

T3 - EJERCICIOS

David Arnal García

2021 - Aprendizaje Automático

Ejercicio 1

Demostrar que en cualquier problema de clasificación en C clases, la estimación de máxima verosimilitud de la probabilidad a priori de cada clase c , $1 \leq c \leq C$, es $\frac{n_c}{N}$ donde $N = n_1 + \dots + n_C$ es el número total de datos observados y n_c es el número de datos de la clase c (transparencia 19). Solución:

Modelo:

$$P(C = i) = \hat{p}_i$$

$$\sum_{i=1}^C \hat{p}_i = 1$$

$$\theta = (\hat{p}_1, \dots, \hat{p}_C)^t$$

Verosimilitud y logaritmo de la verosimilitud:

$$P(S | \theta) = \prod_{i=1}^{n_1} \hat{p}_1 \cdots \prod_{j=1}^{n_c} \hat{p}_c = \hat{p}_1^{n_1} \cdot \hat{p}_2^{n_2} \cdots \hat{p}_c^{n_c}$$

$$q_S(\theta) = L_S(\theta) = \log P(S | \theta) = n_1 \cdot \log \hat{p}_1 + \dots + n_c \cdot \log \hat{p}_c$$

Estimación de máxima verosimilitud:

$$\theta^* = \underset{\hat{p}_1, \dots, \hat{p}_c : \hat{p}_1 + \dots + \hat{p}_c = 1}{\operatorname{argmax}} (n_1 \cdot \log \hat{p}_1 + \dots + n_c \cdot \log \hat{p}_c)$$

Lagrangiana:

$$\Lambda(\hat{p}_1, \dots, \hat{p}_c, \beta) = n_1 \cdot \log \hat{p}_1 + \dots + n_c \cdot \log \hat{p}_c + \beta$$

Soluciones óptimas en función del multiplicador de Lagrange:

$$\frac{\partial \Lambda}{\partial \hat{p}_1} = \frac{n_1}{\hat{p}_1} - \beta = 0, \quad \dots \quad, \quad \frac{\partial \Lambda}{\partial \hat{p}_c} = \frac{n_c}{\hat{p}_c} - \beta = 0$$

De donde se obtiene que:

$$\hat{p}_1^*(\beta) = \frac{n_1}{\beta}, \quad \dots \quad, \quad \hat{p}_c^*(\beta) = \frac{n_c}{\beta}$$

Función dual de Lagrange

$$\begin{aligned} \Lambda_D(\beta) &= n_1 \log \left(\frac{n_1}{\beta} \right) + \dots + n_c \log \left(\frac{n_c}{\beta} \right) + \beta \left(1 - \frac{n_1}{\beta} - \dots - \frac{n_c}{\beta} \right) \\ &= \beta - N \log \beta - N + \sum_{i=1}^C n_i \log n_i \end{aligned}$$

Valor óptimo del multiplicador de Lagrange:

$$\frac{d\Lambda_D}{d\beta} = 1 - \frac{N}{\beta} = 0 \Rightarrow \beta^* = N$$

Solución final:

$$\hat{p}_1^* = \hat{p}_1^*(\beta) = \frac{n_1}{N}, \quad \dots \quad \hat{p}_C^* = \hat{p}_C^*(\beta) = \frac{n_C}{N}$$

Ejercicio 2

Mostrar la traza de tres iteraciones de descenso por gradiente para minimizar: $q(\theta) = (\theta_1 - 1)^2 + (\theta_2 - 2) + \theta_1 \theta_2$ **teniendo en cuenta que** $p_k = \frac{1}{2k}$ **y** $\theta(1) = (-1, +1)$ (transparencia 33).

Solución:

En este caso, a partir de la fórmula del algoritmo general de descenso por gradiente, calculamos las derivadas parciales (en nuestro caso, con $k = 2$):

$$\frac{\partial q(\theta)}{\partial \theta_1} = 2(\theta_1 - 1) + \theta_2$$

$$\frac{\partial q(\theta)}{\partial \theta_2} = 2(\theta_2 - 2) + \theta_1$$

Empezamos por $\theta_1 = -1$ y $\theta_2 = 1$ y calcularemos cada componente del vector por separado, sabiendo que:

$$\theta(k+1) = \theta(k) - \frac{1}{2k} \left(2(\theta_1 - 1) + \theta_2 \right)$$

Iteración 1

$$\theta(2)_1 = (-1) - \frac{1}{2 \cdot 1} [2 \cdot (-1 - 1) + 1] = \frac{1}{2}$$

$$\theta(2)_2 = (+1) - \frac{1}{2 \cdot 1} [2 \cdot (1 - 2) - 1] = \frac{5}{2}$$

Iteración 2

$$\theta(3)_1 = \frac{1}{2} - \frac{1}{2 \cdot 2} \left[2 \cdot \left(\frac{1}{2} - 1 \right) + \frac{5}{2} \right] = \frac{1}{8}$$

$$\theta(3)_2 = \frac{5}{2} - \frac{1}{2 \cdot 2} \left[2 \cdot \left(\frac{5}{2} - 2 \right) + \frac{1}{2} \right] = \frac{17}{8}$$

Iteración 3

$$\theta(4)_1 = \frac{1}{8} - \frac{1}{2 \cdot 3} \left[2 \cdot \left(\frac{1}{8} - 1 \right) + \frac{17}{8} \right] = \frac{1}{16}$$

$$\theta(4)_2 = \frac{17}{8} - \frac{1}{2 \cdot 3} \left[2 \cdot \left(\frac{17}{8} - 2 \right) + \frac{1}{8} \right] = \frac{33}{16}$$

Ejercicio 3

Existe una variante de la función de Widrow-Hoff que incluye un término de regularización con el objetivo de que los pesos no se hagan demasiado grandes: $q_S(\theta) = \frac{1}{2} \sum_{n=1}^N (\theta^t x_n - y_n)^2 + \frac{\theta^t \theta}{2}$ Aplicando la técnica de descenso por gradiente, obtener la correspondiente variante del algoritmo de Widrow-Hoff y la correspondiente versión muestra a muestra (transparencia 33). Solución:

A partir del algoritmo general de descenso por gradiente, calculamos el gradiente:

$$\nabla q_S(\theta) = \frac{1}{2} \sum_{n=1}^N (\theta^t x_n - y_n)^2 + \frac{\theta^t \theta}{2} = \sum_{n=1}^N (\theta^t x_n - y_n) x_n + \theta$$

A continuación obtenemos la variante del algoritmo de Widrow-Hoff:

$$\theta(1) = \textit{arbitrario}$$

$$\theta(k+1) = \theta(k) - p_k \left(\sum_{n=1}^N (\theta(k)^t x_n - y_n) x_n + \theta(k) \right) = \theta(k) + p_k \sum_{n=1}^N (y_n - \theta(k)^t x_n) x_n - p_k \theta(k)$$

Por último, obtenemos la versión del algoritmo muestra a muestra:

$$\theta(1) = \textit{arbitrario}$$

$$\theta(k+1) = \theta(k) + p_k (y(k) - \theta(k)^t x(k)) x(k) - p_k \theta(k)$$