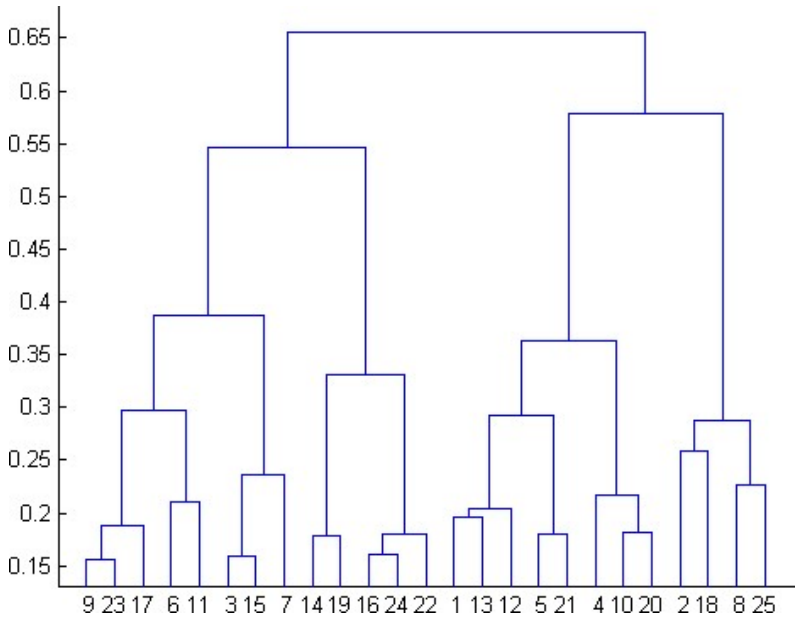
Worksheet 1 Machine Learning

Q1 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 **ANSWER>>(B)**
   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 **ANSWER>>(D)**

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 **ANSWER>>(D)**

4. The most commonly used measure of similarity is the____or its square.
   a) Euclidean distance **ANSWER>>(A)**
   b) city-block distance
   c) Chebyshev's distance
   d) Manhattan distance

Worksheet_1_Machine_Learning

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 **ANSWER>>(B)**
   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 **ANSWER>>(D)**

7. The goal of clustering is to-
   a) Divide the data points into groups **ANSWER>>(A)**
   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 **ANSWER>>(B)**
   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 **ANSWER>>(D)**

10. Which version of the clustering algorithm is most sensitive to outliers?
    a) K-means clustering algorithm **ANSWER>>(A)**
    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 **ANSWER>>(D)**

12. For clustering, we do not require-
    a) Labeled data **ANSWER>>(A)**
    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?

ANSWER>>

Clustering is a broad set of techniques for finding subgroups of observations within set. When we cluster observations, we want observations in the same group to be similar and observations in different groups to be dissimilar. Because there isn't a response variable, this is an unsupervised method, which implies that it seeks to find relationships between

the n observations without being trained by a response variable. Clustering allows us to identify which observations are alike, and potentially categorize them therein. K-means clustering is the simplest and the most commonly used clustering method for splitting a dataset into a set of k groups.
he first step when using k-means clustering is to indicate the number of clusters (k) that will be generated in the final solution. The algorithm starts by randomly selecting k objects from the data set to serve as the initial centers for the clusters. The selected objects are also known as cluster means or centroids. Next, each of the remaining objects is assigned to it's closest centroid, where closest is defined using the Euclidean distance (Eq. 1) between the object and the cluster mean. This step is called "cluster assignment step". After the assignment step, the algorithm computes the new mean value of each cluster. The term cluster "centroid update" is used to design this step. Now that the centers have been recalculated, every observation is checked again to see if it might be closer to a different cluster. All the objects are reassigned again using the updated cluster means. The cluster assignment and centroid update steps are iteratively repeated until the cluster assignments stop changing (i.e until *convergence* is achieved). That is, the clusters formed in the current iteration are the same as those obtained in the previous iteration.

K-means algorithm can be summarized as follows:

1. Specify the number of clusters (K) to be created (by the analyst)
2. Select randomly k objects from the data set as the initial cluster centers or means
3. Assigns each observation to their closest centroid, based on the Euclidean distance between the object and the centroid
4. For each of the k clusters update the cluster centroid by calculating the new mean values of all the data points in the cluster. The centroid of a Kth cluster is a vector of length *p* containing the means of all variables for the observations in the kth cluster; *p* is the number of variables.
5. Iteratively minimize the total within sum of square (Eq. 7). That is, iterate steps 3 and 4 until the cluster assignments stop changing or the maximum number of iterations is reached. By default, the R software uses 10 as the default value for the maximum number of iterations.

14. How is cluster quality measured?

- Suppose you have assessed the clustering tendency of a given data set. You may have also tried to predetermine the number of clusters in the set. You can now apply one or multiple clustering methods to obtain clusterings of the data set.
- We have a few methods to choose from for measuring the quality of a clustering. In general, these methods can be categorized into two groups according to whether ground truth is available. Here, ground truth is the ideal clustering that is often built using human experts.
- If ground truth is available, it can be used by **extrinsic methods**, which compare the clustering against the group truth and measure. If the ground truth is unavailable, we can use **intrinsic methods**, which evaluate the goodness of a clustering by considering how well the clusters are separated. Ground truth can be considered as supervision in the form of cluster labels. Hence, extrinsic methods are also known as supervised methods, while intrinsic methods are unsupervised methods.

Worksheet_1_Machine_Learning

15.What is cluster analysis and its types?

- Clustering is a type of unsupervised learning method of machine learning. In the unsupervised learning method, the inferences are drawn from the data sets which do not contain labelled output variable. It is an exploratory data analysis technique that allows us to analyze the multivariate data sets.
- Clustering is a task of dividing the data sets into a certain number of clusters in such a manner that the data points belonging to a cluster have similar characteristics. Clusters are nothing but the grouping of data points such that the distance between the data points within the clusters is minimal. Clustering is done to segregate the groups with similar traits.

- Clustering itself can be categorized into two types i.e, Hard Clustering and Soft Clustering. In hard clustering, one data point can belong to one cluster only. But in soft clustering, the output provided is a probability likelihood of a data point belonging to each of the pre-defined numbers of clusters.
- Density-Based Clustering
- Hierarchical Clustering
- Fuzzy Clustering
- Partitioning Clustering
- K-Means Clustering
- Grid-Based Clustering
- WaveCluster