WORKSHOP II SOLUTION - POINT 2

DBSCAN

a) In which cases might it be more useful to apply?

It groups data points based on their density, identifying clusters of high density regions, non-planar geometry, unequal cluster sizes and classifying outliers as noise. DBSCAN is effective in discovering arbitrary-shaped clusters in data and is widely used in data mining, spatial data analysis, and machine learning applications.

DBSCAN is robust to noise, meaning it can effectively identify and ignore noise points that do not belong to any cluster. This makes it a useful tool for data cleaning and outlier detection.

b) What are the mathematical fundamentals of it?

DBSCAN requires two parameters: epsilon and minPoints. Epsilon is the radius of the circle to be created around each data point to check the density and minPoints is the minimum number of data points required inside that circle for that data point to be classified as a Core point.

In higher dimensions the circle becomes hypersphere, epsilon becomes the radius of that hypersphere, and minPoints is the minimum number of data points required inside that hypersphere.

DBSCAN creates a circle of epsilon radius around every data point and classifies them into Core point, Border point, and Noise. A data point is a Core point if the circle around it contains at least 'minPoints' number of points. If the number of points is less than minPoints, then it is classified as Border Point, and if there are no other data points around any data point within epsilon radius, then it treated as Noise. The metric used is the distances between nearest points.

c) Is there any relation between DBSCAN and Spectral Clustering? If so, what is it?

The relationship between DBSCAN and Spectral Clustering is that neither uses compactness to cluster the data. In compactness, the points are closer to each other and are compacted towards the center of the cluster.

In these algorithms, although the distance between points is smaller, they are not placed in the same cluster.

Although Spectral Clustering uses connectivity (points in a cluster are immediately close (epsilon distance) or connected) and DBSCAN uses a density-based clustering algorithm (it works under the assumption that clusters are dense regions in space separated by regions of lower density), both are good clustering algorithms for grouping data when points are not closer to each other and are not compact towards the center of the cluster as the following: