Técnicas de optimización, T-3

Vladyslav Mazurkevych

---

[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≤c≤C, es $p_c = \frac{n_c}{N}$ donde $N = n_1 + \ldots + n_C$ 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 Lagrange, transparencias 3.17 y 3.18) .

Empezaremos con la verosimilitud la cual la podemos demostrar como sigue con probabilidades para cada clase:

$$P(S|\theta) = \prod_{i=1}^{n_1} p_1 \cdot \prod_{j=1}^{n_2} p_2 \ldots \prod_{k=1}^{n_c} p_c = p_c^{n_1} \cdot \ldots \cdot p_c^{n_c} \qquad (1)$$

Pasando estos productorios en forma logarítmica y añadiendoles un algoritmo a la probabilidad, obtendremos el logaritmo de la verosimilitud, esto nos será útil para eliminar el posible *vanishing gradient* de los productorios, por lo tanto esta ecuación se reescribiría como la que sigue, recuérdese que la suma de todas las probabilidades es 1:

$$q_s(\theta) = L_s(\theta) = log(P(S|\theta)) = n_1 \cdot log(p_1) + \ldots + n_c \cdot log(p_c) \quad (2)$$

Ahora haciendo la máxima verosimilitud la fórmula nos quedaría como la que sigue:

$$\theta^* = \underset{\substack{p_1 \ldots p_c \\ p_1 + p_2 + \ldots + p_c = 1}}{\operatorname{argmax}} \left(n_1 \cdot \log(p_1) + \ldots + n_c \cdot \log(p_c) \right) \qquad (3)$$

Pasando la ecuación 3 a la Lagraniana obtendremos:

$$\Lambda(p_1, p_2, \ldots, p_c, \beta) = n_1 \cdot \log(p_1) + \ldots + n_c \cdot \log(p_c) + \beta \qquad (4)$$

Así pues nuestras soluciones óptimas en función del multiplicador de Lagrange nos quedaría:

$$\begin{array}{lcl} \frac{\partial \Lambda}{\partial p_1} = \frac{n_1}{p_1} - \beta = 0 & & p_1(\beta) = \frac{n_1}{\beta} \\ \ldots & \Rightarrow & \\ \frac{\partial \Lambda}{\partial p_c} = \frac{n_c}{p_c} - \beta = 0 & & p_c(\beta) = \frac{n_c}{\beta} \end{array} \qquad (5)$$

Por lo tanto la función de Lagrange quedaría como:

$$\begin{aligned} \Lambda_D(\beta) &= n_1 \cdot \log\left(\tfrac{n_1}{\beta}\right) + \ldots + n_c \cdot \log\left(\tfrac{n_c}{\beta}\right) + \beta \left(1 - \tfrac{n_1}{\beta} - \ldots - \tfrac{n_c}{\beta}\right) \\ &= \beta - N \cdot \log(\beta) - N + \sum_{c=1}^{C} n_c \log(n_c) \end{aligned} \qquad (6)$$

Quedando así el valor óptimo del multiplicador de Lagrange:

$$\frac{\partial \Lambda_D}{\partial \beta} = 1 - \frac{N}{\beta} = 0 \Rightarrow \beta^* = N \qquad (7)$$

Así pues la solución final nos quedaría como la siguiente:

$$p_1^* = p_1^*(\beta) = \frac{n_1}{N} \; ; \; p_2^* = p_2^*(\beta) = \frac{n_2}{N} \; ; \ldots ; \; p_c^* = p_c^*(\beta) = \frac{n_c}{N} \qquad (8)$$

Así pues quedaría demostrada dicha ecuación para cualquier problema de clasificación.

[2]: Aplicar la técnica de descenso por gradiente a la búsqueda de mínimo de la función: $q(\theta)=(\theta_1-1)^2+(\theta_2-2)^2+\theta_1\theta_2$ teniendo en cuenta que $\rho_k = \frac{1}{2k}$ y $\theta(1)=(-1, +1)$ y hacer una traza de las 3 primeras iteraciones.

La ecuación general para el cálculo del gradiente para los parámetros θ viene dada a continuación.

$$\theta(k+1) = \theta(k) - \rho_1 \nabla q(\theta) \qquad (9)$$

Pero como tenemos un espacio vectorial, en este caso de dos dimensiones, tendremos que aplicar el descenso del gradiente por cada dimensión de nuestro vector. Así pues vamos a sacar las derivadas parciales por cada componente θ que tengamos, con lo que nos quedaría la ecuación (2).

$$\nabla q(\theta) = \left(\frac{\partial q(\theta)}{\partial \theta_1}, \frac{\partial q(\theta)}{\partial \theta_2}\right)^t \qquad (10)$$

Resolviendo las derivadas parciales para la ecuación $q(\theta) = (\theta_1 - 1)^2 + (\theta_2 - 2)^2 + \theta_1 \theta_2$ obtenemos las siguientes funciones que usaremos para calcular el gradiente para cada componente del vector.

$$\frac{\partial q(\theta)}{\partial \theta_1} = 2(\theta_1 - 1) + \theta_2 \qquad (11)$$
$$\frac{\partial q(\theta)}{\partial \theta_2} = 2(\theta_2 - 2) + \theta_1 \qquad (12)$$

Ahora procederemos a hacer las iteraciones calculando el gradiente para cada componente del vector por separado, sabemos que empezamos con $\theta_1 = -1$ y $\theta_2 = 1$, así que simplemente tenemos que sustituir los valores en su ecuación correspondiente e ir iterando obteniendo nuevos valores de θ cada vez. Empezaremos por la iteración k=1, así pues $\rho_k = \frac{1}{2k} = \frac{1}{2*1}$, usaremos la ecuación (1) para cada una de las componentes.

Iteración 1:

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

$\theta(2)_2 = (1) - \frac{1}{2 \cdot 1}[2 \cdot (1 - 2) - 1] = 1 - \frac{1}{2} \cdot (-3) = \frac{2}{2} + \frac{3}{2} = \frac{5}{2}$
Así pues tras la primera iteración, los valores de Thetas quedan: $\theta(2)_1 = \frac{1}{2}$ y $\theta(2)_2 = \frac{5}{2}$

Iteración 2:
$\theta(3)_1 = (\frac{1}{2}) - \frac{1}{2 \cdot 2}[2 \cdot (\frac{1}{2} - 1) + \frac{5}{2}] = \frac{1}{2} - \frac{3}{8} = \frac{1}{8}$
$\theta(3)_2 = (\frac{5}{2}) - \frac{1}{2 \cdot 2}[2 \cdot (\frac{5}{2} - 2) + \frac{1}{2}] = \frac{5}{2} - \frac{3}{8} = \frac{17}{8}$
Así pues tras la segunda iteración, los valores de Thetas quedan: $\theta(3)_1 = \frac{1}{8}$ y $\theta(3)_2 = \frac{17}{8}$

Iteración 3:
$\theta(4)_1 = (\frac{1}{8}) - \frac{1}{2 \cdot 3}[2 \cdot (\frac{1}{8} - 1) + \frac{17}{8}] = \frac{1}{8} - \frac{1}{16} = \frac{1}{16}$
$\theta(4)_2 = (\frac{17}{8}) - \frac{1}{2 \cdot 3}[2 \cdot (\frac{17}{8} - 2) + \frac{1}{8}] = \frac{17}{8} - \frac{1}{16} = \frac{33}{16}$
Así pues tras la tercera iteración, los valores de Thetas quedan: $\theta(4)_1 = \frac{1}{16}$ y $\theta(4)_2 = \frac{33}{16}$
Estas serían las 3 primeras iteraciones del descenso del gradiente, obsérvese cómo fue cambiando la ρ_k.

[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:

$$q_S(\theta) = \frac{1}{2} \sum_{n=1}^{N} (\theta^t x_n - y_n)^2 + \frac{\theta^t \theta}{2} \quad (13)$$

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

La ecuación general para el cálculo del gradiente para los parámetros θ viene dada a continuación.

$$\theta(k+1) = \theta(k) - \rho_1 \nabla q(\theta) \quad (14)$$

Por lo tanto, vamos a hacer la derivada parcial de nuestra ecuación dada, que viene en base de un sumatorio:

$$q_S(\theta) = \frac{1}{2} \sum_{n=1}^{N} (\theta^t x_n - y_n)^2 + \frac{\theta^t \theta}{2} \quad (15)$$

Vamos a expandir la variable θ para ver cada componente por separado, de esta manera la ecuación (7) sobreescrita quedaría como:

$$q_S(\theta) = \frac{1}{2} \sum_{n=1}^{N} (\left[\theta_1, \theta_2, \ldots, \theta_N\right]^t x_n - y_n)^2 + \frac{\left[\theta_1, \theta_2, \ldots, \theta_N\right] \cdot \left[\theta_1, \theta_2, \ldots, \theta_N\right]^t}{2} \quad (16)$$

A continuación pasaremos a derivar la ecuación desplegada, obsérvese que vamos a derivar respecto a las diferentes θ, así pues vamos tener derivadas parciales en el rango de θ_1, θ_2 ,.., θ_N .Con lo que las derivadas parciales nos quedarían de siguiente manera:

$$\nabla q_S(\theta) = \frac{1}{2} \sum_{n=1}^{N} 2(\theta_n x_n - y_n) \cdot x_n + \frac{2\theta_n^{2-1}}{2} \qquad (17)$$

Simplificando:

$$\nabla q_S(\theta) = \frac{1}{2} \sum_{n=1}^{N} 2(\theta_n x_n - y_n) \cdot x_n + \theta_n \qquad (18)$$

Ahora teniendo la ecuación la derivada parcial implementada, podemos proceder a implementar el descenso del gradiente para completar la función de Widrow-Hoff, la cual nos quedaría de siguiente manera:

$$\theta(1) = (\text{Valores iniciales para } \theta)$$
$$\theta(k+1) = \theta(k) - \rho_k \left[\frac{1}{2} \sum_{n=1}^{N} 2(\theta_{k,n} x_n - y_n) \cdot x_n + \theta_{k,n} \right] \qquad (19)$$

Ahora tendríamos una ecuación genérica respecto a θ_k, que sería nuestra primera solución al ejercicio dado, que sería despejar el gradiente de la misma. Por último vamos a proceder a dar una ecuación de la variante muestra a muestra, vamos a transcribir la función respecto a la muestra *k*:

$$\theta(1) = (\text{Valores iniciales para } \theta)$$
$$\theta(k+1) = \theta(k) - \rho_k \left[\frac{2}{2}(\theta_k x_k - y_k) \cdot x_n + \theta_k \right] \qquad (20)$$

Por lo tanto, este sería la variante del algoritmo muestra a muestra, con lo que podríamos simplificarlo un poco más y la solución final a nuestro ejercicio quedaría de la forma que sigue:

$$\theta(1) = (\text{Valores iniciales para } \theta)$$
$$\theta(k+1) = \theta(k) - \rho_k \left[(\theta_k x_k - y_k) \cdot x_n + \theta_k \right] \qquad (21)$$