Regresión Logística

Notas 5^{ta} reunión Club de Ciencia de Datos La Paz

23/06/18

1 Familias exponenciales

Sea $\phi(x)$ un mapeo del vector aleatorio x a un espacio con los estadíticos suficientes.
Tenemos que la fórmula general de las densidades de la familia exponencial.

$$p(x;\theta) = h(x)e^{<\phi(x),\theta>-g(\theta)} \tag{1}$$

dónde:

$h(x)$ es una densidad definida en x.

θ es el vector de parámetros.

$< A, B >$ denota producto matricial entre dos matrices A y B

Definición 1.1 (Función de partición)

$$g(\theta) = \log \int_x h(x)e^{<\phi(x),\theta>}dx \tag{2}$$

1.1 Propiedades de la función de partición

Cuenta con propiedades demasiado útiles a la hora de computar y optimizar los parámetros θ.

1.1.1 Convexidad

Teorema 1.1 El espacio natural de parámetros θ es un conjunto convexo, y la función de partición cumulante $g(\theta)$ es una función convexa. Si la familia es mínima, la función es estríctamente convexa.

Demostración
Consideramos los parámetros θ_1 y $\theta_2 \in \theta$
además, partimos de la hipótesis de que el vector de parámetros θ es un conjunto convexo, $\theta_\lambda = \lambda\theta_1 + (1-\lambda)\theta_2$ ($\lambda \in [0,1]$).
tenemos:

$$\begin{aligned} e^{g(\theta_\lambda)} &= \int_x e^{<\phi(x),\theta_\lambda>}h(x)dx \\ &= \int_x e^{<\phi(x),\lambda\theta_1+(1-\lambda)\theta_2>}h(x)dx \\ &= \int_x e^{<\phi(x),\lambda\theta_1>}e^{<\phi(x),(1-\lambda)\theta_2>}h(x)dx \end{aligned}$$

Aplicando la inecuación de Hölder que establece:

$$\left|\sum_n a_n b_n\right| \leq \left(\sum_k |a_k|^p\right)^{\frac{1}{p}} \left(\sum_k |b_k|^q\right)^{\frac{1}{q}}$$

$$e^{g(\theta_\lambda)} \leq \left(\int_x \left(e^{<\phi(x),\theta_1>} h(x)\right)^{\frac{\lambda}{\lambda}} dx\right)^{\lambda} \left(\int_x \left(e^{<\phi(x),\theta_2>} h(x)\right)^{\frac{(1-\lambda)}{(1-\lambda)}} dx\right)^{(1-\lambda)}$$

$$e^{g(\theta_\lambda)} \leq \left(\int_x \left(e^{<\phi(x),\theta_1>} h(x)\right) dx\right)^{\lambda} \left(\int_x \left(e^{<\phi(x),\theta_2>} h(x)\right) dx\right)^{(1-\lambda)}$$

Aplicando logaritmos en ambas partes:

$$g(\theta_\lambda) \leq \lambda \log \int_x e^{<\phi(x),\theta_1>} h(x)dx + (1-\lambda) \log \int_x e^{<\phi(x),\theta_2>} h(x)dx$$

$$g(\lambda\theta_1 + (1-\lambda)\theta_2) \leq \lambda g(\theta_1) + (1-\lambda) g(\theta_2)$$

Llegando a cumplir finalmente la inecuación de Jensen.

1.1.2 El gradiente permite obtener la esperanza de la distribución $\nabla_\theta g(\theta) = E[\phi(x)]$

Demostración
sea $z(\phi(x);\theta) = \int_x e^{<\phi(x),\theta>} dx$ y $A(\theta) = \log z(\phi(x);\theta) \rightarrow z(\phi(x);\theta) = e^{A(\theta)}$

$$\begin{aligned}
g(\theta) &= \log z(\phi(x);\theta) \\
\frac{\partial g(\theta)}{\partial \theta} &= \frac{1}{z(\phi(x);\theta)} \frac{\partial z(\phi(x);\theta)}{\partial \theta} \\
&= \frac{\frac{\partial}{\partial \theta} \int_x e^{<\phi(x),\theta>} dx}{\int_x e^{<\phi(x),\theta>} dx} \\
&= \frac{\int_x \phi(x) e^{<\phi(x),\theta>} dx}{\int_x e^{<\phi(x),\theta>} dx} \\
&= \frac{\int_x \phi(x) e^{<\phi(x),\theta>} dx}{e^{A(\theta)}} \\
&= \int_x \phi(x) e^{<\phi(x),\theta> - A(\theta)} dx \\
&= \int_x \phi(x) p(x;\theta) dx \\
&= E[\phi(x)]
\end{aligned}$$

1.1.3 La Hessiana permite obtener la varianza de la distribución $\nabla^2_\theta g(\theta) = Var[\phi(x)]$

Demostración

$$\begin{aligned}
\frac{\partial^2 g(\theta)}{\partial \theta^2} &= \frac{\partial}{\partial \theta} \int_x \phi(x) h(x) e^{<\phi(x),\theta> - A(\theta)} dx \\
&= \int_x \phi(x) \left[\phi(x) - \frac{\partial A(\theta)}{\partial \theta} \right] e^{<\phi(x),\theta> - A(\theta)} h(x) dx \\
&= \int_x \phi(x) \left[\phi(x) - \nabla_\theta g(\theta) \right] p(x;\theta) dx \\
&= \int_x \phi(x)^2 p(x;\theta) dx - E[\phi(x)] \int_x \phi(x) p(x;\theta) dx \\
&= E[\phi(x)^2] - E[\phi(x)]^2 \\
&= Var[\phi(x)]
\end{aligned}$$

1.2 Distribución de Bernoulli

Esta distibución tiene la forma:

$$p(x) = p^x (1-p)^{(1-x)} \tag{3}$$

La idea es llevar esta distribución a su forma de familia exponencial, para identificar los términos que la componen y sobre todo su función de partición.

$$\begin{aligned}
p(x) &= e^{\log\left|p^x (1-p)^{1-x}\right|} \\
&= e^{x \log|p| + (1-x)\log|1-p|} \\
&= e^{x \log|p| + \log|1-p| - x\log|1-p|} \\
&= e^{\log\left|\frac{p}{1-p}\right| x + \log|1-p|} \\
&= e^{\log\left|\frac{p}{1-p}\right| x + [\log|1| - \log|1-p|]} \\
&= e^{\log\left|\frac{p}{1-p}\right| x - \log\left|\frac{1}{1-p}\right|} \\
&= e^{\log\left|\frac{p}{1-p}\right| x - \log\left|\frac{1-p}{1-p} + \frac{p}{1-p}\right|} \\
&= e^{\log\left|\frac{p}{1-p}\right| x - \log\left|1 + e^{\log\left|\frac{p}{1-p}\right|}\right|} \\
&= e^{zx - \log\left|1+e^{\theta}\right|} \\
p(x) &= e^{zx - g(z)}
\end{aligned} \tag{4}$$

El θ de la forma original lo denotaremos como z para evitar futuras confusiones.

De donde podemos identificar:

$$z = \log\left|\frac{p}{1-p}\right|$$

$$\phi(x) = x$$

$$g(z) = \log\left|1 + e^{\theta}\right|$$

$$h(x) = 1$$

Si resolvemos para p:

$$\begin{aligned}
z &= \log\left|\frac{p}{1-p}\right| \\
e^z &= \frac{p}{1-p} \\
e^z - pe^z &= p \\
e^z &= p + pe^z \\
e^z &= p(1+e^z) \\
p &= \frac{e^z}{1+e^z} * \frac{\frac{1}{e^z}}{\frac{1}{e^z}} \\
p &= \frac{1}{\frac{1}{e^z} + \frac{e^z}{e^z}} \\
p &= \frac{1}{1+e^{-z}} \qquad (5)
\end{aligned}$$

Obtenemos así la función sigmoide o logística.
Por notación, definimos $z = \theta^T x$

1.2.1 Esperanza

$$\begin{aligned}
\frac{\partial g(z)}{\partial z} &= \frac{1}{1+e^z} * e^z \\
&= \frac{1}{1+e^{-z}} \\
&= \mu \\
&= \sigma(z) \\
E[x] &= \sigma(\theta^T x) \qquad (6)
\end{aligned}$$

1.2.2 Varianza

$$\begin{aligned}\frac{\partial^2 g(z)}{\partial z^2} &= \frac{\partial \mu}{\partial z} \\ &= \frac{e^{-z}}{(1+e^{-z})^2} \\ &= \frac{1}{1+e^{-z}} * \frac{e^{-z}}{1+e^{-z}} \\ &= \left(1 - \frac{1}{1+e^{-z}}\right) * \frac{1}{1+e^{-z}} \\ &= (1-\mu) * \mu \\ &= \sigma(z) * (1-\sigma(z))\end{aligned}$$

$$Var[x] = \sigma(\theta^T x) * (1 - \sigma(\theta^T x)) \tag{7}$$

2 Negative log likehood

Siempre que tenemos una función $p(x;\theta)$ parametrizada por algún parámetro θ, podemos usar datos para encontrar el $\hat{\theta}$ que maximice la probabilidad de que esos datos hayan sido generados por una distribución con éste parámetros.
Definición: Dado un modelo $p(x;\theta)$ parametrizado por θ, el estimador de máxima verosimilitud es:

$$\hat{\theta}[\phi(x)] = \arg\max_x \ p(\phi(x), \theta)$$

dónde:

$$p(\phi(x);\theta) = \prod_{i=1}^{m} p(\phi(x),\theta) = \prod_{i=1}^{m} h(x) e^{<\phi(x),\theta> - g(\theta)}$$

para la familia exponencial.
Desde la perspectiva computacional, es difícil calcular una serie de multiplicaciones con valores decimales pequeños, por lo cual se utiliza una alternativa, que es sacar su logaritmo, ya que maximizar el logaritmo de una función, es equivalente a maximizar la función en sí.

$$\log|p(\phi(x);\theta)| = \sum_{i=1}^{m} \log\left|h(x) e^{<\phi(x),\theta> - g(\theta)}\right|$$

Ya que los algoritmos de optimización a utilizar están más que nada pensados para minimizar funciones, debemos llevar nuestra ecuación a un problema de minimización. Mediande la equivalencia de que maximizar un número positivo, es igual a minimizar un número negativo, convertimos el problema en:

$$\hat{\theta}[\phi(x)] = \arg\min_x - \log|p(\phi(x),\theta)|$$

3 Regresión logística

Es un clasificador probabilístico, en el que ajustamos un modelo de la forma $p(y/x;\theta)$. Dado que es un clasificador binario, definimos que la variable $y_i \sim Ber(\hat{y}_i)$

$$p(y/x;\theta) = Ber(y/\mu(x))$$

Modelamos z como una combinación lineal de los parámetros θ y los atributos x.

$$\log\left|\frac{\hat{y}_i}{1-\hat{y}_i}\right| = \theta_0 + \theta_1 x_1 + ... + \theta_n x_n$$

Sabemos por (5), que si resolvemos, tendremos.

$$p[y_i = 1] = p_i = \frac{1}{1 + e^{\theta_0+\theta_1 x_1+...+\theta_n x_n}}$$

$$p_i = \frac{1}{1+e^{-z}} = \sigma(z)$$

Entonces, el modelo se convierte en:

$$p(y/x;\theta) = Ber(y/\sigma(\theta^T x)) \tag{8}$$

3.1 Estimador de máxima verosimilitud caso binario

Con el objetivo de encontrar los valores óptimos de θ, debemos obtener primero su negative log likehood.
Dado que nuestro $p(y/x;\theta)$ es una distribución de Bernoulli, entonces por notacion definimos $p = \hat{y} = \sigma(\theta^T x)$, que si sustituimos.

$$\hat{\theta} = \arg\max_x \ p(y/x;\theta)$$

$$= \prod_{i=1}^{m} [\hat{y_i}^{[y_i=1]}(1-\hat{y_i})^{[y_i=0]}]$$

$$= \prod_{i=1}^{m} [\hat{y_i}^{y_i}(1-\hat{y_i})^{1-y_i}]$$

Si lo convertimos a negative log likehood.

$$\hat{\theta} = -\log \prod_{i=1}^{m} [\hat{y_i}^{y_i}(1-\hat{y_i})^{1-y_i}]$$

Definimos nuestro Loss aplicando propiedades de logarítmos.

$$\mathscr{L} = -\sum_{i=1}^{m} [y_i \log|\hat{y_i}| + (1-y_i)\log|1-\hat{y_i}|] \tag{9}$$

Obtenemos así la llamada Cross Entropy Loss.

Finalmente para obtener el costo total que usaremos para optimizar nuestra función, mediamos diviendo entre m.

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} [y_i \log|\hat{y_i}| + (1-y_i)\log|1-\hat{y_i}|] \tag{10}$$

3.1.1 Gradiente

Reescribimos

1er término

$$\begin{aligned}\log|\hat{y}_i| &= \log\left|\frac{1}{1+e^{-z}}\right| \\ &= -\log\left|1+e^{-z}\right|\end{aligned}$$

2do término

$$\begin{aligned}\log|1-\hat{y}_i| &= \log\left|1-\frac{1}{1+e^{-z}}\right| \\ &= \log\left|e^{-z}\right| - \log\left|1+e^{-z}\right| \\ &= -z - \log\left|1+e^{-z}\right|\end{aligned}$$

Sustituyendo en $J(\theta)$

$$\begin{aligned}J(\theta) &= -\frac{1}{m}\sum_{i=1}^{m}\left[y_i\left(-\log\left|1+e^{-z}\right|\right)+(1-y_i)\left(-z-\log\left|1+e^{-z}\right|\right)\right] \\ &= -\frac{1}{m}\sum_{i=1}^{m}\left[\cancel{-y_i\log\left|1+e^{-z}\right|} - z - \log\left|1+e^{-z}\right| + zy_i + \cancel{y_i\log\left|1+e^{-z}\right|}\right] \\ &= -\frac{1}{m}\sum_{i=1}^{m}\left[zy_i - z - \log\left|1+e^{-z}\right|\right] \\ &= -\frac{1}{m}\sum_{i=1}^{m}\left[zy_i - \left(\log e^{z} + \log\left|1+e^{-z}\right|\right)\right] \\ &= -\frac{1}{m}\sum_{i=1}^{m}\left[zy_i - \log\left|e^{z}\left(1+e^{-z}\right)\right|\right] \\ &= -\frac{1}{m}\sum_{i=1}^{m}\left[zy_i - \log\left|1+e^{z}\right|\right]\end{aligned}$$

Ahora debemos calcular la derivada.

1er término

$$\begin{aligned}\frac{\partial}{\partial\theta_j}y_i\theta^T x_i &= y_i\frac{\partial}{\partial\theta_j}\sum_{j0}^{n}\theta_j x_i^j \\ &= y_i x_i^j\end{aligned}$$

2do término

$$\begin{aligned}\frac{\partial}{\partial\theta_j}\log\left|1+e^{\theta^T x_i}\right| &= \frac{\frac{\partial}{\partial\theta_j}\left(1+e^{\theta^T x_i}\right)}{1+e^{\theta^T x_i}}\\ &= \frac{\frac{\partial}{\partial\theta_j}\left(\theta^T x_i\right)e^{\theta^T x_i}}{1+e^{\theta^T x_i}}\\ &= \frac{x_i e^{\theta^T x_i}}{1+e^{\theta^T x_i}}\\ &= x_i^j\hat{y}_i\end{aligned}$$

donde:

i representa el i-ésima observación (cada observación es un vector).

j represnta el j-ésimo atributo del i-ésimo vector.

$$\begin{aligned}\frac{\partial J(\theta)}{\partial\theta_j} &= -\sum_{i=1}^{m}\left[y_i x_i^j - \hat{y}_i x_i^j\right]\\ &= \sum_{i=1}^{m}\left[\hat{y}_i - y_i\right]x_i^j\end{aligned}$$

3.1.2 Hessiana

$$\begin{aligned}\frac{\partial^2 J(\theta)}{\partial\theta_j^2} &= \sum_{i=1}^{m}\left[x_i^j\frac{\partial}{\partial\theta}\hat{y}_i - \frac{\partial}{\partial\theta}x_i^j y_i\right]\\ &= \sum_{i=1}^{m}\left[x_i^j\frac{\partial}{\partial\theta}\sigma(\theta^T x_i) - \cancelto{0}{\frac{\partial}{\partial\theta}x_i^j y_i}\right] && \text{Por } \hat{y}_i = \sigma(\theta^T x_i)\\ &= \sum_{i=1}^{m}\left[x_i^j * x_i^j\left(\sigma(\theta^T x_i) * (1-\sigma(\theta^T x_i))\right)\right] && \text{Por (7) y regla de la cadena para } \theta_j\\ &= \sum_{i=1}^{m}\left[(\hat{y}_i * (1-\hat{y}_i))\right]x_i^j * x_i^j\end{aligned}$$

3.1.3 Formas matriciales de las derivadas

$$\nabla_\theta J(\theta) = \sum_{i=1}^{m}\left[\hat{y}_i - y_i\right]x_i \tag{11}$$

$$\nabla_\theta^2 J(\theta) = \sum_{i=1}^{m}\left[(\hat{y}_i * (1-\hat{y}_i))\right]x_i^T x_i \tag{12}$$

3.2 Distribución Multinomial (caso multiclase)

Tiene la forma

$$p(x_1, x_2, ..., x_k) = \frac{n!}{x_1!x_2!...x_k!} * p_1^{x_1} p_2^{x_2} ... p_k^{x_k} \tag{13}$$

la cual ahora tiene K resultados posibles.

Podemos reescribirla como:

$$\begin{aligned} p(x_1, x_2, ..., x_k) &= \frac{n!}{x_1!x_2!...x_c!} * \prod_{k=1}^{C} p_k^{x_k} \\ &= \frac{n!}{x_1!x_2!...x_c!} * exp\left\{\sum_{k=1}^{C} x_k \log p_k\right\} \\ &= h(x) * exp\left\{\sum_{k=1}^{C} x_k \log p_k\right\} \end{aligned}$$

de donde podemos observar que:

C simboliza el número de clases en vez de K, para evitar confusiones.

$h(x) = \frac{n!}{x_1!x_2!...x_c!}$

Ahora procedemos a llevar la distribución a su forma de familia exponencial.
Para esto, utilizaremos la igualdad $x_c = 1 - \sum_{k=1}^{C-1} x_k$.

$$p(x_1, x_2, ..., x_k) = h(x) * exp\left\{\sum_{k=1}^{C} x_k \log p_k\right\}$$

$$\begin{aligned} \frac{\log |p(x_1, x_2, ..., x_k)|}{h(x)} &= \sum_{k=1}^{C} x_k \log |p_k| \\ &= \sum_{k=1}^{C-1} [x_k \log |p_k|] + x_c \log |p_c| \\ &= \sum_{k=1}^{C-1} [x_k \log |p_k|] + \left(1 - \sum_{k=1}^{C-1} x_k\right) \log |p_c| \\ &= \sum_{k=1}^{C-1} [x_k \log |p_k|] - \sum_{k=1}^{C-1} [x_k \log |p_c|] + \log |p_c| \\ &= \sum_{k=1}^{C-1} [x_k \log |p_k| - x_k \log |p_c|] + \log |p_c| \end{aligned}$$

$$= \sum_{k=1}^{C-1} \left[\log \left| \frac{p_k}{p_c} \right| x_k \right] + \log |p_c|$$

$$p(x_1, x_2, ..., x_k) = h(x) e^{\sum_{k=1}^{C-1} \left[\log \left| \frac{p_k}{p_c} \right| x_k \right] + \log |p_c|}$$

El θ de la forma original lo denotaremos como z para evitar futuras confusiones.

De donde podemos identificar:

$z_k = \log \left| \frac{p_k}{p_c} \right|$

$\phi(x) = x_k$

$g(z_k) = -\log |p_c|$

Si resolvemos para p_c:

$$\begin{aligned}
z_k &= \log \left| \frac{p_k}{p_c} \right| \\
e^{z_k} &= \frac{p_k}{p_c} \\
p_k &= p_c e^{z_k} && //\sum_{k=1}^{C-1} \\
\sum_{k=1}^{C-1} p_k &= p_c \sum_{k=1}^{C-1} e^{z_k} \\
1 - p_c &= p_c \sum_{k=1}^{C-1} e^{z_k} \\
1 &= p_c + p_c \sum_{k=1}^{C-1} e^{z_k} \\
p_c &= \frac{1}{1 + \sum_{k=1}^{C-1} e^{z_k}} \\
p_c &= \frac{1}{\sum_{k=1}^{C} e^{z_k}} && \text{Por equivalencia } \sum_{k=1}^{C} e^{z_k} = 1 + \sum_{k=1}^{C-1} e^{z_k}
\end{aligned}$$

Si:

$$\begin{aligned}
g(z) &= -\log |p_c| \\
&= -\log \left| \frac{1}{\sum_{k=1}^{C} e^{z_k}} \right| \\
&= \log \left| \sum_{k=1}^{C} e^{z_k} \right|
\end{aligned}$$

Derivando:

$$\begin{aligned}\nabla_{\theta_k} g(z) &= \frac{1}{\sum_{k=1}^{C} e^{z_k}} * e^{z_k} \\ &= \frac{e^{z_k}}{\sum_{j=1}^{\mathcal{K}} e^{z_j}} \qquad \text{Re escribimos los índices } \mathcal{K}, \text{ y } j \\ &= \mu \\ &= softmax(z)\end{aligned}$$

$$E[x] = softmax(\theta^T x) = \frac{e^{z_k}}{\sum_{j=1}^{\mathcal{K}} e^{z_j}} \tag{14}$$

Obtenemos así la función Softmax o logit multinomial.

3.3 Softmax

La idea más simple de clasificación multiclase, es entrenar $\mathcal{K}$ regresiones logísticas binarias diferentes que discriminen entre la clase k contra todas las demás, para cada una de las posibles respuestas.

Por otra parte, al utilizar la Softmax podemos obtener un modelo generalizado a multiclases que nos da un mayor rendimiento desde el punto de vista computacional.

La función Softmax, permite "comprimir" un vector $\mathcal{K}$-dimensional, $\hat{y}$, a probabilidades, con la característica de que:

$$\sum_{k=1}^{\mathcal{K}} \hat{y_k} = 1$$

Podemos desarrollarlo de una manera más vistosa.

$$\begin{aligned}p(y = k/x; z^{(k)}) &= softmax(z^{(k)}) \\ &= \frac{e^{z^{(k)}}}{\sum_{j=1}^{\mathcal{K}} e^{z^{(k)}} x} \\ &= \frac{e^{\theta^{(k)^T} x}}{\sum_{j=1}^{\mathcal{K}} e^{\theta^{(j)^T}} x}\end{aligned}$$

Si desarrollamos.

$$\hat{y} = \begin{bmatrix} p(y = 1/x; z^{(k)}) \\ p(y = 2/x; z^{(k)}) \\ \vdots \\ p(y = \mathcal{K}/x; z^{(k)}) \end{bmatrix} = \frac{1}{\sum_{j=1}^{\mathcal{K}} e^{\theta^{(j)^T} x}} \begin{bmatrix} e^{\theta^{(1)^T}} x \\ e^{\theta^{(2)^T}} x \\ \vdots \\ e^{\theta^{(\mathcal{K})^T}} x \end{bmatrix}$$

3.3.1 Loss Function

Seguimos utilizando la entropía cruzada, pero con una generalización a multiclases.

$$\mathscr{L} = -\sum_{i=1}^{m}\sum_{k=1}^{\mathcal{K}}[y_i^k \log|\hat{y_i}^k|] \tag{15}$$

En base a esto, definimos el costo como:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{\mathcal{K}}[y_i^k \log|\hat{y_i}^k|] \tag{16}$$

3.3.2 Gradiente

$$\nabla_{\theta_k} J(\theta) = \sum_{i=1}^{m}\left[\hat{y_i}^k - y_i^k\right] x_i \tag{17}$$