

Apuntes Semana 4

Andrés Sánchez Rojas
 Escuela de Ingeniería en Computación
 Instituto Tecnológico de Costa Rica
 26/8/2025

Abstract—La clase comenzó con un quiz de 4 preguntas relacionadas a la materia vista en clases anteriores, luego el profesor nos explicó las respuestas del quiz antes de comenzar con la materia de la clase. Durante la clase vimos el algoritmo de KNN, hicimos un repaso de derivadas y pasamos a ver cómo se construye y optimiza un modelo de regresión lineal.

I. QUIZ

- 1) 1. Anote y describa las 3 propiedades de la norma. 30pts
 R//

- a) Positividad: $\|x\| \geq 0$ y $\|x\| = 0$ si y solo si $x = 0$.
- b) Homogeneidad: $\|\alpha x\| = |\alpha| \|x\|$ para todo escalar α .
- c) Desigualdad triangular: $\|x + y\| \leq \|x\| + \|y\|$.

- 2) 2. Describa los tipos de aprendizaje supervised, unsupervised y one-shot learning. 30 pts
 R//

- a) Supervised: Se utiliza un conjunto de datos con características y etiquetas. Las etiquetas sirven para validar y corregir las aproximaciones del sistema.
- b) Unsupervised: No hay etiquetas con las que evaluar o corregir, se usa en algoritmos de cluster para agrupar valores.
- c) One-Shot Learning: Se le da un ejemplo al modelo y luego debe resolver un ejercicio similar

- 3) 3. Si u y v son dos vectores colineales con magnitudes de 5 y 6 respectivamente. ¿Desarrolle Cuál es el resultado del producto punto entre u y v ?
 R//

- a) $\|u\| = 5$, $\|v\| = 6$.
- b) $u \cdot v = \|u\| \|v\| \cos \theta$.
- c) $\cos 0 = 1$
- d) $5 \cdot 6 \cdot 1$.
- e) $u \cdot v = 30$

4. ¿Quién propone las Redes Generativas Adversarias
 R// Ian Goodfellow

II. K-NEAREST NEIGHBORS (KNN)

Se tiene un conjunto de datos tiquetados y se le quiere asignar una etiqueta a un dato basado en otros datos similares a este. Estos datos similares son los K vecinos más cercanos. Una vez que se tiene a los vecinos más cercanos se revisa las etiquetas de estos en una "votación" la etiqueta más común en estos K vecinos se le asigna al dato nuevo. Este K es un hiperparámetro y normalmente es un número impar para evitar empates.

Ventajas:

- Es sencillo de implementar
- Sirve para regresión y clasificación

Desventajas:

- Es muy costoso
- Features irrelevantes pueden distorsionar las distancias
- No es muy consistente ya que la clasificación puede variar dependiendo del K usado

III. REGRESIÓN LINEAL

Método estadístico que intenta hallar la relación entre una variable dependiente y un conjunto de variables independientes.

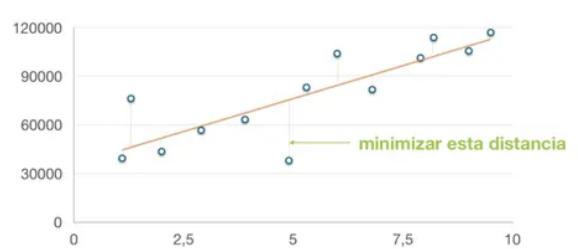


Fig. 1. Ejemplo de regresión lineal.

A. ¿Qué queremos hacer?

Buscamos construir un modelo

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

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

Lo que queremos es encontrar los valores de w y b óptimos para nuestro modelo. Es importante recordar que no tiene que ser perfecto (mínimo absoluto) pero debemos buscar que sea óptimo (mínimo local) para las necesidades que tengamos.

B. Loss Function

Esta función nos permite calcular qué tan bueno es nuestro modelo. Con esta función calculamos el error cometido por el modelo en cada muestra. La función de pérdida penaliza más los errores grandes por el error cuadrático.

$$(f_{w,b}(x_i) - y_i)^2$$

C. Cost Function

Es el error promedio del loss function sobre todo el dataset. Nuestro objetivo es minimizarla ajustando los parámetros w y b . Si tenemos un L grande quiere decir que el modelo da valores muy distintos a las etiquetas. Un L pequeño indica lo opuesto.

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

D. Repaso de derivadas

Propiedades de las derivadas:

$$\begin{aligned} \frac{d}{dx}[k] &= 0, \\ \frac{d}{dx}[x] &= 1, \\ \frac{d}{dx}[x^n] &= n x^{n-1}, \\ \frac{d}{dx}[f(x) + g(x)] &= f'(x) + g'(x), \\ \frac{d}{dx}[f(x) - g(x)] &= f'(x) - g'(x), \\ \frac{d}{dx}[k f(x)] &= k f'(x), \\ \frac{d}{dx}[f(x) g(x)] &= f'(x) g(x) + f(x) g'(x), \\ \frac{d}{dx}\left[\frac{f(x)}{g(x)}\right] &= \frac{f'(x) g(x) - f(x) g'(x)}{[g(x)]^2}, \\ \frac{d}{dx}[f(g(x))] &= f'(g(x)) \cdot g'(x). \end{aligned}$$

Ejemplo de derivada parcial:

Sea $f(x, y) = 2x + 3y$,

Al calcular $\frac{\partial f}{\partial y}$, tratamos x como constante,

$$\frac{\partial f}{\partial y} = 3$$

E. Función Convexa vs No Convexa

La función convexa sólo tiene un mínimo absoluto mientras que la no convexa puede tener múltiples mínimos locales.

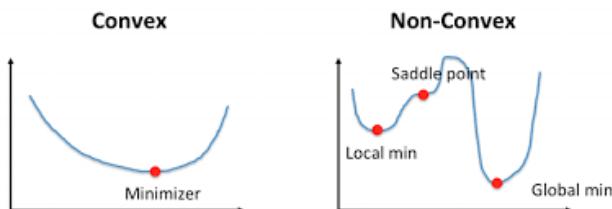


Fig. 2. Ejemplo de una función convexa y una no convexa.

F. Descenso de Gradiente

El profe puso un ejemplo para explicar este concepto. Estamos en la cima de una montaña con los ojos vendados y debemos encontrar la ruta más corta al punto más bajo posible. El proceso para esto sería:

- Buscar la dirección de mayor pendiente hacia abajo
- Descender por ese camino hacia abajo
- En cada paso repetimos el proceso.

Tenemos la función:

$$x_{\text{nuevo}} = x_{\text{antiguo}} - \alpha \nabla f(x_t)$$

α es la taza de aprendizaje que es un hiperparámetro y $\nabla f(x_t)$ es el gradiente o la derivada. Debemos tener cuidado al definir el α . Si se utiliza un α muy grande el algoritmo probablemente va a saltarse el punto óptimo muchas veces. Un α muy pequeño nos va a forzar a hacer muchas iteraciones. Debemos pensar bien en el learning rate que se utilizará pero se recomienda que sea relativamente pequeño para no saltarnos el punto óptimo o usar el Early Stopping Method. Este consiste en definir un valor razonable de L y detener la función cuando se llega a ese valor de L .



Fig. 3. Ilustración de descenso de gradiente.

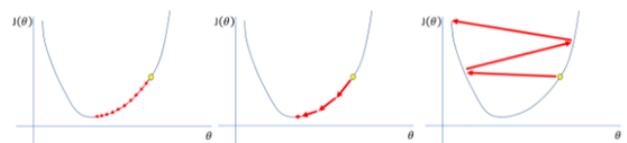


Fig. 4. Ejemplo de descenso de gradiente con diferentes learning rates

G. ¿Por qué utilizamos MSE (Mean Squared Error) y no MAE(Mean Absolute Error)

- El MSE penaliza más los errores grandes, el MAE los penaliza de manera lineal
- MAE no tiene una derivada continua ya que no es derivable en 0

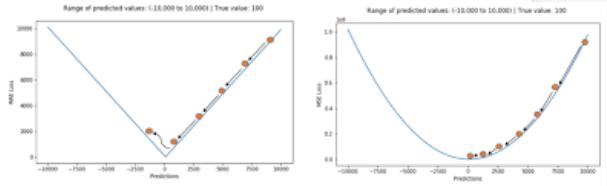


Fig. 5. Ilustración de MAE vs MSE

REFERENCES

- [1] Steven Pacheco Portuguez, *Clase sobre Regresión Lineal*, Tecnológico de Costa Rica, 2025.