DBSCAN Clustering

Aaryan
CO21BTECH11001

DBSCAN (**D**ensity-**B**ased **S**patial **C**lustering of **A**pplications with **N**oise) is an unsupervised learning algorithm that will take a set of points and make clusters of points with similar properties. It is based on the density-based clustering, and it will mark the outliers also which do not lie in any of the cluster or set.

Input:

Set of data points: $X = \{x_1, x_2, \dots, x_m\}$.
ϵ: Specifies how close points should be to each other to be considered a part of a cluster.
minPts: Minimum number of points required to form a cluster.

Algorithm:

1. Start with an arbitrary starting point (say x_i) that has not been visited.

2. Determine the ε-neighborhood of x_i .
$$N_\epsilon(x_i) = \{x \in X; \left||x_i - x|\right| \leq \epsilon\}$$

3. If $|N_\epsilon(x_i)| \geq$ minPts i.e., if the number of points in $N_\epsilon(x_i)$ is more than minPts, then the clustering process starts, and x_i is marked as visited else x_i is labeled as noise.

4. If a point is found to be a part of the cluster, then its ε-neighborhood is also the part of the cluster and the above procedure from step 2 is repeated for all ε-neighborhood points. This is repeated until all

points in the cluster is determined.

5. A new unvisited point is retrieved and processed, leading to the discovery of a further cluster or noise.

6. This process continues until all points are marked as visited.

After applying DBSCAN on a dataset, we get three types of points:

1. **Core Point:** A data point is a core point if it has a minimum number of neighboring data points (minPts) at an epsilon distance from it.
2. **Border Point:** A data point that has less than the minimum number of data points needed but has at least one core point in the neighborhood.
3. **Noise Point:** A data point that is not a core point or a border point is considered noise or an outlier.

Border points are also included in the cluster corresponding to the Core point nearest to them.

Questions:

1. Number of clusters to be formed are given as a parameter in DBSCAN.
(a) True
(b) False
Ans. False

2. k-Means doesn't work well for nested clusters while DBSCAN does.
(a) True
(b) False
Ans. True

3. What happens if inappropriate ϵ is taken?
   Ans.
   If too small epsilon is taken, then a large part of the data will not be clustered.
   If too large epsilon is taken, then clusters will merge and the majority of objects will be in the same cluster.

4. The worst case time complexity of DBSCAN is:
   (N – number of data points)
   (a) $O(NlogN)$
   (b) $O(N^2)$
   (c) $O(N)$
   (d) $O(N^{1.5})$

5. What is the major disadvantage of using DBSCAN?
   Ans. It is sensitive to parameters (ϵ,minPts) i.e., it's hard to determine the correct set of parameters.