Apuntes Semana 6 - Regresión Logística

Apuntes del martes 25 de marzo

Mariana Fernández Martínez - 2021104026

Abstract - In this paper we reviewed the topic of logistic regression with the aggregation of the cost function, derivative of this function and composition of functions.
Keywords: functions, logistic regression, derivate.

I. RESPUESTA DEL QUIZ 3

1. Definición de overfitting y underfitting.
    a. **Overfitting:** Bajo error en training, alto error en testing. El modelo no es capaz de generalizar.
    b. **Underfitting:** Alto error en training, alto error en testing. El modelo no está aprendiendo nada de los datos.

2. Dos técnicas para evitar el alto sesgo y alta varianza.
    a. **Alta varianza:**
        i. Reducir la complejidad del modelo.
        ii. Reducir la dimensionalidad.
    b. **Alta sesgo:**
        i. Aumentar la complejidad del modelo si este es muy simple.
        ii. Revisar que los features seleccionados tengan capacidad de predicción.

3. Descripción del k-Fold Cross-Validation.
    a. Se divide el conjunto en K partes.
    b. Se entrena el modelo con K-1 partes, reservando una para validación.
    c. Se continúa el proceso de rotación de particiones usadas para el entrenamiento y validación. Es decir, se realiza una iteración.
    d. Se toma el promedio del rendimiento del modelo.
    e. Es útil cuando se manejan pocos datos y se desea validar el modelo.

4. Derivada parcial de L respecto a W.

$$L = \frac{1}{N}\sum_{i=1}^{N}((w \cdot x_i + b) - y_i)^2$$

$$= \frac{1}{N}\sum_{i=1}^{N} 2((w \cdot x_i + b) - y_i) \cdot x_i$$

$$= \frac{2}{N}\sum_{i=1}^{N}((w \cdot x_i + b) - y_i) \cdot x_i$$

II. DUDAS DE LA TAREA 1

De acuerdo a varias preguntas realizadas en clase se pueden tomar las siguientes recomendaciones:

- Creación de más features.
- Tener claro que el objetivo es el resultado que genere el modelo.
- Realizar pruebas para probar cuales features son razonables para el modelo.
- Es válido utilizar la biblioteca scikit-learn.
- Utilizar hiperparámetros desde 0.00001 en adelante.
- Para el testing utilizar el accuracy.

III. EXPLICACIÓN DEL PRIMER PROYECTO

El proyecto consiste en aplicar diferentes técnicas de clasificación de datos para dos conjuntos de datos, comparando un algoritmo KNN contra uno de regresión logística. Se va a realizar en Jupyter Notebook el cual debe contar con todas las pruebas realizadas a los modelos. Se pueden generar dos Jupyter si lo considera necesario.

Además, se creará una documentación a parte en LaTex y formato IEEE. Los entregables de este proyecto son el código fuente en LaTex, la documentación en PDF y el Jupyter Notebook. Las instrucciones a detalle se encuentran en la sección de documentos del Tec Digital.

IV. REGRESIÓN LOGÍSTICA

A. Diferencia entre Regresión Lineal y Regresión Logística

- La regresión lineal se trata de recibir un valor continuo en la función. Como en el ejemplo de las calabazas, se pretende predecir el precio de estas en base al tamaño.

- La regresión logística se trata de clasificar entre dos eventos o un sample y se tiene una salida binaria (0, 1).

B. ¿Qué es la Regresión Logística?

Es un algoritmo de clasificación binaria. Predice la probabilidad de ocurrencia de un evento binario. Se basa en la distribución de Bernoulli:

$$P(X = k) = p^k \cdot (1 - p)^{1-k}$$

Describe un valor *k* que puede ser cero o uno y p la probabilidad que ocurra el evento. Además, la relación de los features y pesos se da por la regresión lineal.

C. Sigmoid (Standar Logistic Function)

Es una función descrita de la siguiente manera:

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

Puede transformar cualquier valor real en un valor dentro del intervalo [0, 1]. Entre las características de esta se encuentran:

- Tiene comportamiento no lineal
- Codominio es [0, 1]
- El valor *x* puede ser cualquier número, incluso un resultado de otra función (composición de funciones).

D. Linealidad

$$\sigma(x) = \frac{1}{1 + e^{-x}} \rightarrow \text{Resultado no lineal}$$

$$fw, b(x) = wx + b \rightarrow \text{Resultado lineal}$$

La combinación de una función sigmoid y una línea genera un resultado no lineal:

$$\sigma(fw, b(x)) = \frac{1}{1 + e^{-(fw,b(x))}}$$

E. ¿Por qué queremos hacerlo así?

- Calcular una función lineal muy simple en términos computables.
- Es un método simple para mantener la relación entre variables y pesos.
- Obtener comportamiento no lineal con una función sencilla como Sigmoid.
- Permite modelar problemas con mayor complejidad como un clasificador.

F. Clasificador

Para crear un clasificador, se puede definir un umbral de la siguiente forma:

- y >= 0.5 = 1
- y < 0.5 = 0

El umbral se puede cambiar conforme a los datasets (0.6, 0.7, …), depende del problema. Por ejemplo: una moneda cargada.

G. Optimización

Para actualizar los pesos *w* y *b* en la regresión lineal se debe conseguir una función de pérdida ***L*** que sirve para probabilidades ya que *w* y *b* se encuentran muy dentro de la función sigmod. Para generar la función ***L***, primero se deriva la sigmoid.

H. Derivada Sigmod

Esta función se define por:

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

Para derivarla respecto a x, se utiliza la regla del cociente:

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

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

Derivando cada término:

$$\sigma'(x) = \frac{0 \cdot (1 + e^{-x}) - 1 \cdot (0 - e^{-x})}{(1 + e^{-x})^2}$$

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

Reescribiendo e^x como:

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

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

Factorizar:

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

La fórmula final sería:

$$\sigma'(x) = \sigma(x) \cdot (1 - \sigma(x))$$

I. Verosimilitud vs MSE

- No se utiliza MSE que es la métrica estándar para evaluar el modelo ya que mide la diferencia entre valores y no es adecuada para la regresión logística.
- Se necesita una función de costo relacionada a probabilidades, que sería la verosimilitud.
- La verosimilitud se define por $L(\theta|X) = P(X|\theta)$ y representa la probabilidad condicional de observar todos los samples de acuerdo a los parámetros que se están seleccionando

Máxima verosimilitud: El objetivo es encontrar los parámetros del modelo que maximizan la verosimilitud. Es decir, encontrar la mayor probabilidad posible para observar los datos.

MSE vs Maximize Likelihood: En la regresión lineal se minimiza el MSE, ahora en la regresión logística se quiere maximizar la verosimilitud, por lo tanto se quieren optimizar los parámetros *w* y *b*.

K. Función de costo

El objetivo de esta función es que la decisión sobre un evento sea la más probable posible. Para lograr esto se realizan dos pasos:

- Maximizar la verosimilitud de todo el training set.
- De acuerdo a los parámetros *w* y *b*, se calcula la probabilidad de cada observación:

$$L = \prod fw, b(x_i)^{y_i}(1 - fw, b(x_i))^{(1-y_i)},\ i = 1 \ldots N$$

Caso 1 y_i ***= 1***

$$f_{w,b}(x)^{y_i}(1 - f_{w,b}(x))^{1-y_i}$$

$$= f_{w,b}(x)^1(1 - f_{w,b}(x))^{1-1}$$

$$= f_{w,b}(x)^1(1 - f_{w,b}(x))^0$$

Caso 2 y_i ***= 0***

$$f_{w,b}(x)^{y_i}(1 - f_{w,b}(x))^{1-y_i}$$

$$= f_{w,b}(x)^0(1 - f_{w,b}(x))^{1-0}$$

$$= f_{w,b}(x)^0(1 - f_{w,b}(x))^1$$

$$= (1)(1 - f_{w,b}(x))$$

$$= (1 - f_{w,b}(x))$$

Derivada de la Función de Costo

$$L = \prod fw, b(x_i)^{y_i}(1 - fw, b(x_i))^{(1-y_i)},\ i = 1 \ldots N$$

Para calcular la verosimilitud se necesita hallar la probabilidad de observar cada sample. No obstante, la derivada de una multiplicación es compleja por lo que se buscará una equivalencia donde no se tenga que multiplicar: Logaritmo (estrictamente creciente).

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

Aplicación de logaritmo a la verosimilitud

$$L = \prod_{i=1}^{N} f_{w,b}(x)^{y_i}(1 - f_{w,b}(x))^{1-y_i}$$

$$\ln(L) = \sum_{i=1}^{N} \ln(f_{w,b}(x)^{y_i}) + \ln((1 - f_{w,b}(x))^{1-y_i})$$

$$\ln(L) = \sum_{i=1}^{N} y_i \cdot \ln(f_{w,b}(x)) + (1 - y_i) \cdot \ln(1 - f_{w,b}(x))$$

- Es más fácil de computar (NaN).
- Es una función más sencilla de derivar
- Log-Likelihood.

Para minimizar maximizando lo que se debe hacer es darle vuelta a la función y aplicar un -1, de la siguiente manera:

$$L = \frac{1}{N} \prod_{i=1}^{N} f_{w,b}(x)^{y_i}(1 - f_{w,b}(x))^{1-y_i}$$

$$L = -\frac{1}{N} \prod_{i=1}^{N} f_{w,b}(x)^{y_i}(1 - f_{w,b}(x))^{1-y_i}$$

Composición de funciones

Para actualizar los parámetros *w* y *b* se necesitan las derivadas parciales $\frac{\partial L}{\partial w}$ y $\frac{\partial L}{\partial b}$ por lo que se realiza una composición de funciones de la forma:

$$L = y_i \cdot \ln(f_{w,b}(x)) + (1 - y_i) \cdot \ln(1 - f_{w,b}(x))$$

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

$$z(x) = wx + b$$

El resultado de combinar ambas es:

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

Luego se procede a calcular las derivadas parciales:

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

Después de varios cálculos, se llega al resultado:

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

$a(z(x))$ = Mi modelo

Actualización de parámetros:

- $w = w - \alpha \frac{\partial L}{\partial w}$
- $b = b - \alpha \frac{\partial L}{\partial b}$

→ α es un hiperparámetro (learning rate).
→ Aplicar gradiente descendiente.