MACHINE LEARNING ASSIGNMENT – 1

Q1 to Q12 have only one correct answer. Choose the correct option to answer your question.

1. What is the most appropriate no. of clusters for the data points represented by the following dendrogram:
c) 6

2. In which of the following cases will K-Means clustering fail to give good results?
1. Data points with outliers
2. Data points with different densities
3. Data points with round shapes
4. Data points with non-convex shapes

Options:
d) 1, 2 and 4

3. The most important part of_____ is selecting the variables on which clustering is based.
d) formulating the clustering problem

4. The most commonly used measure of similarity is the____or its square.
a) Euclidean distance

5._____ is a clustering procedure where all objects start out in one giant cluster. Clusters are formed by dividing this cluster into smaller and smaller clusters.
b) Divisive clustering

6. Which of the following is required by K-means clustering?
d) All answers are correct

7. The goal of clustering is to
a) Divide the data points into groups

8. Clustering is a
b) Unsupervised learning

9. Which of the following clustering algorithms suffers from the problem of convergence at local optima?
d) All of the above

10. Which version of the clustering algorithm is most sensitive to outliers?
a) K-means clustering algorithm

11. Which of the following is a bad characteristic of a dataset for clustering analysis.
d) All of the above

12. For clustering, we do not require
a) Labeled data

Q13 to Q15 are subjective answers type questions, Answers them in their own words briefly.
13. How is cluster analysis calculated?
→ 1) calculate the distances
2) link the clusters
3) choose a solution by selecting the right number of clusters.
14. How is cluster quality measured?
→ To measure the quality of a clustering, we can use the average silhouette coefficient value of all objects in the data set. The silhouette coefficient and other intrinsic measures can also be used in the elbow method to heuristically derive the number of clusters in a data set by replacing the sum of within-cluster variances.

15. What is cluster analysis and its types?
→ Cluster analysis is an exploratory analysis that tries to identify structures within the data. Cluster analysis is also called segmentation analysis or taxonomy analysis. More specifically, it tries to identify homogenous groups of cases if the grouping is not previously known. Because it is exploratory, it does not make any distinction between dependent and independent variables. The different cluster analysis methods that SPSS offers can handle binary, nominal, ordinal, and scale (interval or ratio) data.
Cluster analysis is often used in conjunction with other analyses (such as discriminant analysis). The researcher must be able to interpret the cluster analysis based on their understanding of the data to determine if the results produced by the analysis are actually meaningful.

Centroid Clustering

This is one of the more common methodologies used in cluster analysis. In centroid cluster analysis you choose the number of clusters that you want to classify. For example, if you're a pet store owner you may choose to segment your customer list by people who bought dog and/or cat products.

The algorithm will start by randomly selecting centroids (cluster centers) to group the data points into the two pre-defined clusters. A line is then drawn separating the data points into the two clusters based on their proximity to the centroids. The algorithm will then reposition the centroid relative to all the points within each cluster. The centroids and points in a cluster will adjust

through all iteratations, resulting in optimized clusters. The result of this analysis is the segmentation of your data into the two clusters. In this example, the data set will be segmented into customers who are own dogs and cats.

Density Clustering

Density clustering groups data points by how densely populated they are. To group closely related data points, this algorithm leverages the understanding that the more dense the data points...the more related they are. To determine this, the algorithm will select a random point then start measuring the distance between each point around it. For most density algorithms a predetermined distance between data points is selected to benchmark how closely points need to be to one another to be considered related.. Then, the algorithm will identify all other points that are within the allowed distance of relevance. This process will continue to iterate by selecting different random data points to start with until the best clusters can be identified.

Distribution Clustering

Distribution clustering identifies the probability that a point belongs to a cluster. Around each possible centroid The algorithm defines the density distributions for each cluster, quantifying the probability of belonging based on those distributions The algorithm optimizes the characteristics of the distributions to best represent the data.

These maps look a lot like targets at an archery range. In the event that a data point hits the bulls eye on the map, then the probability of that person/object belonging to that cluster is 100%. Each ring around the bulls eye represents lessening percentage or certainty.

Distribution clustering is a great technique to assign outliers to clusters, where as density clustering will not assign an outlier to acluster.

Connectivity Clustering

Unlike the other three techniques of clustering analysis reviewed above, connectivity clustering initially recognizes each data point as its own cluster. The primary premise of this technique is that points closer to each other are more related. The iterative process of this algorithm is to continually incorporate a data point or group of data points with other data points and/or groups until all points are engulfed into one big cluster. The critical input for this type of algorithm is determining where to stop the grouping from getting bigger.