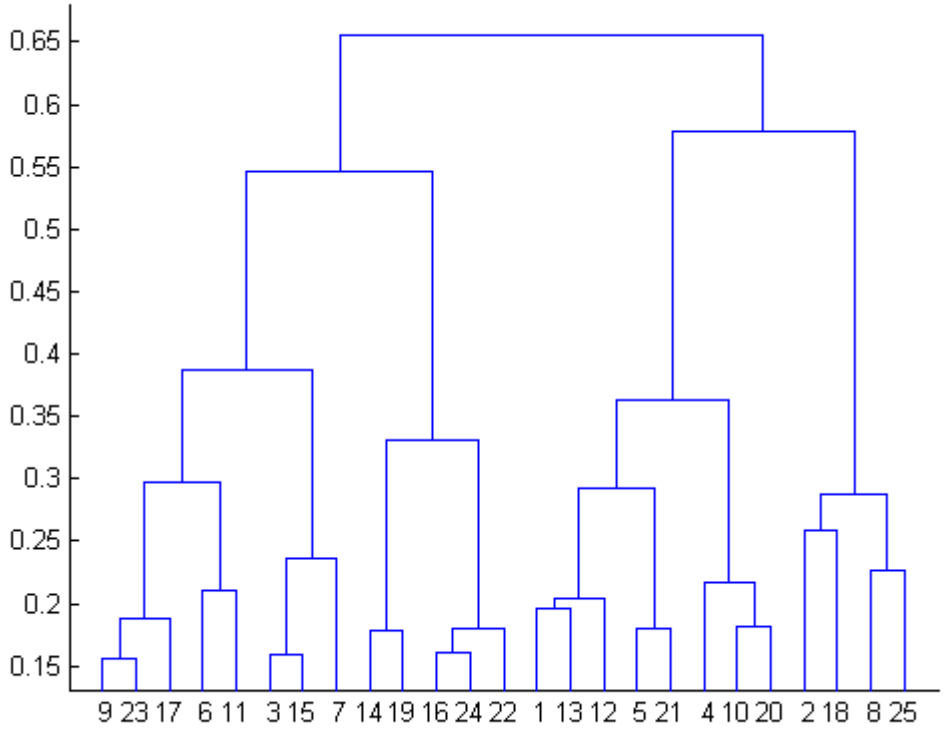
MACHINE LEARNING – WORKSHEET (CLUSTERING)

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:



a. 2
b. 4
c. 6
d. 8

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:
a. 1 and 2
b. 2 and 3
c. 2 and 4
d. 1, 2 and 4
e. 1, 2, 3 and 4

3. The most important part of _____ is selecting the variables on which clustering is based.
   a. interpreting and profiling clusters
   b. selecting a clustering procedure
   c. assessing the validity of clustering
   d. formulating the clustering problem

4. The most commonly used measure of similarity is the _____ or its square.
   a. euclidean distance
   b. city-block distance
   c. Chebyshev's distance
   d. Manhattan 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.
   a. Non-hierarchical clustering
   b. Divisive clustering
   c. Agglomerative clustering
   d. K-means clustering

6. Which of the following is required by K-means clustering?
   a. defined distance metric
   b. number of clusters
   c. initial guess as to cluster centroids
   d. all answers are correct

7. The goal of clustering is to-
   a. Divide the data points into groups
   b. Classify the data point into different classes
   c. Predict the output values of input data points
   d. All of the above

8. Clustering is a-
   a. Supervised learning
   b. Unsupervised learning
   c. Reinforcement learning
   d. None

9. Which of the following clustering algorithms suffers from the problem of convergence at local optima?
   a. K- Means clustering
   b. Hierarchical clustering
   c. Diverse clustering
   d. All of the above

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

c. K-medians clustering algorithm
d. None

11. Which of the following is a bad characteristic of a dataset for clustering analysis-
a. Data points with outliers
b. Data points with different densities
c. Data points with non-convex shapes
d. All of the above

12. For clustering, we do not require-
a. Labeled data
b. Unlabeled data
c. Numerical data
d. Categorical data

Q13 to Q15 are subjective answers type questions, Answers them in their own words briefly.

13. How is cluster analysis calculated?
14. How is cluster quality measured?
15. What is cluster analysis and its types?

13 answer: cluster analysis following three basic steps:
1) calculate the distances,
2) link the clusters and
3) choose a solution by selecting the right number of clusters. First, we have to select the variables upon which we base our clusters.

14 answer: To measure a cluster's fitness within a clustering, we can compute the average silhouette coefficient value of all objects in the cluster. To measure the quality of a clustering, we can use the average silhouette coefficient value of all objects in the data set.

15 answer: cluster analysis is four types:
1. Centroid Clustering

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.

2.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.

3.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.

4.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.