23 Oct-2023

# DBSCAN

- DBSCAN (Density based Spatial clustering of Applications with Noise)
- clusters are dense regions in the data space, separated by regions of the lower density of points.
- The DBSCAN algorithm is based on this intuitive notion of "clusters" and noise.
- The key idea is that for each point of a cluster, the neighborhood of a given radius has to contain at least a minimum number of points.

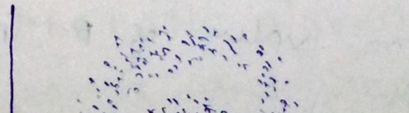
Why DBSAN?

partitioning methods(K-mean, PAM clustering) and hierarchical clustering work for finding spherical-shaped clusters or convex clusters. In other words. they are suitable only for compact and well-separated clusters.

Moreover, they are also severely affected by the presence of noise and outliers in the data.

Real life data may contain irregularities like:

1. clusters can be of arbitrary shape such as those shown in the fig.
2. Data may contain noise.



in the above fig. shows data set containing non- convex shape clusters and outliers. Given such data, the k-mean algorithm has difficulties in identifying these clusters with arbitrary shape.

Parameters Required for DBSCAN Algo:

1. eps:

- it defines the neighborhood around a data point. i.e. if the distance between two points is lower or equal to 'eps' then they are considered neighbours
- if the eps value is chosen too small then a large part of the data will be considered as an outlier. if it is choosen very large then the clusters will merge and the majority of the data points will be in the same clusters.
- One way to find the eps value is based on the k-distance graph.

2. MinPts:

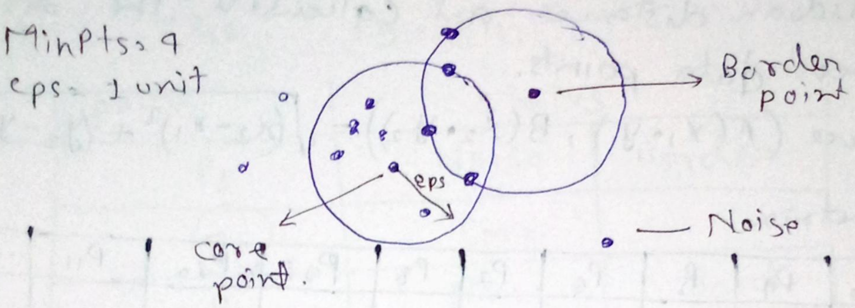
- minimum number of neighbors (data points) within eps radios.
- The larger the dataset, the larger the value of MinPts must be choosen.
- As a general rule, the minimum MinPts can be derived from the number of dimensions D in the dataset as. MinPts >= D+1.
- The minimum value of MinPts must be choosen at least 3.

Three Types of Data points in DBSCAN:

Core point: A point is core point if it has more than MinPts points within eps.

Border point:
A point which has fewer than MinPts within eps but it is in the neighborhood of a core point.

Noise or outlier:
A point which is not a core point or border point.



Steps used in DBSCAN algorithm:

1. Find all the neighbors points within eps and identify the core points or visited with more than MinPts neighbors.
2. For each core point if it is not already assigned to a cluster, create a new cluster.
3. Find recursively all its density- connected points. and assign them to the same cluster as the core point.
4. iterate through the remaining unvisited points in the dataset. Those points that do not belong to any cluster are noise.

Question: Apply DBSCAN algorithm to the given data points and create a cluster with minPts = 4 and epsilon $(\varepsilon) = 1.9$

Data points.

| | | | |
|---|---|---|---|
| P1 : | (3, 4) | P2 : | (4, 6) |
| P3 : | (5, 5) | P4 : | (6, 4) |
| P5 : | (7, 3) | P6 : | (6, 2) |
| P7 : | (7, 2) | P8 : | (8, 4) |
| P9 : | (3, 3) | P10 : | (2, 6) |
| P11 : | (3, 5) | P12 : | (2, 4) |

use euclidean distance and calculate the distance b/w each data points.

$$\text{Distance}\left(A(x_1, y_1), B(x_2, y_2)\right) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Distance matrix

| | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 | P12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 0 | | | | | | | | | | | |
| P2 | 1.41 | 0 | | | | | | | | | | |
| P3 | 2.83 | 1.41 | 0 | | | | | | | | | |
| P4 | 4.24 | 2.83 | 1.41 | 0 | | | | | | | | |
| P5 | 5.66 | 4.24 | 2.83 | 1.41 | 0 | | | | | | | |
| P6 | 5.83 | 4.47 | 3.16 | 2.00 | 1.41 | 0 | | | | | | |
| P7 | 6.40 | 5.00 | 3.61 | 2.24 | 1.00 | 1.00 | 0 | | | | | |
| P8 | 5.83 | 4.47 | 3.16 | 2.00 | 1.41 | 2.83 | 2.24 | 0 | | | | |
| P9 | 4.00 | 3.16 | 2.83 | 3.16 | 4.00 | 3.16 | 4.12 | 5.10 | 0 | | | |
| P10 | 1.41 | 2.00 | 3.16 | 4.47 | 5.83 | 5.66 | 6.40 | 6.32 | 3.16 | 0 | | |
| P11 | 2.00 | 1.41 | 2.00 | 3.16 | 4.47 | 4.24 | 5.00 | 5.10 | 2.00 | 1.41 | 0 | |
| P12 | 3.16 | 2.83 | 3.16 | 4.00 | 5.10 | 4.47 | 5.39 | 6.00 | 1.41 | 2.00 | 1.41 | 0 |

the minimum distance of any point from P1 with epsilon(ε) = 0.19

P1 : P2 , P10

P2 : P1 , P3 , P11

P3 : P2 , P4

P4 : P3 , P5

$P_5$ : $P_4$ , $P_6$ , $P_7$ , $P_8$

$P_6$ : $P_5$ , $P_7$

$P_7$ : $P_5$ , $P_6$

$P_8$ : $P_5$

$P_9$ : $P_{12}$

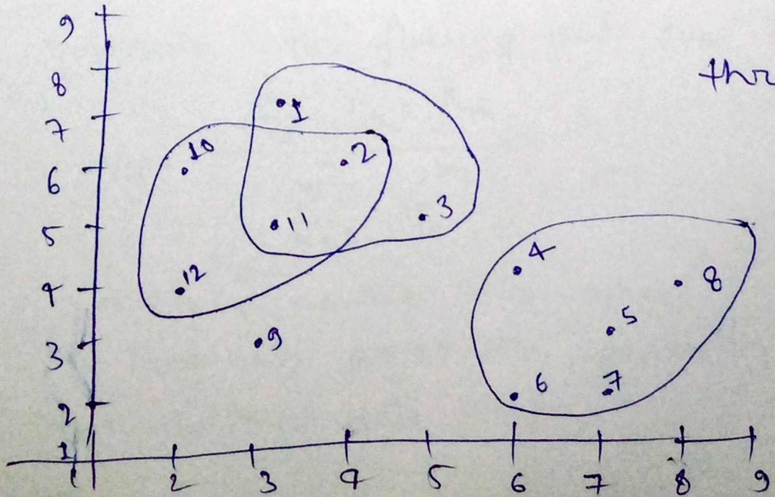
$P_{10}$ , $P_1$ , $P_{11}$

$P_{11}$ : $P_2$ , $P_{10}$ , $P_{12}$

$P_{12}$ : $P_9$ , $P_{11}$

| point. | status | |
|---|---|---|
| $P_1$ | Noise | Border |
| $P_2$ | Core | |
| $P_3$ | Noise | Border |
| $P_4$ | Noise | Border |
| $P_5$ | Core | |
| $P_6$ | Noise | Border |
| $P_7$ | Noise | Border |
| $P_8$ | Noise | Border |
| $P_9$ | Noise | |
| $P_{10}$ | Noise | Border |
| $P_{11}$ | Core | |
| $P_{12}$ | Noise | Border |

outlier (Not part of cluster



three cluster.