DBSCAN

DBSCAN

DBSCAN

Density-based spatial clustering of applications with noise (DBSCAN).

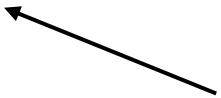
Es un algoritmo de agrupamiento basado en densidad.

El clustering basado en densidad localiza regiones de alta densidad que están separadas una de la otra por regiones de baja densidad.

Densidad, en este contexto, se define como el número de puntos dentro de un radio especificado.

DBSCAN es especialmente eficaz para tareas como la identificación de clases en un contexto espacial.

Una cualidad de DBSCAN es que puede encontrar cualquier forma arbitraria de cluster sin verse afectado por el ruido.



Outliers!!

DBSCAN

DBSCAN trabaja en la idea de que si un punto en particular pertenece a un cluster, debería estar cerca de un montón de otros puntos en ese cluster.

Funciona en base a 2 parámetros: Radio y Puntos Mínimos.

El radio, R, determina una longitud específica que, si incluye suficientes puntos dentro de él, lo llamamos de “Área densa”.

Los Puntos Mínimos, M, determina la cantidad mínima de datos que queremos alrededor de otra observación para definir un cluster.

Por ejemplo, definamos el radio de 2 unidades y fijemos que el punto mínimo sea de 6, incluyendo el punto de interés.

DBSCAN

Para ver cómo funciona DBSCAN, tenemos que determinar el tipo de puntos.

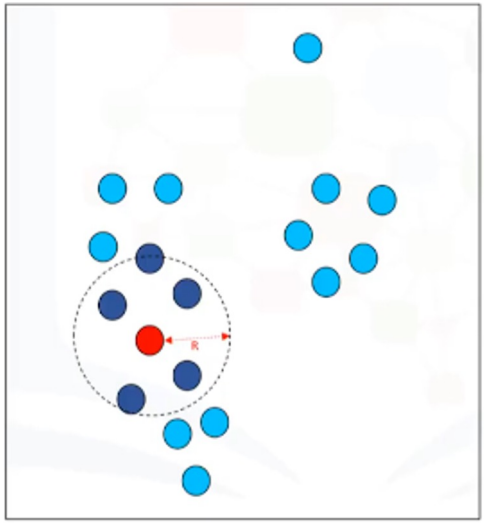
Cada punto de nuestro conjunto de datos puede ser un punto central (core), fronterizo (border) o atípico (outlier).

Toda la idea detrás del algoritmo DBSCAN es visitar cada punto, y encontrar su tipo. A continuación, agrupamos los puntos como clusters en función de sus tipos.

Dado un punto al azar. Primero revisamos si un punto de datos es core.

Un punto de datos es core si, dentro de la R-vecindad de este punto, hay al menos M puntos.

Por ejemplo, ya que hay 6 puntos en el vecindario de 2 centímetros del punto rojo, marcamos este punto como un punto central.



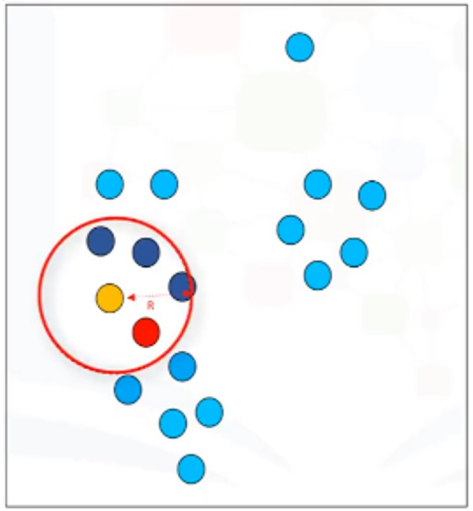
DBSCAN

Veamos otro punto. Este sólo tiene 5 puntos en su vecindario, incluyendo el punto amarillo.

Entonces, se trata de un punto “border”. Un punto de datos es un punto border si cumple estas condiciones:

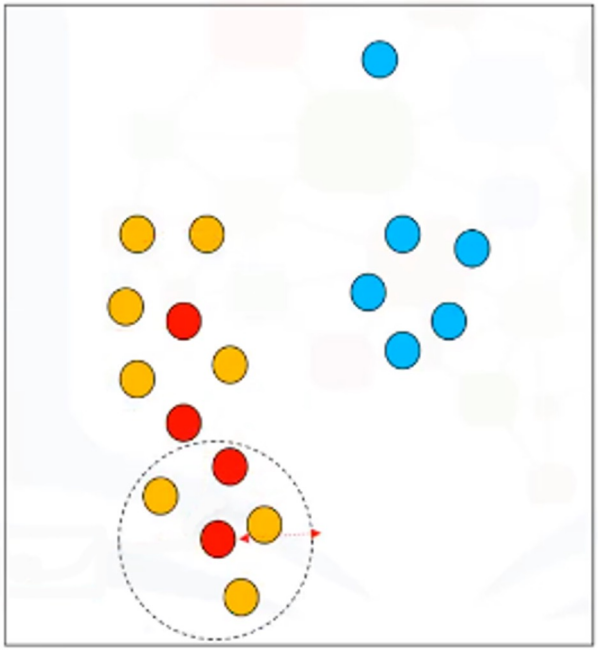
1. Su vecindario contiene menos de M puntos de datos
2. Se puede acceder desde algún punto central. Es decir, que dentro de su R-distancia hay un punto central

Esto significa que a pesar de que el punto amarillo se encuentra dentro del vecindario de 2 centímetros del punto rojo, no es por sí mismo un punto central, porque no tiene por lo menos 6 puntos en su vecindario.



DBSCAN

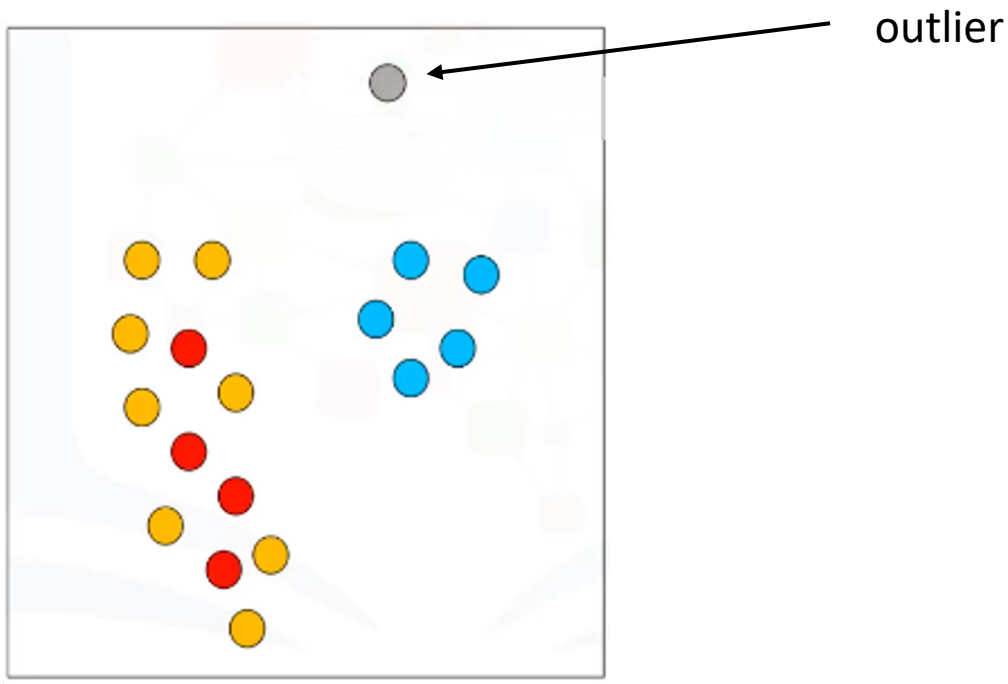
Seguimos con el siguiente punto. Como se puede ver, también es un punto core.
Y todos los puntos en torno a él, que no son puntos core son puntos border.



DBSCAN

Tomemos este punto. Puede ver que no es un punto core, ni tampoco es un punto border. Así que, lo etiquetaremos como un valor outlier.

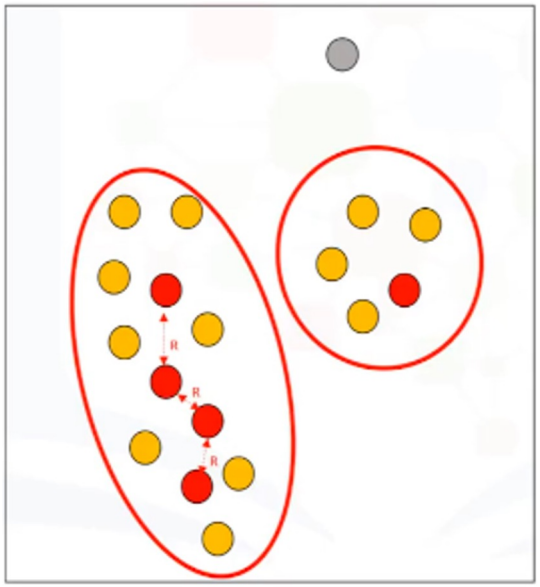
Un valor outlier es un punto que: no es un punto core, y tampoco está lo suficientemente cerca como para ser accesible desde un punto core.



DBSCAN

Seguimos y visitamos todos los puntos del conjunto de datos y los etiquetamos como Core, Border, o Outlier.

El siguiente paso es conectar los puntos core que son vecinos, y colocarlos en el mismo cluster. Por lo tanto, un cluster se forma con al menos un punto core, además de todos los puntos core accesibles, además de todas sus puntos border.



Fuente: https://www.statdeveloper.com/clustering-dbscan/

DBSCAN

DBSCAN es uno de los principales algoritmos de clustering basados en densidad.

Existen mejoras como:

1. OPTICS
2. EnDBSCAN
3. SNN

En scikit Learn:

sklearn.cluster.DBSCAN

DBSCAN

Ventajas:

- No necesita de la especificación del número de clusters deseado como lo requiere k-means.
- Puede encontrar clusters con formas geométricas arbitrarias.
- Tiene noción del ruido, y es robusto detectando outliers.
- Requiere solo de dos parámetros.

Desventajas:

- No es enteramente determinista: los puntos borde que son alcanzables desde más de un cluster pueden etiquetarse en cualquiera de estos.
- La calidad de DBSCAN depende de la noción de distancia. ¿Por que es malo?
- No puede agrupar conjuntos de datos bien con grandes diferencias en las densidades