WEEK-4 REPORT P-3

DASARI SRINITH (CS21BTECH11015)

DBSCAN:

Theory:

- Density-based spatial clustering of applications with noise (DBSCAN) is a clustering method.
- 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.
- DBSCAN is used cause , the real-life data may contain noise or the clusters can be of arbitrary shape
- DBSCAN algorithm requires two parameters:
    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 neighbors.
    2. **MinPts** : Minimum number of neighbors (data points) within eps radius.As a general rule, MinPts $\geq$ D +1 , where D is the number of dimentions.
- Based on the eps , the datapoints are also classified into 3 types ,
    1. **Core Point** : A point is a core point if it has more than MinPts points within eps.
    2. **Border Point** : A point which has fewer than MinPts within eps but it is in the neighborhood of a core point.
    3. **Noise or outlier** : A point which is not a core point or border point.
- So , the DBSCAN algorithm works as follows,
    1. For all the points in the database , it classifies them into core or non-core points.
    2. Then it randomly picks a core point and assign it to the first cluster.
    3. Next , the core points closer to the first cluster are added to it .
    4. Then , the core points closer to the growing first cluster are added to it.
    5. Finally , the non-core points closer to the points of the final cluster are added to it.
- After forming of all the clusters , the left points are called outliners , or noise points.

1. Is scaling required for DBSCAN?

   Solution:
   If you run DBSCAN on geographic data, and distances are in meters, you probably don't want to normalize anything, but set your epsilon threshold in meters.A non-uniform scaling does distort distances.

2. Why DBSCAN is better than k-means?

   Solution:
   K-means has difficulty with non-globular clusters and clusters of multiple sizes but DBSCAN is used to handle clusters of multiple sizes and structures and is not powerfully influenced by noise or outliers.

3. What is the drawback of the DBSCAN clustering algorithm??

   Solution:
   DBSCAN cannot cluster data-sets with large differences in densities.

4. Is DBSCAN fast or slow?

   Solution:
   DBSCAN is very slow for large datasets and can use a lot of memory, especially in higher dimensions.

5. What is the time complexity of DBSCAN?

   Solution: DBSCAN Algorithm scans whole dataset only one time and needs to calculate the distance of any pair of objects in the dataset. Hence, the computational complexity of the whole algorithms O(n2), where n is the number of degrees in the data set.