(12) DBSCAN (Density Based Spatial Clustering of Applications with Noise.

→ It is an unsupervised machine learning algorithm. this algorithm defines clusters as continuous regions of high density.

Key words :-

⊙ Epsilon:- (EPS)

this is the distance till which we look for the neighbouring points.

⊙ min-points :-

the min no. of points specified by the user

⊙ core points:-

If the no. of points inside the epsilon radius of a point is greater than or equal the min-points then its called a core point.

⊙ Border point:-

If the no. of points inside the epsilon radius of a point is less than the min points and it lies within the eps radius region of a corepoint, then its called border point.

⊙ Noise:-

A point which is neither core nor a border point is a noise point.

Algorithm steps

① The algorithm starts with a random point in the dataset which has not been visited yet and its neighbouring points are identified based on the eps value.

② If the point contains ≥ points than min points then the cluster formation starts and this point becomes a core point,
else its considered as Noise !

→ The point to note here is that a point initially classified as noise can later become a border point if its in the eps radius of a core point.

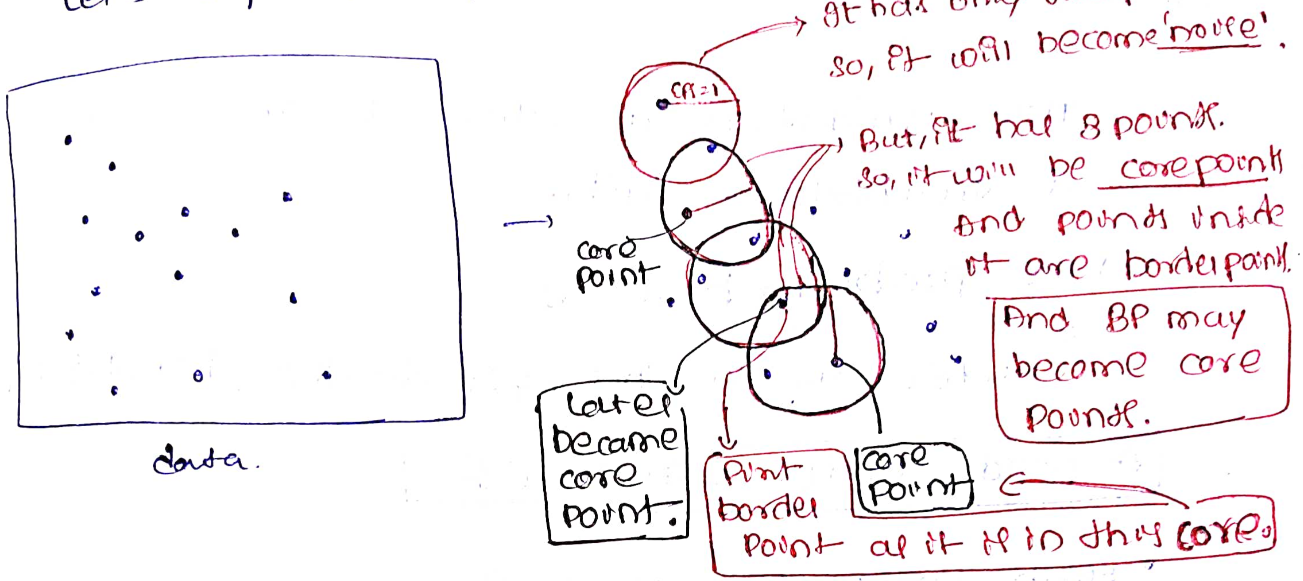
③ If the point is a core point, then all its neighbours become a part of cluster. If the points in the neighbourhood turn out to be core points then their neighbours are also part of cluster.

④ Repeat the steps above until all points are classified into different clusters or noises.

This Algo works well if all clusters are dense enough, and they are well separated by low dense regions.

Eg :-

Let's say eps=1, min points = 3.



It has only two points in it. So, it will become 'noise'.

But, it has 8 points. So, it will be core point and points inside it are border point.

And BP may become core points.

Later became core point.

Point border point as it is in this core.

core point

→ In short, DBSCAN is a very simple yet powerful Algorithm, capable of identifying any no. of clusters of any shape; it is robust to outliers, and it has just two hyper parameters. (eps & min samples)

→ However, if the density varies significantly across the clusters, it can be impossible for it to capture all the clusters properly. Moreover its computational complexity is roughly O(m log m), making it pretty close to linear with regard to the no. of instances.