Instrucciones práctica tema 8 – CLUSTERING

1-Implementar el algoritmo kmeans con los siguientes parámetros:

- Distancia: Euclidea/manhatan
- Numero de clusters
- Dataset

  dataset: https://shorturl.at/eqstC

Como ejemplo utilizaremos de entradas un conjunto de datos en el que se analizan rasgos de la personalidad de usuarios de Twitter. 140 personalidades pertenecientes a diferentes áreas: deporte, cantantes, actores, etc.

Características de entrada:

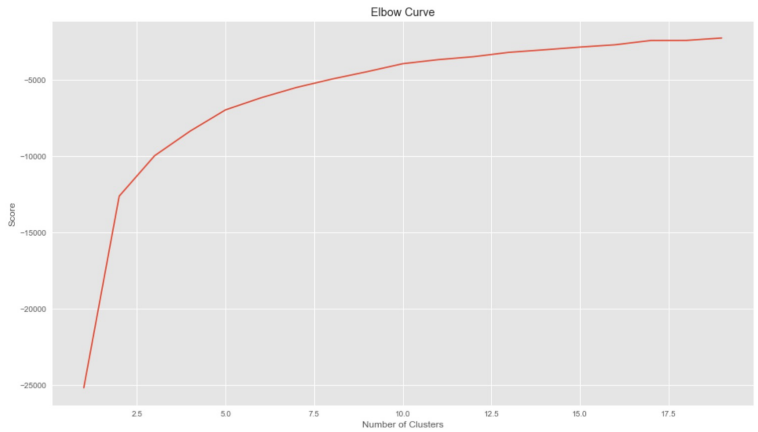
usuario
"op" = Openness to experience – grado de apertura mental a nuevas experiencias
"co" =Conscientiousness – grado de orden, prolijidad, organización
"ex" = Extraversion – grado de timidez, solitario o participación ante el grupo social
"ag" = Agreeableness – grado de empatía con los demás, temperamento
"ne" = Neuroticism, – grado de neuroticismo, nervioso, irritabilidad, seguridad
Wordcount – Cantidad promedio de palabras usadas en sus tweets
Categoria – Actividad laboral del usuario (actor, cantante, etc.)

Utilizaremos el algoritmo K-means para que agrupe estos usuarios por sus similitudes en la personalidad. Si bien tenemos 8 columnas de entrada, sólo utilizaremos *op, ex y ag*

2- Realizar una pequeña experimentación que dadas las dos distancias, calcule y pinte en base a la puntuación de las medidas intra/intercluster (Davies-Bouldin Index)cuál es la mejor elección de clusters.

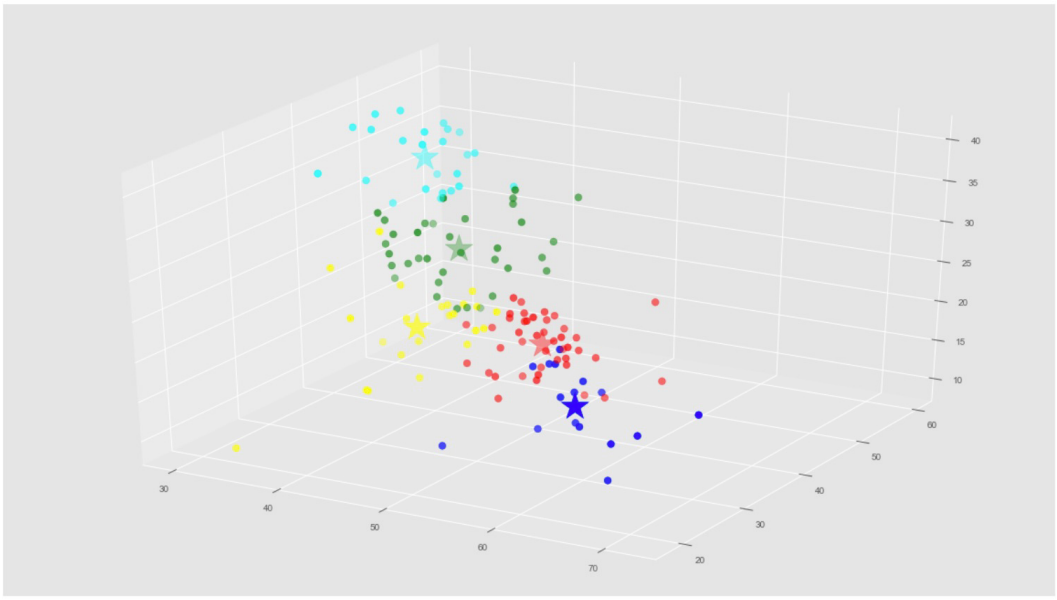
La similitud R para el grupo i con su grupo j más cercano se define como (Si + Sj)/Dij, donde Si es la distancia promedio de cada punto en el grupo i a su centroide. Dij es la distancia entre el centroide del grupo i y j

3- Mostrar el punto de codo, es decir donde la gráfica tiene un punto de inflexión .



4- Elegir a partir de ahora num. Cluster = 5.

Dibujar por pares en distintos colores los clústeres con sus centroides (*op/ex op/ag ex/ag) o en una sóla gráfica 3D (utiliza orto símbolo para los centroides)*



5- Muestra los usuarios de cada cluster.

Rúbrica:

- 5- Puntos 1-2
- 6-7 Puntos 1-2-3
- 8 1-2-3-4-5 Analisis de los resultados
- 9-10- Utilización de otros atributos y medidas de calidad. Explicación de las formas de los clústeres en función de diferentes medidas.