Discriminantes Logísticos

Discriminantes Logísticos

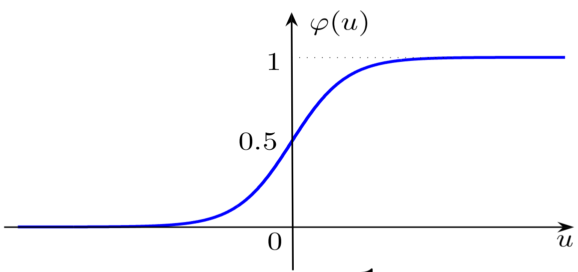
- Uma transformação linear seguida de uma não-linearidade (degrau - também conhecido como função de activação)

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

$$\mathbf{u} = \begin{bmatrix} u_1 \\ \ldots \\ u_c \end{bmatrix} = \mathbf{W}^\top \mathbf{x}, \qquad \hat{\mathbf{y}} = \varphi(\mathbf{u}) = \begin{bmatrix} \varphi(u_1) \\ \ldots \\ \varphi(u_c) \end{bmatrix}$$

- $\varphi(\cdot)$ é uma de duas funções:

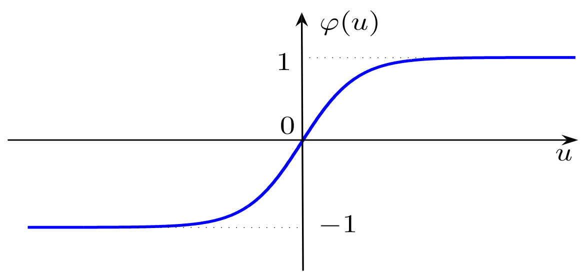
sigmóide



$$\varphi(u) = \frac{1}{1 + e^{-u}}$$

ou

tangente hiperbólica



$$\varphi(u) = \tanh(u)$$

Discriminantes Logísticos

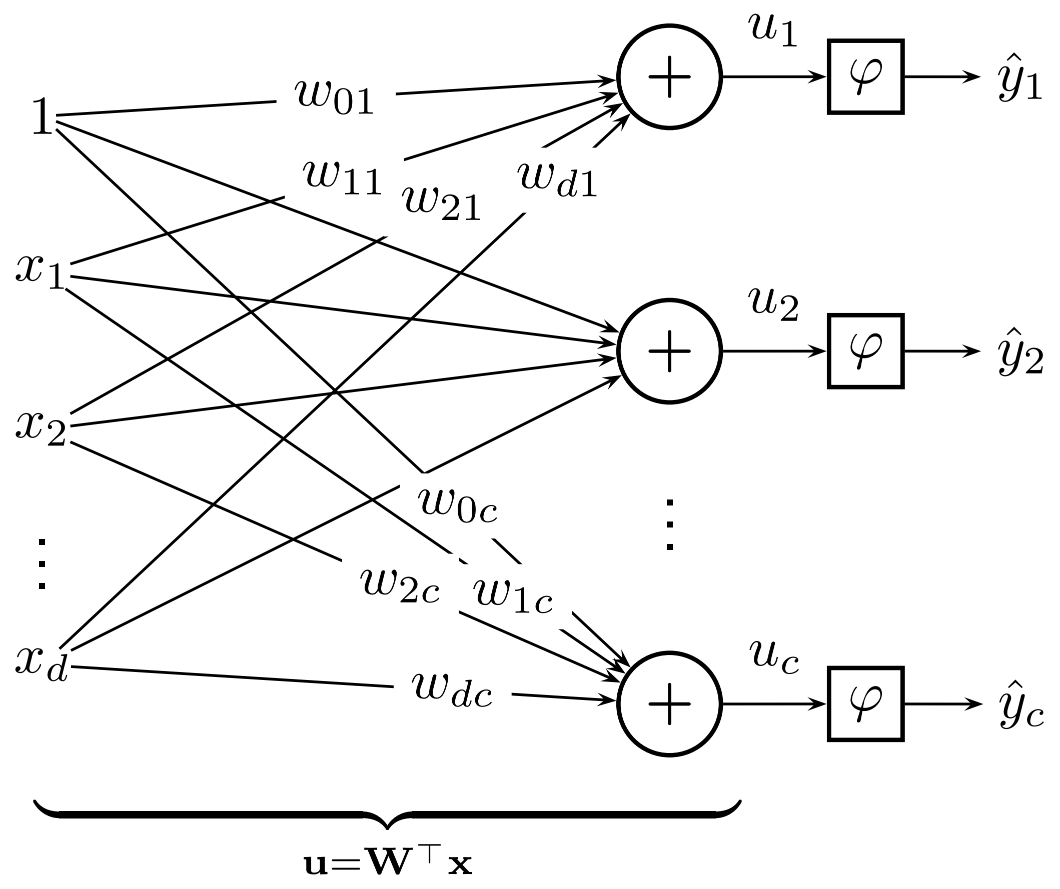
- Transformação $\hat{\mathbf{y}} = \varphi\left(\mathbf{W}^\top \mathbf{x}\right)$

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_c \end{bmatrix} = \begin{bmatrix} \varphi(u_1) \\ \varphi(u_2) \\ \vdots \\ \varphi(u_c) \end{bmatrix} = \varphi\left(\underbrace{\begin{bmatrix} w_{01} & w_{11} & w_{21} & \cdots & w_{d1} \\ w_{02} & w_{12} & w_{22} & \cdots & w_{d2} \\ \vdots & & & & \vdots \\ w_{0c} & w_{1c} & w_{2c} & \cdots & w_{dc} \end{bmatrix}}_{\mathbf{W}^\top} \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \right)$$

- $\mathbf{x}$ vector de entrada
- Cada coluna $\mathbf{w}_k$ de $\mathbf{W}$ (linha k de $\mathbf{W}^\top$) só afecta a saída $\hat{y}_k$

Discriminantes Logísticos

- Transformação $\hat{\mathbf{y}} = \varphi\left(\mathbf{W}^\top \mathbf{x}\right)$



Discriminantes Logísticos

Fase de Treino

- Objectivo: estimar a matriz $\mathbf{W}$
- Dados disponíveis:
  - Conjunto de N vectores $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ (dados - entradas)
  - Conjunto $\mathcal{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ (classificações - saídas desejadas)

Saídas desejadas:

$$\text{se} \quad \mathbf{x} \in \varpi_k, \quad \mathbf{y} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \longleftarrow \text{linha } k \qquad \text{ou} \quad \mathbf{y} = \begin{bmatrix} -1 \\ \vdots \\ -1 \\ +1 \\ -1 \\ \vdots \\ -1 \end{bmatrix}$$

com $\varphi(u) = \frac{1}{1+e^{-u}}$ com $\varphi(u) = \tanh(u)$

Discriminantes Logísticos

Fase de Treino

- Objectivo: estimar a matriz $\mathbf{W}$
- Dados disponíveis:
  - Conjunto de N vectores $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ (dados - entradas)
  - Conjunto $\mathcal{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ (classificações - saídas desejadas)
- Minimizar erro quadrático médio: $\mathcal{E}(\mathbf{W}) = \dfrac{1}{N} \displaystyle\sum_{n=1}^{N} \|\mathbf{y}_n - \hat{\mathbf{y}}_n\|^2$
- Adaptação da matriz $\mathbf{W}$: máxima descida de gradiente

$$\mathbf{W}(i+1) = \mathbf{W}(i) + \Delta\mathbf{W}(i)$$

com $\quad \Delta\mathbf{W}(i) = -\eta \frac{\partial \mathcal{E}(\mathbf{W}(i))}{\partial \mathbf{W}}$

ou $\quad \Delta\mathbf{W}(i) = -\eta \mathbf{z}(i) = -\eta \left(\frac{\partial \mathcal{E}(\mathbf{W}(i))}{\partial \mathbf{W}} + \alpha \mathbf{z}(i-1) \right)$

Discriminantes Logísticos

Fase de Classificação

- Objectivo: classificar um vector $\mathbf{x}$
- Transformar: $\hat{\mathbf{y}} = \varphi(\mathbf{W}^\top \mathbf{x})$
- Classificar: $\mathbf{x} \in \varpi_k, \;\; \text{se} \;\; y_k = \arg\max_i (\hat{y}_i) \quad \text{com } i = 1, \ldots, c$

Discriminantes Logísticos

Exemplo: Dados 3D, e 2 classes

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \end{bmatrix} = \begin{bmatrix} \varphi(u_1) \\ \varphi(u_2) \end{bmatrix} \qquad \mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \mathbf{W}^\top \mathbf{x} = \begin{bmatrix} w_{01} & w_{11} & w_{21} & w_{31} \\ w_{02} & w_{12} & w_{22} & w_{32} \end{bmatrix} \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

$$u_1 = \mathbf{w}_1^\top \mathbf{x} = w_{01} + w_{11}x_1 + w_{21}x_2 + w_{31}x_3$$

$\mathbf{w}_1$: 1ª coluna da matriz $\mathbf{W}$

$$u_2 = \mathbf{w}_2^\top \mathbf{x} = w_{02} + w_{12}x_1 + w_{22}x_2 + w_{32}x_3$$

$\mathbf{w}_2$: 2ª coluna da matriz $\mathbf{W}$

$$\mathbf{W} = \begin{bmatrix} w_{01} & w_{02} \\ w_{11} & w_{12} \\ w_{21} & w_{22} \\ w_{31} & w_{32} \end{bmatrix} = \begin{bmatrix} \mathbf{w}_1 & \mathbf{w}_2 \end{bmatrix}$$

$$\begin{array}{cc} \underbrace{\phantom{w_{01}}}_{\mathbf{w}_1} & \underbrace{\phantom{w_{02}}}_{\mathbf{w}_2} \end{array}$$

Discriminantes Logísticos

Exemplo: Dados 3D, e 2 classes

$$\mathcal{E}(\mathbf{W}) = \frac{1}{N}\sum_{n=1}^{N} \|\mathbf{y}_n - \hat{\mathbf{y}}_n\|^2$$

$$\|\mathbf{y} - \hat{\mathbf{y}}\|^2 = (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 = \left(y_1 - \varphi(\mathbf{w}_1^\top \mathbf{x})\right)^2 + \left(y_2 - \varphi(\mathbf{w}_2^\top \mathbf{x})\right)^2$$

$$\frac{\partial \mathcal{E}(\mathbf{W})}{\partial \mathbf{W}} = \frac{1}{N}\sum_{n=1}^{N} \frac{\partial \|\mathbf{y}_n - \hat{\mathbf{y}}_n\|^2}{\partial \mathbf{W}}$$

$$\frac{\partial \|\mathbf{y} - \hat{\mathbf{y}}\|^2}{\partial \mathbf{W}} = \begin{bmatrix} \underbrace{\begin{matrix} \frac{\partial \|\mathbf{y}-\hat{\mathbf{y}}\|^2}{\partial w_{01}} \\ \frac{\partial \|\mathbf{y}-\hat{\mathbf{y}}\|^2}{\partial w_{11}} \\ \frac{\partial \|\mathbf{y}-\hat{\mathbf{y}}\|^2}{\partial w_{21}} \\ \frac{\partial \|\mathbf{y}-\hat{\mathbf{y}}\|^2}{\partial w_{31}} \end{matrix}}_{\frac{\partial \|\mathbf{y}-\hat{\mathbf{y}}\|^2}{\partial \mathbf{w}_1}} & \underbrace{\begin{matrix} \frac{\partial \|\mathbf{y}-\hat{\mathbf{y}}\|^2}{\partial w_{02}} \\ \frac{\partial \|\mathbf{y}-\hat{\mathbf{y}}\|^2}{\partial w_{12}} \\ \frac{\partial \|\mathbf{y}-\hat{\mathbf{y}}\|^2}{\partial w_{22}} \\ \frac{\partial \|\mathbf{y}-\hat{\mathbf{y}}\|^2}{\partial w_{32}} \end{matrix}}_{\frac{\partial \|\mathbf{y}-\hat{\mathbf{y}}\|^2}{\partial \mathbf{w}_2}} \end{bmatrix} = \left[\frac{\partial \|\mathbf{y} - \hat{\mathbf{y}}\|^2}{\partial \mathbf{w}_1} \quad \frac{\partial \|\mathbf{y} - \hat{\mathbf{y}}\|^2}{\partial \mathbf{w}_2} \right]$$

Discriminantes Logísticos

Exemplo: Dados 3D, e 2 classes

- Derivada da sigmóide:

$$\varphi(u) = \frac{1}{1+e^{-u}}$$

$$\frac{\partial \varphi(u)}{\partial u} = \frac{e^{-u}}{(1+e^{-u})^2} = \varphi(u) - \varphi(u)^2$$

- Derivada da tangente hiperbólica:

$$\varphi(u) = \tanh(u) = \frac{e^{u} - e^{-u}}{e^{u} + e^{-u}}$$

$$\frac{\partial \varphi(u)}{\partial u} = 1 - \tanh^2(u) = 1 - \varphi(u)^2$$

Discriminantes Logísticos

Exemplo: Dados 3D, e 2 classes

$$\mathcal{E}(\mathbf{W}) = \frac{1}{N}\sum_{n=1}^{N} \|\mathbf{y}_n - \hat{\mathbf{y}}_n\|^2$$

$$\|\mathbf{y} - \hat{\mathbf{y}}\|^2 = \quad (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 = (y_1 - \varphi(u_1))^2 + (y_2 - \varphi(u_2))^2$$

$$\frac{\partial\|\mathbf{y} - \hat{\mathbf{y}}\|^2}{\partial \mathbf{w}_1} = \frac{\partial\|\mathbf{y} - \hat{\mathbf{y}}\|^2}{\partial \hat{y}_1} \times \frac{\partial \hat{y}_1}{\partial u_1} \times \frac{\partial u_1}{\partial \mathbf{w}_1} = -2\,(y_1 - \hat{y}_1)\left(\hat{y}_1 - \hat{y}_1^2\right)\mathbf{x}$$

$$\frac{\partial\|\mathbf{y} - \hat{\mathbf{y}}\|^2}{\partial \mathbf{w}_2} = \frac{\partial\|\mathbf{y} - \hat{\mathbf{y}}\|^2}{\partial \hat{y}_2} \times \frac{\partial \hat{y}_2}{\partial u_2} \times \frac{\partial u_2}{\partial \mathbf{w}_2} = -2\,(y_2 - \hat{y}_2)\left(\hat{y}_2 - \hat{y}_2^2\right)\mathbf{x}$$

$$\frac{\partial\|\mathbf{y} - \hat{\mathbf{y}}\|^2}{\partial \mathbf{W}} = \left[\frac{\partial\|\mathbf{y} - \hat{\mathbf{y}}\|^2}{\partial \mathbf{w}_1} \quad \frac{\partial\|\mathbf{y} - \hat{\mathbf{y}}\|^2}{\partial \mathbf{w}_2}\right]$$

$$\frac{\partial \mathcal{E}(\mathbf{W})}{\partial \mathbf{W}} = \frac{1}{N}\sum_{n=1}^{N} \frac{\partial\|\mathbf{y}_n - \hat{\mathbf{y}}_n\|^2}{\partial \mathbf{W}}$$

Discriminantes Logísticos

Exemplo: Dados 3D, e 2 classes

$$\mathcal{E}(\mathbf{W}) = \frac{1}{N}\sum_{n=1}^{N}\|\mathbf{y}_n - \hat{\mathbf{y}}_n\|^2 = \frac{1}{N}\sum_{n=1}\|\mathbf{e}_n\|^2 \quad \text{com } \mathbf{e} = \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \end{bmatrix}$$

$$\frac{\partial\|\mathbf{y} - \hat{\mathbf{y}}\|^2}{\partial \mathbf{w}_1} = -2\,(y_1 - \hat{y}_1)\left(\hat{y}_1 - \hat{y}_1^2\right)\mathbf{x} = -2\mathbf{x}\left(e_1\frac{\partial\varphi(u_1)}{\partial u_1}\right)$$

$$\frac{\partial\|\mathbf{y} - \hat{\mathbf{y}}\|^2}{\partial \mathbf{w}_2} = -2\,(y_2 - \hat{y}_2)\left(\hat{y}_2 - \hat{y}_2^2\right)\mathbf{x} = -2\mathbf{x}\left(e_2\frac{\partial\varphi(u_2)}{\partial u_2}\right)$$

$$\frac{\partial\|\mathbf{y}-\hat{\mathbf{y}}\|^2}{\partial \mathbf{W}} = -2\begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_3 \end{bmatrix}\begin{bmatrix} e_1(\hat{y}_1 - \hat{y}_1^2) & e_2(\hat{y}_2 - \hat{y}_2^2) \end{bmatrix}$$

$$\frac{\partial\mathcal{E}(\mathbf{W})}{\partial \mathbf{W}} = \frac{1}{N}\sum_{n=1}^{N}\frac{\partial\|\mathbf{y}_n - \hat{\mathbf{y}}_n\|^2}{\partial \mathbf{W}}$$

Discriminantes Logísticos

Generalização: Dados a d-dimensões, e com c classes

- Transformação $\hat{\mathbf{y}} = \varphi\left(\mathbf{W}^\top \mathbf{x}\right) = \begin{bmatrix} \varphi(u_1) & \dots & \varphi(u_c) \end{bmatrix}^\top$
- Erro quadrático médio: $\mathcal{E}(\mathbf{W}) = \dfrac{1}{N} \displaystyle\sum_{n=1}^{N} \|\mathbf{y}_n - \hat{\mathbf{y}}_n\|^2$
- Derivada do erro: $\dfrac{\partial \mathcal{E}(\mathbf{W})}{\partial \mathbf{W}} = \dfrac{1}{N} \displaystyle\sum_{n=1}^{N} \dfrac{\partial \|\mathbf{y}_n - \hat{\mathbf{y}}_n\|^2}{\partial \mathbf{W}}$

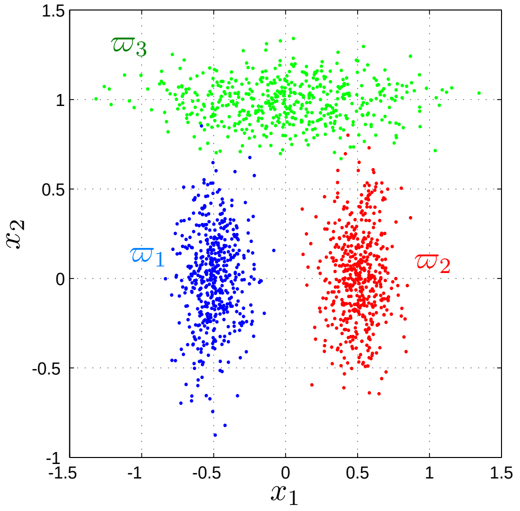
com $$\frac{\partial \|\mathbf{y} - \hat{\mathbf{y}}\|^2}{\partial \mathbf{W}} = -2 \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \begin{bmatrix} e_1(\hat{y}_1 - \hat{y}_1^2) & e_2(\hat{y}_2 - \hat{y}_2^2) & \dots & e_c(\hat{y}_c - \hat{y}_c^2) \end{bmatrix}$$

e com $$\varphi(u) = \frac{1}{1 + e^{-u}}$$

Discriminantes Logísticos

Outro exemplo:

- Três classes
  (dados 2D - 500pts/classe)



Código Python:
(com `import numpy.random as rd`)

- Geração dos dados

```
x1=np.dot(np.diag((0.125,0.25)),rd.randn(2,500))
x1[0,:]=x1[0,:]-0.5
x2=np.dot(np.diag((0.125,0.25)),rd.randn(2,500))
x2[0,:]=x2[0,:]+0.5
x3=np.dot(np.diag((0.5,0.125)),rd.randn(2,500))
x3[1,:]=x3[1,:]+1
X=np.mat((np.ones(1500),np.hstack((x1,x2,x3))))
```

- Saídas desejadas (classes)

```
Y=np.array(-np.zeros((3,1500)))
Y[0,0:500]=1.  Y[1,500:1000]=1.  Y[2,1000:1500]=1.
```

Discriminantes Logísticos

Outro exemplo:

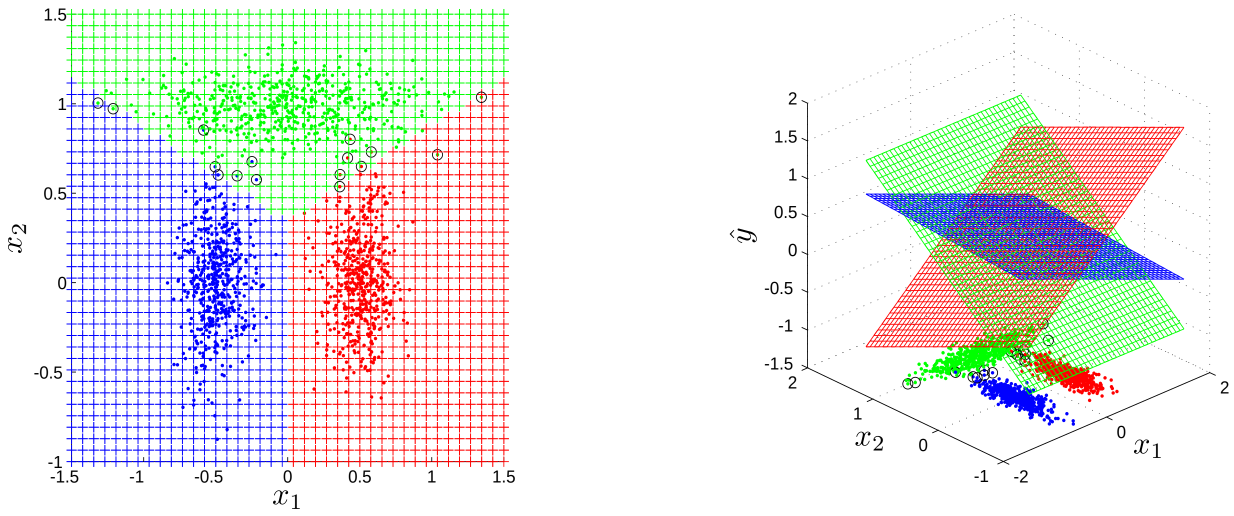
- Discriminantes Lineares: solução analítica $\mathbf{W} = \left(\mathbf{X}\mathbf{X}^\top\right)^{-1}\mathbf{X}\mathbf{Y}^\top$

```
W =np.dot(sp.linalg.pinv(np.dot(X,X.T)),npdot(X,Y.T))=
```

$$\begin{array}{ccc} +0.4735 & +0.4680 & +0.0585 \\ -0.6682 & +0.6741 & -0.0059 \\ \underbrace{-0.4117}_{\mathbf{w}_1} & \underbrace{-0.4123}_{\mathbf{w}_2} & \underbrace{+0.8240}_{\mathbf{w}_3} \end{array}$$

Três planos (1 por classe)

$$\hat{y}_1 = \mathbf{w}_1^\top \mathbf{x}$$
$$\hat{y}_2 = \mathbf{w}_2^\top \mathbf{x}$$
$$\hat{y}_3 = \mathbf{w}_3^\top \mathbf{x}$$



Discriminantes Logísticos

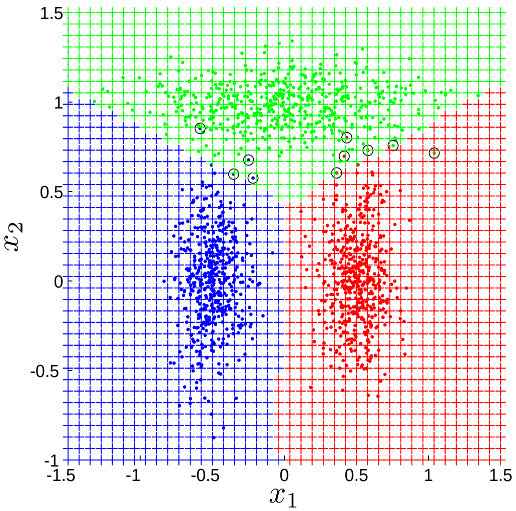
Outro exemplo:

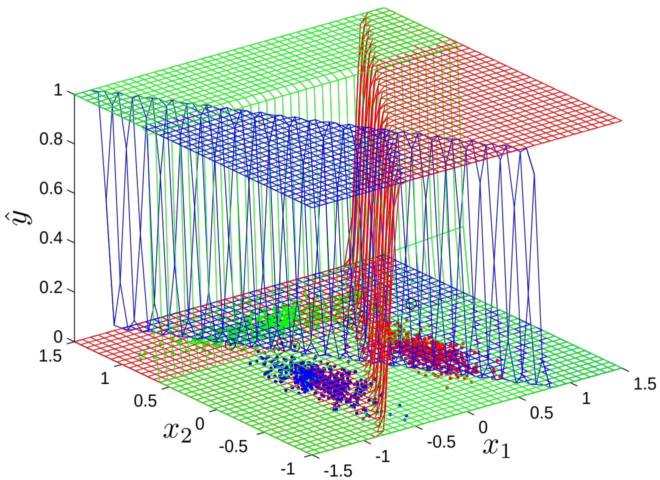
- Discriminantes Logísticos: descida de gradiente
  Pseudo-Código:
  - Calcular saídas:
    ```
    Yh=(1+np.exp(-np.dot(W.T,X)))**(-1)
    ```
  - Calcular gradiente:
    ```
    gW=(-2*np.dot((Y-Yh)*(Yh-Yh**2),X.T)).T
    ```
  - Adaptar pesos:
    ```
    W=W-eta*gW;
    ```

Discriminantes Logísticos

Outro exemplo:

- Discriminantes Logísticos
  - Resultados ligeiramente melhores do que os obtidos com discriminantes lineares.

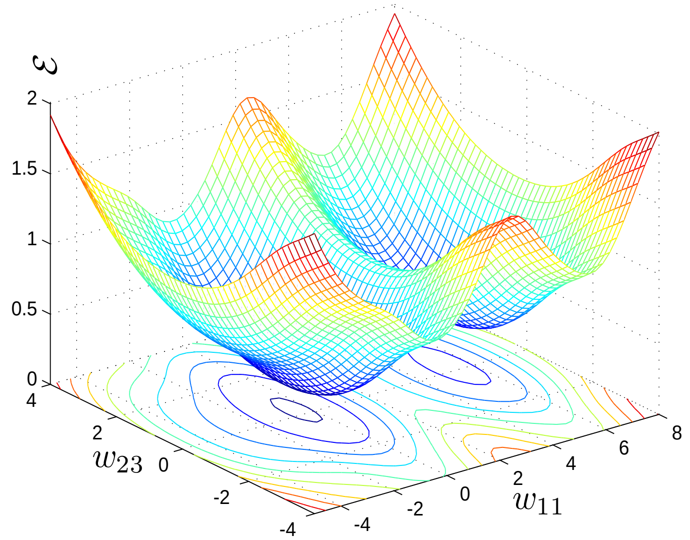




Discriminantes Logísticos

Outro exemplo:

- Discriminantes Logísticos
  - Resultados ligeiramente melhores do que os obtidos com discriminantes lineares.
  - Função do erro não é uma parábola. Podemos ficar presos em mínimos locais!



Função do erro em termos de dois pesos, w_{11} e w_{23} da matriz $\mathbf{W}$. Os restantes pesos foram fixados nos valores óptimos.

Discriminantes Logísticos

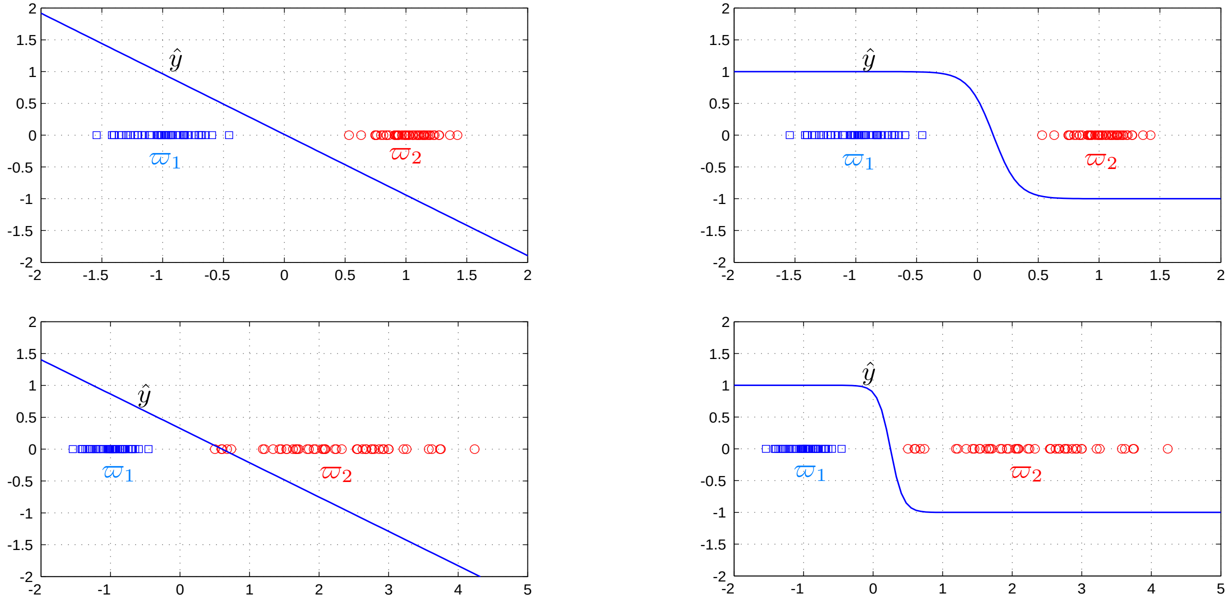
Vantagens relativamente a discriminantes lineares

- Saídas $\hat{\mathbf{y}}$ com valores limitados (podem-se aproximar mais facilmente das saídas desejadas $\mathbf{y}$).
  - Classes com pontos espalhados já não causam problemas.

    2 Exemplos: pontos 1D, duas classes

1º

2º



$$\hat{y} = w_0 + w_1 x \qquad \hat{y} = \tanh(w_0 + w_1 x)$$

Discriminantes Logísticos

Vantagens relativamente a discriminantes lineares

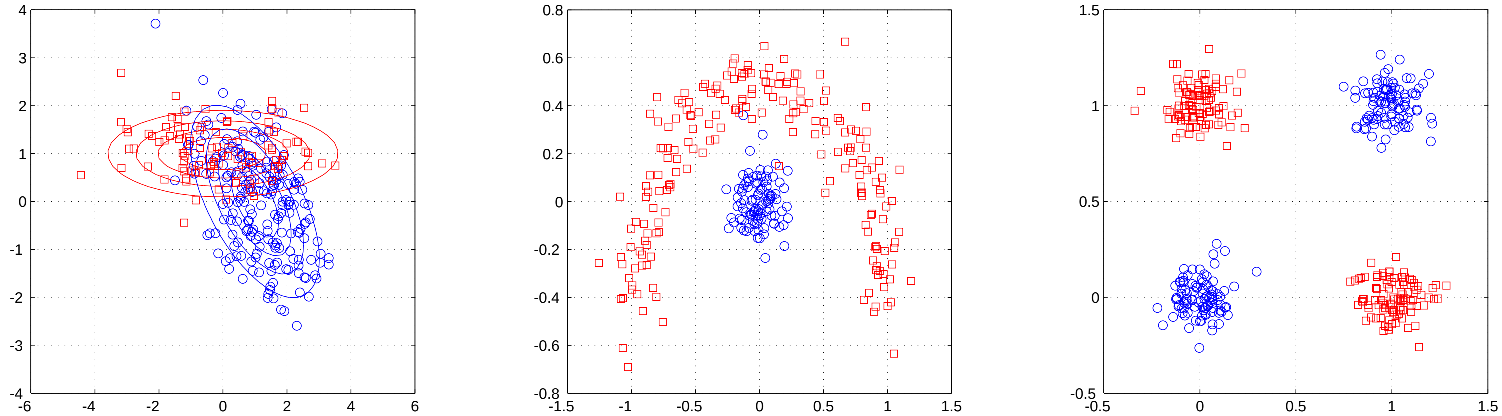
- Saídas $\hat{\mathbf{y}}$ com valores limitados (podem-se aproximar mais facilmente das saídas desejadas $\mathbf{y}$).
- Saídas podem ser convertidas em probabilidades. Podemos obter um grau de confiança sobre os resultados de classificação.

Discriminantes Logísticos

Limitações:

- Fronteiras de decisão lineares
- Problemas quando os dados de uma classe estão repartidos por regiões distintas

  Exemplos



- Diferentes inicializações podem obter diferentes resultados