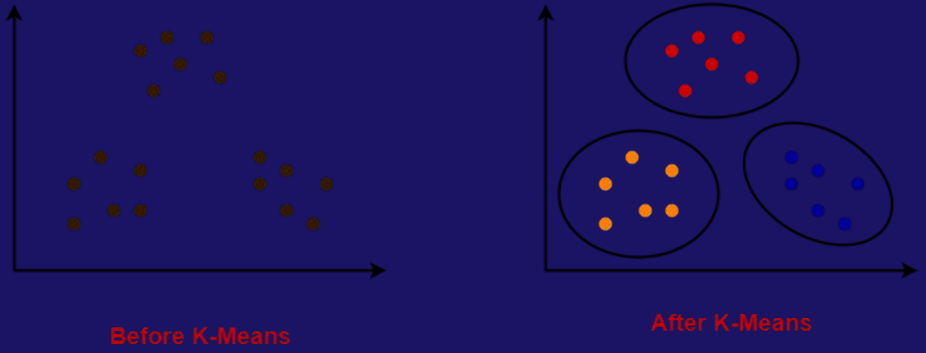
K-MEDIAS

Carlos de Jesús Morales Tovar 1857712
Alejandra Maldonado Ramírez 18844656
Saul Angel Torres Guerrero 1842161
Jesús Alejandro Espinosa Orrante 1941500
Saúl Andrés Rivera Castillo 1857810

Introducción

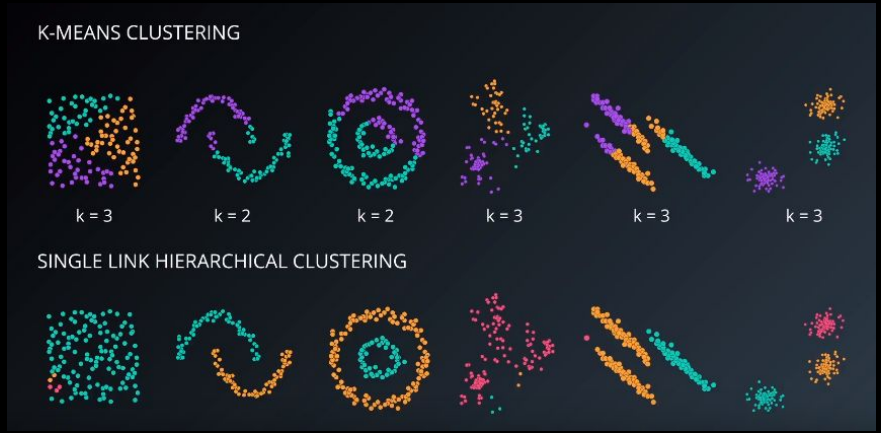
- ¿Qué es?

K-medias es un método de agrupamiento, que también se conoce como clustering, es un algoritmo no supervisado lo que quiere decir que no tiene variable dependiente. Lo que trata de buscar con este método en las observaciones son grupos con características similares, las observaciones en cada grupo tienen que ser similares pero diferentes a los demás grupos.



Aplicaciones

1) Detectar células cancerosas
2) Agrupamiento de palabras
3) Separamiento de personas reales de los bots en redes
4) Determinar el comportamiento de votación de una comunidad
5) Segmentar grupos de personas respecto a sus intereses
6) Clasificación de dígitos



Ventajas y Desventajas

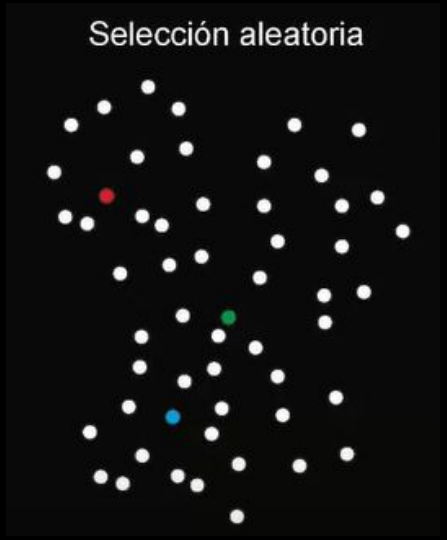
Ventajas

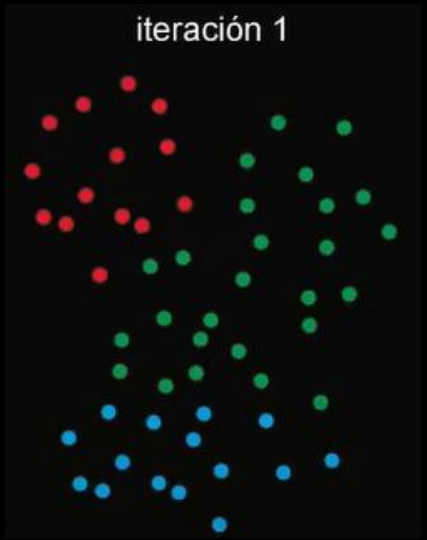
- Es un método rápido
- Ocupa poco almacenamiento

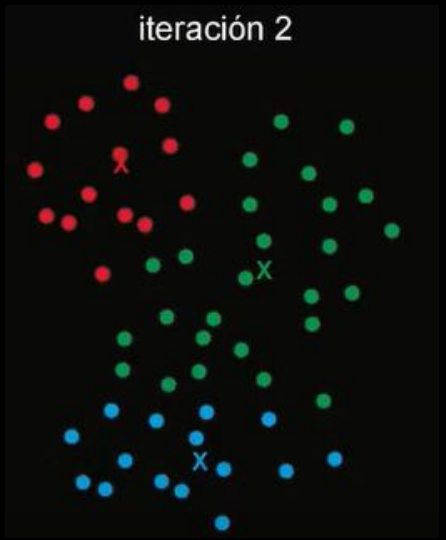
Desventajas

- Se tiene que comprobar cada número de clusters (grupos)
- Débil si hay outliers (datos muy diferentes)

¿Cómo actúa K-Medias?





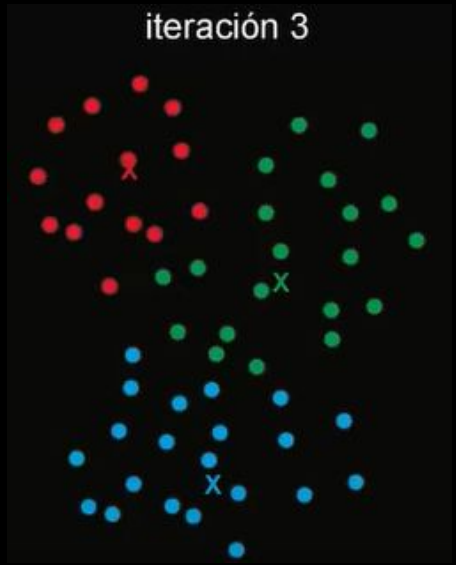


1) Se hace una selección aleatoria de observaciones, tantas como clusters se hayan definido

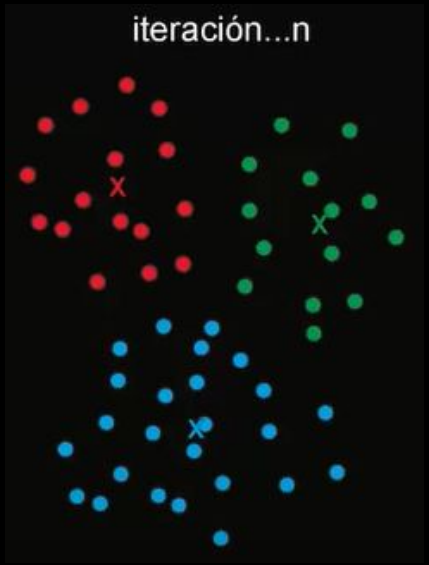
2) Asignar cada observación al punto más cercano

3) Se calculan los centroides de cada uno de los grupos creados

¿Cómo actúa K-Medias?



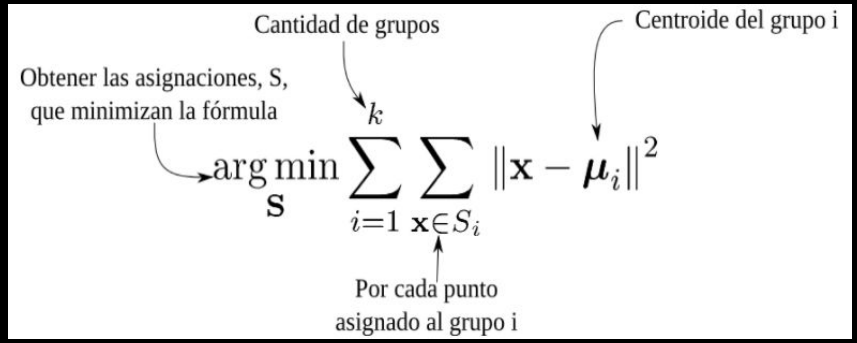
4) Volver a reasignar las observaciones en función de los nuevos centroides



5) Estas iteraciones se repetirán tantas veces sean necesarias

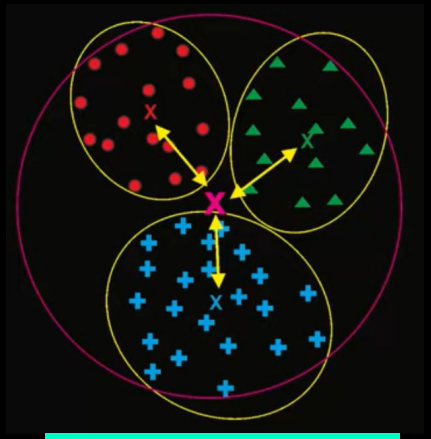
¿Por qué funciona K-Medias?

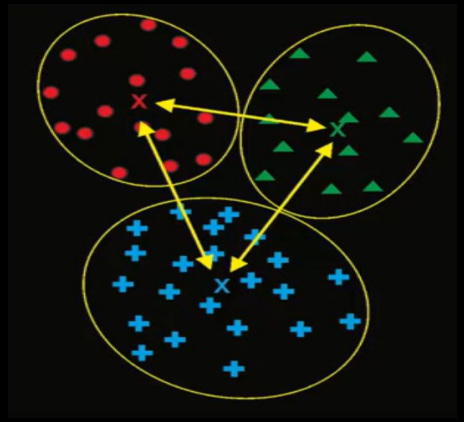
Una medida para indicar cuán bien los centroides representan a los miembros de su grupo es la suma de los errores al cuadrado. K-medias, en cada iteración, intenta reducir el valor de la suma de los errores al cuadrado. La medida consiste en la sumatoria de las distancias al cuadrado de cada observación al centroide de su grupo:

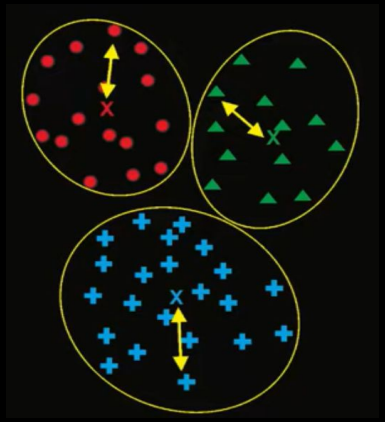


Hallar un mínimo de la función, a pesar de que no se trate del mínimo absoluto, garantiza un agrupamiento en el que los grupos son poco dispersos y se encuentran separados entre sí. El algoritmo es significativamente sensible a los centroides que se seleccionan inicialmente de manera aleatoria.

Evaluación del modelo





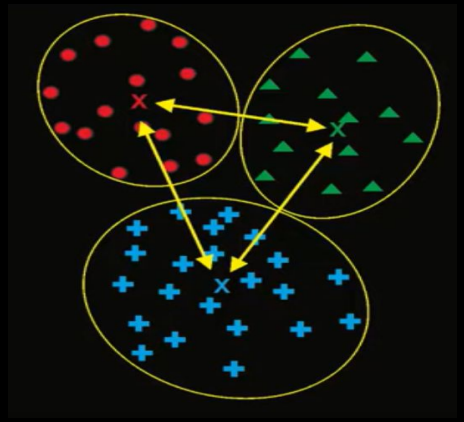


1) La inercia total es la inercia de los grupos respecto a su centroide de acuerdo a sus observaciones

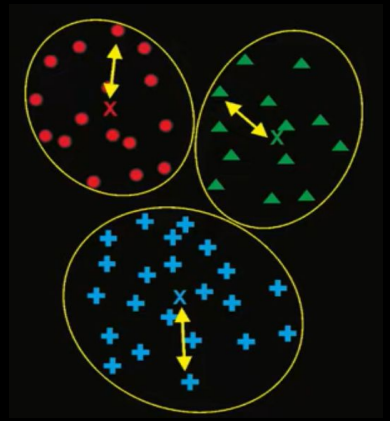
2) La inercia entre grupos procura tomar todos los datos para ser más preciso al momento de realizar las iteraciones.

3) La inercia dentro de los grupos o intra-grupos nos indica las inercias individuales de cada uno de ellos.

Evaluación del modelo



Se busca **maximizar** la variación **inter-cluster**, esto nos asegura una heterogeneidad entre los grupos



y **minimizar** la **intra-cluster**

PYTHON

- Pandas
- Numpy
- Matplotlib.pyplot
- Sklearn.cluster

Librerías

```
from sklearn.cluster import KMeans
```

Comandos

- **k_means=KMeans(n_clusters=n)**
(Definir el número de clusters a formar)
- **k_means.fit(X)**
(Ajustar los datos de entrada al algoritmo de k-medias con n clusters)
- **centroides=k_means.cluster_centers_**
(Obtener los centroides de cada cluster)
- **etiquetas=k_means.labels_**
(Obtener las etiquetas que indican el cluster al cual pertenece cada dato)

R

R ya tiene una potente función incorporada para realizar agrupaciones de K-Medias. La función se llama **kmeans**.

Datos generados por kmeans:

`kmeans(X, k)$clusters`	grupo al cual ha sido clasificado cada individuo
`kmeans(X, k)$centers`	centro de cada grupo
`kmeans(X, k)$withinss`	suma de cuadrados dentro de cada grupo
`kmeans(X, k)$totss`	suma de cuadrados total
`kmeans(X, k)$tot.withinss`	suma de cuadrados total dentro
`kmeans(X, k)$betweenss`	suma de cuadrados entre grupos
`kmeans(X, k)$size`	tamaño de los grupos

Preguntas

- **¿Qué trata de buscar este método en las observaciones?**
  Grupos con características similares
- **Menciona una desventaja que conlleva usar este algoritmo**
  Comprobar cada numero de clusters
- **¿Qué se hace para asegurar heterogeneidad entre los grupos o clusters?**
  Se maximiza la variación inter-cluster
- **Menciona una aplicación en la cual el método K-Medias sea útil**
  Segmentar grupos de personas respecto a sus intereses
- **¿Qué intenta K-medias en cada iteración?**
  Reducir el valor de la suma de los errores al cuadrado

Ejemplo

Bibliografía

- *kmeans*. (s. f.). Unioviedo. Recuperado 5 de septiembre de 2021, de https://www.unioviedo.es/compnum/laboratorios_py/kmeans/kmeans.html
- *Técnicas de clasificación iteratica y jerárquica*. (2018, febrero). Recuperado 5 de septiembre de 2021, de http://ares.inf.um.es/OORteam/pub/clas/VI_clasificacion.html
- Caride, R. (2017, 18 febrero). *Ejemplo básico algoritmo K-means con R studio* [Vídeo]. YouTube. https://www.youtube.com/watch?v=w_aUCJHRv0Y&feature=youtu.be
- Definición | K-medias. (2016). K-medias. http://163.10.22.82/OAS/Agrupamiento_Kmedias/definicin.html