Aprendizaje Automático y Minería de Datos

Clasificación

Cristina Tîrnăucă

Dept. Matesco, Universidad de Cantabria

Fac. Ciencias – Grado en Ing. Informática

El proceso de minería de datos incluirá fases de modelado a partir de observaciones (datos) sobre una realidad compleja y existente.

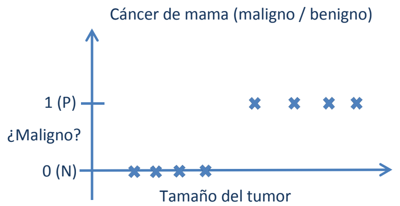
Taxonomía:

- Modelos descriptivos:
  - Segmentación,
  - Asociación.
- Modelos predictivos:
  - Regresión,
  - Clasificación,

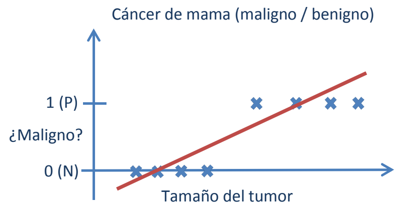
- Modelos supervisados,
- Modelos no supervisados,

(Nociones mutuamente no excluyentes.)

¿Utilizar regresión lineal para clasificación?



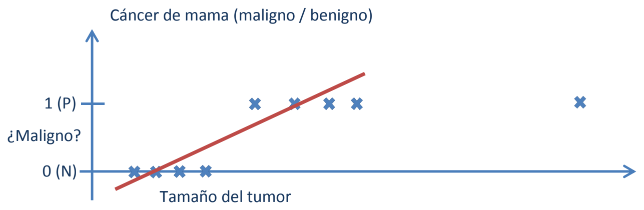
¿Utilizar regresión lineal para clasificación?



Idea: establecer el umbral de clasificación para $h_\theta(x)$ a 0.5:

- si $h_\theta(x) \geq 0.5$, predecir "y=1"
- si $h_\theta(x) < 0.5$, predecir "y=0"

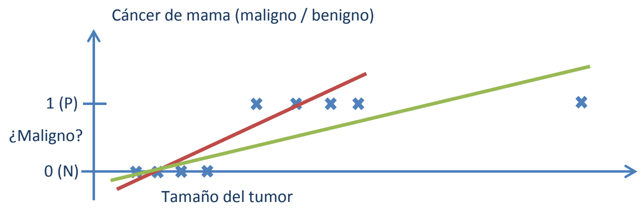
¿Utilizar regresión lineal para clasificación?



Idea: establecer el umbral de clasificación para $h_\theta(x)$ a 0.5:

- si $h_\theta(x) \geq 0.5$, predecir "y=1"
- si $h_\theta(x) < 0.5$, predecir "y=0"

¿Utilizar regresión lineal para clasificación?



Idea: establecer el umbral de clasificación para $h_\theta(x)$ a 0.5:

- si $h_\theta(x) \geq 0.5$, predecir "y=1"
- si $h_\theta(x) < 0.5$, predecir "y=0"

Problema: $y = 0$ o $y = 1$ (clasificación)
Pero $h_\theta(x)$ puede ser > 1 o < 0
Regresión logística: $0 \leq h_\theta(x) \leq 1$

Modelo de regresión logística

Queremos: $0 \leq h_\theta(x) \leq 1$

$$h_\theta(x) = g(\theta_0 + \theta_1 * x)$$

Función sigmoide (función logística): $g(z) = \frac{1}{1+e^{-z}}$

Predecimos "$y = 1$" si $h_\theta(x) \geq 0.5$
"$y = 0$" si $h_\theta(x) < 0.5$

Umbral de decisión

$$h_\theta(x) = g(\theta_0 + \theta_1 * x_1 + \theta_2 * x_2)$$

Por ejemplo, si $\theta = (-3, 1, 1)^t$.

Predecimos "$y = 1$" si $-3 + x_1 + x_2 \geq 0$
"$y = 0$" si $-3 + x_1 + x_2 < 0$

Umbral de decisión no lineal

$h_\theta(x) = g(\theta_0 + \theta_1 * x_1 + \theta_2 * x_2 + \theta_3 * x_1^2 + \theta_4 * x_2^2)$

Por ejemplo, si $\theta = (-1, 0, 0, 1, 1)^t$.

Predecimos "$y = 1$" si $1 + x_1^2 + x_2^2 \geq 0$
"$y = 0$" si $1 + x_1^2 + x_2^2 < 0$

... y podemos complicar aún más $h_\theta(x) =$
$g(\theta_0+\theta_1*x_1+\theta_2*x_2+\theta_3*x_1^2+\theta_4*x_1^2*x_2+\theta_5*x_1^2*x_2^2+\theta_6*x_1^3*x_2+\ldots)$

Función de coste

Regresión lineal: $J(\theta) = \frac{1}{m}\sum_{i=1}^m \frac{1}{2}(h_\theta(x^{(i)}) - y^{(i)})^2$

$$Cost(h_\theta(x^{(i)}), y^{(i)}) = \frac{1}{2}(h_\theta(x^{(i)}) - y^{(i)})^2$$

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

Función de coste para la regresión logística:

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

En una linea,
$Cost(h_\theta(x), y) = -y * \log(h_\theta(x)) - (1 - y) * \log(1 - h_\theta(x))$

Algorítmo del gradiente descendente, I

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

Objetivo: buscar θ tal que $J(\theta)$ sea mínimo.

repetir hasta convergencia $\{$
$\theta_j = \theta_j - \alpha\frac{\partial}{\partial\theta_j}J(\theta)$
$\}$

Algorítmo del gradiente descendente, II

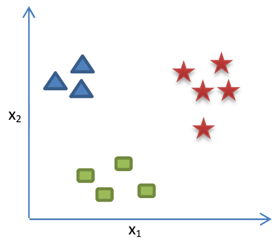
$$J(\theta) = -\frac{1}{m}[\sum_{i=1}^m y^{(i)} * \log(h_\theta(x^{(i)})) + (1 - y^{(i)}) * \log(1 - h_\theta(x^{(i)}))]$$

Objetivo: buscar θ tal que $J(\theta)$ sea mínimo.

repetir hasta convergencia $\{$
$\theta_j = \theta_j - \frac{\alpha}{m}\sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)}) * x_j^{(i)}$
$\}$

Clasificación multi-clase

One-vs-all



$h_\theta^{(1)}(x)$

$h_\theta^{(2)}(x)$

$h_\theta^{(3)}(x)$

Para predecir, eliges i tal que $h_\theta^{(i)}(x) =$ máximo.