Compara y contrasta dos algoritmos diferentes de agrupamiento. ¿Cuándo usaría uno y cuándo la otra?

k means es util cuando se tiene una idea clara del numero de clusters que se desea crear y cuando los datos tienen una distribucion uniforme en el espacio.

dbscan es util cuando no se tiene una idea clara del numero de clusters y cuando los datos tienen una distribucion no uniforme en el espacio.

Explica las diferencias entre k-medias y agrupamiento jerárquico

k means agrupa los clusters separados por centroides, mientras que el agrupamiento jerarquico construye una jerarquia de clusters a partir de los datos. esto lo hace de dos modos: aglomerativo o divisivo.

aglomerativo comienza con cada punto como un cluster y fusiona los clusters mas cercano entre si para formar cluster mas grandes.

divisivo comienza con todos los puntos en un solo cluster y divide el cluster en dos o mas cluster pequeños.

Explica brevemente cómo funciona el algoritmo de agrupamiento k-medias.

selecciona el numero de cluster a crear, luego inicializa los centroides aleatorios para cada uno de los clusters y asigna cada punto de datos al cluster mas cercano segun la distancia euclidiana. finalmente recalcula los centroides de cada cluster como la media de todos los puntos de datos que pertenecen a ese cluster

¿Cuál es un caso común para el agrupamiento k-medias?

Segmentacion de mercado en grupos de clientes homogeneos segun sus preferencias, comportamientos y caracteristicas.

¿Por qué es difícil identificar el número “ideal” de clústeres en un conjunto de datos utilizando la agrupación k-medias?

Los factores principales son

No existe un enfoque unico para determinar el número óptimo de clusters que se deben utilizar para un conjunto de datos determinado, algunos metodos como el de codo o silueta no son infalibles.

La eleccion del numero de clusters puede ser subjetiva y depende del conocimiento previo del analista de datos, diferentes analistas pueden llegar a conclusiones diferentes sobre el numero optimo de clusters.

La calidad de los resultados puede depender de la calidad de los datos, como valores atipicos o caracteristicas que no son relevantes.

Los resultados pueden ser sensibles a la inicializacion de los centroides puesto que son aleatorios, por lo tanto se pueden necesitar multiples ejecuciones del algoritmo con diferentes inicializaciones para obtener una mejor comprension del numero optimo de clusters.

¿Cuál es una heurística para seleccionar "k" para la agrupación de k-medias?

Existen diversas formas para definir K:

Metodo del codo, este implica trazar la suma de las distancias cuadradas intracluster SSW en funcion del numero de clusters. El numero optimo de clusters se considera aquiel en el que la mejora de la SSW disminuye drasticamente y se forma un codo en el grafico.

Metodo de la silueta, este metodo implica calcular el valor de la silueta para diferentes valores de K, el valor de la silueta mide la calidad de la agrupacion, donde los valores mas cercanos a 1 indican una buena separacion entre los clusters.

Conocimiento del dominio. En algunos casos el numero optimo de cluster puede ser conocido de antemano en funcion del conocimiento del dominio y la naturaleza de los datos.

¿Cuál es la diferencia entre aprendizaje supervisado y no supervisado?

El aprendizaje supervisado se refiere a un enfoque de aprendizaje automatico en el que el modelo se entrena utilizando un conjunto de datos etiquetado. est utilizado para clasificar o para regresiones.

El aprendizaje no supervisado implica entrenar un modelo utilizando un conjunto de datos no etiquetado, es decir, los datos no tienen una variable objetivo conocida. El modelo debe identificar patrones y relaciones en los datos por si mismo.

¿Cuál es la diferencia entre k-medias y k vecinos más cercanos?

Kmeans es un algoritmo de agrupamiento no supervisado que se utiliza para dividir un conjunto de datos en grupos o clusters basados en similitudes en las caracteristicas.

KNN es un algoritmo de clasificacion supervisado que se utiliza para predecir la clase de un punto de datos basado en la clase de sus vecinos cercanos.