MACHINE LEARNING WORKSHEET – 3

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?
   Ans. d. 1, 2 and 3

8. Which of the following are true?
   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.

13. What is the importance of clustering?
    Ans. Clustering refers to the grouping together of objects with similar characteristics. The entire data sets present are many for a particular problem, and it is impossible to analyse them individually; hence, clustering makes it easy to handle and gather insightful data from it.
    Clustering allows researchers to identify and define patterns between data elements. Revealing these patterns between data points helps to distinguish and outline structures which might not have been apparent before, but which give significant meaning to the data once they are discovered. Once a clearly defined structure emerges from the dataset at hand, informed decision-making becomes much easier.

When cluster analysis is performed as part of market research, specific groups can be identified within a population. The analysis of these groups can then determine how likely a population cluster is to purchase products or services. If these groups are defined clearly, a marketing team can then target varying cluster with tailored, targeted communication.

14. How can I improve my clustering performance?
    Ans. K-means clustering algorithm can be significantly improved by using a better initialization technique, and by repeating (re-starting) the algorithm.
    When the data has overlapping clusters, k-means can improve the results of the initialization technique. Initialization can be key for the performance of k-means. The k-means++ algorithm is a simple and widely applied technique.
    Also as a pre-processing stage of data mining and machine learning, dimension reduction not only decreases computational complexity, but also significantly improves the accuracy of the learned models from large data sets. Therefore, 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.