NOMBRE: Beltrán Domínguez, Victoria
GRUPO: 4CO11

APRENDIZAJE AUTOMÁTICO

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, $1 \le c \le C$, es $\hat{p}_c = n_c/N$ donde $N=n_1+...+n_C$ es el número total de datos observados y n_c es el número de datos de la clase c. (ver el último ejemplo de aplicación de la técnica de los multiplicadores de Langrange, transparencias 3.17 y 3.18)**

A partir del problema que se nos plantea, podemos definir los datos sobre los que vamos a partir:

- Existe un número C de clases diferentes.
- Un conjunto de muestras donde cada muestra pertenece a una clase c, tal que 1<=c<=C.
- N representa el número total de muestras
- n_c representa el número total de muestras de la clase c.

Así, con los siguientes datos, el criterio de máxima verosimilitud sería:

$$\theta^* = arg_\theta \max L_s(\theta) = \log P(S/\theta) = n_1 * \log p_1 + n_2 * \log p_2 + \ldots + n_c \log p_c$$

Aplicamos ahora la técnica de los multiplicadores de Lagrange. Para ello, seguiremos los pasos descritos a continuación:

1. Definimos los multiplicadores de Lagrange y Lagrangiana:

$$\Lambda(p_1, p_2, \ldots, p_C, \beta)_{p_1+p_2+\cdots+p_c=1} = n_1 * \log p_1 + n_2 * \log p_2 + \ldots + n_c \log p_c + \beta\,(1 - p_1 - p_2 - \ldots - p_c)$$

2. Obtener minimizador θ^* de la Lagrangiana en función de β:

$$\frac{\partial \Lambda}{\partial p_1} = \frac{n_1}{p_1} - \beta = 0 \;\rightarrow\; p_1^*(\beta) = \frac{n_1}{\beta}$$

$$\frac{\partial \Lambda}{\partial p_2} = \frac{n_2}{p_2} - \beta = 0 \;\rightarrow\; p_2^*(\beta) = \frac{n_2}{\beta}$$

…

$$\frac{\partial \Lambda}{\partial p_c} = \frac{n_c}{p_c} - \beta = 0 \;\rightarrow\; p_c^*(\beta) = \frac{n_c}{\beta}$$

3. Obtener función dual de Lagrange:

$\Lambda_D(\beta) = n_1 * \log \frac{n_1}{\beta} + n_2 * \log \frac{n_2}{\beta} + \ldots + n_c \log \frac{n_c}{\beta} + \beta\,(1 - \frac{n_1}{\beta} - \frac{n_2}{\beta} - \ldots - \frac{n_c}{\beta})$

4. Optimizar $\Lambda D(\beta)$:

$\frac{\partial \Lambda_D}{\partial \beta} = -\frac{n_1}{\beta} - \frac{n_2}{\beta} - \ldots - \frac{n_c}{\beta} + 1 = 0 \rightarrow -n_1 - n_2 - \cdots - n_c = \beta \rightarrow \beta = n_1 + n_2 + \cdots + n_c = N$

5. Solución final:

$$p_1^* = p_1^*(\beta) = \frac{n_1}{N}$$

$$p_2^* = p_2^*(\beta) = \frac{n_2}{N}$$

$$\ldots$$

$$p_C^* = p_C^*(\beta) = \frac{n_C}{N}$$

Como podemos observar, los valores calculados son los esperados y queda demostrado que la estimación de máxima verosimilitud de la probabilidad a priori de cada clase c, $1 \leq c \leq C$, es $p_c = n_c/N$.

2. **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:**

$$q_S(\boldsymbol{\theta}) = \frac{1}{2} \sum_{n=1}^{N} \left(\boldsymbol{\theta}^t \boldsymbol{x}_n - y_n \right)^2 + \frac{\boldsymbol{\theta}^t \boldsymbol{\theta}}{2}$$

Aplicando la técnica de descenso por gradiente, obtener la correspondiente variante del algoritmo de Widrow-Hoff y la correspondiente versión muestra a muestra.

Dada la función $q_s(\theta)$, aplicamos la técnica de descenso por gradiente. Para ello, vamos a derivar $q_s(\theta)$ respecto a θ.
Por simplificación, descomponemos la derivada en dos partes:
La primera parte, la parte izquierda de la suma se deriva como se muestra a continuación:

$$Vq_s'(\theta) = V\frac{1}{2} \sum_{n=1}^{N} (\theta^t x_n - y_n)^2 = \sum_{n=1}^{N} (\theta^t x_n - y_n) x_n$$

La derivada de la segunda parte, quedaría como sigue:

$$V\frac{\theta^t \theta}{2}$$

Para resolver esta derivada, resolvemos primero solo la derivada del numerador de la fracción:

$\theta^t\theta$ = θ_1^2+ θ_2^2 + … + θ_n^2 (siendo n el número de dimensiones de θ)

$\frac{\partial\theta^t\theta}{\partial\theta_1} = 2\theta_1$ $\qquad$ $\frac{\partial\theta^t\theta}{\partial\theta_2} = 2\theta_2$ $\quad \dots \quad$ $\frac{\partial\theta^t\theta}{\partial\theta_n} = 2\theta_n$

Con este razonamiento, llegamos a la conclusión de que $V\theta^t\theta = 2\theta$

Y por lo tanto: $V\,\frac{\theta^t\theta}{2} = \theta$

Sintetizando, la derivada resultante de la función general es:

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

Así pues, esto resulta en una variante del algoritmo de Windrow-Hoff donde:

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

Además, la versión muestra a muestra es equivalente a:

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