# Apuntes de clase

Luis Felipe Calderón Pérez
Escuela de Ingeniería en Computación
Tecnológico de Costa Rica
Cartago, Costa Rica
2021048663
26-08-2025

***Resumen*—Este documento presenta los apuntes de la cuarta semana del curso de inteligencia artificial. Primeramente se dan las respuestas del primer quiz. Se repasa la tarea de clasificación, el algoritmo de los K-nearest Neighbors. Se introduce el tema de la regresión líneal, función de pérdida, mínimos locales, mínimos globales, el descendo del gradiente. Además, se repasaron las derivadas y se terminó con la pregunta de porque escoger MSE y no MAE.**

***Index Terms*—IA, derivadas, descenso del gradiente, regresión lineal**

## I. Preguntas y respuestas del primer quiz

1. Anote y describa las tres propiedades de la norma

   **Respuesta:**

   - Positividad: $\|x\| \geq 0$ y $\|x\| = 0$ si y solo si $x = 0$.
   - Homogeneidad: $\|\alpha x\| = |\alpha| \cdot \|x\|$ para cualquier escalar $\alpha$.
   - Desigualdad triangular: $\|x + y\| \leq \|x\| + \|y\|$.

2. Describa los tipos de aprendizaje supervised, unsupervised y one-shot learning.

   **Respuesta:**

   - Supervised: El modelo aprende a partir de datos que incluyen etiquetas, las cuales sirven como referencia durante el entrenamiento.
   - Unsupervised: : El modelo trabaja con datos sin etiquetas y se encarga de encontrar patrones en los datos ocultos.
   - One-shot: Basta con mostrarle una única vez como realizar la tarea para que el módelo pueda reproducirla.

3. Si u y v son dos vectores colineales con magnitudes 5 y 6 respectivamente.¿Desarrolle cuál es el resultado del producto punto entre u y v?

   **Respuesta:**

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

$$u \cdot v = \ 5 \cdot \ 6 \cdot \cos(0)$$

$$\therefore u \cdot v = 30$$

4. ¿Quién propone las Redes Generativas Adversarias?.

   **Respuesta:** Ian Goodfellow

## II. Clase

### *II-A. Clasificación*

*II-A1. K-nearest Neighbors:* Algoritmo en donde un conjunto de datos etiquetados recibe un nuevo dato. A ese nuevo dato se le calcula la distancia con sus datos vecinos, una vez se encuentra a los vecinos más cercanos, se realiza una votación, para determinar a que categoria o tipo pertenece. En este algoritmo el hiperparámetro es el k.

1. Ventajas
   - No requiere fase de entrenamiento.
   - Fácil de implementar.
   - Flexible: regresión y clasificación.

2. Desventajas
   - Poco eficiente.
   - Las features irrelevantes distorcionan las distancias entre los datos.
   - Puede ser costoso a nivel computacional.
   - Dependiendo del K usado, cambia la clasificación del dato ingresado.

**Nota:** Se requiere normalización o estandarización en caso de que se dispare o haya gran diferencia en las distancia entre los datos.

### *II-B. Regresión líneal*

Es un algoritmo usado para encontrar un modelo en el conjunto de los números reales, para predecir valores contiguos.

1. Existen 2 tipos de variables:
   - Variables Independientes, representan los features que introducimos en el modelo.
   - Variables dependientes, representan las etiquetas o el objetivo que deseamos predecir.

Cuadro I
RELACIÓN HORAS DE ESTUDIO CON NOTAS

| Horas de estudio (x) | Nota (y) |
|---|---|
| 1 | 50 |
| 2 | 55 |
| 3 | 65 |
| 4 | 70 |
| 5 | 75 |

En el ejemplo anterior la regresión líneal que corresponde es:

$$y = 5x + 45$$

En donde 5x representa la inclinación de la función y 45 el intercept o en donde corta el eje y.

El modelo debe cumplir con la siguiente función:

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

donde $w$ y $x$ son vectores, y su producto punto genera un escalar; $b$ es un escalar. Los valores de $w$ y $b$ afectan directamente los resultados, por lo que debemos encontrar los valores óptimos de ambos para obtener un modelo óptimo.

**Nota:** Se van a trabajar desde modelos simples de regresión líneal, hasta multiple linear.

Figura 1. Regresión líneal simple

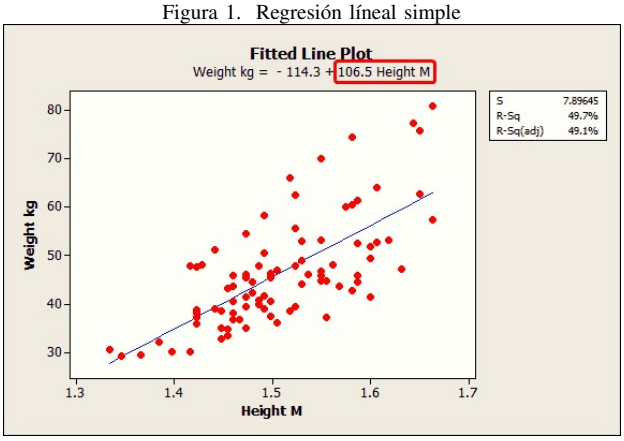
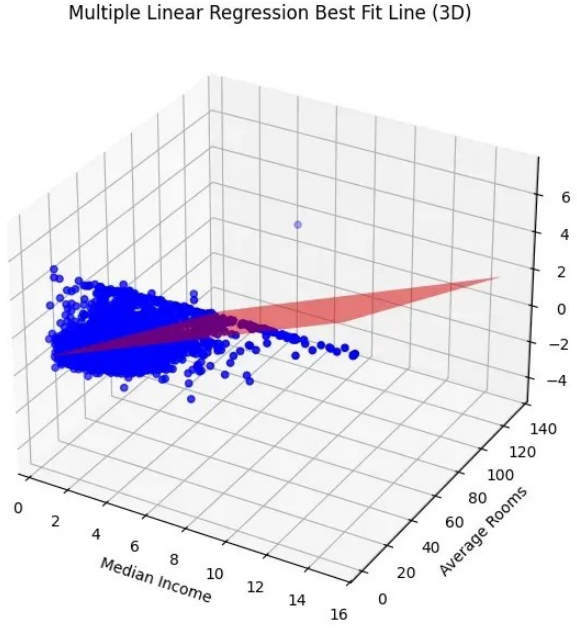



Figura 2. Multiple linear




### *II-C. Función de pérdida*

La función de pérdida mide el error cometido por el modelo en cada muestra (*sample*). Una de sus caracteristicas es el error cuadrático, que penaliza de manera más fuerte los errores

grandes. La función de pérdida de una muestra se denota como $\mathcal{L}_i = (f_{w,b}(\mathbf{x}_i) - y_i)^2$.

Para evaluar el desempeño del modelo en todo el conjunto de datos, se calcula la función de costo (cost function), que es el promedio de la función de pérdida sobre todas las $N$ muestras:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \left( f_{w,b}(\mathbf{x}_i) - y_i \right)^2, \tag{1}$$

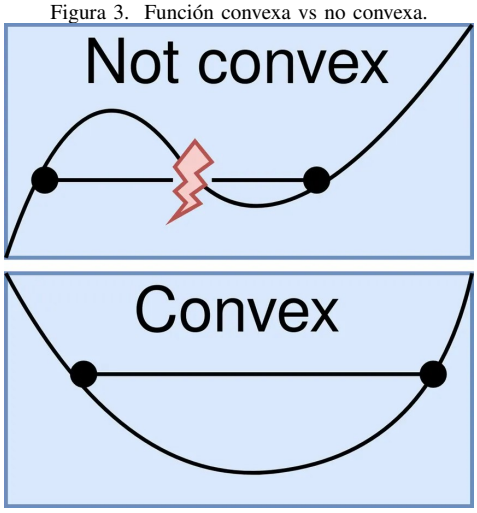
donde $f_{w,b}(\mathbf{x}_i)$ es la predicción del modelo para la muestra $i$, $y_i$ es el valor real, y $N$ es el número total de muestras.

Si logramos minimizar $\mathcal{L}$, reducimos la discrepancia entre las predicciones del modelo y los valores reales, obteniendo así un mejor ajuste. Otra forma de mejorar el modelo es ajustando los valores de $w$ y $b$, evitando el *underfitting*, o modificando el conjunto de datos (*dataset*).

### *II-D. Función convexa vs no convexa*

Al realizar regresiones lineales, como parte de la fórmula está elevada al cuadrado, sabemos que es posible encontrar una solución óptima (mínimo local). Sería ideal siempre que la función sea convexa, porque a diferencia de la no convexa es fácil identificar un mínimo global.

Figura 3. Función convexa vs no convexa.




### *II-E. Descenso del gradiente*

Se propone una analogía sobre que se esta en una montaña muy elevada, se tiene los ojos vendados y la meta es bajar con la menor cantidad de esfuerzo y los más rápido posible. Y la solución es desde el punto inicial, calcular el lado que tiene más pendiente, dar un paso y repetir ese mismo proceso hasta llegar a un punto mínimo de altura.

Matemáticamente, esto se formaliza mediante la regla de actualización:

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

donde $\alpha$ es la tasa de aprendizaje (*learning rate*) y $\nabla f(x_t)$ representa el gradiente de la función en el punto $x_{antiguo}$.

El valor de $\alpha$ es crítico: un valor demasiado grande puede provocar que el algoritmo oscile y no converja, mientras que un valor demasiado pequeño ocasiona una convergencia muy lenta. Para mitigar estos problemas, se suelen emplear estrategias como la búsqueda de una tasa de aprendizaje óptima o el *early stopping*, que detiene el entrenamiento cuando la función de pérdida deja de mejorar significativamente o alcanza un valor aceptable.

**Nota:** Los términos derivada, pendiente y gradiente son equivalentes.

$$\text{Derivada de una constante: } \frac{d}{dx}[c] = 0$$
$$\text{Derivada de una variable: } \frac{d}{dx}[x] = 1$$
$$\text{Derivada de constante por variable: } \frac{d}{dx}[c \cdot x] = c$$
$$\text{Regla de la potencia: } \frac{d}{dx}[x^n] = nx^{n-1}$$
$$\text{Derivada de una suma: } \frac{d}{dx}[f(x) + g(x)] = f'(x) + g'(x)$$
$$\text{Regla del producto: } \frac{d}{dx}[f(x)g(x)] = f'(x)g(x) + f(x)g'(x)$$
$$\text{Derivadas parciales: } \frac{\partial f}{\partial x_i} = \text{derivada de } f \text{ respecto a } x_i$$
$$\text{Ejemplo de parciales: } f(x,y) = x^2y + 3xy^2, \quad \frac{\partial f}{\partial x} = 2xy + 3y^2, \quad \frac{\partial f}{\partial y} = x^2 + 6xy$$

*Pregunta final*

Se concluye la clase con la siguiente pregunta, ¿Porque escoger MSE y no MAE?

**Respuesta:** MAE no es derivable en 0 y nos lleva a errores de cálculo

## REFERENCIAS

[1] S. A. P. Portuguez, “Apuntes de la clase de inteligencia artificial,” Cartago, Costa Rica, agosto 2025, clase del 26 de agosto de 2025.