Equipo 5:
Edgar Castillo Ramírez A00827826
Roel De la Rosa Castillo A01197595
Rodrigo Montelongo Pinales A00827757
Juan Pablo Yáñez González A00829598
Héctor San Román Caraza A01422876
Miguel Alejandro Salas Reyna A00827219

Inteligencia artificial avanzada para la ciencia de datos II (Gpo 502)
Tarea I: Vectorización
Profesor: Daniel Otero Fadul
7 de octubre 2022

1.- Función de costo:

Como vimos en clase, la función de costo del modelo de regresión logística es la siguiente:

$$C(w) = -\frac{1}{n}\sum_{i=1}^{n} y_i log(h(x_i)) + (1 - y_i)\, log(1 - h(x_i))$$

Demuestra que la siguiente expresión es equivalente:

$$C(w) = -\frac{1}{n}y^T log(h(Xw)) + (1 - y)^T log(1 - h(Xw))$$

R.- Después de un proceso de investigación, se puede confirmar que ambas ecuaciones son equivalentes, por los siguientes motivos:

- La primera ecuación representa la función de costo del modelo de regresión logística sin regularización.

- Por otro lado, la segunda ecuación representa la misma función de costo del modelo de regresión, pero de una forma vectorizada.

- Es decir, la diferencia entre estas ecuaciones es que en la primera los datos se manejan uno por uno, buscando realizar una suma. Mientras que en la segunda, se almacenan todos estos valores en un vector.

Analizando parte por parte que es lo que cambia, primeramente desaparece la sumatoria. Esto es porque ya no es necesario ir elemento por elemento, ya que todo estará en los vectores. Esto se ve reflejado también en y_i(valor de y en la posición i), que pasa a ser y^T(vector de valores de y).

La parte de $log(h(x_i))$, pasa a ser $log(h(Xw))$, donde se utiliza el vector de valores de X en sustitución a un elemento individual. Sin embargo, viene acompañado de w, que representa las estimaciones de los parámetros del modelo (esto es por lo que se hace el gradiente descendente, ya que no se pueden conseguir mediante una expresión).

Recordando que es a lo que se refiere $h(Xw)$, es la hipótesis que en el caso de la regresión logística está definida con una función sigmoide, ya que la función de costo tiende a valores entre cero y uno. A la hora de representar X, sería una matriz, mientras que w es un vector.

La última parte es $(1 - y)^T log(1 - h(Xw))$, pero realmente se repite lo mencionado anteriormente.

2.- Gradiente de la función de costo:

Tomando como punto de partida la ecuación vectorizada del primer inciso, demuestra que el gradiente de dicha función está dado por:

$$\nabla C(w) = \frac{1}{n} X^T (h(Xw) - y)$$

R.- Primeramente, cabe recordar que la forma general de la función del gradiente es:

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

Al analizar el proceso matemático que se sigue para obtener el gradiente descendiente de la función de costo sumarizada:

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

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

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

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

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

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

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

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

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

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

Se obtiene como resultado la ecuación (con las variables utilizadas para esta tarea):

$$\nabla C(w) = \frac{1}{n} \sum_{i=1}^{n} [h(x_i) - y_i] * x_i$$

Por lo que si se busca vectorizar como en el primer inciso, se deberían realizar los siguientes cambios:

- Desaparece la sumatoria, como sucede en el ejercicio anterior, ya que ahora se usarán vectores con todos los datos, no elemento por elemento.

- Las variables x_i y y_i pasan a ser solamente X y y, ya que ahora serán una matriz transpuesta y un vector respectivamente..

- Para la parte de la hipótesis no hay cambios, se mantiene como una función sigmoide de Xw

Por lo tanto, $\nabla C(w) = \frac{1}{n} X^T (h(Xw) - y)$ es el gradiente de la función de costo.