MACHINE LEARNING
Assignment 3

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

   Biological network analysis

   Market trend prediction

   Topic modelling

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

   None of the mentioned.

3. **Netflix's movie recommendation system uses**

   Reinforcement learning and unsupervised learning

4. **The final output of Hierarchical clustering is:-**

   The tree representing how close the data points are to each other

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

   None of the mentioned

6. **Which is the following is wrong?**

   k-nearest neighbor is same as k-means

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

   Single-link Complete-link Average-link

8. **Which of the following are true?**

   Clustering analysis is not negatively affected by heteroscedasticity but the results are negatively impacted by multicollinearity of features/ variables used in clustering as the correlated feature/ variable will carry extra weight on the distance calculation than desired.

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?**

   Two clusters will be formed. Since the number of vertical lines intersecting the red horizontal line at y=2 in the dendrogram are 2, therefore, two clusters will be formed.

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

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:

Solution: (A)

For the single link or MIN version of hierarchical clustering, the proximity of two clusters is defined to be the minimum of the distance between any two points in the different clusters. For instance, from the table, we see that the distance between points 3 and 6 is 0.11, and that is the height at which they are joined into one cluster in the dendrogram. As another example, the distance between clusters {3, 6} and {2, 5} is given by dist({3, 6}, {2, 5}) = min(dist(3, 2), dist(6, 2), dist(3, 5), dist(6, 5)) = min(0.1483, 0.2540, 0.2843, 0.3921) = 0.1483.

(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.

Solution: (B)

For the single link or MAX version of hierarchical clustering, the proximity of two clusters is defined to be the maximum of the distance between any two points in the different clusters. Similarly, here points 3 and 6 are merged first. However, {3, 6} is merged with {4}, instead of {2, 5}. This is because the dist({3, 6}, {4}) = max(dist(3, 4), dist(6, 4)) = max(0.1513, 0.2216) = 0.2216, which is smaller than dist({3, 6}, {2, 5}) = max(dist(3, 2), dist(6, 2), dist(3, 5), dist(6, 5)) = max(0.1483, 0.2540, 0.2843, 0.3921) = 0.3921 and dist({3, 6}, {1}) = max(dist(3, 1), dist(6, 1)) = max(0.2218, 0.2347) = 0.2347.

(13.) What is the importance of clustering?

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining. Clustering in data mining **helps in the discovery of information by**

classifying the files on the internet. It is also used in detection applications. FOR EXAMPLE :-Fraud in a credit card can be easily detected using clustering in data mining.

14). How can I improve my clustering performance?

Graph-based clustering performance can easily be improved by applying ICA blind source separation during the graph Laplacian embedding step. Applying unsupervised feature learning to input data using either RICA or SFT, improves clustering performance.