MACHINE LEARNING HANDS ON!!!

Ph.D Ana Rocío
Cárdenas Maita

CONTENIDO

- Configuración de Python Notebook
- Cargar el dataset de prueba
- Algoritmo de regresión lineal
- Algoritmo de k-means
- Discutir los resultados

OBJETIVO

Entender el procedimento de entrenamiento de un modelo de aprendizaje automático.
Ver las diferencias entre el aprendizaje supervisado y no supervisado.

REGRESIÓN LINEAL

La regresión lineal es un método estadístico utilizado para modelar la relación entre una variable dependiente (respuesta) y una o más variables independientes (predictoras).

Tipos:

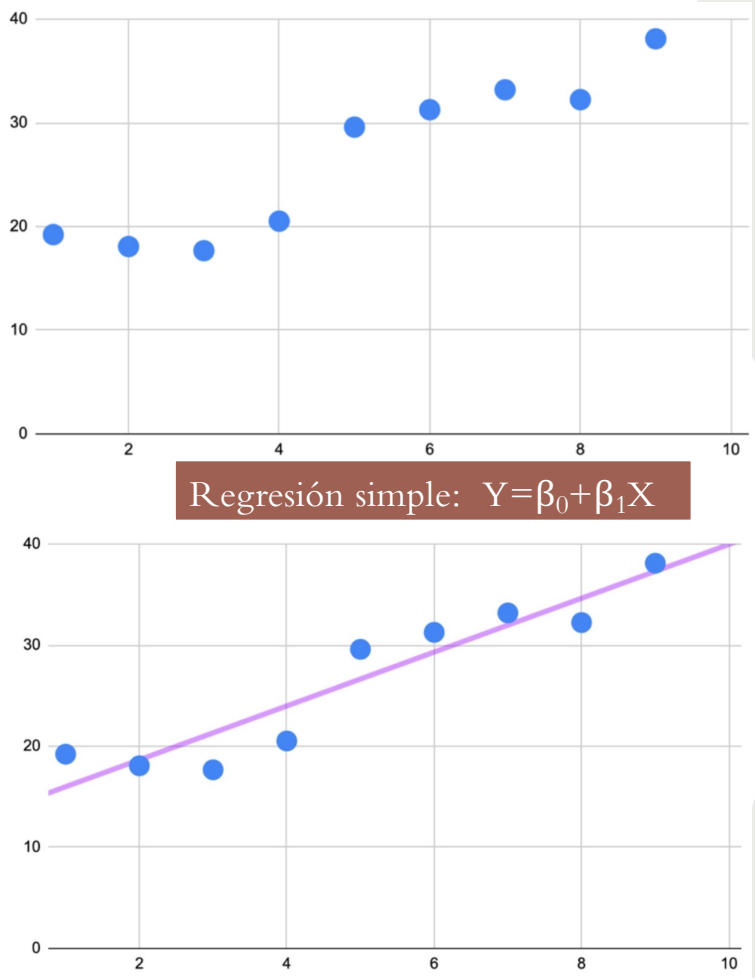
- ***Regresión lineal simple:*** Solo una variable independiente.
  Ejemplo: predecir el peso según la altura.
- ***Regresión lineal múltiple:*** Dos o más variables independientes.
  Ejemplo: predecir el precio de una casa usando tamaño, número de habitaciones, ubicación, etc.

REGRESIÓN LINEAL

$Y=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_nX_n+\varepsilon$

- Y: variable dependiente (lo que queremos predecir)
- $X_1, X_2, \ldots, X_n$: variables independientes
- β_0: intercepto
- $\beta_1, \beta_2, \ldots$: coeficientes que indican el efecto de cada X
- ε: error aleatorio

[Libro: Machine Learning for absolute beginners, Pg. 51]



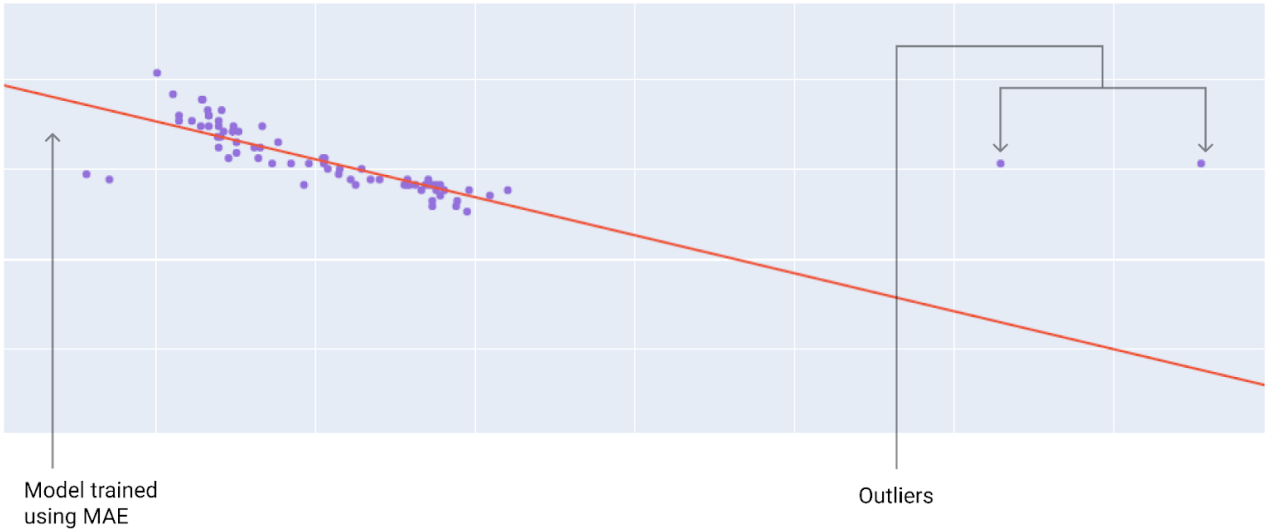
REGRESIÓN LINEAL

Interpretación:

- Los coeficientes indican cuánto cambia Y en promedio cuando una X_i aumenta una unidad.
- Se evalúa la calidad del modelo usando métricas como:
  - R^2: proporción de la varianza explicada por el modelo, ósea qué % de la variabilidad de y se explica por x. Cuanto más cerca de 1, mejor.
  - p-valores: significancia estadística de los coeficientes. Indica si hay evidencia suficiente para afirmar que un coeficiente es significativamente diferente de cero
  - MAE: Promedio del error absoluto. Más interpretable. El modelo está más lejos de los valores atípicos, pero más cerca de la mayoría de los otros puntos de datos.
  - MSE: Penaliza más los errores grandes. Sensible a outliers. El modelo está más cerca de los valores atípicos, pero más lejos de la mayoría de los otros puntos de datos.

REGRESIÓN LINEAL

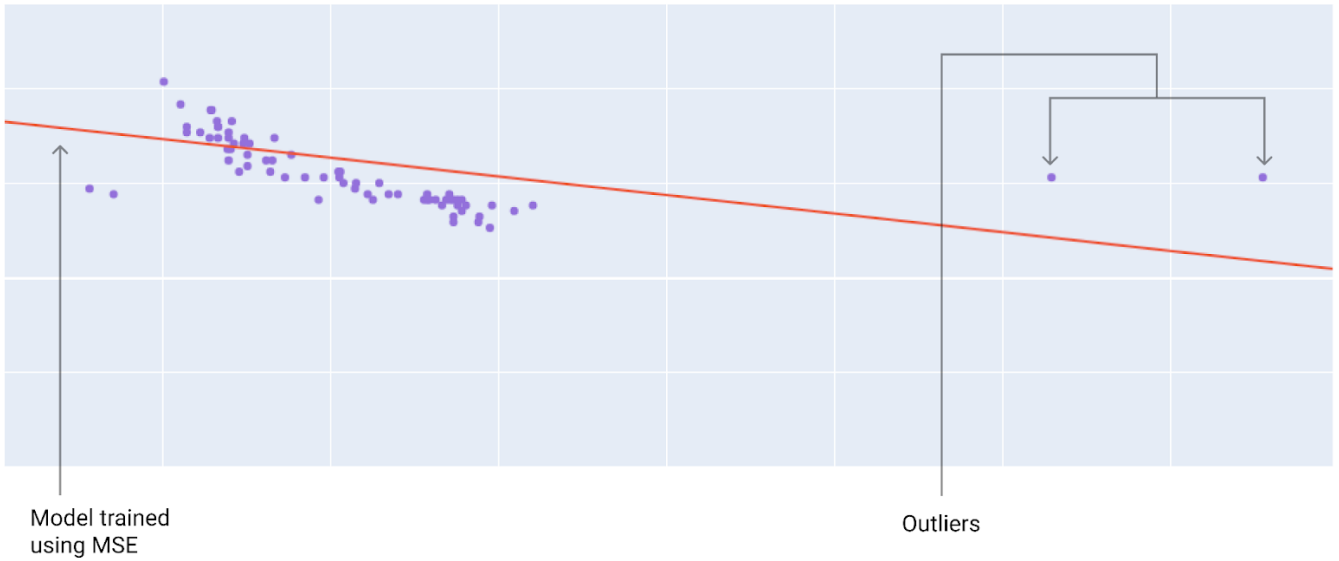
- MAE: Promedio del error absoluto. Más interpretable. El modelo está más lejos de los valores atípicos, pero más cerca de la mayoría de los otros puntos de datos.



[https://developers.google.com/machine-learning/crash-course/linear-regression/loss?hl=es-419]

REGRESIÓN LINEAL

- MSE: Penaliza más los errores grandes. Sensible a outliers. El modelo está más cerca de los valores atípicos, pero más lejos de la mayoría de los otros puntos de datos.



[https://developers.google.com/machine-learning/crash-course/linear-regression/loss?hl=es-419]

NOTEBOOK

AGRUPAMENTO POR PARTIÇÃO

ALGORITMO K-MEANS

- El algoritmo k-means busca una partición que minimice la suma de errores al cuadrado (SSE) entre los objetos de un conjunto de datos y el centroide de sus respectivos grupos
- La SSE se define como:

$$\mathrm{SSE} = \sum_{i=1}^{k} \sum_{x_j \in C_i} d(x_j, \bar{x}_{C_i})^2$$

- donde d(·,·) es la distancia euclidiana y (x_{Ci}) es el centroide de un grupo C_i, calculado como:

$$\bar{x}_{C_i} = \frac{1}{|C_i|} \sum_{x_j \in C_i} x_j$$

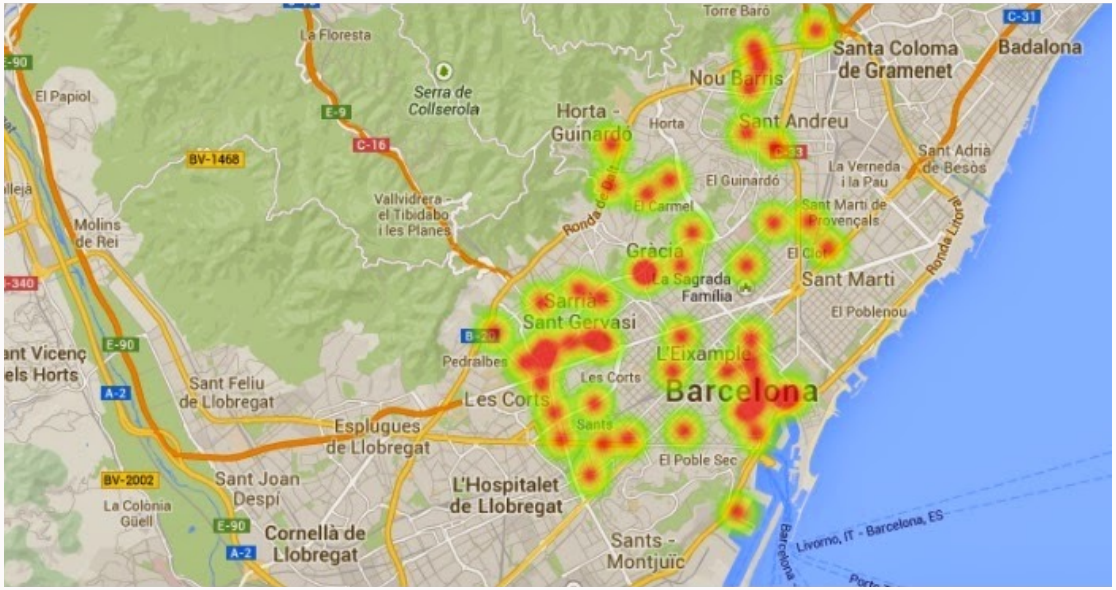
ALGORITMO

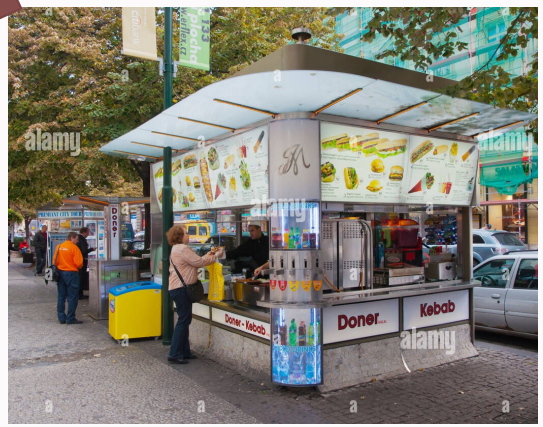
Algoritmo 2: *k-means*.

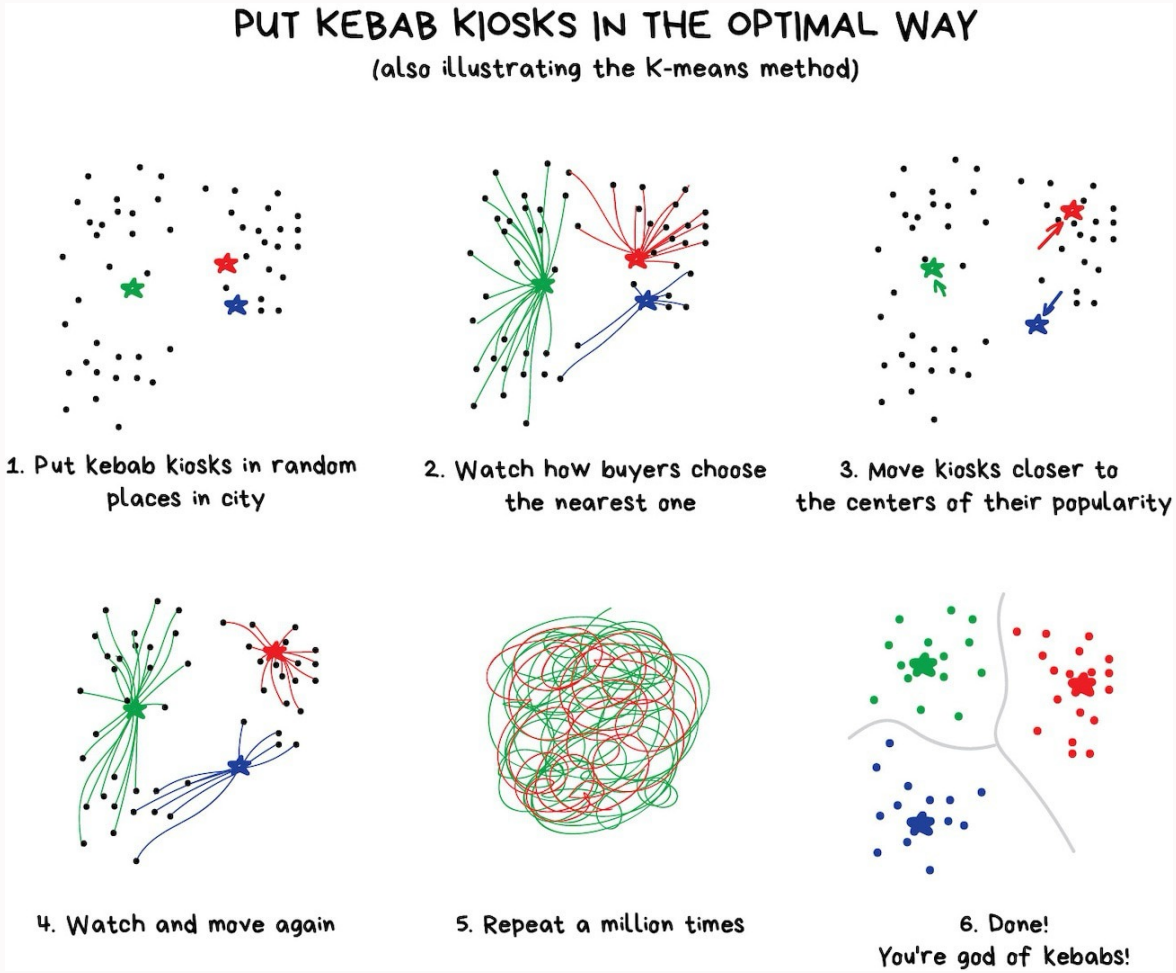
Entrada: conjunto de dados $X \in \mathbb{R}^{n \times m}$ e o número de grupos k
Saída : agrupamento particional de X em k grupos

1 gerar k centróides aleatoriamente;
2 **repita**
3 | calcular a distância entre cada objeto x_j e cada centróide $\bar{x}_{C_i}$;
4 | atribuir cada objeto x_j ao grupo C_i com centróide mais próximo;
5 | recalcular o centróide de cada grupo conforme a Equação (10);
6 **até** *que um critério pré-definido seja atingido ou que os objetos não mudem de grupo*;

EXEMPLO


Torre Baró
La Floresta
Serra de Collserola
El Papiol
Nou Barris
Santa Coloma de Gramenet
Badalona
Horta - Guinardó
Horta
Sant Andreu
Sant Adrià de Besòs
La Verneda i la Pau
El Guinardó
El Carmel
Vallvidrera - el Tibidabo i les Planes
Molins de Rei
Sant Martí de Provençals
Gràcia
El Clot
La Sagrada Família
Sant Martí
Sarrià
Sant Gervasi
El Poblenou
Pedralbes
L'Eixample
Les Corts
Barcelona
Sant Feliu de Llobregat
Les Corts
Esplugues de Llobregat
Sants
El Poble Sec
Sant Joan Despí
L'Hospitalet de Llobregat
La Colònia Güell
Cornellà de Llobregat
Sants - Montjuïc
Livorno, IT - Barcelona, ES
Ronda de Dalt
Ronda Litoral
Barcelona
BV-1468
BV-2002
C-16
C-31
E-9
E-90
A-2
B-20


Doner
Kebab
alamy

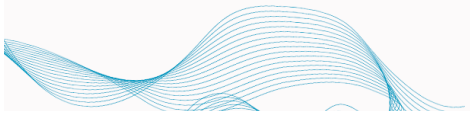

PUT KEBAB KIOSKS IN THE OPTIMAL WAY
(also illustrating the K-means method)
1. Put kebab kiosks in random places in city
2. Watch how buyers choose the nearest one
3. Move kiosks closer to the centers of their popularity
4. Watch and move again
5. Repeat a million times
6. Done!
You're god of kebabs!

EXEMPLO

¿CÓMO ELEGIR K DE K-MEDIAS?

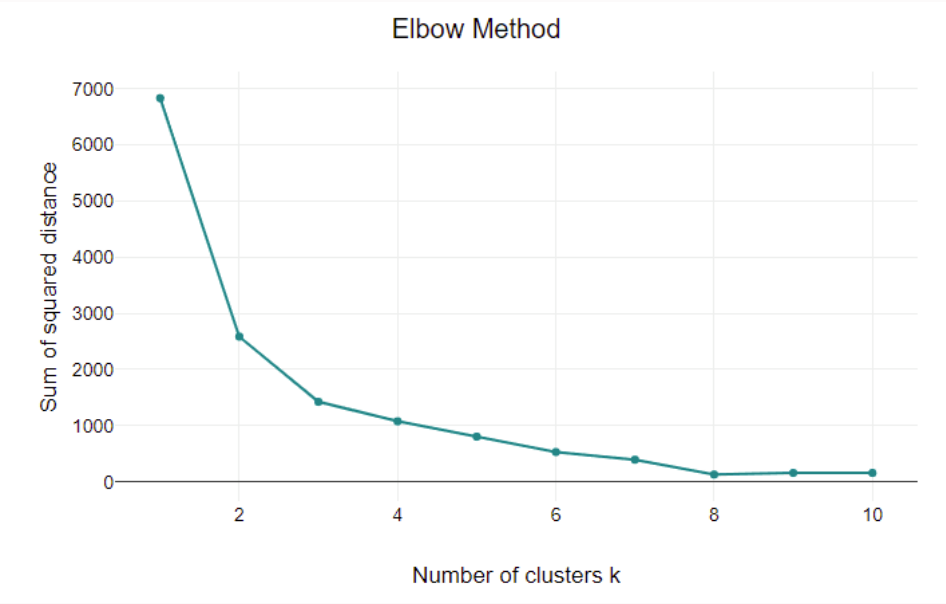
- No existe un método particular para determinar el valor exacto de k.
- Una métrica popular que se usa comúnmente para comparar resultados en numerosos valores k es la distancia promedio en el medio del centroide del clúster y sus puntos de datos.
- Dado que el aumento de los clústeres minimizará la distancia entre los puntos de datos, el aumento del número de clústeres reducirá la distancia entre los puntos de datos a la vez.
  - Expandir k disminuirá la métrica y puede hacerla tan baja como cero, siempre que k sea similar a la cantidad de puntos de datos.
  - Por lo tanto, no puede utilizar esta métrica como un único destino.

  Alternativamente, puede trazar la distancia promedio desde el centroide como una función de k, donde la tasa de disminución cambia bruscamente. Puede dar una respuesta aproximada a k.
  - Es decir, observe cuándo las métricas disminuyen (o crecen) sistemáticamente como k y entonces se debe buscar una discontinuidad en la curva general.



CURVA ELBOW (CODO)

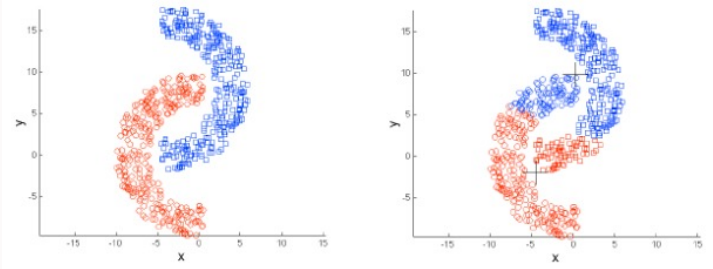
- Con cada nuevo clúster, la varianza total en cada clúster se hace cada vez más pequeña.
- En el caso extremo, cuando hay tantos clústeres como puntos, el resultado es cero.
- Sin embargo, en la mayoría de los casos, la reducción de la variación total se reduce a partir de cierto punto.
- Este punto se utiliza como el número de clúster óptimo.



[https://datatab.net/statistics-calculator/cluster]

PROBLEMAS

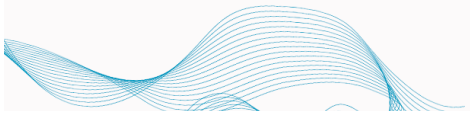
- Es muy susceptible a problemas cuando los cúmulos son de diferentes formatos (generalmente no globulares)



- Dificultad para definir el valor de k.
- Limitado a atributos numéricos.
- Cada elemento debe pertenecer a un solo clúster.
- Partición rígida (sin superposición).

FORMAS DE MEJORAR SU RENDIMIENTO

- Actualización incremental.
- El cálculo de los nuevos centroides no requiere volver a calcular todo de nuevo
- **K-medianas**
- reemplaza los promedios con medianas, es menos sensible a los valores atípicos
- tiene una mayor complejidad computacional debido al paso de clasificación
- **K-medoides**
- reemplaza cada centroide por un objeto representativo del clúster
- Un medoide es el objeto más cercano (en promedio) a los otros objetos del clúster
- Es menos sensible a los valores atípicos, se puede aplicar a atributos categóricos
- Tiene complejidad cuadrática



VEAMOS EL NOTEBOOK

Google Research

PhawAI

Universidad Católica San Pablo | Departamento de Ciencia de la Computación

University of Pittsburgh

TaReCDa

RICE

UTMACH

GRACIAS