Fundamentos de Aprendizaje de Máquina

PhD Jorge Rudas

INFOTEP
INSTITUCIÓN DE EDUCACIÓN SUPERIOR

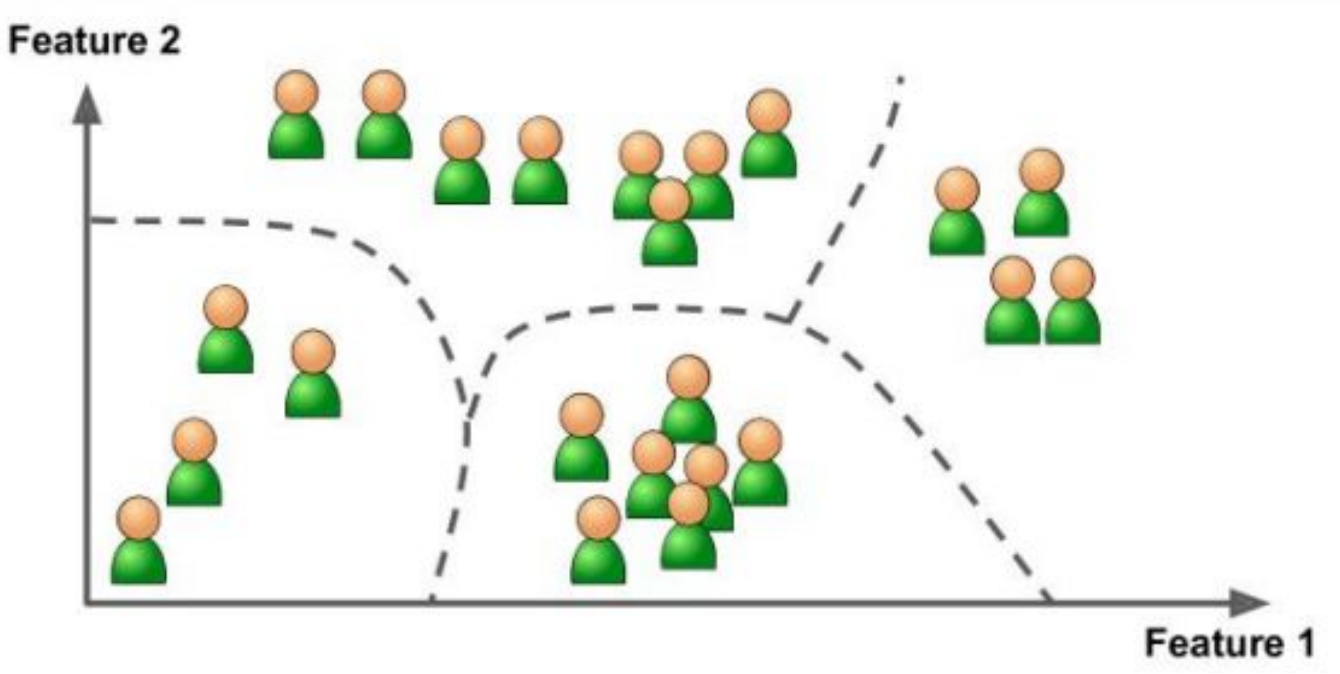
https://github.com/jrudascas/ml-infotep

Aprendizaje NO supervisado

- En el aprendizaje no-supervisado los datos de entrenamiento no están etiquetados. El sistema trata de aprender una organización de los datos
    - **Clustering:** encontrar grupos de objetos o clusters donde los objetos de un mismo cluster sean similares entre sí y distintos de los objetos de otros clusters.
    - **Reducción de dimensionalidad:** reducir la cantidad de atributos de mis objetos combinando atributos redundantes o muy correlacionados (ej: el kilometraje de un auto con los años que tiene)
        - Reducir dimensionalidad se puede hacer antes de entrenar un algoritmo de ML para que funcione de forma más eficiente.

La gran limitación de estas técnicas es que son difíciles de evaluar, pues a diferencia del aprendizaje supervisado, aquí no sabemos cual es la respuesta correcta.

Clustering



El resultado de un proceso de clustering donde se encontraron 4 clusters. Los clusters no eran conocidos antes del entrenamiento.

¿Qué es análisis de clusters?

- Técnica para encontrar grupos de objetos tal que los objetos en un grupo sean similares (o relacionados) entre sí y que sean diferentes (o no relacionados) a los objetos en otros grupos.

- Es una técnica de aprendizaje no-supervisado (no requiere etiquetas para los datos).

Un clustering es una colección de clusters

Tipos de clusterings:

- **Clustering Particional:** Divide los datos en subconjuntos sin traslape (clusters), tal que cada dato está en un solo subconjunto

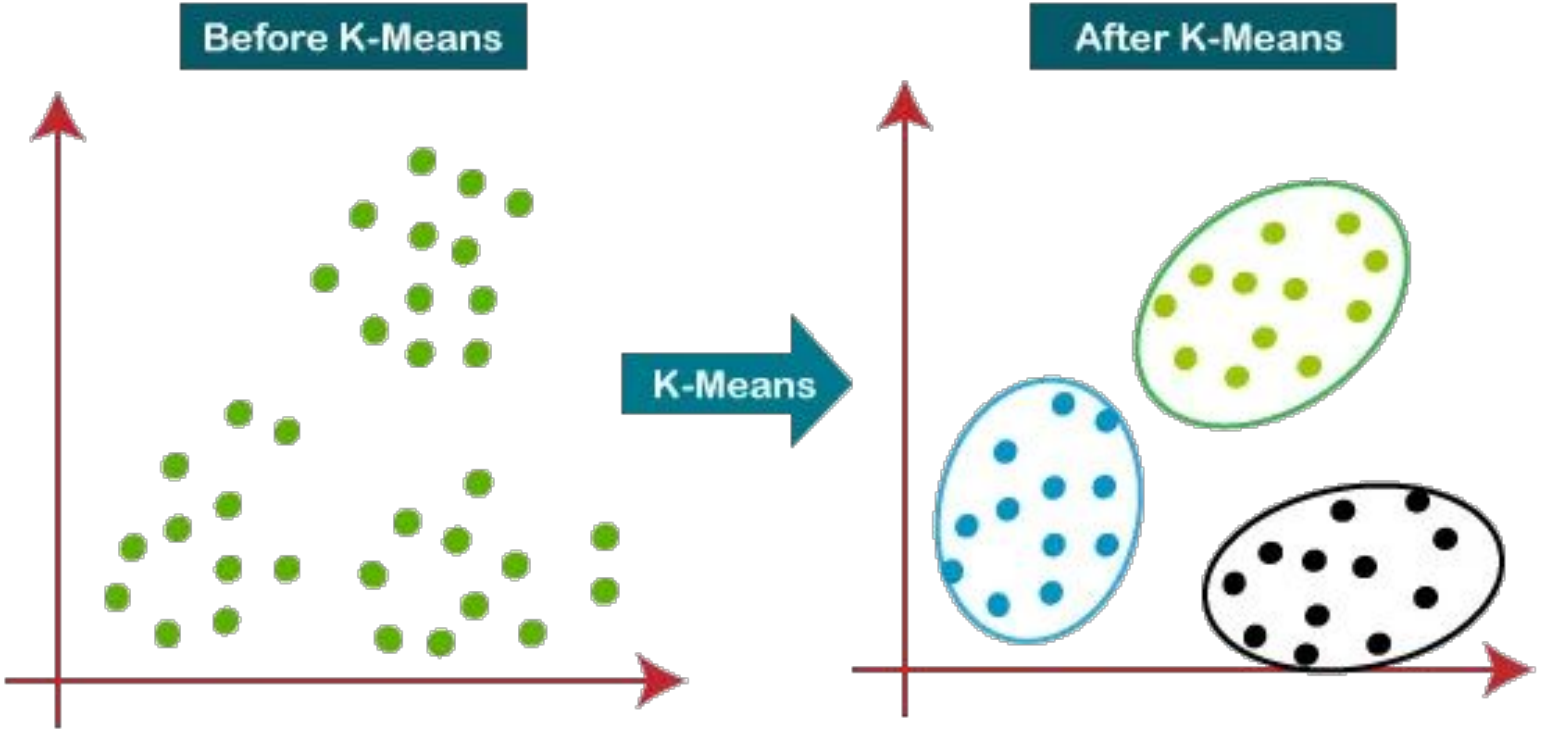
- **Clustering Probabilístico o Difuso:** Cada objeto pertenece a cada cluster con un peso de pertenencia entre 0 y 1

- **Clustering Jerárquico:** Un conjunto de clusters anidados, organizados como un árbol

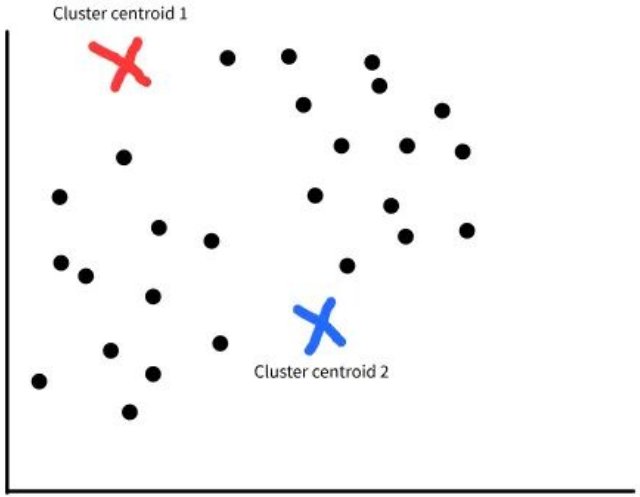
Métodos de clustering

- **K-means**
- **Método jerárquico aglomerativo**
- **DBSCAN**
- **Mixture of Gaussians y algoritmo EM**
- **...**

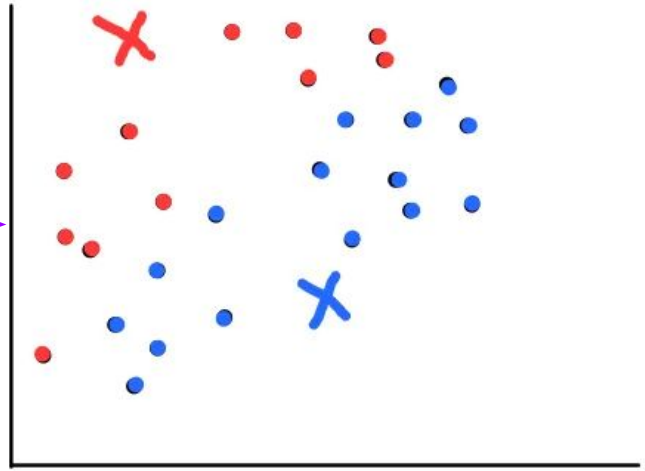
K-means


Before K-Means
K-Means
After K-Means

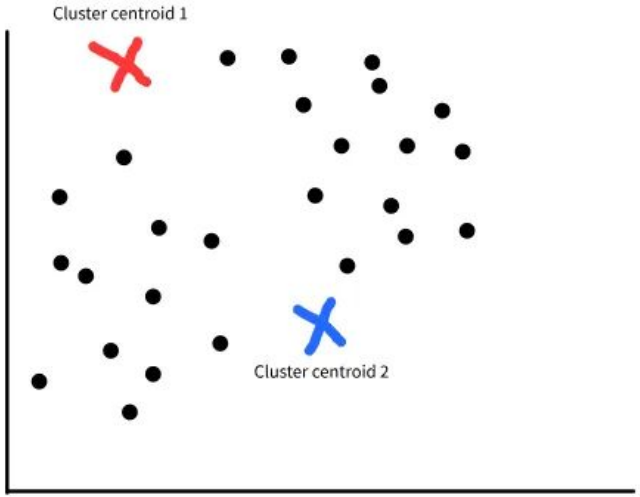
Paso 1: Define un k y genera k centroides de forma aleatoria



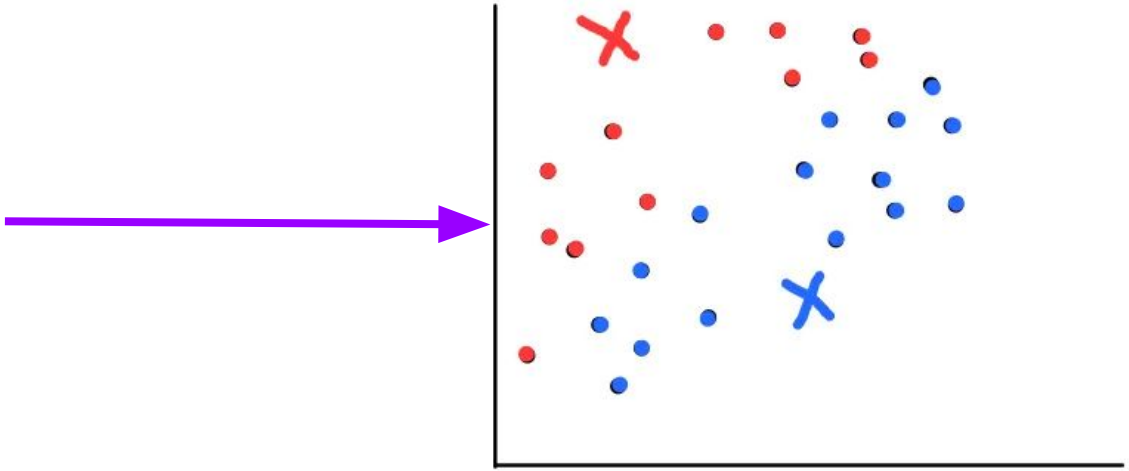
Paso 2: Calcule para cada muestra cuales es el centroide más cercano y asignele la etiqueta de que pertenece a dicho centroide



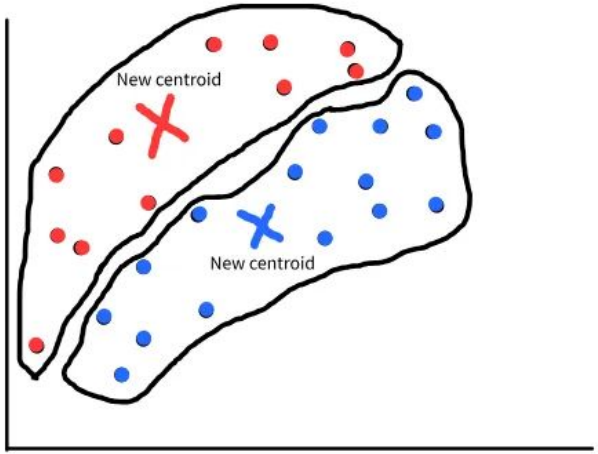
Paso 1: Define un ***k*** y genera ***k centroides*** de forma aleatoria



Paso 2: Calcule para cada muestra cuales es el centroide ***más cercano*** y asignele la etiqueta de que pertenece a dicho centroide



Paso 4: *Recalcule el centroide* para cada nuevo grupo



Repita el paso # 2 hasta ...

```
1. Randomly initialize the k cluster centroids
repeat{
  2. Assign the data points to cluster centroids
  3. Move cluster centroids
}
```

Taller grupal en clase

Dataset
https://www.datos.gov.co/Transporte/Accidentes-de-transito-Palmira-2020/mg8y-amuh/about_data

- Utilice las columnas (hora, jornada, dia de la semana, barrio, hipótesis, clase de siniestro, clase vehículo, marca y matrícula). Preprocesar si es necesario.
- Implemente un algoritmo de k-mean y pruebe sus resultados para k = 2, 3, 4 y 5. Socialice sus resultados. (No se permite el uso de bibliotecas).
- Evalúe los resultados realizando 10, 100, 1000 iteraciones de su algoritmo.

Ejemplo cuantización de colores en imágenes
https://github.com/dkarunakaran/kmeans_clustering/blob/master/kmeans.ipynb

Pruebe con una imágen de alguna locación en Ciénaga

Proponga una estrategia para determinar el mejor valor de k

Se pueden mejorar los resultados si se hace una selección NO aleatoria de los centroides

Homework

Animación 2D