**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?  
a. Biological network analysis  
b. Market trend prediction  
c. Topic modeling  
d. All of the above  
Ans→ d. All of the above.

2. On which data type, we cannot perform cluster analysis?  
a. Time series data  
b. Text data  
c. Multimedia data  
d. None  
Ans→ d. None

3. Netflix’s movie recommendation system uses  
a. Supervised learning b. Unsupervised learning  
c. Reinforcement learning and Unsupervised learning  
d. All of the above  
Ans→ c. Reinforcement learning and Unsupervised learning

4. The final output of Hierarchical clustering is  
a. The number of cluster centroids  
b. The tree representing how close the data points are to each other  
c. A map defining the similar data points into individual groups  
d. All of the above

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?  
a. A distance metric  
b. Initial number of clusters  
c. Initial guess as to cluster centroids  
d. None  
Ans→ d. None

6. Which is the following is wrong?  
a. k-means clustering is a vector quantization method  
b. k-means clustering tries to group n observations into k clusters  
c. k-nearest neighbour is same as k-means  
d. None  
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:  
a.1 and 2  
b. 1 and 3  
c. 2 and 3  
d. 1, 2 and 3  
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:  
a. 1 only  
b. 2 only  
c. 1 and 2  
d. None of them  
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?  
a. 2  
b. 4  
c. 3  
d. 5  
Ans→ a. 2

10. For which of the following tasks might clustering be a suitable approach?  
a. Given sales data from a large number of products in a supermarket, estimate  
future sales for each of these products.  
b. Given a database of information about your users, automatically group them  
into different market segments.  
c. Predicting whether stock price of a company will increase tomorrow.  
d. Given historical weather records, predict if tomorrow's weather will be sunny  
or rainy.  
Ans→b

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→ D  
Ward method is a centroid method. Centroid method calculates the  
proximity between two clusters by calculating the distance between the  
centroids of clusters. For Ward’s method, the proximity between two  
clusters is defined as the increase in the squared error that results when  
two clusters are merged. The results of applying Ward’s method to the  
sample data set of six points. The resulting clustering is somewhat different  
from those produced by MIN, MAX, and group average.

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 important in data analysis and data mining applications  
[1]. 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.

1. Identification: Know what you are dealing with. Identifying is like putting  
   all the pieces on the table, mapping out the situation, and sorting them  
   using patterns.  
   2. Analysis: Analyze these patterns to make your clusters more focused and  
   accurate.  
   3. Strategy: Create differentiated strategies for each of the clusters, with  
   specific objectives, actions, and goals.

14. How can I improve my clustering performance?  
Ans→ k-means is a very simple and ubiquitous clustering algorithm. But  
quite often it does not work on your problem, for example because the  
initialization is bad. I ran into a similar problem recently, where I applied k-  
means to a smaller number of files in my data sets and everything worked  
fine, but when I ran it on many more samples it just wasn’t reliably getting  
good results.  
Fortunately, there is an improved initialization method, k-means++, which  
can help to alleviate this problem. In this article we will cover the following:  
Why initialization is so important for k-means  
An intuitive description into the k-means++ algorithm  
Implementation of k-means++ in R  
A common, but wrong variant of k-means++