Apuntes Semana 6 - 25/03/2025

Victoria Sandí Barrantes c.2022146536

***Abstract*—En el presente documento se mencionan las respuestas del quiz 3. Además, se profundiza en los conceptos fundamentales de regresión logística, incluyendo la función sigmoide, verosimilitud y optimización de parámetros. Se abordan temas como overfitting/underfitting, técnicas de validación cruzada y el desarrollo matemático completo de las derivadas necesarias para la actualización de pesos en modelos de clasificación binaria.**

I. Respuestas Quiz 3

A. Describa qué es overfitting y underfitting

Overfitting: el modelo se ajusta demasiado a los datos de entrenamiento, causando que no generalice.

Underfitting: el modelo no aprender de los datos de entrenamiento.

B. Describa dos técnicas para evitar el alto sesgo y la alta varianza

1) Alto sesgo:

- Aumentar la complejidad del modelo.
- Hacer un cambio de features ya que pueden no ser los adecuados.

2) Alta varianza:

- Disminuir la complejidad del modelo.
- Conseguir más ejemplos para el training en caso de ser posible.

C. Describa k-Fold Cross-Validation

Técnica de validación donde se divide el dataset en k subconjuntos, el modelo se entrena k veces, usando en cada iteración k-1 subconjuntos para el entrenamiento y dejando el restante para validación (se cambia cada iteración).

D. Desarrolle la derivada parcial de L con respecto a W

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

$$2((wx_i + b) - y_i) \cdot ((wx_i + b) - y_i)'$$

$$(2(wx_i + b) - y_i) \cdot x_i$$

II. Preguntas sobre la tarea 1

A. Justificación del uso de un plot

Para justificar un plot externo hay que nombrar las figuras referenciadas y las razones para la decisión.

B. ¿Se pueden obviar unos features de la matriz de correlación?

Si consideran que no son necesarios se pueden quitar, pero preferiblemente la mayor cantidad de features posibles.

C. ¿Hay que hacer regresión lineal con los features anque no tenga sentido?

Sí, hay que comparar el dataset original y el dataset con los features que se elijan.

D. ¿Cuantás veces hay que hacer la regresión lineal con los n features?

A prueba y error.

E. Uso de Scikit-learn

No se puede usar el modelo de regresión lineal, todo lo de ingeniería de datos sí.

F. ¿El alpha debe ser 0.001?

A criterio de los estudiantes, sin embargo, entre 0.005 y 0.009 suele estar bien.

III. Proyecto 1

Fecha de entrega: 22 de abril de 2025
Valor: 20%
Tema: Clasificación
Especificaciones extra:

- Más de 600 samples
- De 5 features en adelante

IV. Repaso

A. Regresión Logística

1) Regresión Lineal vs Regresión Logística: La regresión lineal busca predecir valores de una función continua, a diferencia de la regresión logística que usando la función sigmoide clasifica entre dos eventos (salida binaria).

2) Características de la regresión logística:

- Clasificador binario.
- Predice la probabilidad de ocurrencia de un evento.
- Se basa en la distribución Bernoulli:

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

Con k 0 o 1 y p la probabilidad de que ocurra el evento.

B. Sigmoid (Standard Logistic Function)

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

- Tiene comportamiento lineal.
- Codominio $[0, 1]$
- x puede ser cualquier número, incluso el resultado de otra función (composición de funciones)

C. Linealidad

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

- Resultado no lineal: $\sigma(x) = \frac{1}{1+e^{-x}}$
- Resultado lineal: $f_{w,b}(x) = wx + b$
- Resultado de la combinación de la sigmoide con una lineal (no lineal): $\sigma(f_{w,b}(x)) = \frac{1}{1+e^{-(f_{w,b}(x))}}$

D. Ventajas de la combinación de la sigmoide con una lineal

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

E. Clasificador

- Si se quiere realizar un clasificador, se puede definir un umbral.
- $y \geq 0.5 = 1$.
- $y < 0.5 = 0$.
- El umbral se puede cambiar (0.6, 0.7, ...), depende del problema.
- Ejemplo: moneda cargada

F. Características de la Regresión Logística

- $f_{w,b}(x) = \frac{1}{1+e^{-(wx+b)}}$
- Su resultado es no lineal.
- Probabilidad de que un evento suceda.
- Binario [0,1].
- La relación entre los features y pesos se da por la regresión lineal.

G. Optimización

Para optimizar los pesos w y b de la regresión lineal se debe conseguir una función de pérdida que sirve para las probabilidades, para estose deriva la función Sigmoid.

1) Derivada Sigmoid:

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

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

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

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

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

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

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

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

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

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

H. Verosimilitud vs MSE

- No vamos a usar MSE.
- Necesitamos una función de costo relacionada a probabilidades.
- Verosimilitud: $L(\theta|X) = P(X|\theta)$
- Representa la probabilidad condicional de observar X dado el parámetro θ.
- Probabilidad de observar todos los datos: $L(\theta|X) = P(x_1, x_2, ..., x_n|\theta) = P(x_1|\theta) \cdot P(x_2|\theta), ..., P(x_n|\theta)$.
- MSE vs Maximize Likelihood: En la regresión lineal se minimiza el MSE, a diferencia de la regresión logística donde se busca maximizar la verosimilitud optimizando los parámetros w y b.

I. Función de Costo

- Se busca que la decisión sobre un evento sea la más probable posible.
- Se maximiza la verosimilitud de todo el training set.
- Se acuerdo a los parámetros w y b, se calcula la probabilidad de cada observación: $L = \Pi\ f_{w,\mathrm{b}}(x_i)^{y_i}\big(1 - f_{w,\mathrm{b}}(x_i)\big)^{(1-y_i)}, \quad i = 1 \ldots N$

J. Casos particulares

Caso 1: $y_1 = 1$

$$f_{w,b}(x_i)^{y_i}(1 - f_{w,b}(x_i))^{1-y_i}$$
$$f_{w,b}(x_i)^1(1 - f_{w,b}(x_i))^0$$
$$f_{w,b}(x_i)$$

Caso 2: $y_1 = 0$

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

K. Propiedades logaritmo

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

L. Derivada de la Función de Costo

$$L = \prod f_{w,b}(\hat{x}_i)^{y_i}(1 - f_{w,b}(\hat{x}_i))^{(1-y_i)}, \quad i = 1...N$$

Para calcular la verosimilitud se necesita encontrar la probabilidad de observar cada sample. Debido a que la derivada de un producto es compleja, se busca una equivalencia donde no se tenga que multiplicar, aplicando logaritmo (función estrictamente creciente).

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

$$\ln(L) = \sum \ln\left(f_{w,b}(x_i)^{y_i}\right) + \ln\left((1 - f_{w,b}(x_i))\right)^{1-y_i}$$

$$\ln(L) = \sum y_i \ln\left(f_{w,b}(x_i)\right) + (1 - y_i)\ln\left(1 - f_{w,b}(x_i)\right)$$

Ventajas:

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

Para convertir el problema de maximización a minimización, se multiplica por -1:

$$\ln(L) = \frac{1}{N}\sum y_i \ln\left(f_{w,b}(x_i)\right) + (1 - y_i)\ln\left(1 - f_{w,b}(x_i)\right)$$

$$\ln(L) = -\frac{1}{N}\sum y_i \ln\left(f_{w,b}(x_i)\right) + (1 - y_i)\ln\left(1 - f_{w,b}(x_i)\right)$$

Para actualizar w y b se necesita calcular las derivadas parciales $\frac{\partial \mathcal{L}}{\partial w}$ y $\frac{\partial \mathcal{L}}{\partial b}$.

$$L = y_i \cdot \ln(f_{w,b}(x)) + (1 - y_i) \cdot \ln(1 - f_{w,b}(x))$$
$$a(x) = \frac{1}{1 + e^{-x}}$$
$$z(x) = wx + b$$

Sustituyendo tenemos:

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

Luego las derivadas parciales:

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

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

Con $a(z(x))$ como nuestro modelo.

M. Actualización de parámetros

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

donde α es el learning rate (hiperparámetro).

V. Conclusion

Este documento ha presentado un resumen de los fundamentos teóricos y matemáticos de la regresión logística, con aspectos clave como la función sigmoide, la maximización de verosimilitud, y el desarrollo detallado de las derivadas parciales necesarias para la optimización de parámetros. Los conceptos presentados forman una base sólida para la implementación práctica de algoritmos de clasificación binaria y su aplicación en el Proyecto 1.