

MACHINE LEARNING-WORKSHEET

SET 3

ANSWERS

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 of 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. 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)

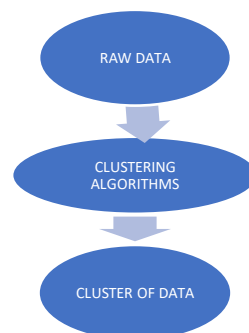
12. Given, six points with the following attributes: Which of the following clustering representations and dendrogram depicts the use of MAX or Complete link proximity function in hierarchical clustering.?

Ans:- (B)

13. What is the importance of clustering?

Ans:- **Clustering** is an unsupervised machine learning method of identifying and grouping similar data points in larger datasets without concern for the specific outcome. Clustering (sometimes called cluster analysis) is usually used to classify data into structures that are more easily understood and manipulated.

Clustering is important in data analysis and data mining applications. 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 (clusters). A good clustering algorithm is able to identify clusters irrespective of their shapes. The stages involved in clustering algorithm are as follows:



Clustering helps to find out the patterns in the data without taking any particular variable into consideration. This will help to find the relation between the features of the data. After clustering we could treat the records based on its groups.

14. How can I improve my clustering performance?

Ans:- Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).

It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.

Improving the clustering performing rather a subject and domain specific tasks. But there are some steps that we could perform to ensure betterment in performance .

- Firstly, the more the data the better the results. Also, data cleaning is necessary before using it for analysis.
- Secondly, using an appropriate clustering algorithm is also very important during cluster analysis. Choosing appropriate algorithm for finding the optimal number of cluster such as elbow method, gap stats method, silhouette method .