MACHINE LEARNING

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
   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
   Answer: None

3. Netflix's movie recommendation system uses-
   Answer: Reinforcement learning and Unsupervised learning

4. The final output of Hierarchical clustering is-
   Answer: The tree representing how close the data points are to each other

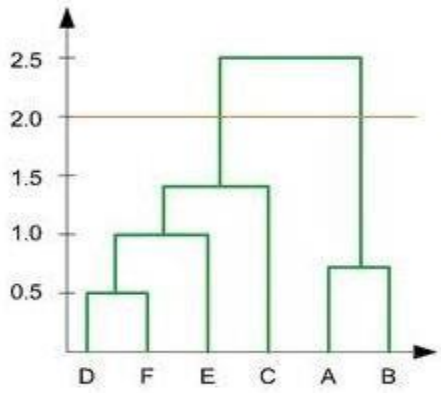
5. Which of the step is not required for K-means clustering?
   Answer: None

6. Which is the following is wrong?
   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-linkOptions:
   a. 1 and 2
   b. 1 and 3
   c. 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:
   Answer: 1 only (the 1st Answere is true)

9. In the figure above, if you draw a horizontal line on y-axis for y=2. What will be the number of clustersformed?



Answer – 2 Clusters will be formed.

10. For which of the following tasks might clustering be a suitable approach?
    Answer : Given sales data from a large number of products in a supermarket, estimate future sales for eachof these products.

11. Given, six points with the following attributes:

point	x coordinate	y coordinate
p1	0.4005	0.5306
p2	0.2148	0.3854
p3	0.3457	0.3156
p4	0.2652	0.1875
p5	0.0789	0.4139
p6	0.4548	0.3022

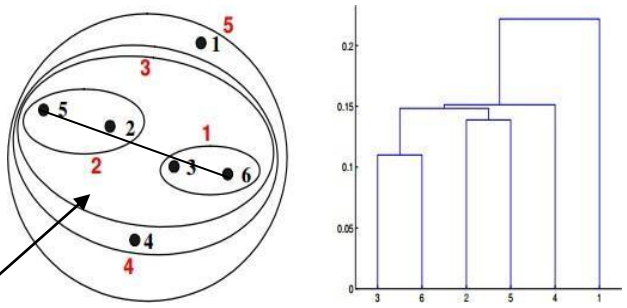
Table : X-Y coordinates of six points.

	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

Table : Distance Matrix for Six Points

Which of the following clustering representations and dendrogram depicts the use of MIN or Single link proximity function in hierarchical clustering:

a.



Answer: A

In the single link or MIN version of hierarchical clustering, the proximity of two clusters is defined to be the minimum of the distance between any two points in the different clusters. For instance, from the table, we see that the distance between points 3 and 6 is 0.11, and that is the height at which they are joined into one cluster in the dendrogram.

1. Given, six points with the following attributes:

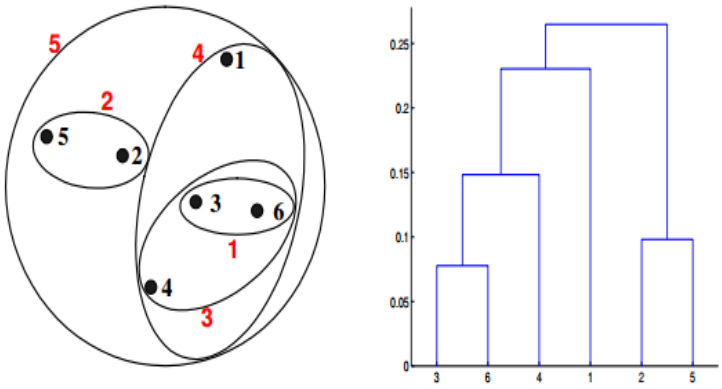
point	x coordinate	y coordinate
p1	0.4005	0.5306
p2	0.2148	0.3854
p3	0.3457	0.3156
p4	0.2652	0.1875
p5	0.0789	0.4139
p6	0.4548	0.3022

Table : X-Y coordinates of six points.

	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

Table : Distance Matrix for Six Points

Which of the following clustering representations and dendrogram depicts the use of MAX or Completelink proximity function in hierarchical clustering.



Answer : B

For the single link or MAX version of hierarchical clustering, the proximity of two clusters is defined to be the maximum of the distance between any two points in the different clusters. Similarly, here points 3 and 6 are merged first.

Q13 to Q14 are subjective answers type questions, Answers them in their own words briefly

2. What is the importance of clustering?

Answer: Clustering is an unsupervised learning technique that groups similar data points together. It is a powerful tool for a variety of applications, including:

Data Exploration: Clustering can be used to explore a data set and discover patterns or hidden structures that may not be immediately apparent.

Anomaly Detection: Clustering can be used to identify data points that do not fit well with any of the identified clusters, which can indicate an anomaly or outlier.

Dimensionality Reduction: Clustering can be used to reduce the dimensionality of a data set by grouping similar features together.

Image Segmentation: Clustering can be used to segment images into different regions based on their color or texture.

Customer Segmentation: Clustering can be used in marketing to group customers with similar characteristics together, which can help with target marketing and customer retention.

Text mining: Clustering can be used to classify and group similar documents, news articles, or customer feedback.

Recommender systems: Clustering can be used to group similar items together and recommend items from the same cluster to users

Overall, clustering is a versatile technique that can be used to extract valuable insights from data, making it an important tool for data analysis and machine learning

3. How can I improve my clustering performance?
   Answer: There are several ways to improve clustering performance, including:

   Selecting a more appropriate distance metric: Different distance metrics can produce different results depending on the data set and the type of clustering algorithm being used.

   Scaling the data: Scaling the data can help ensure that all features have an equal influence on the clustering results.

   Choosing the right number of clusters: The optimal number of clusters for a data set can be determined using techniques such as the elbow method or silhouette analysis.

   Using ensemble methods: Combining the results of multiple clustering algorithms can often improve performance.

   Using a more advanced algorithm: Some advanced clustering algorithms, such as density-based or hierarchical clustering, may be more appropriate for certain data sets and may produce better results than traditional k-means.

   Using domain knowledge: Incorporating domain-specific knowledge into the clustering process can improve the interpretability and accuracy of the results..