MACHINE LEARNING – WORKSHEET

(CLUSTERING)

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

d. All of the above

ANS. c. Reinforcement 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 neighbor is same as k-means

d. None

ANS. c. 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?

1. Single-link

2. Complete-link

3. 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?

1. Clustering analysis is negatively affected by multicollinearity of features

2. 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. 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 Q15 are subjective answers type questions, Answers them in their own words briefly

13. What is the importance of clustering?

ANS. Clustering is important because:

1. It is useful for exploring data. Thus, helps in performing expolatory datz analysis.

1. Helps in generating summary of the data.

1. Detects the outliers.

1. Finds the duplictes.

1. It works as a pre-processing step.

14. How do you cluster a profile?

ANS. Profiling involves generating descriptions of the clusters with reference to the input variables you used for the cluster analysis. to cluster a profile:

A. Graphically represent your clusters according to your input variables.
B. Score your clusters in a table so that you can measure and compare them on each input variable with regards to numerical or descriptive values.
C. variables should be described in a type of 'story' about the category or customer base. The output of this step is a clearly described set of clusters with a focus placed on the input variables. If you have access to a wider set of data, you can use other loyalty data to supplement the cluster profile even if it was not used in the original cluster analysis.
D. At last, needs to build a 'story' or profile around each cluster so, that the obtained information can be used.

15. How can I improve my clustering performance?

ANS. 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. Techniques like DipScaling and DipTransformation—which enhance the data set by rescaling and transforming its features and thus emphasizing and accentuating its structure. If the structure is sufficiently clear, clustering algorithms will perform far better.