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
   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

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

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

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

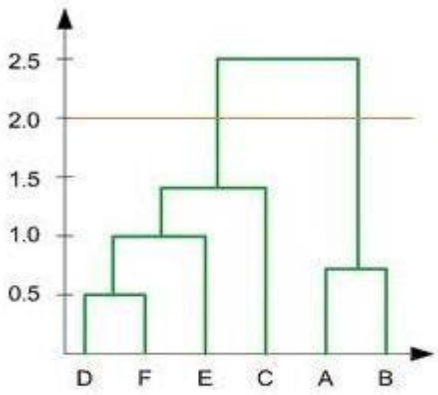
6. Which is the following is wrong?
   a. k-means clustering is a vector quantization method
   b. k-means clustering tries to group n observations into k clusters
   c. k-nearest neighbour is same as k-means
   d. None

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

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

MACHINE LEARNING

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

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.

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.

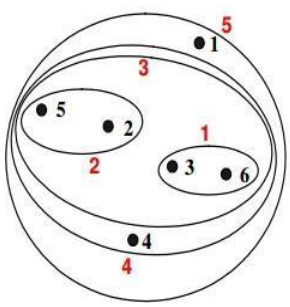
	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

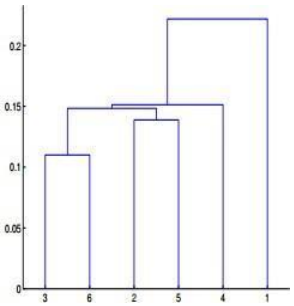
Table : Distance Matrix for Six Points

MACHINE LEARNING

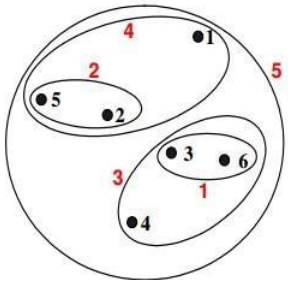
Which of the following clustering representations and dendrogram depicts the use of MIN or Single link proximity function in hierarchical clustering:

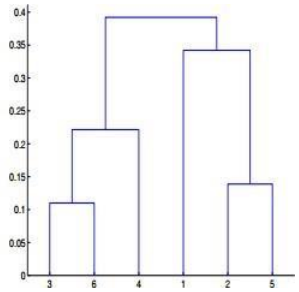
a.



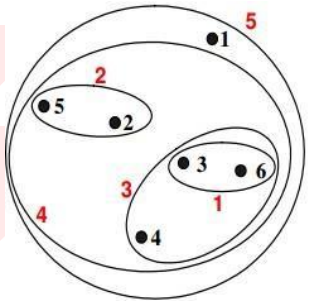


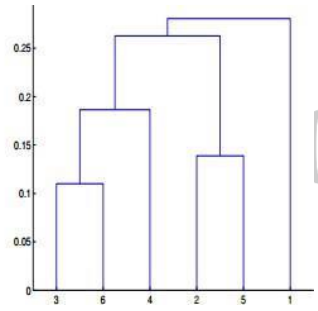
b.



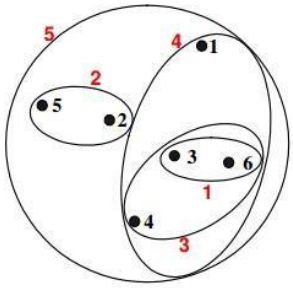


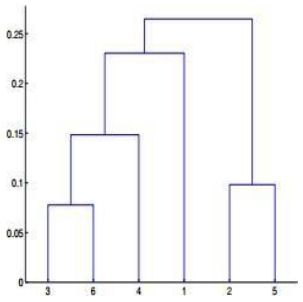
c.





d.





MACHINE LEARNING

12. 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

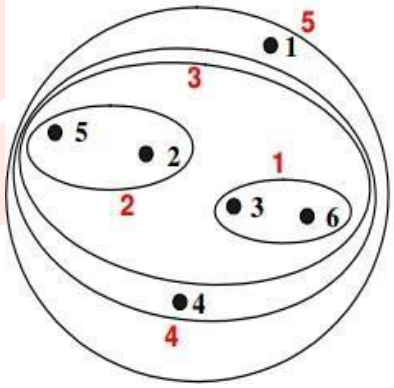
Table : X-Y coordinates of six points.

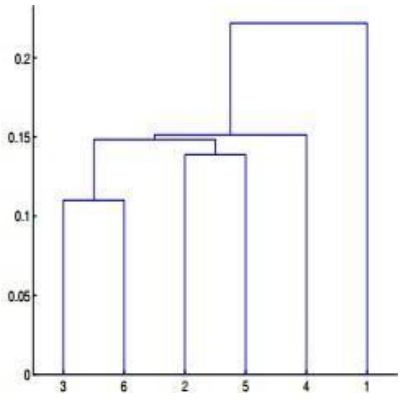
	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

Table : Distance Matrix for Six Points

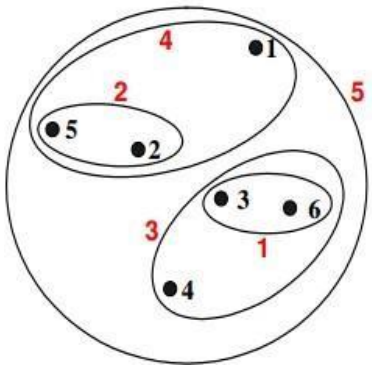
Which of the following clustering representations and dendrogram depicts the use of MAX or Complete link proximity function in hierarchical clustering.

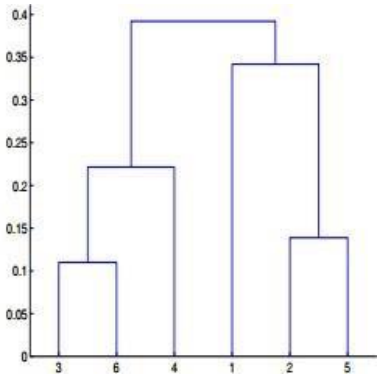
a.





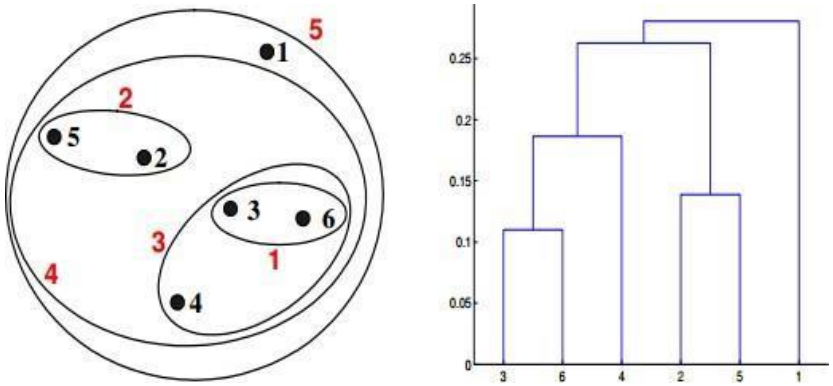
b.



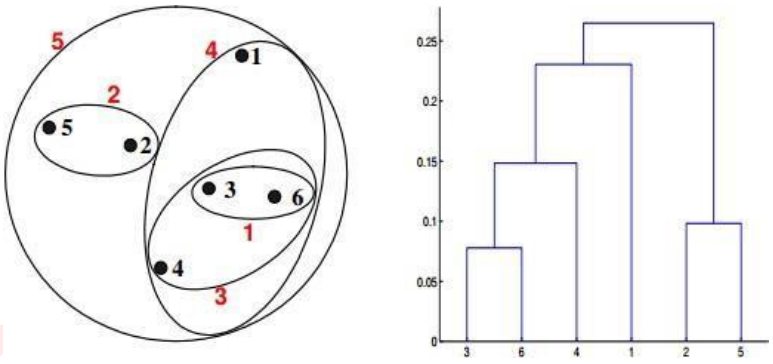


MACHINE LEARNING

c.



d.



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

13. What is the importance of clustering?

Answer: Clustering is an unsupervised machine learning technique that is used to partition a dataset into groups, or clusters, of similar observations. Clustering is important for several reasons:

1. Data Exploration: Clustering can help to explore and understand the underlying structure of a dataset. It can help to identify patterns and groups of similar observations, which can provide insights into the data and lead to new hypotheses.

2. Dimensionality Reduction: Clustering can be used as a dimensionality reduction technique by grouping similar observations together, which can make it easier to visualize and analyze the data.

3. Anomaly Detection: Clustering can be used to identify observations that do not belong to any cluster, which can be useful for identifying outliers or anomalies in the data.

4. Image and Text Segmentation: Clustering is used in image and text segmentation to group similar regions or words together. This can be used in image recognition and natural language processing.

5. Customer Segmentation: Clustering can be used in customer segmentation to group customers with similar characteristics together. This can be used to target marketing campaigns and improve customer service.

6. Market Segmentation: Clustering can be used in market segmentation to group potential customers with similar characteristics together. This can be used to understand market trends, develop new products and services, and target different market segments.

7. Recommender Systems: Clustering can be used in recommendation systems to group similar users or items together. This can be used to make personalized recommendations.

MACHINE LEARNING

14. How can I improve my clustering performance?

Answer: There are several ways to improve the performance of a clustering algorithm:

1. Feature Selection: Selecting the most relevant features can improve the performance of a clustering algorithm by reducing noise and dimensionality of the data.
2. Pre-processing: Data pre-processing such as normalization, scaling, and outlier removal can improve the performance of a clustering algorithm by removing noise and ensuring that the data is in a suitable format for the algorithm.
3. Selecting the right clustering algorithm: Different clustering algorithms have different strengths and weaknesses. Selecting the right algorithm for your data and problem can improve the performance of the clustering.
4. Hyperparameter tuning: Clustering algorithms have several hyperparameters that can be adjusted to improve performance. Tuning these parameters can improve the performance of the clustering algorithm.
5. Evaluating the clustering: It is important to evaluate the clustering performance by using appropriate evaluation metrics such as silhouette score, Davies-Bouldin index, or adjusted Rand index. This can help to identify which clustering algorithm and parameter settings produce the best results.
6. Ensemble Clustering: Combining multiple clustering algorithms together can improve the performance of the clustering.
7. Handling missing data: Handling missing data is important for improving clustering performance. One way to handle missing data is to use imputation methods such as k-nearest neighbors imputation or mean imputation to fill in the missing values.