# Machine Learning Assignment – 1

1. What is the most appropriate no. of clusters for the data points represented by the following dendrogram:
Ans. B) 4

2. In which of the following cases will K-Means clustering fail to give good results?
Ans. D) 1, 2 and 4

3. The most important part of____ is selecting the variables on which clustering is based.
Ans. D) formulating the clustering problem

4. The most commonly used measure of similarity is the or its square.
Ans. 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.
Ans. B) Divisive clustering

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

7. The goal of clustering is to -
Ans. A) Divide the data points into groups

8. Clustering is a -
Ans. A) Unsupervised learning

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

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

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

12. For clustering, we do not require-
Ans. A) Labelled data

13. How is cluster analysis calculated?
Ans.

i) Decide number of clusters
ii) Make a guess of K cluster centre locations
iii) Each data point attaches to its nearest centroid which leads to forming clusters for each centroid and the cluster contains all the datapoints of that centroid.
iv) Each cluster finds itself a new centroid.
v) Then the process from “datapoint assigning to nearest centroid” to “finding new centre” repeats itself till there comes a point where the centroids wont change anymore.

14. How is cluster quality measured?
Ans.
Cluster quality can be measured using Silhouette score

15. What is cluster analysis and its types?
Ans.
Cluster analysis is a method in which we find similar objects or datatypes to form a cluster with. It is an unsupervised machine learning-based algorithm that acts on unlabelled data. Its is a multivariate data mining technique for grouping objects.
There are various methods of cluster analysis –

i) Hierarchical cluster analysis – In this method we clusters are made from datapoints, and then again, the clusters are clustered with the ones it is most similar to. This method is known as **Agglomerative** method. Agglomerative clustering starts with single objects and starts grouping them into clusters.
ii) K-Means - In K-Means, the similarity measure is the distance between the data points. Smaller the distance, more similar the points are.
iii) Density-Based Clustering - In this method, the clusters are created based upon the density of the data points. The regions that have become dense due to the huge number of data points residing in that region are considered as clusters.
iv) DBSCAN – It is ‘Density-Based Spatial Clustering of Applications with Noise’, it clusters datapoints based on the distance metric between them. It can differentiate in clusters based on their shape and size, from the large data which is containing noise and outliers.
v) OPTICS – It is ‘Ordering Points to Identify Clustering Structure’, along with metric distance it also considers core distance and reachability distance.