Apuntes Semana 5, Martes 5 de marzo

Jose Julian Gutierrez Badilla
Instituto Tecnologico de Costa Rica
Profesor: Steven Pacheco
IC-6200 Inteligencia Artificial

I. Desarrollo

A. Regresión Lineal

- Genera un modelo que puede predecir valores continuos
- Salida de valores continuos

B. Regresión Logística

- Resultado es no lineal
- Se utiliza la funcion Sigmon para transformar cualquier valor real en un numero dentro del rango [0,1]: $\sigma(x) = \frac{1}{1+e^{-x}}$
- Se utiliza para clasificación binaria
- Asimismo, se utiliza la Distribución de Bernoulli para predecir la probabilidad de la ocurrencia de un evento binario: $P(X = x) = p^x(1-p)^{1-x}$
- Se observa el sample, y se clasifica (etiqueta 0 y etiqueta 1)
- La relacion de los features y pesos se da por la regresion lineal

C. Sigmoide (Función Logística Estándar)

- Tiene un comportamiento no lineal
- Codominio [0, 1]
- Note que x puede ser cualquier número. Incluso el resultado de una función (composición de funciones)
- Se puede combinar con regresión lineal
- Función: $\frac{1}{1+e^{-x}}$

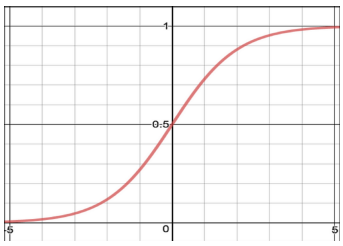

Fig. 1: Funcion Sigmond. Obtenido de [2]

D. Combinar regresión Logística con regresión Lineal. Por qué queremos hacerlo así?

- Calcular una función lineal es muy simple
- Es un método simple para mantener la relación entre variables y pesos
- Obtener un comportamiento no lineal con una función sencilla como Sigmond
- Permite modelar problemas con complejidad mayor

E. Clasificador

- Recordemos que la en la regresión logística, se utiliza la función Sigmon para modelar la probabilidad de que una observación pertenezca a una clase particular. Esta función transforma una entrada en un valor entre 0 y 1.
- Si queremos realizar un Clasificador, se puede definir un umbral, por ejemplo: podemos definir que para todo valor $y \geq 0.5 = 1$ y $y < 0.5 = 0$. En otras palabras, para hacer una clasificación binaria, se necesita este umbral ya que la salida de la función es cualquier valor entre 0 y 1.

F. Optimización

- Se necesita optimizar los pesos de w y b de la regresión lineal: En la regresión logística, queremos encontrar los pesos w y el sesgo b que mejor se ajusten a los datos de entrenamiento. Estos pesos determinan la relación entre las características de entrada y la salida del modelo.
- Para actualizar los pesos necesitamos encontrar la dirección en la que la función de pérdida disminuye. Esto implica calcular el gradiente de la función de pérdida con respecto a los pesos y luego ajustar los pesos en la dirección opuesta al gradiente para minimizar la función de pérdida.

G. Lost Function

$L = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \cdot \ln(f_{w,b}(x_i)) + (1-y_i)\cdot \ln(1-f_{w,b}(x_i))\right]$

- L es la función de pérdida.
- N es el número total de muestras en el conjunto de datos.
- y_i es la etiqueta de la observación i.
- $f_{w,b}(x_i)$ es la predicción del modelo para la observación i, dada por la función sigmond con pesos w y sesgo b
- $\ln$ para el logaritmo natural

H. Composición de Funciones

- $\frac{\partial L}{\partial w} = \frac{\partial L}{\partial a}\cdot\frac{\partial a}{\partial z}\cdot\frac{\partial z}{\partial w}$
- Ademas:
- $\frac{\partial L}{\partial b} = \frac{\partial L}{\partial a}\cdot\frac{\partial a}{\partial z}\cdot\frac{\partial z}{\partial b}$

Dadas estas ecuaciones es necesario resolver:

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

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

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

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

Por otro lado: $\frac{\partial L}{\partial z} = \frac{\partial L}{\partial a} \cdot \frac{\partial a}{\partial z} = -\frac{y_i}{a(z(x))} + \frac{1-y_i}{1-a(z(x))} \cdot a(z(x)) \cdot (1 - a(z(x)))$

.

.

.

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

Finalmente note que:

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

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

I. Verosimilitud vs MSE

- No usamos MSE como función de costo. Esto se debe a que el MSE es adecuado para problemas de regresión donde las salidas son valores continuos. En la regresión logística, estamos tratando de predecir probabilidades, no valores continuos.
- En lugar del MSE, en la regresión logística se utiliza la verosimilitud como función de costo. La verosimilitud es una medida de qué tan probable es observar los datos dados los parámetros del modelo. En la regresión logística, maximizamos la verosimilitud, lo que significa que queremos maximizar la probabilidad de observar los datos de entrenamiento dados los parámetros del modelo.
- $L(\theta|X) = P(X|\theta)$. La verosimilitud $(L(\theta|X)$ mide la posibilidad de observar los datos X dados los parámetros del modelo θ

J. Proyecto

- Primer conjunto de datos: Instituto Nacional de Diabetes y Enfermedades Digestivas y Renales (si un paciente tiene o no diabetes)
- El segundo conjunto de datos lo elige el grupo (solo conjuntos de datos tabulares, IRIS no se puede ya que fue utilizado para la tarea)
- Si se tienen muchos features, se debe transformar a binario (1 o 0)
- Debe ser un informe con un análisis fuerte (por ejemplo por que el dataset esta balanceado o no). Se debe justificar en los valores faltantes si se utilizo la media, se rellenó con 0, o no y porqué
- Análisis de overfitting
- Utilizar las librerias para los modelos
- Describir el accuracy del modelo
- Se pueden obtener conclusiones con el training set, ya que entre más conclusiones se tengan mejor.
- Formato científico (abstract, referencias en la introducción a libros, artículos, bibliografía, etc.).
- Fecha de entrega: 4 de abril

REFERENCES

[1] IBM. (2022). Regresión logística. [Online]. Disponible en: https://www.ibm.com/es-es/topics/logistic-regression

[2] Ichi Pro. (s/f). ¿Qué es la función sigmoidea? Cómo se implementa en regresión logística. [Online]. Disponible en: https://ichi.pro/es/que-es-la-funcion-sigmoidea-como-se-implementa-en-regresion-logistica-77981969140323