Apuntes Semana 4 – 11/03/2025

Marco Rivera Serrano

***Abstract*—En esta semana, se realiza un repaso sobre el algoritmo K-Nearest-Neighbors y algunas de sus ventajas y desventajas a la hora de realizar su implementación. Por otro lado, se introduce el tema de "Regresión lineal" donde se explica su funcionamiento y conceptos importantes a tener en cuenta cuando se trabajan con varias dimensiones. Uno de esos conceptos es el Descenso del Gradiente para poder encontrar el mínimo de la función y lograr que el modelo realice predicciones acertadas. Como el gradiente utilia derivadas, también se realiza un repaso de las más comunes que se pueden encontrar así como las derivadas multivariables conocidas como derivadas parciales.**

I. Respuestas quiz 1

A. Si dos vectores u y v son colineales y su magnitud es de 6 y 5 respectivamente ¿Cuál es el valor de su producto punto?

Por definición, el producto punto es:

$$u \cdot v = |u||v| \cos\theta$$

donde θ es el ángulo entre los vectores. Entonces:

$$\begin{aligned} &= 6 \cdot 5 \cdot \cos(0) \\ &= 30 \cdot 1 \\ &= 30 \end{aligned}$$

B. Si el producto punto de un vector u consigo mismo es de 25, calcule la distancia L_2

La distancia L_2 o norma euclidiana de un vector u se calcula:

$$||u|| = \sqrt{u \cdot u}$$

Entonces:

$$||u|| = \sqrt{25} = 5$$

C. Defina los siguientes tipos de entrenamiento: Supervised, Unsupervised y Reinforced

1) Supervised: Es aquel donde el set de datos incluye las etiquetas $\{x_i, y_i\}_i^N$.

2) Unsupervised: Por el contrario, en este tipo de entrenamiento, solo se tienen los datos de entrada x_i.

3) Reinforced: En este tipo de entrenamiento, se "premia" al modelo cuando realiza predicciones acertadas.

D. Explique que es GridSearch

Es una técnica para machine learning la cual consiste en encontrar la mejor combinación de hiperparámetros para un modelo.

II. Repaso de K-Nearest Neighbors

Para este algoritmo se utiliza la distancia L2 para encontrar la cantidad de vecinos y así, poder asignar la clase para una nueva entrada según la distancia del radio. Dicha distancia es un hiperparámerto.

A. Ventajas

- Fácil de implementar.
- No requiere un entrenamiento previo.

B. Desventajas

- Pesado a nivel computacional.
- Los datos deben estar cargados en memoria.
- Hay features irrelevantes que afectan a a la distancia.
- Poco eficiente con volúmenes de datos grandes.

III. Regresión Lineal

Este algoritmo permite calcular la pendiente de una recta dado un set de datos y predice nuevas entradas. Dado que es un algoritmo supervisado, requiere las etiquetas dentro del set de datos.

$$(x_i, y_i)_{i=1}^N$$

Donde:

- N es el tamaño de la colección.
- x_i es un vector D-dimensional.
- y_i son las etiquetas y es un valor que pertenece al conjunto de los reales $\mathbb{R}$.

En sí, es un método estadístico que intenta hallar la relación entre una variable dependeinte y y un conjunto de variables independientes x.

IV. ¿Qué queremos hacer?

La idea principal es construir un modelo que siga la siguiente estructura:

$$f_{w,b}(x) = wx + b$$

Donde:

- w es un vector D-dimensional.
- b es un número real.
- y es el modelo $f_{w,b}(x)$.

Cabe resaltar que la operación wx es un producto escalar y w permite darle relevancia a los features. Es decir, se tiene una combinación lineal de lo features donde los parámetros son w y b. Para poder hacer predicciones

acertadas, se necesitan encontrar valores óptimos de w y b. Un aspecto importante es que la regresión lineal puede ser una línea, plano o hiperplano según las dimensiones de los datos utilizados. Sin embargo, ser óptimo no es lo mismo a que el modelo sea perfecto.

A. Datos outlayers

Son aquellos datos que se salen de la distribución de datos y afectan al resultado de la recta final. Existen diversos algortimos para eliminar dichos datos como el rango intercuartílico (IQR).

V. Minimizar Min Square Error (MSE)

Tenemos la función:

$$L = \frac{1}{N}\sum(f_{w,b}(x_i) - y_i)^2, i = 1, ...N$$

Conocida como "Loss function" y es una medida de penalidad la cual mide el error de forma cuadrática. Por otro lado, la "cost function" tiene la siguiente estructura:

$$\frac{1}{N}\sum(f_{w,b}(x_i) - y_i)^2, i = 1, ...N$$

Ambas son iguales y su objetivo es encontrar el promedio de los errores entre la predicción y la etiqueta. Entonces, ¿cuál es el objetivo de minimizar?.

En general, se busca el menor error posible para L y que el modelo sea preciso en sus predicciones. Si el valor de L es muy grande, el modelo falla en hacer predicciones. Por otro lado, si el valor de L es pequeño, el modelo hace predicciones acertadas.

A. Hallar valores de parámetros

Como se puede apreciar el la figura 1, las funciones convexas, tienen un único mínimo global. Mientras que las funciones no convexas, tienen mínimos locales y un mínimo global.

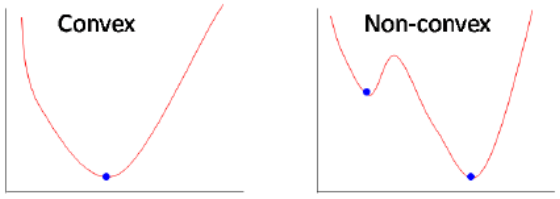

Fig. 1. Ejemplo de funciones convexas y no convexas

La definición formal de una función convexa es $f(x)$ tiene un mínimo local si $x = c, f(x) \geqslant f(c)$ para cada x que está en un intervalo alrededor de $x = c$.

La idea es utilizar la función L para encontrar ese mínimo global para aproximar mejor el modelo. Gracias a la naturaleza de una función convexa, garantiza que el mínimo sea global y que no existan mínimos locales. La función L será convexa gracias a que está elevado al cuadrado.

VI. Descenso de Gradiente

A. Analogía del escalador

Si se quiere llegar a la llanura más próxima con los ojos cerrados, utilizamos la planta del pie para poder medir la inclinación del suelo y saber hacia donde debemos caminar para llegar a la llanura. A esto se le conoce como descenso del gradiente.

Recordando de cálculo diferencial e integral, la derivada de una función otorga la pendiente y dirección. Para poder hallar el mínimo, necesitamos derivar la función L la cual está compuesta por $f_{w,b}$ que es el modelo. Eso implica que $f_{w,b}$ debe ser derivable. La idea es encontrar los mejores parámetros de w y b para que minimicen el resultado de la función L. Se puede asociar con un problema de espacios de búsqueda. Por otra parte, se utiliza MSE y no MAE (Min Absolute Error), porque existe un punto donde no es derivable.

VII. Repaso de derivadas

A. Constante

$$f(x) = k$$
$$f'(x) = 0$$

B. Constante por función

$$f(x) = kx$$
$$f'(x) = k$$

Ejemplo:

$$f(x) = 2x$$
$$f'(x) = 2$$

C. Función de grado n

$$f(x) = x^n$$
$$f'(x) = n \cdot x^{n-1}$$

Ejemplo:

$$f(x) = x^2$$
$$f'(x) = 2x$$

D. Suma de funciones

$$f(x) = u(x) + v(x)$$
$$f'(x) = u'(x) + v'(x)$$

Ejemplo:

$$f(x) = 2x + 3x$$
$$f'(x) = 2 + 3$$
$$= 5$$

E. Multiplicación de funciones

$$f(x) = u(x) \cdot v(x)$$
$$f'(x) = u'(x) \cdot v(x) + u(x) \cdot v'(x)$$

Ejemplo:

$$f(x) = 2x \cdot 3x$$
$$f'(x) = 2 \cdot 3x + 2x \cdot 3$$
$$= 6x + 6x$$
$$= 12x$$

F. Función más constante

$$f(x) = u(x) + z, \text{ con z constante}$$
$$f'(x) = u'(x) + 0$$

Z puede ser otra variable diferente a x. Ejemplo:

$$f(x) = 2x + 5$$
$$= 2x + z$$
$$= 2 + 0$$
$$= 2$$

G. Derivadas multivariable

$$f(x, y) = u(x) + v(y)$$
$$\frac{\partial f}{\partial x} = u'(x) + 0$$
$$\frac{\partial f}{\partial v} = v'(y) + 0$$

Ejemplo:

$$f(x, y) = 2x + 3y \qquad f(x, y) = 2x + 3y$$
$$= u(x) = 2x, z = 3y \qquad = v(x) = 3y, z = 2x$$
$$f(x, y) = 2x + z \qquad f(x, y) = 3y + z$$
$$\frac{\partial f}{\partial u} = 2 + 0 \qquad \frac{\partial f}{\partial v} = 3 + 0$$
$$= 2 \qquad = 3$$

VIII. Derivadas parciales

Las derivadas paciales son una generalización de las derivadas vistas en cálculo diferencial. En sí, es la tasa de cambio de una función respecto a sus variables independientes manteniendo las demás variables como constantes.

Su notación es:

$$f(x, y) = 2x + 3y$$
$$\frac{\partial f}{\partial x} = 2$$
$$\frac{\partial f}{\partial y} = 3$$