MACHINE LEARNING
ASSIGNMENT – 3

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?
ANS: d. All of the above

2. On which data type, we cannot perform cluster analysis?
ANS: d. None

3. Netflix's movie recommendation system uses-
ANS: c. Reinforcement learning and Unsupervised learning

4. The final output of Hierarchical clustering is-
ANS: b. The tree representing how close the data points are to each other

5. Which of the step is not required for K-means clustering?
ANS: d. None

6. Which is the following is wrong?
ANS: c. k-nearest neighbour is same as k-means

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
Options:
ANS: 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
Options:
ANS: 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?
ANS: a. 2

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

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

13. What is the importance of clustering?
ANS: Clustering is the task of grouping a set of objects in such a way that objects in the same group called cluster are similar to each other than to those in other groups/ clusters. Clustering is very important in data analysis and data mining It is the task 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. A good clustering algorithm is able to identity clusters irrespective of their shapes. It's used for Organizing data into clusters to show the internal structure of the data

14. How can I improve my clustering performance?
ANS: Clustering analysis is one of the main analytical methods in data mining. K-means is the most popular and partition-based clustering algorithm. But it is computationally expensive and the quality of resulting clusters heavily depend on the selection of initial centroid and the dimension of the data. Principal Component Analysis (PCA) is an important approach to unsupervised dimensionality reduction technique. The k-means++ algorithm uses a different initialization. Instead of random initialization, we only choose the first center randomly. All centers are sampled, but with a probability that is proportional to their squared distance from all current centers.