**BDSCAN – ALGORITMO DE TIPO NO SUPERVISADO**

Presentado por: Lorena Lugo

Es una de las técnicas de clústeres más populares y se refiere a la agrupación espacial de aplicaciones con ruido basada en la densidad, que puede ser utilizado para identificar clústeres de cualquier forma en un conjunto de datos con valores atípicos.

Fue propuesto en los años 90’s, por un grupo de personas enfocadas en bases de datos y minería de datos y se deriva de un método intuitivo de agrupamiento humano, y la idea consiste en que para cada punto de un clúster, la vecindad de un radio dado tiene que contener al menos un número mínimo de puntos.

El algoritmo básicamente requiere de dos parámetros:

1. **Épsilon (eps):** Especifica lo cerca que deben estar los puntos entre sí para ser considerados parte de un clúster, es decir que, si la distancia entre dos puntos es menor o igual a este valor de épsilon, estos dos puntos se consideran vecinos.
2. **Puntos mínimos (minPts):** El número mínimo de puntos para formar una región densa.

Hay tres tipos de puntos de datos:

* **Punto de núcleo:** Un punto, es un punto de núcleo si tiene más de un número especificado de puntos mínimos dentro de un radio de épsilon a su alrededor.

Siempre pertenece a una región densa

* **Punto de borde:** Un punto, es un punto de borde si tiene menos de puntos mínimos dentro de épsilon, pero se encuentra en la vecindad de un punto de núcleo.
* **Punto de ruido**: Cualquier punto que no sea un punto de núcleo o de borde.

También existen dos conceptos que son necesarios para entender el algoritmo DBSCAN:

* **Borde de densidad:** Si p y q son puntos centrales y la distancia entre p y q es menor o igual a épsilon entonces podemos conectar el vértice de p y q en un gráfico y llamarlo borde de densidad.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

* **Densidad de puntos conectados:** Se dice que dos puntos p y q son puntos de conexión de densidad si tanto p como q son puntos de núcleo y existe una trayectoria formada por los bordes de densidad que conectan con el punto p y q.

Imagen en blanco y negro

Descripción generada automáticamente con confianza media

**PASOS DEL ALGORTIMO DBSCAN**

**ELECCIÓN DE PUNTOS MÍNIMOS**

1. Una de las cosas más importantes que hace los puntos mínimos es eliminar los puntos atípicos.
2. La regla para tomar puntos mínimos es tomar siempre sus mínimos para ser mayor o igual a la dimensionalidad del conjunto de datos.
3. Usualmente, las personas que trabajan con DBSCAN toman el punto mínimo dos veces de la dimensionalidad de los datos.
4. Si el conjunto de datos es más ruidoso, se debe elegir u n valor mayor de puntos mínimos.
5. Al elegir los puntos mínimos, depende mucho del conocimiento de dominio.

**DETERMINAR ÉPSILON**

El método consiste en calcular las distancias vecinas más cercanas en una matriz de puntos. Se debe calcular el promedio de las distancias de cada punto a sus vecinos más cercanos. El valor de k será especificado por el usuario y corresponde a los puntos mínimos.

Posteriormente, las k distancias se trazan en orden ascendente, el objetivo es determinar la rodilla, que corresponde al parámetro épsilon óptimo. Una rodilla corresponde a un umbral en el que se produce un cambio brusco a lo largo de la curva de distancia k:

Imagen que contiene agua, barco, tabla, grande

Descripción generada automáticamente

**VENTAJAS DEL ALGORITMO DBSCAN**

* A diferencia de K-means, DBSCAN no requiere que el usuario especifique el número de clústeres que se generarán.
* DBSCAN puede encontrar cualquier forma de clúster, el clúster no tiene que ser circular.
* Puede manejar bien el ruido y los valores atípicos.
* Es excelente para separar clústeres de alta densidad frente a clústeres de baja densidad dentro de un conjunto de datos determinados.

**DESVENTAJAS DEL ALGORITMO**

* Si la base de datos tiene puntos de datos que forman clústeres de densidad variable, entonces DBSCAN no puede agrupar bien los puntos de datos, ya que el agrupamiento depende del parámetro épsilon y puntos mínimos, no se pueden seleccionar por separado para todos los clústeres.
* Si los datos y características no son bien entendidos por un experto en dominios, entonces configurar épsilon y los puntos mínimos podría ser complicado y podría necesitar comparaciones para varias iteraciones con diferentes valores de épsilon y puntos mínimos.
* Es extremadamente sensible a los hiperparámetros. Un cambio en los hiperparámetros puede llevar a un cambio drástico en los resultados.

**APLICACIONES REALES DE DBSCAN – EJEMPLOS**

* **Recomendación de productos a usuarios que usan comercio electrónico:**

Se pueden mejorar las ventas de un establecimiento que use comercio electrónico recomendando productos relevantes a sus clientes. Esto se puede realizar basándose en un conjunto de datos para predecir cuál sería un producto de interés para un cliente específico.

Usando el algoritmo DBSCAN se pueden encontrar los clústeres basados en los productos más comprados por los usuarios. Usando estos grupos, se pueden encontrar similitudes entre los clientes, por ejemplo, si el usuario A compró una bicicleta, un uniforme deportivo y un par de tenis y el usuario B compró una bicicleta y un par de tenis, se le puede recomendar un uniforme deportivo.

* **Investigaciones para enfermedades cancerígenas**

Los investigadores utilizaban DBSCAN para segregar genes de un con junto de datos de genes que tenían la posibilidad de padecer cáncer.