ASSIGNMENT – 3

MACHINE LEARNING

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

1. Which of the following is an application of clustering?

d. All of the above

2. On which data type, we cannot perform cluster analysis?

a. Time series data

3. Netflix's movie recommendation system uses

a. Supervised learning

4. The final output of Hierarchical clustering is

d. All of the above

5. Which of the step is not required for K-means clustering?

d. None

6. Which is the following is wrong?

d. None

7. Which of the following metrics, do we have for finding dissimilarity between two clusters in hierarchical clustering?

i. Single-link ii. Complete-link iii. Average-link

d. 1, 2 and 3

8. Which of the following are true? i. Clustering analysis is negatively affected by multicollinearity of features ii. Clustering analysis is negatively affected by heteroscedasticity.

a. 1 only

9. In the figure above, if you draw a horizontal line on y-axis for y=2. What will be the number of clusters formed?

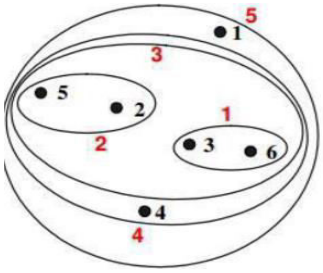
a. 2

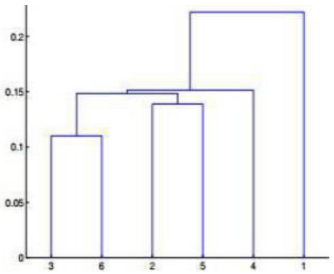
10. For which of the following tasks might clustering be a suitable approach?

b.Given a database of information about your users, automatically group them into different market segments.

11. Given, six points with the following attributes Which of the following clustering representations and dendrogram depicts the use of MIN or Single link proximity function in hierarchical clustering:

a.

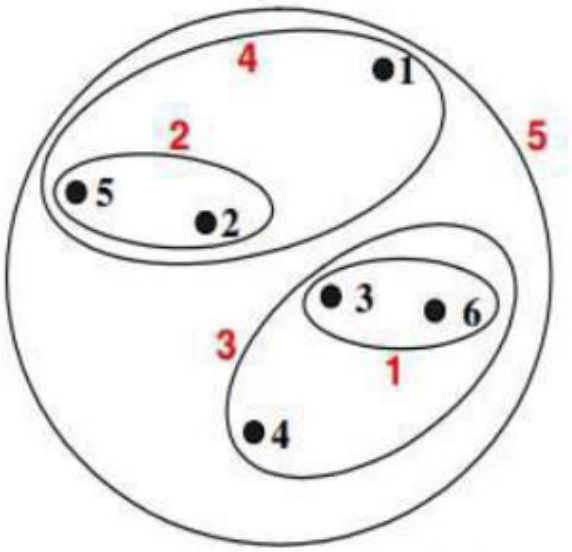


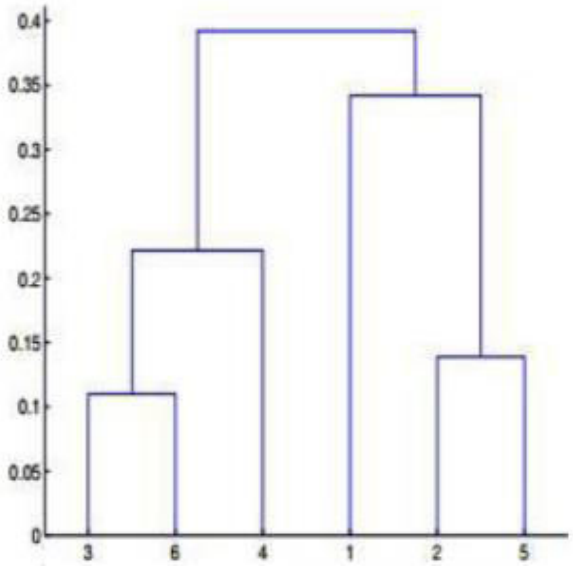


.

12. Given, six points with the following attributes: Which of the following clustering representations and dendrogram depicts the use of MAX or Complete link proximity function in hierarchical clustering.

b.





Q13 to Q14 are subjective answers type questions, Answers them in their own words briefly

13. What is the importance of clustering?

The process of grouping a set of objects so that objects in the same group are more similar to each other than to those in other groups. The process of clustering is important in data analysis and data mining applications. There are various types of clustering methods; they are Hierarchical methods Partitioning methods Density-based Model-based clustering Grid-based model Clustering Intelligence Servers provides the following benefits:

- **Simplified management:**
- **Increased resource availability:**
- **Greater scalability:**
- **Strategic resource usage:**
- **Increased performance:**

14. How can I improve my clustering performance.

An efficient method to improve the clustering performance for high dimensional data by Principal

Component Analysis and modified K-means. PCA is a classical multivariate data analysis method that is useful in linear feature extraction. Without class labels it can compress the most information in the original data space into a few new features, i.e., principal components. Handling high dimensional data using clustering techniques obviously a difficult task in terms of higher number of variables involved. In order to improve the efficiency, the noisy and outlier data may be removed and minimize the execution time and we have to reduce the no. of variables in the original data set. The central idea of PCA is to reduce the dimensionality of the data set consisting of a large number of variables. It is a statistical technique for determining key variables in a high dimensional data set that explain the differences in the observations and can be used to simplify the analysis and visualization of high dimensional data set. K-means is a prototype-based, simple partitioned clustering technique which attempts to find a user-specified k number of clusters. These clusters are represented by their centroids. A cluster centroid is typically the mean of the points in the cluster. This algorithm is simple to implement and run, relatively fast, easy to adapt, and common in practice.

The algorithm consist of two separate phases: the first phase is to select k centers randomly, where the value of k is fixed in advance. The next phase is to assign each data object to the nearest center. Euclidean distance is generally considered to determine the distance between each data object and the cluster centers. When all the data objects are included in some clusters, recalculating the average of the clusters. This iterative process continues repeatedly until the criterion function becomes minimum.