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

Albert Ribes

24 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

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

   En nuestro caso queremos modelar $y(x_i)$ como $w^T x_i + w_0$. Para simplificar la notación añadiremos a x_i el elemento 1 al principio y juntaremos el vector w con w_0 para definir nuestra función como $y(x_i) = w^T x_i$

   Definiremos la función log-verosimilitud de manera que dependa de w, e intentaremos maximizar esa verosimilitud.

   Haciendo una sustitución de la fórmula anterior, podemos definir log-verosimilitud como:

$$l = \ln \prod_{i=1}^{n} (w^T x_i)^{t_i}(1 - w^T x_i)^{(1-t_i)}$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} \ln(w^T x_i)^{t_i} + \ln(1 - w^T x_i)^{(1-t_i)} \\
&= \sum_{i=1}^{n} t_i \ln w^T x_i + \sum_{i=1}^{n} (1 - t_i) \ln(1 - w^T x_i)
\end{aligned}
$$

Para encontrar la función log-verosimilitud hemos de encontrar los valores de w que maximicen esta probabilidad, y para ello haremos la derivada respecto de w_j para cada $j \in [0, d]$ de su logaritmo natural y la igualaremos a 0:

$$
\begin{aligned}
\frac{\partial l}{w_j} &= 0 \\
\frac{\partial}{\partial w_j} \sum_{i=1}^{n} t_i \ln(w^T x_i) + \sum_{i=1}^{n} (1 - t_i) \ln(1 - w^T x_i) &= 0 \\
\sum_{i=1}^{n} t_i \frac{1}{w^T x_i} x_j + \sum_{i=1}^{n} (1 - t_i) \frac{1}{1 - w^T x_i} x_j &= 0 \\
x_j \sum_{i=1}^{n} \frac{t_i}{w^T x_i} - x_j \sum_{i=1}^{n} \frac{t_i - 1}{1 - w^T x_i} &= 0 \\
\sum_{i=1}^{n} \frac{t_i}{w^T x_i} &= \sum_{i=1}^{n} \frac{t_i - 1}{1 - w^T x_i}
\end{aligned}
$$

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

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