

# Regresión logística

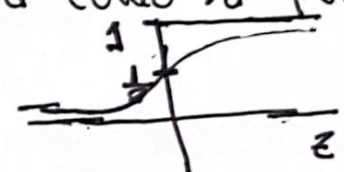
1

Supongamos que clasificamos dos clases la clase 0 y la clase 1 mediante probabilidades, y que dado el patrón  $x$  la probabilidad que pertenezca a la clase 1 sea  $P$  y por lo tanto la probabilidad que pertenezca a la clase 0 es  $1-P$ . Consideremos una variable auxiliar  $z$  para evaluar  $P$  de una manera sencilla.

$$\ln\left(\frac{P}{1-P}\right) = z, \text{ despejando } P \text{ tenemos } \frac{P}{1-P} = e^z,$$

$$P = e^z - P e^z, \quad P + P e^z = e^z, \quad (1 + e^z)P = e^z, \quad P = \frac{e^z}{1 + e^z},$$

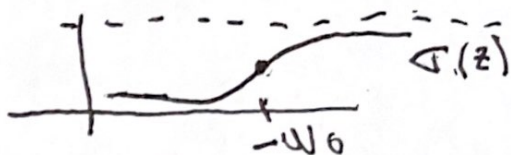
$P = \frac{1}{1 + e^{-z}}$  es decir en términos de  $z$   $P$  se comporta como la función logística

$$P = \sigma(z) = \frac{1}{1 + e^{-z}} \quad \text{y} \quad 1-P = \frac{1}{1 + e^z}$$


Si tenemos un conjunto de  $L$  datos etiquetados  $(x^l, y^l)_{l=1}^L$  esto se puede escribir como

$$\ln\left[\frac{P(y^l=1|x^l, w)}{P(y^l=0|x^l, w)}\right] = w_0 + \sum_{i=1}^n w_i x_i^l, \text{ donde } x \text{ es el patrón y } w \text{ los parámetros del clasificador.}$$

En este caso  $z = w_0 + \sum_{i=1}^n w_i x_i^l$  es la variable de decisión y  $w_0$  es el desplazamiento que permite ubicar la frontera de las clases en cualquier punto  $-w_0$  del eje  $z$ .



Entonces dados los datos  $D = \{(x^l, y^l) : x^l \in \mathbb{R}^n, y^l \in \{0, 1\}\}$  el problema consiste en hallar  $w \in \mathbb{R}^{n+1}$  vector de pesos tal que maximice la probabilidad de clasificar adecuadamente los datos  $D$ .

$$w_* = \arg\max_w P(D|w)$$



2

Suponiendo independencia en los atributos del patrón  $x$ , la ecuación de búsqueda de  $w$  queda como sigue.

$$w_* = \arg \max_w \prod_l P(y^l | x^l, w), \text{ donde } P(D|w) = \prod_l P(y^l | x^l, w)$$

haciendo

$$g(w) = \log \prod_l P(y^l | x^l, w) = \sum_l \ln P(y^l | x^l, w)$$

Esto se puede escribir como

$$\begin{aligned} g(w) &= \sum_l y^l \ln P(y^l=1 | x^l, w) + (1-y^l) \ln P(y^l=0 | x^l, w) \\ &= \sum_l y^l \ln P(y^l=1 | x^l, w) + \ln P(y^l=0 | x^l, w) - y^l \ln P(y^l=0 | x^l, w) \\ &= \sum_l y^l [\ln P(y^l=1 | x^l, w) - \ln P(y^l=0 | x^l, w)] + \ln P(y^l=0 | x^l, w) \\ &= \sum_l y^l \left[ \ln \frac{P(y^l=1 | x^l, w)}{P(y^l=0 | x^l, w)} \right] + \ln P(y^l=0 | x^l, w), \text{ finalmente} \end{aligned}$$

$$g(w) = \sum_l y^l (w_0 + \sum_{i=1}^n w_i x_i^l) - \ln (1 + e^{w_0 + \sum_{i=1}^n w_i x_i^l})$$

Para maximizar  $g(w)$  tenemos que calcular su gradiente

$$\begin{aligned} \frac{\partial g(w)}{\partial w_j} &= \sum_l \left[ y^l x_j^l - \left[ \frac{1}{1 + e^{w_0 + \sum_{i=1}^n w_i x_i^l}} \right] e^{w_0 + \sum_{i=1}^n w_i x_i^l} x_j^l \right] \\ &= \sum_l x_j^l \left[ y^l - \frac{e^{w_0 + \sum_{i=1}^n w_i x_i^l}}{1 + e^{w_0 + \sum_{i=1}^n w_i x_i^l}} \right], \text{ finalmente} \end{aligned}$$

Para  $j=0, 1, 2, \dots, n$

$$\frac{\partial g(w)}{\partial w_j} = \sum_l x_j^l [y^l - \hat{p}(y^l=1 | x^l, w)], \text{ es decir el gradiente}$$

error de predicción es proporcional al error de predicción.

$\nabla g = (\frac{\partial g}{\partial w_0}, \frac{\partial g}{\partial w_1}, \dots, \frac{\partial g}{\partial w_n})$  y un máximo local

se encuentra iterando en la dirección del  $\nabla g(w)$   
 $w = w + \eta \nabla g(w)$ , donde  $0 < \eta < 1$  es el paso de aprendizaje.