PAGE No.
DATE / /

DBSCAN (Density-Based spatial clustering of Application with Noise)

invented by Ester, Kriegel, Sander, Xu in 1996.

* here Noise means outliers.

* when to use DBSCAN
  -> when data is huge.
  -> when data is not in linear fashion.
  -> Not Sensitive to Noise.

* in K-means it have some drawback are
  (1) we have to specify the no. of K.
  (2) K-means does not perform well on finding non-convex (non-spherical shapes of clusters.
  (3) K-means is Sensitive to noise data.

Hierarchical clustering drawback.

-> Not suitable for big dataset.
-> high computational complexity
-> Need clusters (linkage) that affects the clustering result.
-> sensitivity to Noise

DBSCAN intuition

it discovers clusters of Non spherical shape the DBSCAN Clustering method can represent cluster of arbitary shape and to handle Noise.

DBSCAN is like human intuition it means DBSCAN will find out cluster just like humans way.

based on these 2 parameters DBSCAN will create clusters.

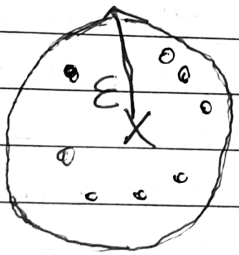
DBSCAN parameters :

(1) epsilon distance (ϵ) (user-define)

$$N_\epsilon(X) = Bd(X, \epsilon) = \{Y \mid \delta(X,Y) \leq \epsilon\}$$

(a) epsilon distance should be bigger than 0 that because it radious.

it defines the size and borders of each Neighbourhood.



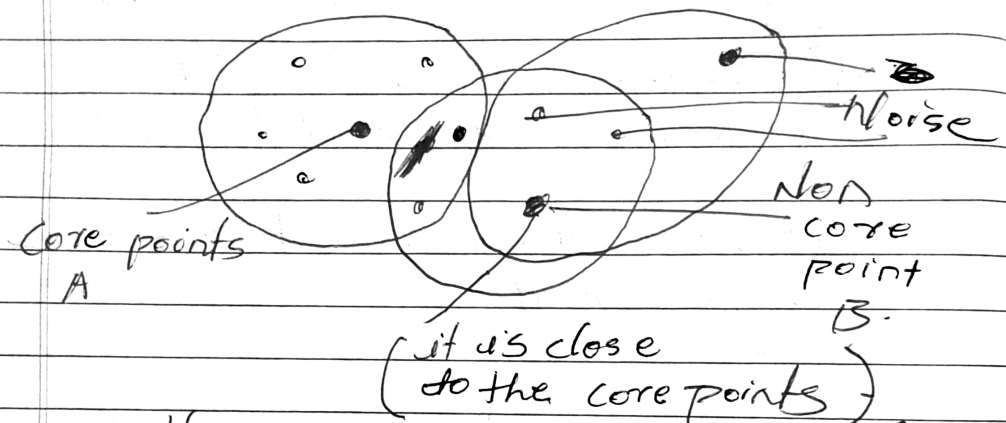
(2) minPTs (minimum points) (user-define)
should be at least 3 points in minPTs
its a density threhold for example
lets say minPTs = 5 so in ϵ neighborhood include at least 5.
then it will consider cluster other wise Not -

PAGE No.
DATE / /

(3) Border point.

to understand this Border point let take one example.

$\varepsilon \longrightarrow$ 5 minPts $\longrightarrow$ 4



in this cas core A has more than 4 points so it core point but B is not a core but it doesnt have more than 5 point but it closer to the Neighbood of core point so it called Border point but it count as Noise. & the Noise will count whe we dont Non core point further to make cluster case you can see the diagram this is also called Directly Density Reachable

* in general ε should be choosen as small as possible.

* metrics for measuring DBSCAN's performee.

For performance.

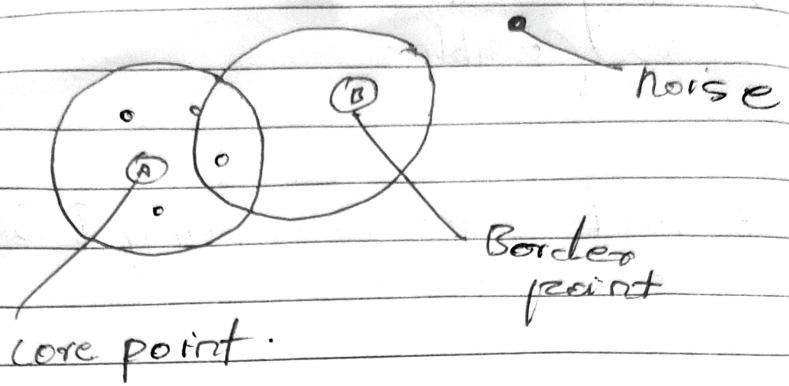
K-means ⟶ inertia

DBSCAN ⟶ Silhouette Score.

/

−1 (worst)
1 (best)
0 (overlapping clusters.)

Note :- when two core point meets two each then it's going to merge to the cluster.



steps in DBSCAN algorithms.

(1) find the Neighbor points.

(2) if core point is not assigned to a cluster create a New cluster.

(3) these who are Non core points but the Neighbor points belongs to the core points then it will take as a cluster but that Non core points will not further be extend to make epsilon distance

(4) iterate the process.

PAGE No.
DATE / /

Pros

1. identify randomly shapes clusters.
2. doesn't need k value
3. Handle Noise

Cons

1. varying densities are problematics
2. input of DBSCAN may be difficult to determine.
3. computational complexity - when the dimensionality is high