1. ¿Cuáles son las desventajas de usar el método del codo para estimar el número de clusters?

Primera

* Sensibilidad a la forma de los datos: El método del codo puede no funcionar bien si los datos tienen formas de cluster no esféricas o de tamaños muy diferentes.
* No siempre es claro: En algunos casos, el codo en la gráfica no es obvio, lo que dificulta determinar el número óptimo de clusters de manera definitiva.
* Dependencia de la métrica: La elección de la métrica de distancia puede afectar los resultados, lo que hace que la estimación de clusters sea subjetiva.

Segunda

* No siempre es claro y preciso para determinar el número óptimo de clusters.
* Puede generar ambigüedad en la elección del "codo" en la gráfica.
* No funciona bien con clusters de formas y tamaños irregulares.

Tercera

* No siempre es claro en la elección del número óptimo de clusters.
* Puede ser subjetivo, dependiendo de la interpretación visual.
* No funciona bien con clusters de formas no esféricas o de tamaños diferentes.

Cuarta

* El método del codo es subjetivo y a menudo requiere interpretación manual.
* No siempre es efectivo cuando los clusters tienen formas y densidades diferentes.
* Puede no funcionar bien en conjuntos de datos con dimensiones muy altas.
* No proporciona una métrica numérica clara para la elección de k.

Definitiva:

* El método del codo puede no funcionar bien si los datos tienen formas de cluster no esféricas o de tamaños muy diferentes, el método asume este tipo de distribución.
* En algunos casos, el codo en la gráfica no es obvio, es muy subjetivo la interpretación visual
* La métrica de distancia puede afectar los resultados, lo que hace que la estimación de clusters sea subjetiva, si se usa una distancia euclideana o manhattan, etc.
* Puede no funcionar bien en conjuntos de datos con dimensiones muy altas, ya que el costo computacional que calcular las distancias se vuelve mas pesado.

1. ¿En qué consiste DBSCAN?

Es un algoritmo de clustering que se basa en la densidad de puntos en el espacio de datos. Funciona de la siguiente manera:

Primera

* Define un punto central y un radio epsilon (ε).
* Busca todos los puntos dentro de ese radio epsilon alrededor del punto central.
* Si el número de puntos dentro de ese radio supera un umbral predefinido, se forma un cluster.
* Repite este proceso para todos los puntos, creando clusters de diferentes formas y tamaños.
* Los puntos que quedan solos o no están lo suficientemente cerca de ningún cluster se consideran ruido.

Segunda

* DBSCAN es un algoritmo de clustering que agrupa puntos de datos en función de su densidad en el espacio.
* Identifica automáticamente el número de clusters sin necesidad de especificar K de antemano.
* Agrupa puntos cercanos en clusters y puede manejar clusters de formas y tamaños variados.
* También detecta puntos ruidosos como outliers.

Tercera

* DBSCAN es un algoritmo de clustering basado en densidad.
* Agrupa puntos de datos en función de la densidad de los mismos en lugar de la distancia euclidiana.
* Identifica clusters de cualquier forma y es capaz de detectar puntos de ruido.

Cuarta

* DBSCAN es un algoritmo de clustering que se basa en la densidad de puntos en el espacio de datos.
* En lugar de requerir que se defina el número de clusters de antemano, DBSCAN detecta clusters automáticamente.
* Agrupa puntos que están cerca unos de otros en regiones densas, ignorando áreas menos densas.
* Identifica puntos ruidosos como outliers.

Definitiva:

* No asume distancias como k-means o k-medoides, si no que agrupa los datos en función de la densidad. Crea radios de convergencia en los datos y los datos que se encuentren dentro de este umbral, pertenecen a dicho cluster.
* Esto permite tener cluster de diferentes tamaños y formas.
* Identifica puntos ruidosos como outliers.

1. ¿Qué algoritmo podemos usar cuando no podemos pre-definir K?

Primero

Hierarchical Clustering es un método que crea una jerarquía de clusters desde los datos originales hasta un solo cluster o puntos individuales. Luego, puedes elegir el número de clusters observando la estructura jerárquica a través de dendrogramas y cortando la jerarquía en el nivel que mejor se adapte a tus necesidades, lo que lo hace adecuado para casos en los que K no es conocido de antemano.

Segundo

DBSCAN se adapta de manera dinámica a la estructura de densidad de los datos, lo que lo hace ideal para datos con clusters de formas y tamaños desconocidos o variables.

Tercero

* Cuando no podemos predefinir el número de clusters (K), DBSCAN es una excelente opción debido a su capacidad para detectar clusters de forma automática basándose en la densidad de los datos.
* Otra opción es el "cluster espectral," que puede encontrar clusters sin requerir una predefinición de K, aunque su aplicación y resultados pueden ser sensibles a los parámetros y la calidad de los datos.

Cuarto

* Cuando no podemos predefinir el número de clusters (K), algoritmos como DBSCAN son una excelente opción, ya que no requieren una especificación previa de K.
* Otro enfoque útil es el clustering espectral, que identifica clusters basados en la estructura de similitud entre los datos sin necesidad de definir un valor de K.

Estos algoritmos son particularmente útiles cuando no tenemos información previa sobre la cantidad de clusters en nuestros datos y cuando los clusters tienen formas y tamaños irregulares.

Definitiva:

DBSCAN como no asume forma de los datos si no que se adapta densidad de los datos, lo que lo hace ideal para datos con clusters de formas y tamaños desconocidos o variables.

# Guion

Las desventajas de usar el método del codo es que este no funciona bien para datos con forma no esféricas, el método por defecto asume una forma esférica (hiperesferica si queremos generalizar), al ser un método visual es muy subjetivo el punto de inflexión del codo, ya que el ojo humano se puede equivocar, no funciona bien para datos de alta dimensionalidad debido a que los cálculos de distancia se vuelven muy pesados computacionalmente y tambien puede ser subjetivo según el método de distancias que se utilice, da un valor diferente si es euclideana, manhattan, etc.

El DBSCAN es un método que no asume distancias entre los datos como k-means o k-medoides, si no que agrupa losdatos en función de la misma densidad de los datos, este método crea radios deconvergencia en los datos y los puntos que se encuentren dentro de este umbral de convergencia pertenecen a dicho cluster, esto permite que los cluster tengan tamañaos y forma diferentes e identifica puntos de mucho ruido como atípicos.

Si no se puede definir un numero de cluster claro, el mejor método de los que vimos es el DBSCAN, ya como no asume formas en los datos, si no que se adapta a la densidad de los mismo, lo hace ideal para obtener datos con cluster de formas y tamaños desconocidos. Ya existen otro como el cluster jerarquico, mean shift o cluster aglomerativo, que tampoco necesitan un valor previo de k.