Procesado de Información Biológica

Sesión 7

Mónica Rojas Martínez

Universidad El Bosque

Contenido

- Introducción a la clasificación
- Regresión logística
  - Hipotesis
  - Modelo
  - Función de coste
  - Algorimto
- Clasificación múltiples clases
  - Todos vs. uno

Clasificación

› Modelo donde la salida es discreta→ clases

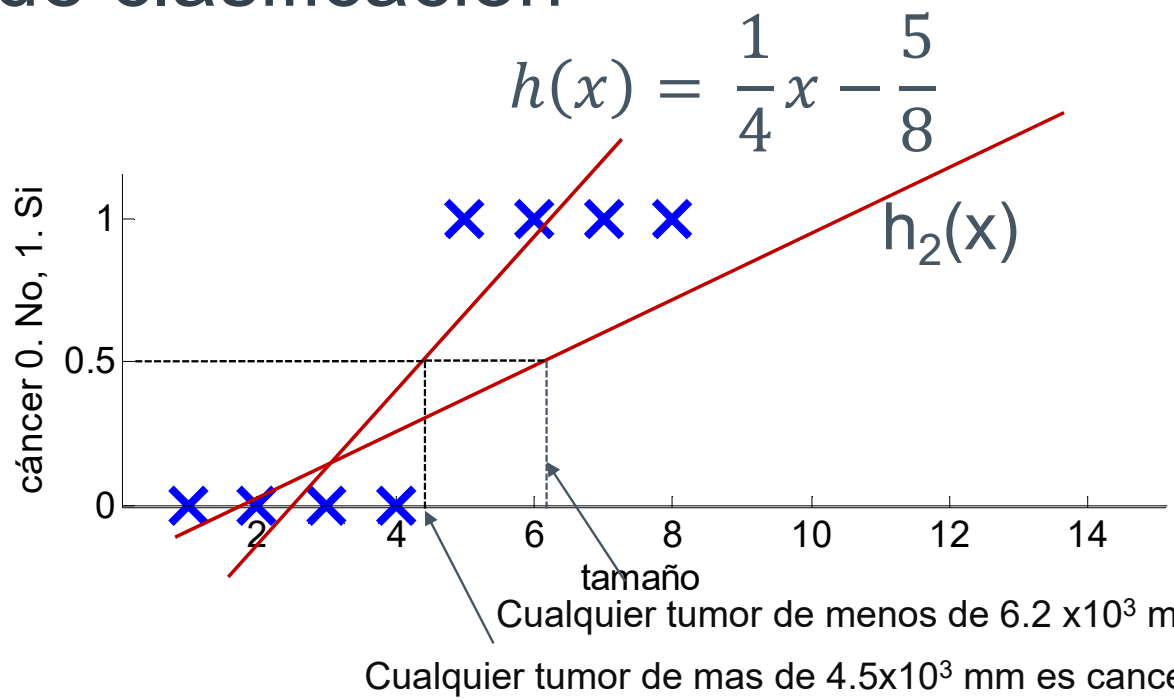
› 2 clases: salida binaria (1,0)

› Múltiples clases

› Ejemplos… enfermos / sanos, riesgo/ seguro, etc.

Problema de clasificación

x (tamaño tumor, cm)	y (cáncer, si-1 / no- 0)
1	0
2	0
3	0
4	0
5	1
6	1
7	1
8	1



~~Para clasificar puedo establecer un umbral sobre h(x):~~
~~si h(x)> 0.5 → cáncer~~
~~si h(x)<0.5 → no cáncer~~

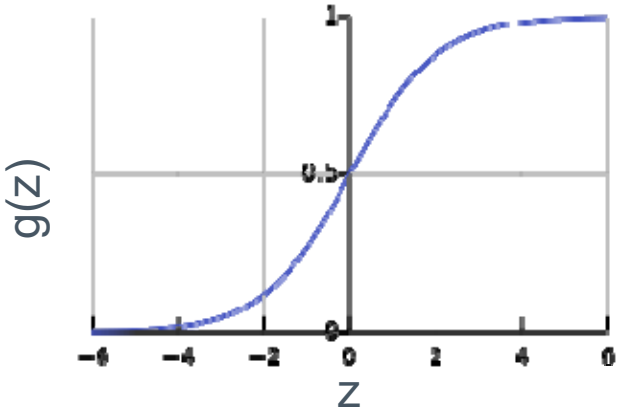
Regresión logística

- Clasificación binaria
- Regresión lineal $h_\theta\ (\mathbf{x}) = \boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}$ → valores continuos (recta o hiperplano)
- Regresión logística → queremos que $0 \leq h_\theta(x) \leq 1$
- En este caso $h_\theta\ (\mathbf{x})$ se puede entender como la probabilidad (de pertenencia a una clase)

Regresión logística- Hipótesis

- Queremos que $0 \leq h_\theta(x) \leq 1$
- Si tenemos $h_\theta(\mathbf{x}) = \theta^T \mathbf{x}$ (regresión lineal) podemos buscar una transformación de $h_\theta(\mathbf{x})$ de manera que su salida esté entre 0 y 1
- Sigmoide (función logística)

$$g(z) = \frac{1}{1 + e^{-z}}, \text{con } z = \theta^T \mathbf{x}$$



- Transformación: $h_\theta(x) = g(z) = \frac{1}{1+e^{-\theta^T x}}$

Regresión Lógistica- Hipótesis (II)

› Interpretación

$$h_\theta(x) = \frac{1}{1+e^{-\theta^T X}} \qquad 0 < h_\theta(x) < 1$$

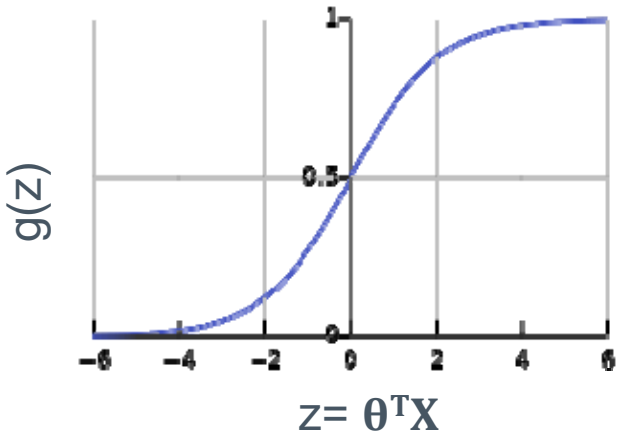
$h_\theta(x)$ es la probabilidad de tener una salida $y = 1$ dada la entrada x

$$h_\theta(x) = P(y = 1/x); \qquad \mathrm{P(y = 0/x) = 1 - P(y = 1/x)}$$

Regresión logística: umbral de decisión

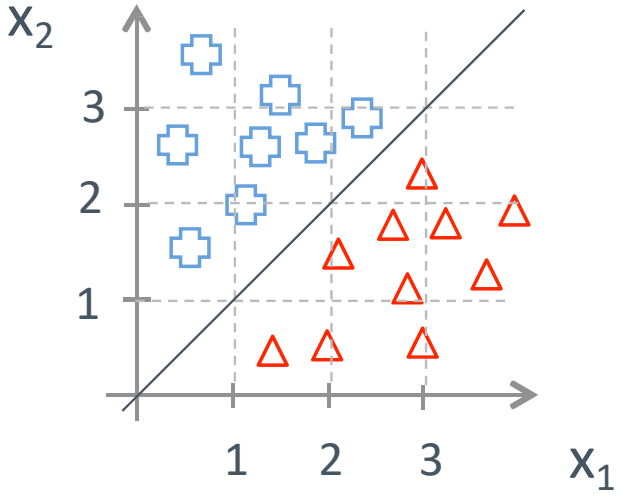
› Si predecimos y= 1 si $h_\theta \geq 0.5$ → $\boldsymbol{\theta^T X}$ >0

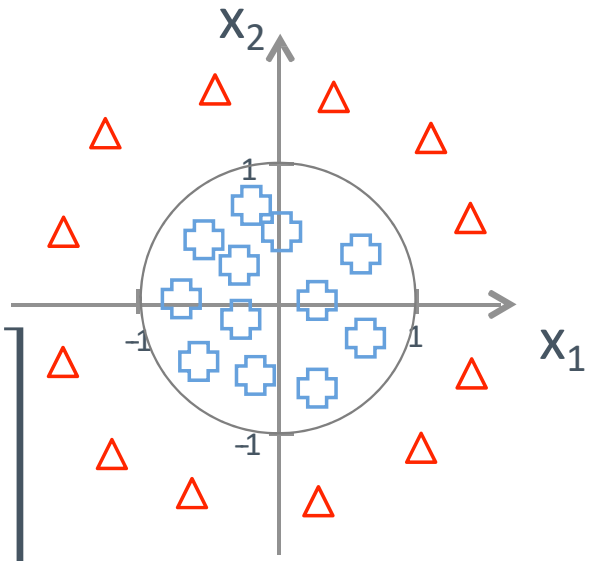
y por tanto y=0 si $h_\theta < 0.5$ es decir $\boldsymbol{\theta^T X}$ <0



Umbral de decisión

$$\theta = \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}$$





$$\theta = \begin{bmatrix} -1 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

Lineal: $h(x) = \theta_o x_o + \theta_1 x_1 + \theta_2 x_2 = x_1 - x_2$,

si $x_1 - x_2 \geq 0, y =$ △ , si $x_1 - x_2 < 0, y =$ ✠

No lineal $h(x) = \theta_o x_o + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 = -1 + x_1^2 + x_2^2$

si $x_1^2 + x_2^2 \geq 1, y =$ △ , si $x_1^2 + x_2^2 < 1, y =$ ✠

Regresión logística- Modelo

Tenemos un modelo con m datos de entrenamiento así:

$$\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \cdots, (x^{(m)}, y^{(m)})\}$$

x son las variables de entrada y y las salidas de la forma:

$$x = \begin{bmatrix} x_0 \\ x_1 \\ \cdots \\ x_n \end{bmatrix} \qquad x_0 = 1, y \in \{0, 1\}$$

Y finalmente una hipótesis de la forma:

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T X}}$$

Problema: Como escogemos los parámetros θ_j de la hipótesis?

Regresión logística- Función de coste (I)

En regresión lineal teníamos una función de coste convexa de la forma:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 = \frac{1}{m}\sum_{i=1}^{m}\mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)})$$

Escoger una función J(θ) convexa para garantizar un único mínimo

Se puede escoger una función del estilo:

$$\mathrm{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

Regresión logística- Función de coste (II)

Simplificación:

$$\text{Cost}(h_\theta(x), y) = -y \log\big(h_\theta(x)\big) - (1-y)\log(1-h_\theta(x))$$

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \text{Cost}(h_\theta(x^{(i)}), y^{(i)})$$

$$J(\theta) = -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1-y^{(i)}) \log (1-h_\theta(x^{(i)}))]$$

Regresión logística- Ajuste

$$J(\theta) = -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log (1 - h_\theta(x^{(i)}))]$$

Para encontrar el mínimo → $\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$

› Gradiente descendente

Repetir hasta que converja {

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

}

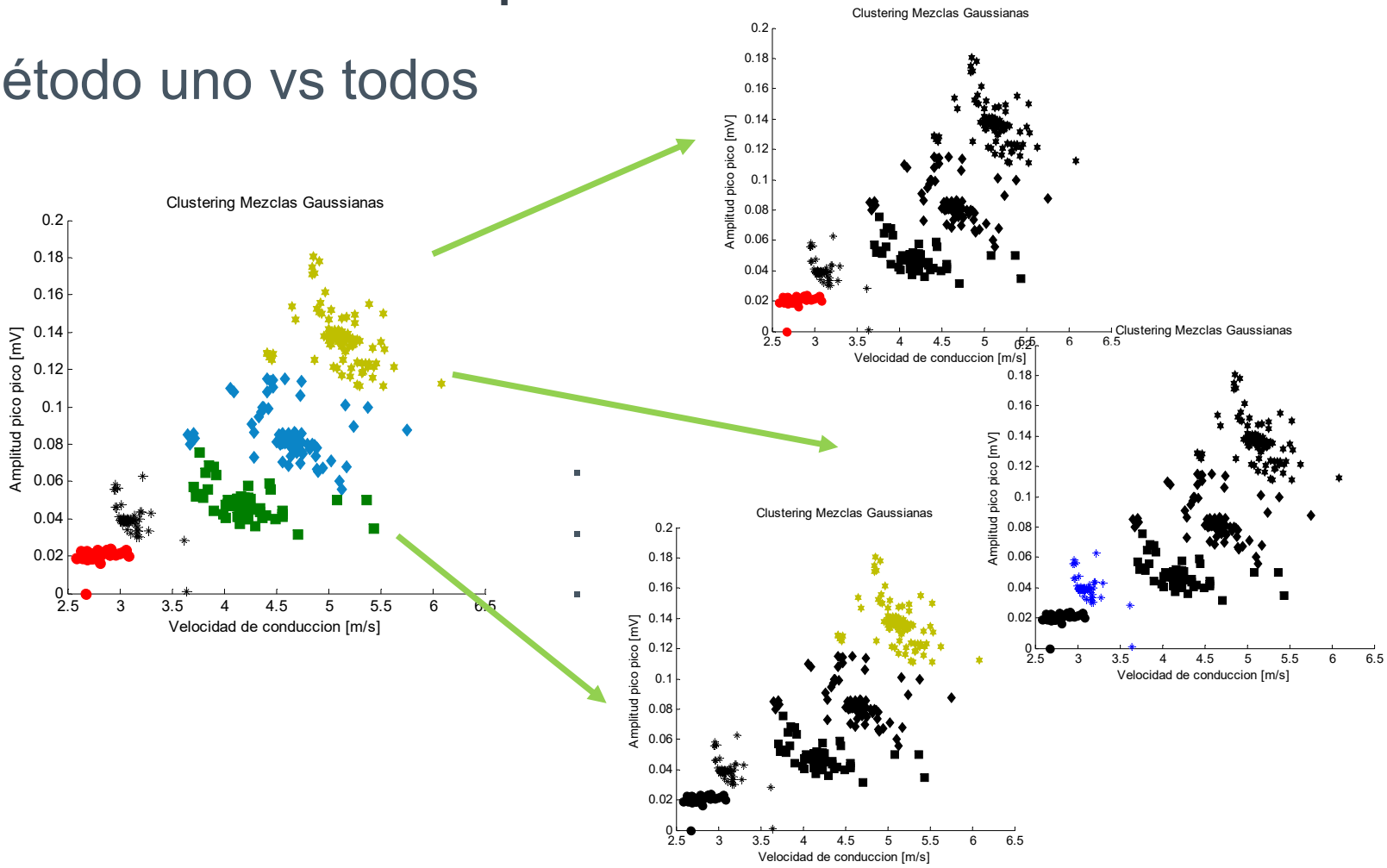
Se puede demostrar que:

$$\theta_j := \theta_j - \alpha \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

La misma que en el caso de regresión lineal

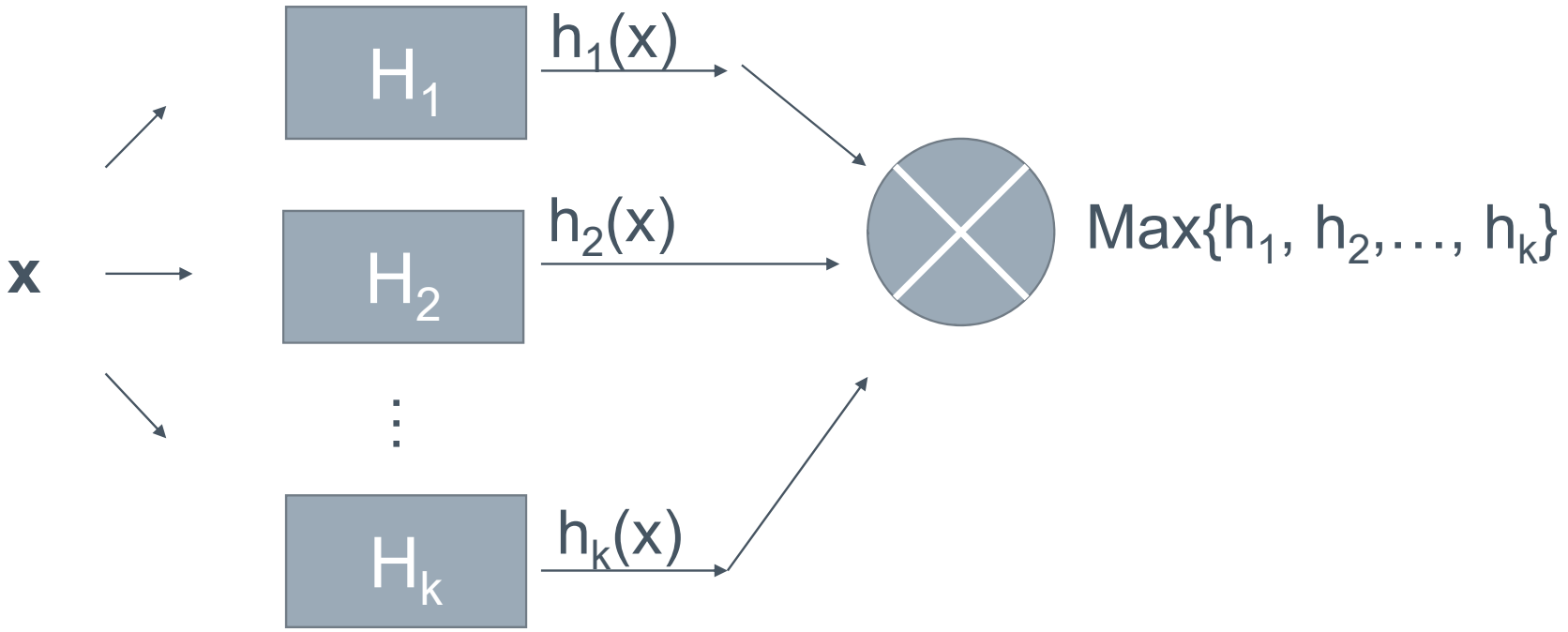
Clasificación múltiples clases

- Método uno vs todos



Múltiples clases- Uno vs. todos

› Modelo



Recordar, h_k es p(y=1/x) para una clase

Regularización (I)

- Evitar el sobreajuste en casos como:
  - Un número muy alto de variables de entrada
  - Modelos de orden superior donde es difícil estimar el orden
- Es una forma de “castigar” los parámetros de la hipótesis para evitar que la función de coste llegue al mínimo

Regularización, Regresión Logística (II)

› Modificacion a la función de coste

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^i \log\left(h\left(x^i\right)\right) + \left(1 - y^i\right)\log\left(1 - h\left(x^i\right)\right)\right] + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2$$

Donde el término de la derecha el es término de regularización y λ es el factor de regularización. Observe que el término de regularización solo se aplica a los parámetros $[\theta_1, \theta_2, \ldots, \theta_n]$ sin tener en cuenta el término θ_o asociado a x_o=1!!!

› Modificación al gradiente

$$\frac{\partial J}{\partial \theta_o} = \frac{1}{m}\sum_{i=1}^{m}\left(h\left(x^i\right) - y^i\right) \quad \text{sii } j = 0$$

$$\frac{\partial J}{\partial \theta_j} = \frac{1}{m}\sum_{i=1}^{m}\left(h\left(x^i\right) - y^i\right) x_j^i + \frac{\lambda}{m}\theta_j \quad \text{sii } j \neq 0$$