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

Gerard Barrachina Josep de Cid Albert Ribes
Kerstin Winter

7 de diciembre 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 $D = \{(x_1, t_1), \ldots, (x_N, t_N)\}$, on $x_n \in \mathbb{R}^d$ 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

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

   Puesto que se trata de un problema de clasificación con dos clases, modelaremos $y(x)$ como $g(w^T x + w_0)$, donde g es la función logística, que se define como $g(z) = \frac{1}{1+exp(-z)}$. Para simplificar la notación añadiremos a x el elemento 1 al principio y juntaremos el vector w con w_0 para definir nuestra función como $y(x) = g(w^T x)$

   La función log-verosimilitud debe maximizar la probabilidad de haber observado los datos D, y debe hacerlo mediante el parámetro w. Para simplificar los cálculos se trabaja con el logaritmo

natural de esa probabilidad, que no afecta en los parámetros que la maximizan. De este modo podemos definir la función log-verosimilitud como:

$$\begin{aligned} l(D,w) &= \ln \prod_{i=1}^{N} g(w^T x_i)^{t_i}(1-g(w^T x_i))^{(1-t_i)} \\ &= \sum_{i=1}^{N} \ln(g(w^T x_i))^{t_i} + \ln(1-g(w^T x_i))^{(1-t_i)} \\ &= \sum_{i=1}^{N} t_i \ln g(w^T x_i) + (1-t_i)\ln(1-g(w^T x_i)) \end{aligned}$$

Entonces una buena función de error sería menos log-verosimilitud, i.e:

$$\begin{aligned} E(D;w) &= -l(D;w) \\ &= -\sum_{i=1}^{N} t_i \ln g(w^T x_i) + (1-t_i)\ln(1-g(w^T x_i)) \end{aligned}$$

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

   Puesto que ahora tenemos más clases, será necesario un cambio en la notación.

   Definimos $y_k(x)$ como la probabilidad de que el dato x pertenezca a la clase k, i.e: $y_k(x) \equiv P(C_k|x)$.

   Ahora $t_n \in \{1, 2, \ldots, K\}$, y por lo tanto definiremos la matriz

$$T = \begin{array}{c} \\ x_1 \\ x_2 \\ \vdots \\ x_N \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \ldots & C_K \end{array} \\ \begin{bmatrix} t_{11} & t_{12} & \ldots & t_{1K} \\ t_{21} & t_{22} & \ldots & t_{2K} \\ \vdots & \vdots & \vdots & \vdots \\ t_{N1} & t_{N2} & \ldots & t_{Nk} \end{bmatrix} \end{array}$$

   Donde t_{ij} es 1 si $x_i \in C_j$ y es 0 si $x_i \notin C_j$.

   Para no confundir la notación, redefinimos la muestra de datos como $D = \{(x_1, z_1), (x_2, z_2), \ldots, (x_N, z_N)\}$, donde $z_n \in \{1, 2, \ldots, K\}$

   Ahora ya no tenemos un solo vector w, sino que tenemos uno por cada clase. Definimos entonces la matriz

$$W = \begin{array}{c} \\ C_1 \\ C_2 \\ \vdots \\ C_K \end{array} \begin{array}{c} \begin{array}{cccc} w_0 & w_1 & \ldots & w_d \end{array} \\ \begin{bmatrix} w_{10} & w_{11} & \ldots & w_{1d} \\ w_{20} & w_{21} & \ldots & w_{2d} \\ \vdots & \vdots & \vdots & \vdots \\ w_{K0} & w_{K1} & \ldots & w_{Kd} \end{bmatrix} \end{array}$$

Entonces el modelo que definimos es

$$y_k(x) = g(W_k x)$$

Donde W_k es la fila k de la matriz W

Puesto que ahora es un problema de clasificación con más de 2 clases la función logística ya no sirve, pues podría ocurrir que la suma de probabilidades de pertenencia a cada una de las clases para un dato no sumara 1. Hay que usar su equivalente para más de 2 clases, que es la función "softmax", y que para este caso concreto se define como:

$$g(W_k x) = \frac{exp(W_k x)}{\sum_{j=1}^{K} exp(W_j x)}$$

De este modo se asegura que $\sum_{j=1}^{K} g\big(y_j(x)\big) = 1$

Ahora podemos definir

$$P(z|x) = y_z(x)$$

Que por mantener la notación del apartado anterior también se podría definir como

$$P(z|x) = \prod_{k=1}^{K} y_k(x)^{t_{iz}}$$

Entonces la probabilidad de haber observado los datos D es

$$\begin{aligned} P(z_1|x_1)P(z_2|x_2)\ldots P(z_N|x_N) &= \prod_{i=1}^{N} P(z_i|x_i) \\ &= \prod_{i=1}^{N} y_{z_i}(x_i) \\ &= \prod_{i=1}^{N}\prod_{k=1}^{K} y_k(x_i)^{t_{iz_i}} \end{aligned}$$

Y la función log-verosimilitud se puede definir como

$$\begin{aligned} l(D;W) &= \sum_{i=1}^{N} \ln y_{z_i}(x_i) \\ &= \sum_{i=1}^{N} \ln g(W_{z_i} \cdot x_i) \end{aligned}$$

Entonces la función de error propuesta es

$$
\begin{aligned}
E(D;W) &= -l(D;W) \\
&= -\sum_{i=1}^{N} \ln g(W_{z_i} \cdot x_i)
\end{aligned}
$$