Procesado de Información Biológica

Sesión 7

Mónica Rojas Martínez

Universidad El Bosque

Contenido

› Regresión lineal

› Modelo con múltiples entradas

› Función de coste y algoritmo gradiente descendente

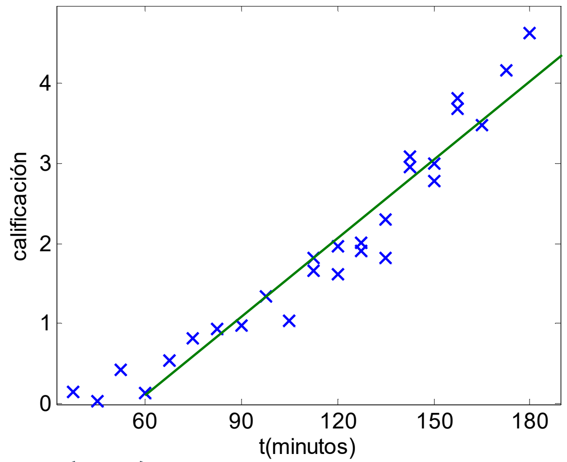
Aprendizaje automático

- Conjunto de métodos para predecir un evento o una salida a partir de un conjunto de ejemplos (entradas)
- Si la variable de salida es continua se habla de una regresión. Ej: predecir el clima durante el día (18.5º a las 7.00 am, 22º a las 12m). Cuales pueden ser las entradas? Cuantas salidas hay? Cuales son dichas salidas y que valores toman?
- Si la variable de salida es discreta se habla de una clasificación. Ej. En una población se quiere diferenciar entre deportistas (clase 1) y sedentarios (clase 0). Cuales pueden ser las entradas? Cuales y cuantas las salidas?

Regresión lineal con una variable

Hasta ahora

	x	y
Obs^{1} →	0,8	74856
Obs^{2} →	1,8	76925
⋮	⋮	
Obs^{m} →	10,8	278624



Hipótesis : $h(x) = \theta_o + \theta_1 x_1$

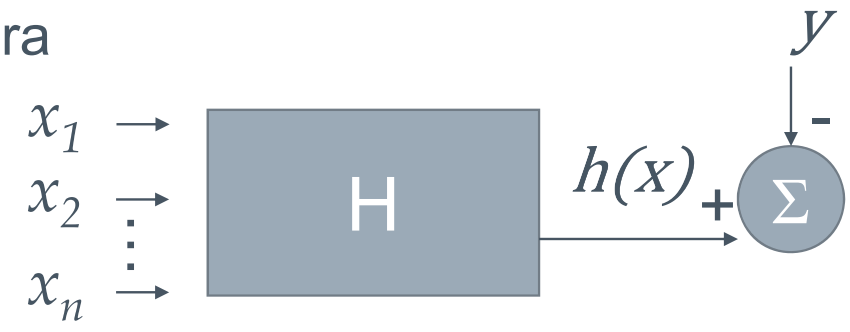
$$h(x) = \boldsymbol{\theta}^T \mathbf{x}$$

Con $\boldsymbol{\theta} = \begin{bmatrix} \theta_o \\ \theta_1 \end{bmatrix}$, $\mathbf{x} = \begin{bmatrix} x_o \\ x_1 \end{bmatrix}$ y $x_o = 1$,

Función de coste: $J(\theta_o, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h(x^i) - y^i)^2$

Regresión lineal con múltiples variables

Ahora



	x_1	x_2	x_n	y
Obs [1] →	0,8	18	560	74856
Obs [2] →	1,8	13	248	76925
⋮	⋮	⋮	⋮	
Obs [m] →	10,8	11	120	278624

Hipótesis : $h(x) = \theta_o + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$

Función de coste: $J(\theta_o, .., \theta_n) = \frac{1}{2m} \sum_{i=1}^{m} (h(x^i) - y^i)^2$

Hipótesis Regresión lineal con múltiples variables

› $h(x) = \theta_o x_o + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$

Si Por conveniencia definimos x_o:=1

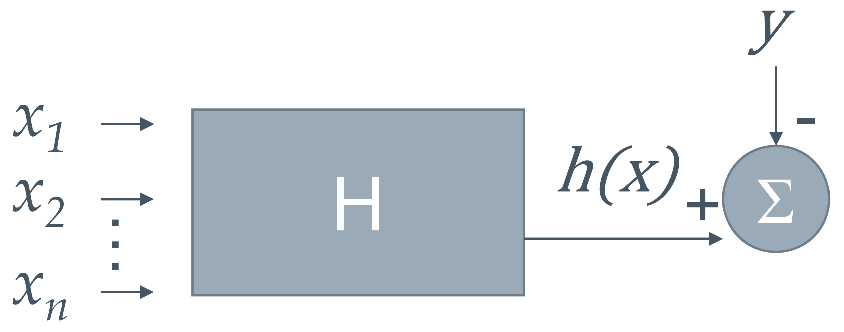
Podemos escribir las entradas y los parámetros de forma vectorial como:

$$\mathbf{x} = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^{n+1} \quad \text{y} \quad \boldsymbol{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix} \in \mathbb{R}^{n+1}$$

→ $\boxed{h(x) = \boldsymbol{\theta}^T \mathbf{x}}$

Regresión con múltiples variables (general)

Un modelo no lineal de múltiples variables considera por ejemplo:



	x_o	x_1	x_2	x_3	x_4	x_5	y
Obs [1]	0.8	1	1	0,6	0,36	0,6	10,1
Obs [2]	0.4	1,1	1,21	1,1	1,21	1,21	11,2
	0.6	1,6	2,56	1,9	3,61	3,04	16,4
⋮		⋮		⋮	⋮		
Obs [m]	0.7	2,3	5,29	3,5	12,3	8,05	20,7

$$h(x) = \theta_o x_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

Con $x_3 = x_1^2,\ x_4 = x_2^2,\ x_5 = x_1 x_2$

Hipótesis (“no lineal”) : $\boldsymbol{h(x) = \theta_o x_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1 x_2}$

Función de coste: $J(\theta_o, .., \theta_n) = \frac{1}{2m} \sum_{i=1}^{m} (h(x^i) - y^i)^2$

Ajuste de parámetros

- A partir del conjunto de datos de entrenamiento, encontrar los valores del vector $\boldsymbol{\theta}$
- Con este propósito se utilizan métodos de optimización sobre una función convexa en sentido amplio (es decir, con un único mínimo absoluto)
- Métodos de optimización: muy variados, ejemplo → gradiente descendente

Gradiente descendente (idea general)