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:

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

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?
- The method of identifying similar groups of Data in dataset is called Clustering.
- Segregate the groups with similar traits and assign them into clusters.
- Specify the desired number of clusters.
- Randomly assign each datapoint to a cluster.
- Compute the cluster centroid.
- Calculate the Euclidian distance among the datapoints and again re assign each point to the closest cluster centroid.
- Repeat the above two steps for every dataset until no improvements are possible.

14. How is cluster quality measured?
- No commonly recognized best suitable measure in practise.
  Majorly there are three categorization of measures
- External – Supervised , employ criteria not inherent to dataset. For external measures, we can consider they are supervised, employ criteria not inherent to the datasets itself. That means we may have some prior or expert knowledge. For example, some ground truth. Then we can comparing the clustering results against the prior or expert specified knowledge, using certain clustering quality measure.
- Internal - unsupervised. Criteria derived from data itself. That means the criteria derived from the data itself. In that case, we will evaluate the goodness of clustering by considering how well the clusters are separated and how compact the clusters are. For example, we can use silhouette coefficient.
- Relative - Directly compare different class rings using those obtained via different parameter setting for the same algorithm.

15. What is cluster analysis and its types?

Cluster analysis which is also called clustering or data segmentation, the essential is getting a set of tape data points. The cluster analysis is to partition them into a set of clusters, or set of groups. They are as similar as possible within the same group and as far apart as possible among different groups. Cluster analysis is unsupervised learning, in the sense there's no predefined classes.

Types of Clusters are as below.

- Hard Clustering – Each Data point either belongs to a cluster completely or not.
- Soft Clustering – Instead of putting each data point into a separate cluster a probability or likelihood of that datapoint to be in any of the clusters.

Cluster analysis Types are as below.

- Hierarchical Clustering methods
- Distribution Clustering.
- Density Clustering
- Centroid Clustering