Tarea 2
Regresión Logística

Sebastián Marroquín Martínez

Noviembre, 2018

1. ¿Qué es la Regresión Logística?

La regresión logística es en realidad un método de clasificación. Este tipo de método introduce una no linealidad adicional sobre un clasificador lineal, esto es: $f(x) = w^T x + b$, usando una función logística (o sigmoide), $\sigma()$.

El clasificador de la Regresión Logística se define a continuación:

$$\sigma(f(x_i)) = \begin{cases} \geq 0{,}5 y_i = +1 \\ < 0{,}5 y_i = -1 \end{cases}$$

Donde, $\sigma(f(x)) = \frac{1}{1+\exp^{-f(x)}}$.

1.1. Función de Costo

Si usamos el error cuadratico como una función de costo para la regresión logística $J(w)$, será no convexa.

Por lo que utilizaremos:

$$Cost(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) if y = 1 \\ -\log(1 - h_\theta(x)) if y = 0 \end{cases}$$

Juntando todo, tenemos que:

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} Cost(h_\theta(x^i), y^i) \tag{1}$$

$$= -\frac{1}{m} [\sum_{i=1}^{m} y^i \log(h_\theta(x^i) + (1 - y^i) \log(1 - h_\theta(x^i)))] \tag{2}$$

La regla de actualización del gradiente nos queda de la siguiente manera:

$$\theta_j := \theta_j - \alpha \sum_{i=1}^{m} (h_\theta(x^i) - y^i) \cdot x_j^i \tag{3}$$

2. Ejercicios - Tarea

2.1. Steepest Descent para la RL

Lo que necesitamos obtener para el Steepest Descent mediante derivadas es lo siguiente:

$$\begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \dots \\ \theta_n \end{pmatrix} = \big(\theta\big) - \frac{\alpha}{m} \cdot \begin{pmatrix} \frac{\partial J}{\partial \theta_0} \\ \\ \frac{\partial J}{\partial \theta_1} \\ \\ \frac{\partial J}{\partial \theta_2} \\ \\ \dots \\ \\ \frac{\partial J}{\partial \theta_n} \end{pmatrix}$$

Recordemos que la forma general del Descenso del Gradiente es la siguiente:

$$\begin{aligned} Repeat : \Big\{ \\ \theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta) \\ \Big\} \end{aligned}$$

Aplicando el método de las derivadas parciales, realizamos lo siguiente:

1. Encontrar la derivada de la función Sigmoide. La derivada de la función Sigmoide es la siguiente:

$$
\begin{aligned}
\frac{d}{dx}\sigma(x) &= \frac{d}{dx}\left(\frac{1}{1+e^{-x}}\right) \\
&= \frac{-(1+e^{-x})'}{(1+e^{-x})^2} \\
&= \frac{e^{-x}}{(1+e^{-x})^2} \\
&= \left(\frac{1}{1+e^{-x}}\right)\left(\frac{e^{-x}}{1+e^{-x}}\right) \\
&= \left(\frac{1}{1+e^{-x}}\right)\left(\frac{1+e^{-x}}{1+e^{-x}} - \frac{1}{1+e^{-x}}\right) \\
&= \sigma(x)\left(\frac{1+e^{-x}}{1+e^{-x}} - \sigma(x)\right) \\
&= \sigma(x)\left(1-\sigma(x)\right)
\end{aligned}
$$

2. Lo primero que debemos de hacer, es obtener la derivada de la función de costo de la Regresión Logística.

$$\frac{\partial J(\theta)}{\partial \theta_j} = \frac{\partial}{\partial \theta_j} \frac{-1}{m} \sum_{i=1}^{m} \left[y^{(i)} \left(\log(h_\theta \left(x^{(i)} \right) \right) + (1 - y^{(i)}) \left(\log(1 - h_\theta \left(x^{(i)} \right) \right) \right]$$

$$= \frac{-1}{m} \sum_{i=1}^{m} \left[y^{(i)} \frac{\partial}{\partial \theta_j} \log \left(h_\theta \left(x^{(i)} \right) \right) + (1 - y^{(i)}) \frac{\partial}{\partial \theta_j} \left(\log(1 - h_\theta \left(x^{(i)} \right) \right) \right]$$

$$= \frac{-1}{m} \sum_{i=1}^{m} \left[y^{(i)} \frac{\frac{\partial}{\partial \theta_j} (h_\theta \left(x^{(i)} \right)}{h_\theta \left(x^{(i)} \right)} + (1 - y^{(i)}) \frac{\frac{\partial}{\partial \theta_j} \left(1 - h_\theta \left(x^{(i)} \right) \right)}{1 - h_\theta \left(x^{(i)} \right)} \right]$$

$$= \frac{-1}{m} \sum_{i=1}^{m} \left[y^{(i)} \frac{\frac{\partial}{\partial \theta_j} \sigma \left(\theta^\top x^{(i)} \right)}{h_\theta \left(x^{(i)} \right)} + (1 - y^{(i)}) \frac{\frac{\partial}{\partial \theta_j} \left(1 - \sigma \left(\theta^\top x^{(i)} \right) \right)}{1 - h_\theta \left(x^{(i)} \right)} \right]$$

$$\frac{-1}{m} \sum_{i=1}^{m} \left[y^{(i)} \left(1 - h_\theta \left(x^{(i)} \right) \right) x_j^{(i)} - \left(1 - y^i \right) \ h_\theta \left(x^{(i)} \right) x_j^{(i)} \right]$$

$$\frac{-1}{m} \sum_{i=1}^{m} \left[y^i - y^i h_\theta \left(x^{(i)} \right) - h_\theta \left(x^{(i)} \right) + y^{(i)} h_\theta \left(x^{(i)} \right) \right] \ x_j^{(i)}$$

$$\frac{-1}{m} \sum_{i=1}^{m} \left[y^{(i)} - h_\theta \left(x^{(i)} \right) \right] \ x_j^{(i)}$$

$$= \frac{1}{m} \sum_{i=1}^{m} \left[h_\theta \left(x^{(i)} \right) - y^{(i)} \right] \ x_j^{(i)}$$

2.2. Regularización?

El ajuste excesivo es un problema común en el aprendizaje automático, donde un modelo se desempeña bien en los datos de entrenamiento, pero no se generaliza bien en datos invisibles (datos de prueba).

Si un modelo sufre un ajuste excesivo, también decimos que el modelo tiene una gran varianza, lo que puede deberse a que haya demasiados parámetros que conduzcan a un modelo demasiado complejo dados los datos subyacentes. De manera similar, nuestro modelo también puede sufrir (alto sesgo), lo que significa que nuestro modelo no es lo suficientemente complejo como para capturar bien el patrón en los datos de entrenamiento y, por lo tanto, también tiene un bajo rendimiento en datos invisibles.

Una forma de encontrar una buena compensación de sesgo-variación es ajustar la complejidad del modelo a través de la regularización. La regularización es un método muy útil para manejar la colinealidad (alta correlación entre características), filtrar el ruido de los datos y, eventualmente, evitar el sobreajuste.

2.2.1. Obtener Steepest Descent con Regularización

Recordamos que la función de costo regularizada para la regresión logística es:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}[-y^{(i)}log(h_\theta(x^{(i)})) - (1-y^{(i)})log(1-h_\theta(x^{(i)}))] + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2$$

Tendremos en cuenta de que no regularizamos el parametro θ_0.

El gradiente para la función de costo es un vector donde el elemento j^{th} se define de la siguiente manera:

$$\frac{\partial J(\theta)}{\partial \theta_0} = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}$$

$$\frac{\partial J(\theta)}{\partial \theta_j} = \Big(\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}\Big) + \frac{\lambda}{m}\theta_j$$

El Steepest Descent cuando se hace la regularización queda de la siguiente manera:

$$
\begin{aligned}
Repeat : \Bigg\{ \\
\theta_0 &:= \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^i) - y^i) \cdot x_0^i \\
\theta_j &:= \theta_j - \alpha [\frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^i) - y^i) \cdot x_j^i + \frac{\lambda}{m} \theta_j] \\
\Bigg\}
\end{aligned}
$$

Por lo que podemos obtener una optimización avanzada, calculando el descenso del gradiente regularizado:

1. Calculamos $J(\theta)$:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}[-y^{(i)}log(h_\theta(x^{(i)})) - (1-y^{(i)})log(1-h_\theta(x^{(i)}))] + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2$$

2. Calculamos $\frac{\partial}{\partial\theta_0}J(\theta)$

$$\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^i) - y^i)x_0^i$$

3. Calculamos $\frac{\partial}{\partial\theta_1}J(\theta)$

$$\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^i) - y^i)x_1^i + \frac{\lambda}{m}\theta_1$$

4. Calculamos $\frac{\partial}{\partial\theta_2}J(\theta)$

$$\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^i) - y^i)x_2^i + \frac{\lambda}{m}\theta_2$$

5. Así hasta calcular el gradiente $n+1$, que sería el $\frac{\partial}{\partial\theta_n}J(\theta)$

$$\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^i) - y^i)x_n^i + \frac{\lambda}{m}\theta_n$$

2.2.2. Obtener Steepest Descent Matricialmente

Sabemos lo siguiente:

$$X\theta = h(x^i)$$

Así pues, esto es:

$$X = \begin{pmatrix} 1 & x_1^1 & x_2^1 & ... & x_n^m \\ 1 & x_1^2 & x_2^2 & ... & x_n^m \\ 1 & x_1^3 & x_2^3 & ... & x_n^m \\ 1 & ... & ... & ... & ... \\ 1 & x_1^m & x_2^m & ... & x_n^m \end{pmatrix}, \theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ ... \\ \theta_n \end{pmatrix}$$

Multiplicando estas dos matrices, tenemos que:

$$\begin{pmatrix} 1 & x_1^1 & x_2^1 & ... & x_n^m \\ 1 & x_1^2 & x_2^2 & ... & x_n^m \\ 1 & x_1^3 & x_2^3 & ... & x_n^m \\ 1 & ... & ... & ... & ... \\ 1 & x_1^m & x_2^m & ... & x_n^m \end{pmatrix} \cdot \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ ... \\ \theta_n \end{pmatrix} = \begin{pmatrix} h(x^1) \\ h(x^2) \\ h(x^3) \\ ... \\ h(x^m) \end{pmatrix}$$

Ahora, deberemos de añadir la matriz de costo de y^i, por lo que, tendremos que restar nuestra hipotesis obtenida, con la matriz de costo. Esto es:

$$\begin{pmatrix} h(x^1) \\ h(x^2) \\ h(x^3) \\ ... \\ h(x^m) \end{pmatrix} - \begin{pmatrix} y^1 \\ y^2 \\ y^3 \\ ... \\ y^m \end{pmatrix} = \begin{pmatrix} h(x^1) - y^1 \\ h(x^2) - y^2 \\ h(x^3) - y^3 \\ ... \\ h(x^m) - y^m \end{pmatrix}$$

Obtenemos la matriz traspuesta de X, para obtener nuestro descenso del gradiente:

$$X^T = \begin{pmatrix} 1 & 1 & 1 & ... & 1 \\ x_1^1 & x_2^1 & x_3^1 & ... & x_n^m \\ x_1^2 & x_2^2 & x_3^2 & ... & x_n^m \\ x_1^3 & x_2^3 & x_3^3 & ... & x_n^m \\ & & & ... & \\ x_1^m & x_2^m & x_3^m & ... & x_n^m \end{pmatrix}$$

$$X^T \cdot (X\theta - y) = \begin{pmatrix} \sum_{i=1}^m (h(x^i) - y^i) \\ \sum_{i=1}^m (h(x^i) - y^i) x_1^i \\ \sum_{i=1}^m (h(x^i) - y^i) x_2^i \\ ... \\ \sum_{i=1}^m (h(x^i) - y^i) x_n^i \end{pmatrix} = \begin{pmatrix} \frac{\partial J}{\partial \theta_0} \\ \frac{\partial J}{\partial \theta_1} \\ \frac{\partial J}{\partial \theta_2} \\ ... \\ \frac{\partial J}{\partial \theta_n} \end{pmatrix} = X^T \cdot (X\theta - y).$$

Por lo que el Steepest Descent queda de la siguiente manera:

$$\theta = \theta - \frac{\alpha}{m}[X^T(X\theta - y)]$$

donde $(X\theta - y) = (h_\theta(x^i) - y^i)$, que es la hipótesis de nuestra función de costo.

2.2.3. Obtener Steepest Descent Matricialmente usando Regularización

De manera similar a la regresión logítica, la función de costo actualizada usando el parámetro de ajuste λ, queda de la siguiente manera:

$$\theta = \theta - \frac{\alpha}{m}[X^T(X\theta - y)] + \frac{\lambda}{2m}\sum_{i=1}^{n}\theta_j^2$$

De manera que nuestro descenso del gradiente queda de la siguiente manera:

1. $$\theta_0 := \theta_0 - \frac{\alpha}{m}[X^T(X\theta - y)]$$
2. ...
3. $$\theta_0 := \theta_j - \frac{\alpha}{m}[X^T(X\theta - y)] + \frac{\lambda}{m}\theta^T\theta, paraj = 1, 2, 3, ..., n \rightarrow \theta_1, ..., \theta_n$$

De manera matricial, añadiendo nuestro parámetro de regularización, tenemos que:

$$\theta := \theta - \frac{\alpha}{m}[X^T(X\theta - y)] + \frac{\lambda}{m}\theta$$

Agrupando algunos términos nos queda:

$$\theta := \theta(1 - \frac{\alpha\lambda}{m}) - \frac{\alpha}{m}X^T(X\theta - y)$$