Problema 2. Funcions d'error per classificació [G]

Albert Ribes

15 de noviembre de 2017

L'objectiu dels models probabilístics discriminatius per classificació és modelar les probabilitats a posteriori $P(C_k|x)$ per a cada classe k. En tasques de classificació binària (dues classes, C_1 i C_2), modelem amb una funció $y(x) = P(C_1|x)$; llavors $1 - y(x) = P(C_2|x)$. Tenim una mostra aleatòria simple D de llargada N del mecanisme $p(t, x)$, que escrivim $D = \{(x_1, t1), \ldots, (x_N, t_N)\}$, on $x_n \in R^d$ i $t_n \in \{0, 1\}$. Prenem la convenció que $t_n = 1$ indica $x_n \in C_1$ i $t_n = 0$ indica $x_n \in C_2$, i modelem:

$$P(t|x) = \begin{cases} y(x) & \text{si } x_n \in C_1 \\ 1 - y(x) & \text{si } x_n \in C2 \end{cases}$$

que pot ser més convenientment expressat com $P(t|x) = y(x)^t(1-y(x))^{1-t}$, $t = \{0, 1\}$. Aquesta és una distribució de Bernoulli, la qual cosa permet d'obtenir una funció d'error amb criteris ben fonamentats.

1. Construïu la funció log-versemblança de la mostra i proposeu una funció d'error a partir d'ella.

   Asumiendo que los datos son independientes e idénticamente distribuidos, la probabilidad de haber observado los datos D es

$$\begin{aligned} P(t_1|x_1)P(t_2|x_2)\ldots P(t_n|x_n) &= \prod_{i=1}^{N} P(t_i|x_i) \\ &= \prod_{i=1}^{n} y(x_i)^{t_i}(1 - y(x_i))^{1-t_i} \end{aligned}$$

   Para encontrar la función log-verosimilitud hemos de encontrar los valores de $y(x_i)$ que maximicen esta probabilidad, y para ello haremos la derivada respecto de $y(x_i)$ para cada $i \in [1, n]$ de su logaritmo natural y la igualaremos a 0:

$$\begin{aligned} l_j &= 0 \\ \frac{\partial}{\partial y(x_j)} \ln\Big[\prod_{i=1}^{n} y(x_i)^{t_i}(1 - y(x_i))^{1-t_i}\Big] &= 0 \end{aligned}$$

$$
\begin{aligned}
\frac{\partial}{\partial y(x_j)}\Big[\sum_{i=1}^{n} t_i \ln y(x_i) + \sum_{i=0}^{n}(1-t_i)\ln(1-y(x_i))\Big] &= 0 \\
t_j\frac{1}{y(x_j)} + (1-t_j)\frac{1}{(1-y(x_j))} &= 0 \\
\frac{t_j}{y(x_j)} &= -\frac{1-t_j}{(1-y(x_j))} \\
(1-y(x_j))t_j &= -y(x_j)(1-t_j) \\
t_j - t_j y(x_j) &= -y(x_j)(1-t_j) \\
t_j &= -y(x_j)(1-t_j) + t_j y(x_j) \\
t_j &= y(x_j)\big(t_j - 1 + t_j\big) \\
\frac{t_j}{2t_j - 1} &= y(x_j)
\end{aligned}
$$

Que, para el caso particular en que $t_i \in \{0,1\}$, se puede escribir como:

$$y(x_j) = t_j$$

Este resultado es lógico: la opción más verosímil es la que ya se ha observado.

2. Generalitzeu el resultat a un número arbitrari $K \geq 2$ de classes.

Para simplificar la notación definimos la matriz

$$
Y_{N\times K} = \begin{bmatrix}
P(C_1|x_1) & P(C_2|x_1) & \dots & P(C_K|x_1) \\
P(C_1|x_2) & P(C_2|x_2) & \dots & P(C_K|x_2) \\
\vdots & \vdots & \ddots & \vdots \\
P(C_1|x_N) & P(C_2|x_N) & \dots & P(C_K|x_N)
\end{bmatrix}
$$

Esta matriz debe cumplir que

$$\forall i \in [1, N], \sum_{j=1}^{K} Y_{ij} = 1$$

Y entonces el modelo se puede definir como

$$P(t_j|x_i) = \prod_{k=1}^{K} Y_{ik}^{unosiesj,0otramente}$$