MACHINE LEARNING WORKSHEET-3

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

Answer : 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

Answer: 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

Answer : 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

Answer : 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

Answer : 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

Answer : 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

Answer : 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

Answer: 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

Answer : 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.

Answer: 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:

Answer : 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.

Answer: B

13. What is the importance of clustering?

Answer : 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).

Machine learning systems can then use cluster IDs to simplify the processing of large datasets. Thus, clustering's output serves as feature data for downstream ML systems.

Clustering is used to identify groups of similar objects in datasets with two or more variable quantities. In practice, this data may be collected from marketing, biomedical, or geospatial databases, among many other places.

Clustering has a myriad of uses in a variety of industries. Some common applications for clustering include the following:

1. market segmentation

2. social network analysis
3. search result grouping
4. medical imaging
5. image segmentation
6. anomaly detection

After clustering, each cluster is assigned a number called a cluster ID. Now, you can condense the entire feature set for an example into its cluster ID. Representing a complex example by a simple cluster ID makes clustering powerful. Extending the idea, clustering data can simplify large datasets.

For example, you can group items by different features as demonstrated in the following examples:

Examples

Group stars by brightness.

Group organisms by genetic information into a taxonomy.

Group documents by topic.

Machine learning systems can then use cluster IDs to simplify the processing of large datasets. Thus, clustering's output serves as feature data for downstream ML systems.

At Google, clustering is used for generalization, data compression, and privacy preservation in products such as YouTube videos, Play apps, and Music tracks.

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.

There are other algorithms that are known, in many situations, to provide better clustering results than k-means. However, k-means is popular for good reasons. First, it is simple to implement. Second, people often prefer to use an extensively studied algorithm whose limitations are known rather than a potentially better, but less studied, algorithm that might have unknown or hidden limitations.

K-means starts by selecting *k* random data points as the initial set of centroids, which is then improved by two subsequent steps. In the assignment step, every point is put into the cluster of the nearest centroid. In the update step, the centroid of every cluster is recalculated as the mean of all data points assigned to the cluster. Together, these two steps constitute one *iteration* of k-means. These steps fine-tune both the cluster borders and the centroid locations. The algorithm is iterated a fixed number of times, or until convergence (no further improvement is obtained).