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?
d. All of the above

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

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

4. The final output of Hierarchical clustering is
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?
d. None

6. Which is the following is wrong?
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:
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: 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?
Answer... The purpose of clustering and classification algorithms is to make sense of and extract value from large sets of structured and unstructured data. If you're working with huge volumes of unstructured data, it only makes sense to try to partition the data into some sort of logical groupings before attempting to analyze it.
Clustering and classification allows you to take a sweeping glance of your data en masse, and then form some logical structures based on what you find there before going deeper into the nuts-and-bolts analysis.

In their simplest form, *clusters* are sets of data points that share similar attributes, and *clustering algorithms* are the methods that group these data points into different clusters based on their similarities. You'll see clustering algorithms used for disease classification in medical science, but you'll also see them used for customer classification in marketing research and for environmental health risk assessment in environmental engineering.

14. How can I improve my clustering performance?
Answer...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. When the data has well separated clusters, the performance of k-means depends completely on the goodness of the initialization. Initialization using simple furthest point heuristic (Maxmin) reduces the clustering error of k-means from 15% to 6%, on average.