Apuntes semana 5, Martes 5 de marzo 2024

1st Sebastián Mora Godínez
Ingeniería en Computadores
Tecnológico de Costa Rica
sebastianmg@estudiantec.cr

I. Clasificación binaria

- Predice la probabilidad de ocurrencia de un evento
- Distribucion de bernoulli
- $P(X = K) = p^k \cdot (1-p)^{1-k}$
- k es el valor de puede tomar 0 o 1 y p es la probabilidad de que ocurra el evento

A. Función sigmoid

La función sigmoid es comúnmente utilizada como clasificador binaria.

$$P(x) = \frac{1}{1 + e^x} \tag{1}$$

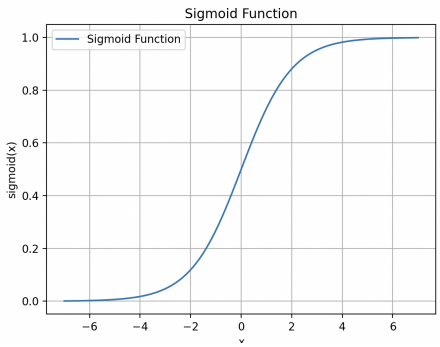

Fig. 1. Gráfica de la función sigmoid. Elaboración propia

a) Características:

- Tiene un comportamiento no lineal
- Codominio de [0, 1]
- x puede ser cualquier otro número, incluso puede ser ocupado por otra función
- Se puede combinar con regresión lineal

B. Linealidad

Como se mencionó anteriormente la función 1 tiene un comportamiento no lineal, mientras que la función

$$f(x) = wx + b \tag{2}$$

tiene un comportamiento lineal. Si combinamos ambas funciones obtenemos un resultado no lineal, pero ¿por que lo queremos así?

Existen varias razones para llegar a combinar ambas funciones dentro de las que se encuentran

- Calcular una función lineal es sencillo (eficiente, computacionalmente no tiene mucha complejidad)
- Es un método sencillo para mantener la relaciones entre variables y pesos.
- Permite obtener un comportamiento no lineal
- Permite modelar problemas con mayor complejidad

C. Clasificador binario

Si queremos realizar un clasificador, podemos definir un umbral, por ejemplo

$$y >= 0.5 = 1$$
$$y < 0.5 = 0$$

El umbral se puede cambiar dependiendo del problema que queremos resolver, como podría ser el caso de una moneda cargada, donde existe la posibilidad que uno de los lados de la moneda tiene más probabilidad de salir que el otro, por lo que es conveniente cambiar el umbral.

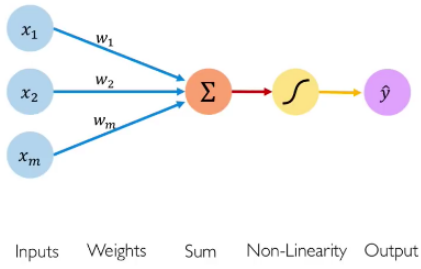

Fig. 2. Clasificador binario. Obtenido de [1]

II. Regresión logística

En la regresión logística se obtiene un resultado no lineal en un codominio de [0, 1]. La relación de los features y pesos se da por la regresión lineal y nos permite conocer la probabilidad de que un evento suceda.

III. Optimización

Para optimizar el modelo de regresión logística, se necesitan los pesos de **w** y **b** de la regresión lineal. Para actualizar dichos pesos se necesita

- Conseguir una función de pérdida L que me sirva para probabilidades
- Derivar la función de pérdida de L con respecto a **w** y **b** para utilizar el descenso del gradiente.

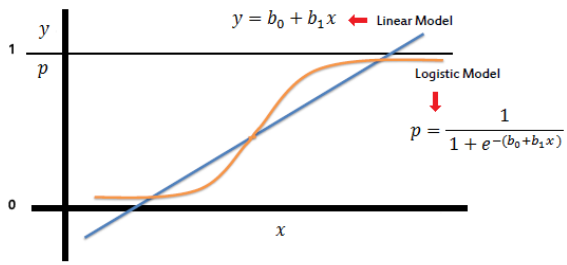

Fig. 3. Regresión lineal vs Regresión logística. Obtenido de [2]

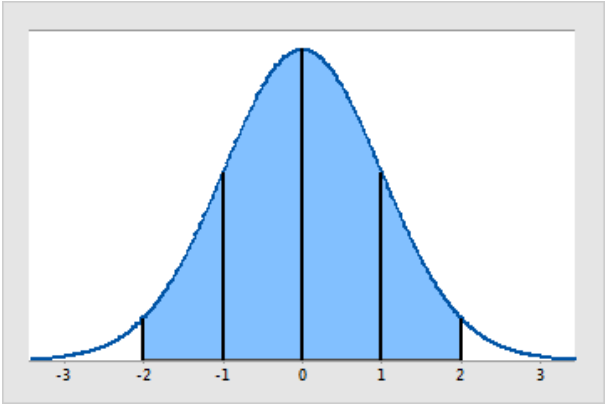

Fig. 4. Gráfica de distribución normal. Obtenido de [4]

A. Derivada de sigmoid

$$\theta' = [\frac{1}{1+e^x}]'$$

$$\theta' = \frac{[1]' \cdot (1+e^x) + 1 \cdot [1+e^x]'}{(1+e^x)^2}$$

$$\theta' = \frac{e^x}{(1+e^x)^2}$$

$$\theta' = \frac{e^x + 1 - 1}{(1+e^x)^2}$$

$$\theta' = \frac{e^x + 1}{(1+e^x)^2} - \frac{1}{(1+e^x)^2}$$

$$\theta' = \frac{1}{1+e^x} - \frac{1}{(1+e^x)^2}$$

$$\theta' = \frac{1}{1+e^x} \cdot (1 - \frac{1}{(1+e^x)})$$

$$\theta' = \theta \cdot (1 - \theta) \tag{3}$$

B. Hallar la función de pérdida

a) Verosimilitud vs MSE: Recordemos que necesitamos una función de costo relacionada a probabilidades, es por esto que no podemos utilizar MSE, si no que mas bien utilizamos la función de verosimilitud que está por la función

$$L(\frac{\theta}{X}) = P(\frac{X}{\theta}) \tag{4}$$

b) Distribución normal: La distribución normal está contemplada por dos parámetros: la media y la desviación estándar. Es un patrón estadístico que aparece cuando un conjunto de datos se distribuye de manera uniforme alrededor de un eje central [3].

Entonces debemos encontrar una distribución que represente los datos que tengamos. La idea es poder observar todos los datos de acuerdo al set de parámetros que tengo (likelihood). Si desplazamos la gráfica podemos aumentar el likelihood.

c) MSE vs likelihood: Mientras que en la regresión lineal minimizamos el MSE, para el caso de la regresión logística vamos a maximizar la verosimilitud.

C. Función de costo - regresión logística

Queremos que la decisión de que un evento sea la más probable posible, para esto maximizamos la verosimilitud de todo el training dataset. De acuerdo con los parámetros **w** y **b**, calculamos la probabilidad de cada observación. A continuación se muestra la función L

$$L = f(x^i)^{y_i}(1 - f(x^i))^{1-y_i}, i = 1, ..., N \tag{5}$$

Tenemos dos casos:

1) Caso y^i = 1:

$$wx + b = 1.458$$

$$f(x_i) = \theta(1.458) = 0.81$$

$$L = f(x_i)^1(1 - f(x_i))^0$$

$$L = f(x_i)$$

2) Caso y^i = 0:

$$wx + b = -1.32$$

$$f(x_i) = \theta(-1.32) = 0.21$$

$$L = f(x_i)^0(1 - f(x_i))^{1-0}$$

$$L = 1 - f(x_i)$$

D. Transformación con logaritmo

Antes de comenzar, recordemos algunas propiedad de los algoritmos

$$ln(a^n) = n \cdot ln(a)$$

$$ln(a \cdot b) = ln(a) + ln(b)$$

$$ln(a^n \cdot b^n) = n \cdot ln(a) + n \cdot ln(b)$$

Ahora aplicamos logaritmo a la verosimilitud, podemos hacer esto debido a que la función logaritmo es estrictamente creciente.

$$ln(L) = \sum ln(f(x^i)^{y_i} + ln((1 - f(x^i))^{1-y_i})$$

$$ln(L) = \sum y_i \cdot ln(f(x^i) + (1 - y_i)ln((1 - f(x^i))$$

Dentro de las ventajas de aplicar logaritmo es que es mucho más fácil de computar (evitamos posibles NaN) y que es una función más sencilla de derivar. Esto se conoce como log-likelihood.

La función L promedio está dada por

$$ln(L) = \frac{1}{n} \cdot y_i \cdot ln(f(x^i)) + (1 - y_i)ln((1 - f(x^i))$$

Mientras que para 1 sample es

$$ln(L) = y_i \cdot ln(f(x^i) + (1 - y_i)ln((1 - f(x^i))$$

En regresión lineal encontramos un algoritmo para minimizar el MSE, ahora imaginemos que tenemos todo el código y queremos maximizar la función, entonces solamente debemos poner el menos a la función.

$$ln(L) = -[y_i \cdot ln(f(x^i) + (1 - y_i)ln((1 - f(x^i))]$$

E. Derivada de la función de costo

Partiendo de

$$ln(L) = -[y_i \cdot ln(f(x^i)) + (1 - y_i)ln((1 - f(x_i))]$$

$$f(x) = a(z(x)) \tag{6}$$

$$a(x) = \frac{1}{1 + e^x} \tag{7}$$

$$z(x) = wx + b \tag{8}$$

sustituyendo, tenemos que

$$ln(L) = -[y_i \cdot ln(a(z(x))) + (1 - y_i) \cdot ln((1 - a(z(x)))]$$

Para obtener la derivada con respecto a **w** se parte de

$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial a} \cdot \frac{\partial a}{\partial z} \cdot \frac{\partial z}{\partial w}$$

$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial z} \cdot \frac{\partial z}{\partial w}$$

Mientras que para obtener la derivada con respecto a **b** se tiene que

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial a} \cdot \frac{\partial a}{\partial z} \cdot \frac{\partial z}{\partial b} \tag{9}$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial z} \cdot \frac{\partial z}{\partial b} \tag{10}$$

Tanto la derivada $\frac{\partial z}{\partial w}$ como $\frac{\partial z}{\partial b}$ son sencillas de calcular, a partir de 8 obtenemos que

$$\frac{\partial z}{\partial w} = x$$

$$\frac{\partial z}{\partial b} = 1$$

Esto significa que solamente tendremos que calcular las derivas $\frac{\partial L}{\partial z}$, la cual se calcula de la siguiente manera

$$\frac{\partial L}{\partial z} = \frac{\partial L}{\partial a} \cdot \frac{\partial a}{\partial z}$$

Por lo que tendremos que calcular $\frac{\partial L}{\partial a}$ y $\frac{\partial a}{\partial z}$ primeramente. Comencemos con $\frac{\partial L}{\partial a}$

$$\frac{\partial L}{\partial a} = -[(y_i \cdot \frac{1}{a(x)} \cdot a(x)') + ((1 - y_i) \cdot \frac{1}{1 - a(x)} \cdot (1 - a(x))']$$

$$\frac{\partial L}{\partial a} = -[(\frac{y_i}{a(x)} \cdot 1) + (\frac{1 - y_i}{1 - a(x)}) \cdot -1]$$

$$\frac{\partial L}{\partial a} = -[(\frac{y_i}{a(x)} \cdot 1) - (\frac{1 - y_i}{1 - a(x)})]$$

$$\frac{\partial L}{\partial a} = \frac{-y_i}{a(x)} \cdot + \frac{1 - y_i}{1 - a(x)}$$

Ahora continuamos con $\frac{\partial a}{\partial z}$, pero antes, recordemos que $a(z(x))$ es nuestro modelo, por lo que podemos utilizar la derivada calculada en 3

$$\frac{\partial a}{\partial z} = a(z(x)) \cdot (1 - a(z(x)))$$

Una vez obtenidos estos resultados, podemos sustituir y continuar el procedimiento para obtener $\frac{\partial L}{\partial z}$

$$\frac{\partial L}{\partial z} = \frac{-y_i}{a(z(x))} + \frac{1 - y_i}{1 - a(z(x))} \cdot [a(z(x)) \cdot (1 - a(z(x)))]$$

$$\frac{\partial L}{\partial z} = \frac{-y_i + a(z(x)) \cdot y_i + a(z(x)) - a(z(x)) \cdot y_i}{a(z(x)) - a(z(x))^2} \cdot [a(z(x)) - a(z(x)))^2]$$

$$\frac{\partial L}{\partial z} = \frac{-y_i + a(z(x)) \cdot y_i + a(z(x)) - a(z(x)) \cdot y_i}{a(z(x)) - a(z(x))^2} \cdot [a(z(x)) - a(z(x)))^2]$$

$$\frac{\partial L}{\partial z} = \frac{-y_i + a(z(x))}{a(z(x)) - a(z(x))^2} \cdot [a(z(x)) - a(z(x)))^2]$$

$$\frac{\partial L}{\partial z} = a(z(x)) - y_i$$

Por lo que sustituyendo en 9 y 10, obtenemos que

$$\frac{\partial L}{\partial w} = (a(z(x)) - y_i) \cdot x$$

$$\frac{\partial L}{\partial b} = (a(z(x)) - y_i) \cdot 1$$

Otra forma de representarlo es

$$\frac{\partial L}{\partial w} = (\hat{y} - y_i) \cdot x_i$$

$$\frac{\partial L}{\partial b} = (\hat{y} - y_i) \cdot 1$$

IV. Actualización de parámetros

Podemos actualizar los parámetros **w** y **b** de la siguiente manera

$$w = w - a \cdot \frac{\partial L}{\partial w}$$

$$b = b - a \cdot \frac{\partial L}{\partial b}$$

donde **a** es un hiperparámetro

V. Proyecto I

A. Exploración y procesamiento de datos

- Análisis fuerte sobre el dato, conocer y entender los datos que vamos a manejar IQR
- escoja una tecnica para valores faltantes (justifciar porque se hizo de esa manera)
- realizar el analisis de overfitting

B. Modelos

- Regresión logística, KNN, Redes neuronales (multilayer perceptor)
- Evalue los modelos utilizando métricas (mínimo accuary, precisión, recall) basado en el conjunto de pruebas seleccionado
- Después de sacar las métricas, sacar una conclusión y justificar las métricas que se van a utilizar

References

[1] S. Kadam, *Neural network part1: Inside a single neuron*, 2020. [Online]. Available: https://medium.com/analytics-vidhya/neural-network-part1-inside-a-single-neuron-fee5e44f1e.

[2] L. Torres, *¿en qué consiste la regresión logística? ¿qué es la regularización?* [Online]. Available: https://www.themachinelearners.com/regresion-logistica-regularizacion/.

[3] P. Rodó, *Distribución normal: Qué es, cómo se calcula y ejemplos*, 2024. [Online]. Available: https://economipedia.com/definiciones/distribucion-normal.html.

[4] Minitab, *¿qué es la distribución normal?* [Online]. Available: https://support.minitab.com/es-mx/minitab/20/help-and-how-to/statistics/basic-statistics/supporting-topics/normality/what-is-the-normal-distribution/.