DBSCAN

THEORY

1) DBSCAN Stands for Density-Based Spatial Clustering of Applications with Noise. It is clus- tering algorithm which groups together points that are close to each other based on a distance measurement (usually Euclidean distance) and a minimum number of points. It also identifies the Noise(outlier) points in the data set.
2) DBSCAN algorithm requires two hyper parameters, eps & MinPts.
3) eps: A data point is considered neighbor of a particular data point if the distance between two points is lower or equal to eps. If the eps value is chosen too small then large part of the data will be considered as outliers. If it is chosen very large then the clusters will merge and the majority of the data points will be in the same clusters. One way to find the eps value is based on the k-distance graph.
4) MinPts: Minimum number of neighbors (data points) within eps radius.
5) There are three types of data points: Core Point, Border Point & Noise/Outlier.

QUIZ

1) What are the advantages of DBSCAN?
2) What are the disadvantages of DBSCAN?

ANSWER

1) DBSCAN can sort data into clusters of varying shapes, is robust to outliers and able to detect the outliers.
2) In some cases, determining an appropriate eps isn't easy & requires domain knowledge.