DBSCAN

VICKY JOHANNA ROJAS BARBOSA

KEY CODE – SENA

BOOT CAMP DATA SCIENCE

BOGOTÁ

2021

DBSCAN

DBSCAN Density Based Spatial Clustering of Applications with Noise (Ester et al., KDD’1996) Es un algoritmo de clúster o agrupamiento basado en la densidad que puede ser utilizado para identificar clústeres de cualquier forma en un conjunto de datos que contiene ruido y valores atípicos, se deriva de un método intuitivo de agrupamiento humano.

**CARACTERÍSTICAS:**

Requiere básicamente 2 parámetros:

* **ÉPSLISON (**eps): Detalla lo cerca que deben estar los puntos entre sí; se consideran puntos vecinos si la distancia es menor o igual al valor eps.
* **PUNTOS MINIMOS** (minPts): número mínimo de puntos para formar una región densa, se necesitan mínimo 5 puntos para formar una región densa.

**Densidad:**

* Número de puntos en un radio específico.

**Punto de núcleo:**

* Se considera punto de núcleo si tiene un numero mayor de puntos especificados.
* Siempre pertenece a una región densa.

**Punto de borde** (border):

* Es cuando tiene menos de los puntos especificados.
* Frontera

**Punto de ruido (**noise):

* Cualquier punto que no sea de núcleo ni de ruido.
* No forma parte de un Cluster.

**Puntos Core:**

* Puntos interiores de un cluster.
* Tienen al menos un número mínimo de puntos MinPts en su vecindario de radio eps.

**METODOLOGÍA:**

* El algoritmo comienza con un punto arbitrario que no ha sido visitado y la información de su vecindario se recupera desde el parámetro de épsilon.
* Si es punto contiene puntos mínimos dentro del vecindario épsilon, se inicia la formación de clústeres. De lo contrario, el punto se etiqueta como ruido. Este punto se puede encontrar más tarde dentro de la vecindad de épsilon de un punto diferente y, por lo tanto, se pude hacer parte del clúster. El concepto de densidad alcanzable y los puntos de conexión de densidad son importantes aquí.
* Si se encuentra que un punto es un punto de núcleo, entonces los puntos dentro del vecindario de épsilon también son parte del grupo. Así que todos los puntos que se encuentran dentro del vecindario de épsilon se agregan, junto con su propio vecindario épsilon, si también son puntos de núcleo.
* El proceso anterior continúa hasta que se encuentra completamente el clúster conectado a la densidad.
* El proceso se reinicia con un nuevo punto que puede ser parte de un nuevo clúster o etiquetado como ruido.

**Usos del algoritmo DBSCAN**

* Minería de datos.
* Agrupación en las aplicaciones.
* Detección de anomalías.
* Acelera las consultas en bases de datos.

**EJEMPLO DE USO ALGORITMO DBSCAN**

En este ejemplo tenemos un min\_sample = 3 y un épsilon = 0.4. Se han formado 2 clusters (rojos y verdes), hay otro punto que es alcanzable por el cluster rojo (azul), y un último punto que es ruido (naranja), porque ninguno de los núcleos con su radio ha conseguido alcanzarlos

Un dibujo de una persona

Descripción generada automáticamente con confianza media