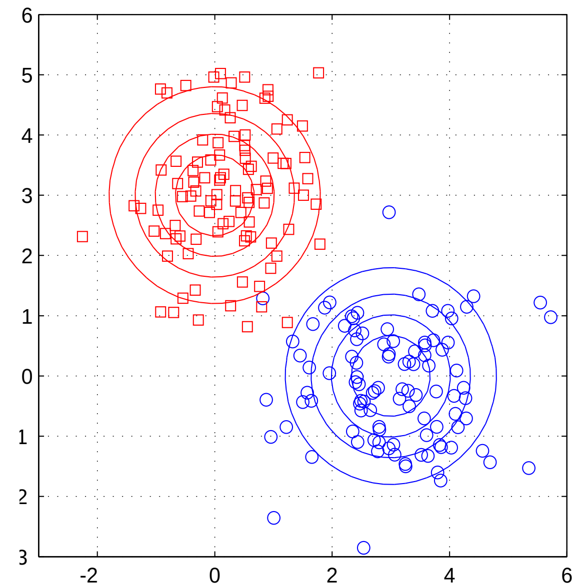
$$\mathbf{R}_{\mathbf{x}} = \sum_{n=1}^{N}\mathbf{x}[n]\mathbf{x}[n]^\top = \mathbf{X}\mathbf{X}^\top$$

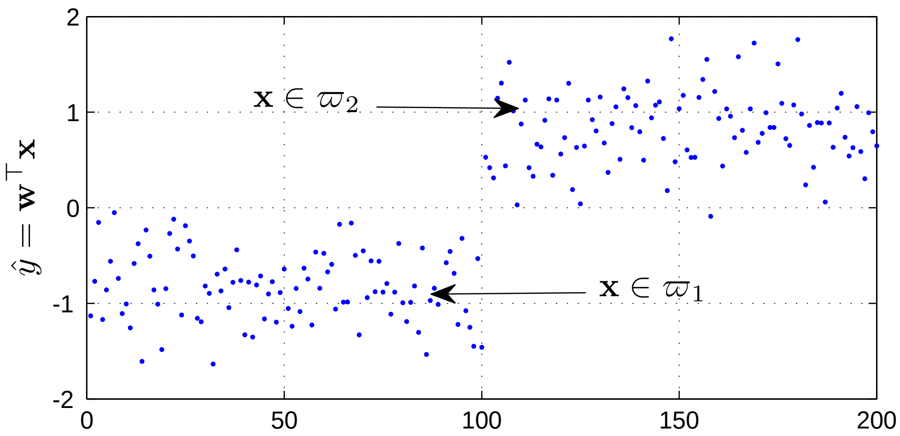
$$\mathbf{r}_{\mathbf{x}y} = \sum_{n=1}^{N}\mathbf{x}[n]y[n] = \mathbf{X}\mathbf{Y}^\top$$

Discriminantes Lineares - duas classes

Exemplo: Duas classes - dados 2D

- 100 pontos da classe ϖ_1 com $p(\mathbf{x}|\varpi_1) = \mathcal{N}\left(\begin{bmatrix} 3 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$
- 100 pontos da classe ϖ_2 com $p(\mathbf{x}|\varpi_2) = \mathcal{N}\left(\begin{bmatrix} 0 \\ 3 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$
- Matriz $\mathbf{X} = [$100 pontos de ϖ_1, 100 pontos de $\varpi_2]$
- Matriz $\mathbf{Y} = [\underbrace{-1, -1, \ldots, -1}_{100\times}, \underbrace{+1, +1, \ldots, +1}_{100\times}]$
- Estimação $\mathbf{w} = \left(\mathbf{X}\mathbf{X}^\top\right)^{-1} \mathbf{X}\mathbf{Y}^\top = \begin{bmatrix} +0.07 \\ -0.29 \\ +0.27 \end{bmatrix}$
- Resultado da classificação: $\hat{\mathbf{Y}} = \mathbf{w}^\top \mathbf{X}$

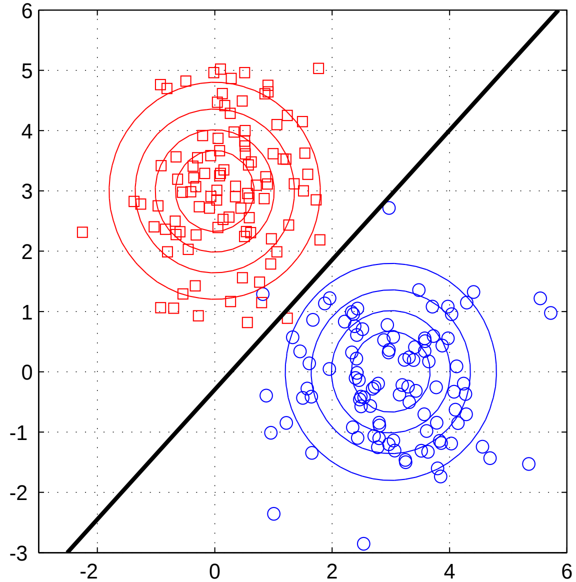




Discriminantes Lineares - duas classes

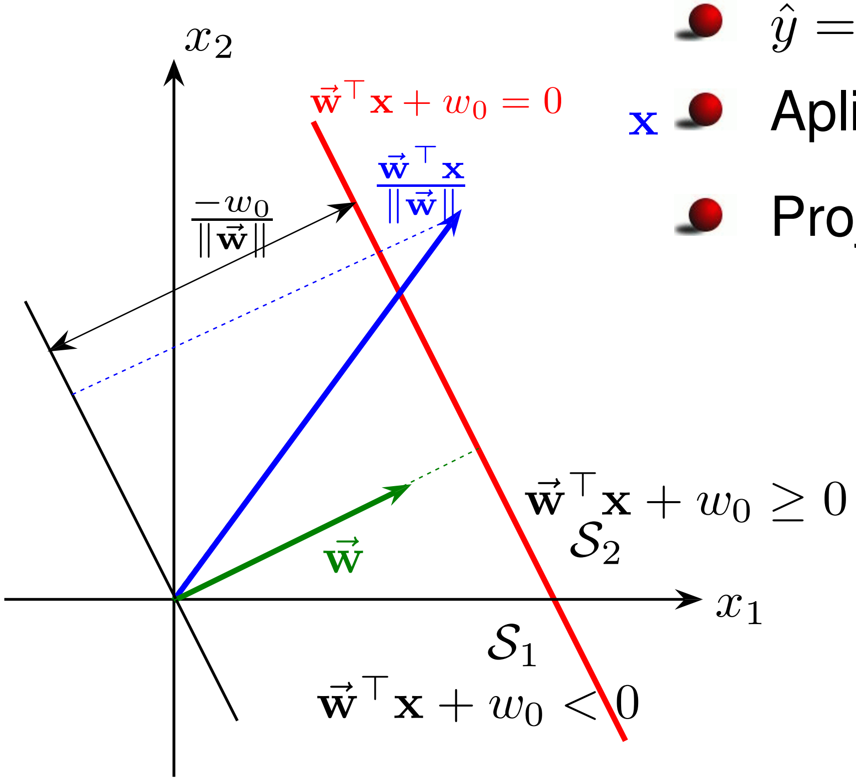
Exemplo: Duas classes - dados 2D

- 100 pontos da classe ϖ_1 com $p(\mathbf{x}|\varpi_1) = \mathcal{N}\left(\left[\begin{smallmatrix}3\\0\end{smallmatrix}\right], \left[\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right]\right)$
- 100 pontos da classe ϖ_2 com $p(\mathbf{x}|\varpi_2) = \mathcal{N}\left(\left[\begin{smallmatrix}0\\3\end{smallmatrix}\right], \left[\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right]\right)$
- Matriz $\mathbf{X} = [100 \text{ pontos de } \varpi_1, 100 \text{ pontos de } \varpi_2]$
- Matriz $\mathbf{Y} = [\underbrace{-1,-1,\ldots,-1}_{100\times},\underbrace{+1,+1,\ldots,+1}_{100\times}]$
- Estimação $\mathbf{w} = \left(\mathbf{X}\mathbf{X}^\top\right)^{-1}\mathbf{X}\mathbf{Y}^\top = \begin{bmatrix}+0.07\\-0.29\\+0.27\end{bmatrix}$
- Resultado da classificação: $\hat{\mathbf{Y}} = \mathbf{w}^\top\mathbf{X}$
- Regra de classificação: $\mathbf{x}\in\varpi_2 \ \overset{\hat{y}}{\underset{0}{\gtrless}} \ \mathbf{x}\in\varpi_1$
- Fronteira de decisão:
  $0 = \mathbf{w}^\top\mathbf{x} = w_0 + w_1x_1 + w_2x_2$



Discriminantes Lineares - duas classes

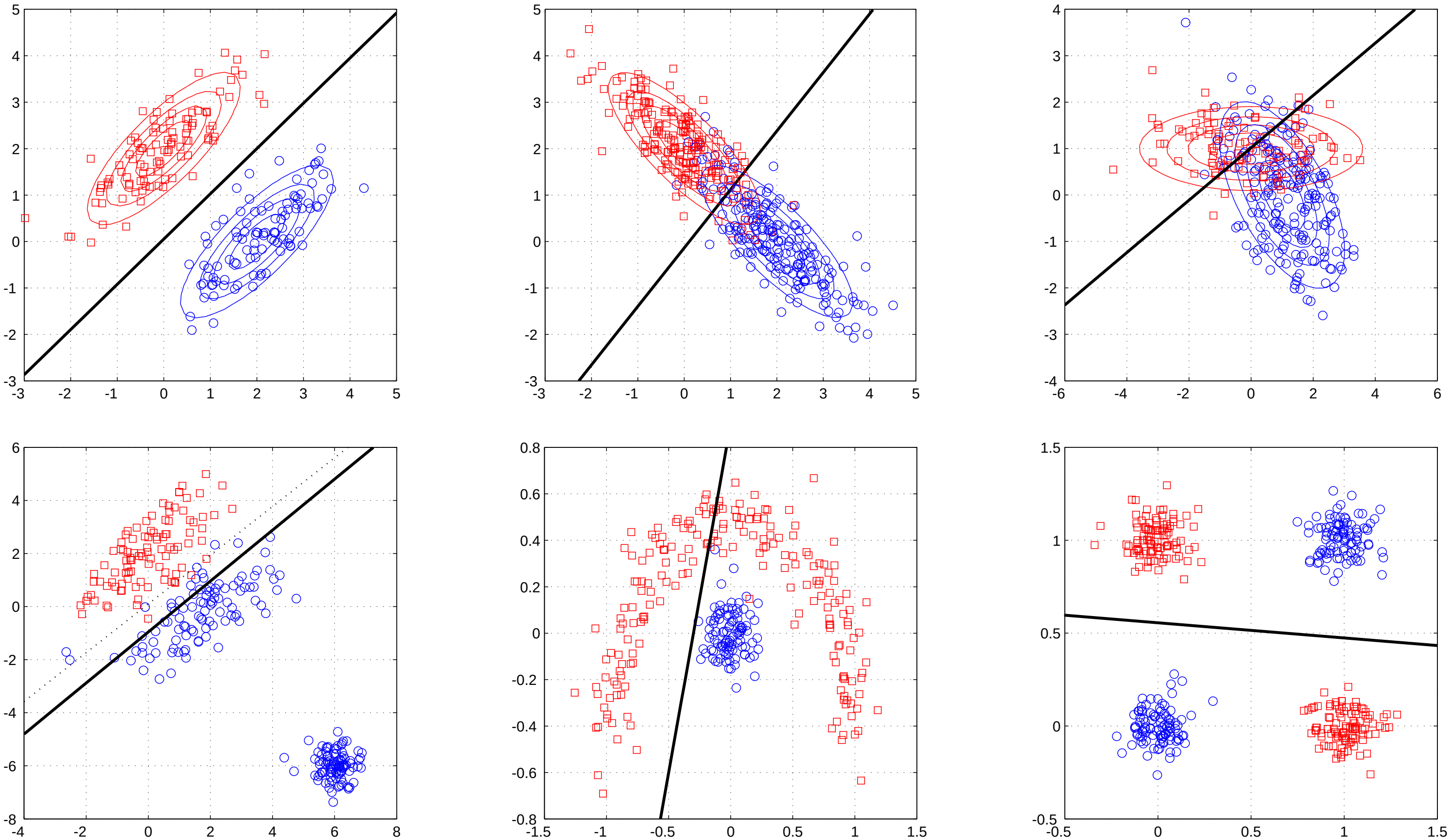
- Dados 2D, 2 classes – discriminantes lineares
- Modelo: $\hat{y} = \mathbf{w}^\top \mathbf{x} = w_0 + w_1 x_1 + w_2 x_2$



- $\hat{y} = \vec{\mathbf{w}}^\top \mathbf{x} + w_0$
- Aplicar *threshold* a $\hat{y}$
- Projecção de $\mathbf{x}$ em $\vec{\mathbf{w}}$ é $\frac{\vec{\mathbf{w}}^\top \mathbf{x}}{\|\vec{\mathbf{w}}\|}$

Discriminantes Lineares

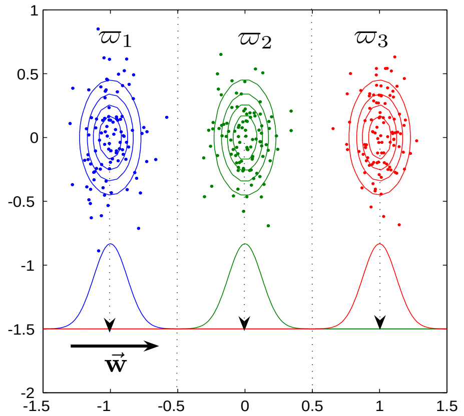
- Funciona bem para gaussianas com a mesma matriz de covariância - as fronteiras de decisão são lineares
- Funciona mal para gaussianas com matrizes de covariância distintas - as fronteiras de decisão são quadráticas. E há mais casos...



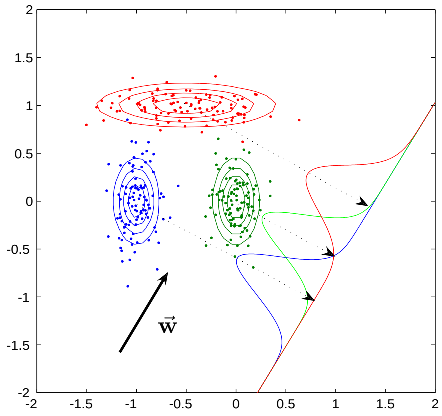
Discriminantes Lineares: Múltiplas Classe

Abordagens:

- Fazer o mesmo que para duas classes mas com $y \in \{0, 1, 2, c-1\}$
- Classificação: vários limiares.
- Projecção: $\hat{y} = \mathbf{w}^\top \mathbf{x}$
- Demasiado limitativo: todos os $\mathbf{x}$ projectados numa recta.



Aqui funciona

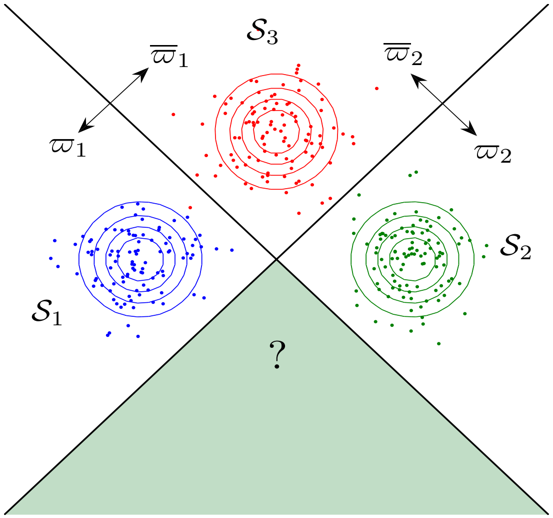


Em muitos casos funcional mal

Discriminantes Lineares: Múltiplas Classe

Abordagens:

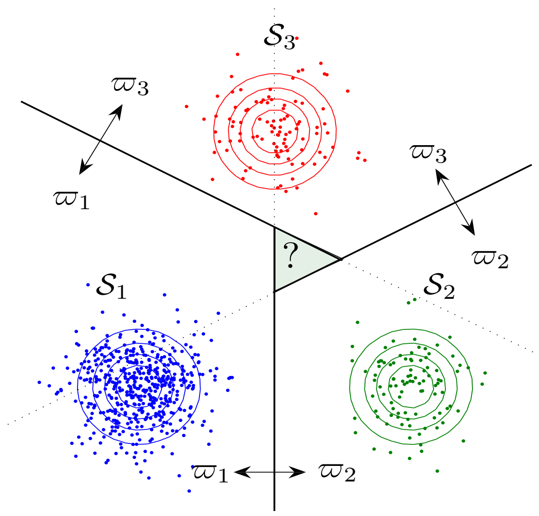
- Um contra todos: projectar $c-1$ classificadores ($c =$ nº total de classes)
  - Escolher uma classe, agrupar as restantes.
  - Treinar (estimar $\mathbf{w}$)
  - Repetir para todas as classes
    (menos a última - se $\mathbf{x} \notin \varpi_k, \;\; k = 1, \ldots, c-1$ então $\mathbf{x} \in \varpi_c$)
- Pode trazer problemas



Discriminantes Lineares: Múltiplas Classe

Abordagens:

- Um contra um: projectar $c(c-1)/2$ classificadores ($c =$ nº total de classes)
  - Escolher um par de classes.
  - Treinar (estimar $\mathbf{w}$)
  - Repetir para todos as pares
- Pode trazer problemas



Discriminantes Lineares: Múltiplas Classe

Solução:

- Usar c classificadores lineares:

$$\hat{y}_k = \mathbf{w}_k^\top \mathbf{x} = w_{0k} + w_{1k}x_1 + \ldots + w_{dk}x_d, \quad k = 1, \ldots, c$$

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 \\ \ldots \\ \hat{y}_c \end{bmatrix} = \mathbf{W}^\top \mathbf{x} = \underbrace{\begin{bmatrix} w_{01} & w_{11} & \ldots & w_{d1} \\ \vdots & \ddots & & \vdots \\ w_{0c} & w_{1c} & \ldots & w_{dc} \end{bmatrix}}_{(c \times d+1)} \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{bmatrix}$$

- O vector $\hat{\mathbf{y}}$ resulta de c projecções de $\mathbf{x}$
- Classificação: $\mathbf{x} \in \varpi_k$ se $\hat{y}_k > \hat{y}_j$ para $j \neq k$ e $j, k = 1, \ldots, c$
- Estimar $\mathbf{W}$ de modo a minimizar a média do erro quadrático entre $\hat{\mathbf{y}}$ e as saídas desejadas $\mathbf{y}$

Discriminantes Lineares: Múltiplas Classe

Solução:

- Transformação $\hat{\mathbf{y}} = \mathbf{W}^\top \mathbf{x}$
- Erro quadrático médio: $\mathcal{E}(\mathbf{W}) = \dfrac{1}{N} \sum_{n=1}^{N} \|\mathbf{y}[n] - \hat{\mathbf{y}}[n]\|^2$
- Saídas desejadas:

$$\text{se} \quad \mathbf{x} \in \varpi_k, \qquad \mathbf{y} = \begin{bmatrix} -1 \\ \vdots \\ -1 \\ +1 \\ -1 \\ \vdots \\ -1 \end{bmatrix} \begin{array}{l} \\ \\ \\ \longleftarrow \text{linha } k \\ \\ \\ \\ \end{array}$$

Discriminantes Lineares: Múltiplas Classe

Solução:

- Transformação $\hat{\mathbf{y}} = \mathbf{W}^\top \mathbf{x}$
- Erro quadrático médio: $\mathcal{E}(\mathbf{W}) = \dfrac{1}{N}\sum_{n=1}^{N} \|\mathbf{y}[n] - \hat{\mathbf{y}}[n]\|^2$
- Pesos óptimos

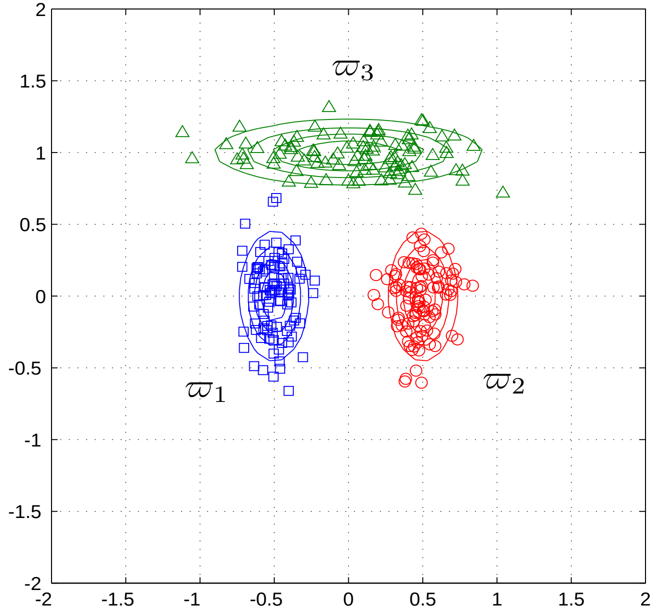
  mais do mesmo - a única diferença é que $\mathbf{Y}$ é uma matriz de $c \times N$:

  $\mathbf{W} = (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{Y}^\top$

Discriminantes Lineares: Múltiplas Classe

Exemplo:

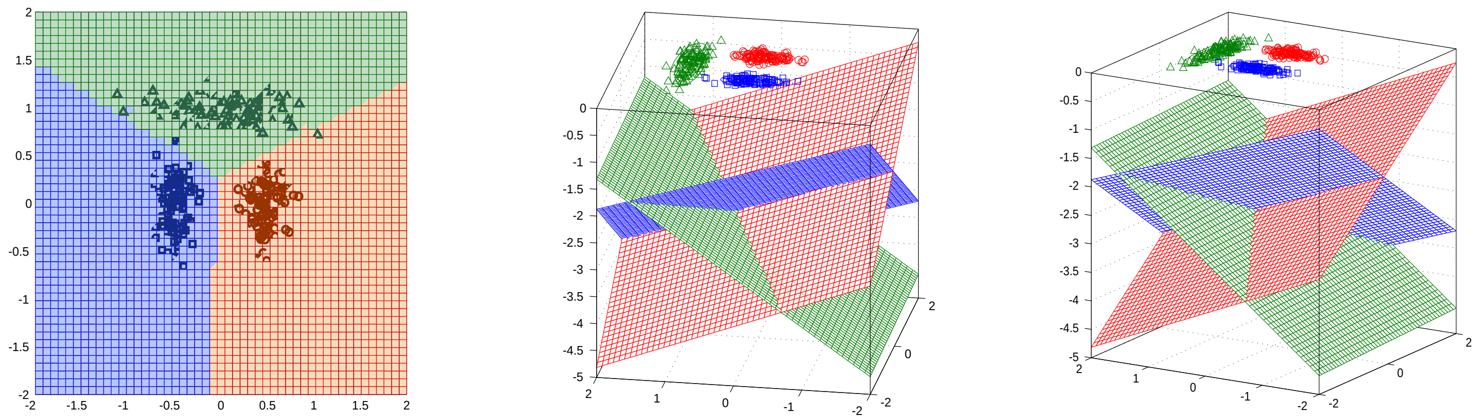
- Três classes (dados 2D - 100 pts/classe)



Discriminantes Lineares: Múltiplas Classe

Exemplo:

- Três classes (dados 2D - 100 pts/classe)
- Vector $\mathbf{x}$ projectado em três planos. Os planos são funções discriminantes do classificador.

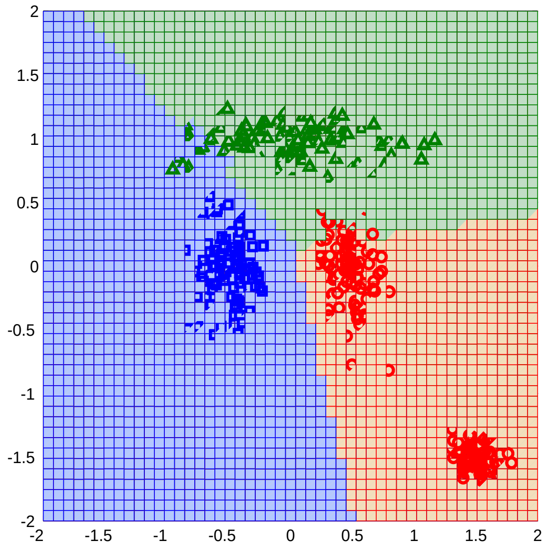


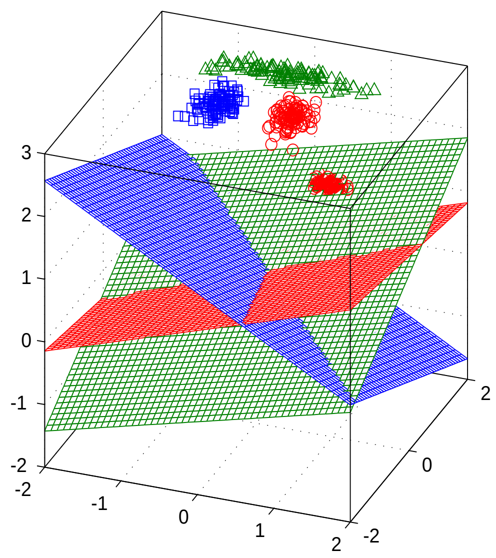
(várias perspectivas)

Discriminantes Lineares: Múltiplas Classe

Exemplo:

- Três classes (dados 2D - 100 pts/classe)
- Vector $\mathbf{x}$ projectado em três planos.
- Não lida bem com *outliers*

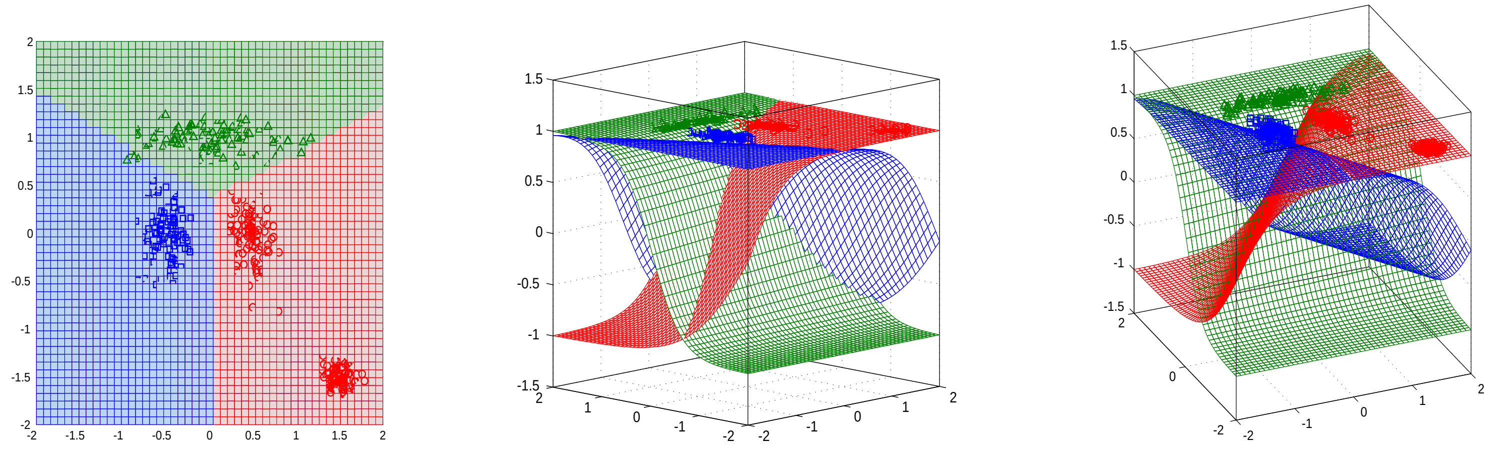




Discriminantes Lineares: Múltiplas Classe

Exemplo:

- Três classes (dados 2D - 100 pts/classe)
- Vector $\mathbf{x}$ projectado em três planos.
- Solução: Discriminantes logísticos



- Não se pode estimar a matriz $\mathbf{W}$ analiticamente - necessário métodos adaptativos.