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 modelling
   d. All of the above

   Answer: 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

   Answer: 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

   Answer: c) Reinforcement learning and Unsupervised learning

4. The final output of Hierarchical clustering is-
   a. The number of cluster centroids
   b. The tree represents 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: b) The tree represents 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: 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

   Answer: c) k-nearest neighbour is same as k-means

MACHINE LEARNING

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: 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

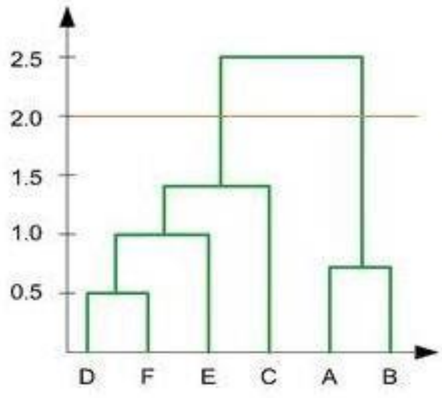
b. 2 only

c. 1 and 2

d. None of them

Answer: 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

Answer: 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.

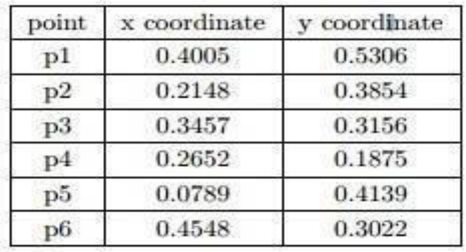
c. Predicting whether the stock price of a company will increase tomorrow.

d. Given historical weather records, predict if tomorrow's weather will be sunny or rainy.

Answer: b) Given a database of information about your users, automatically group them into different market segments.

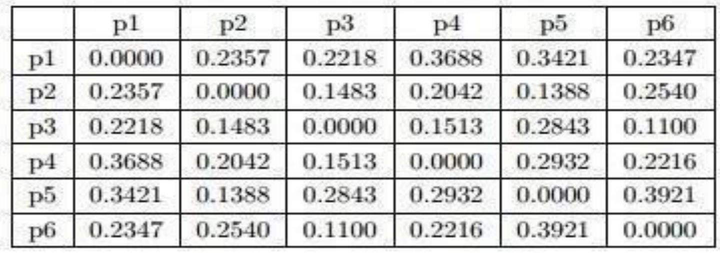
MACHINE LEARNING

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

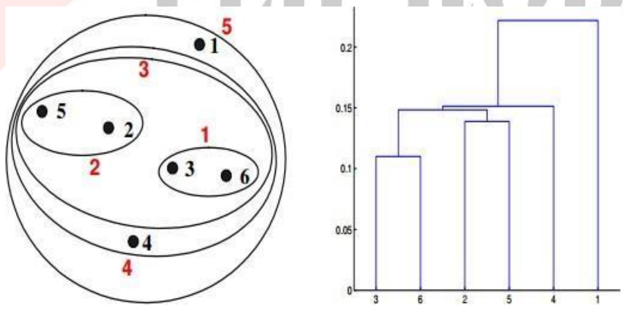
Table : X-Y coordinates of six points.



	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

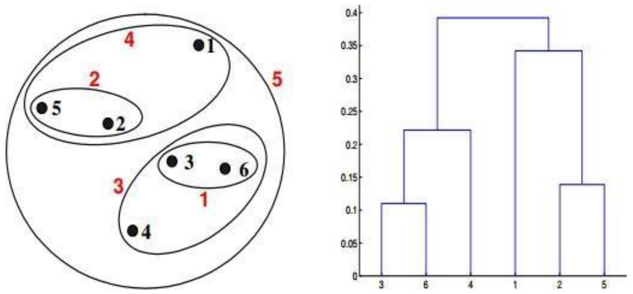
Table : Distance Matrix for Six Points

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

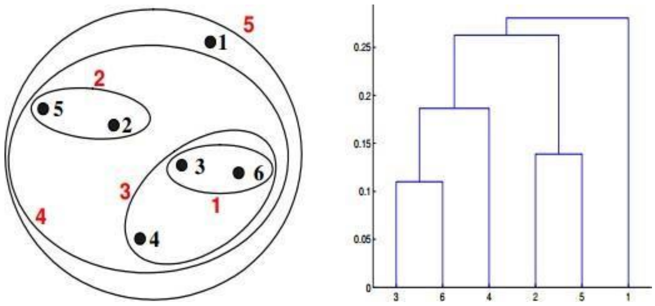


a.

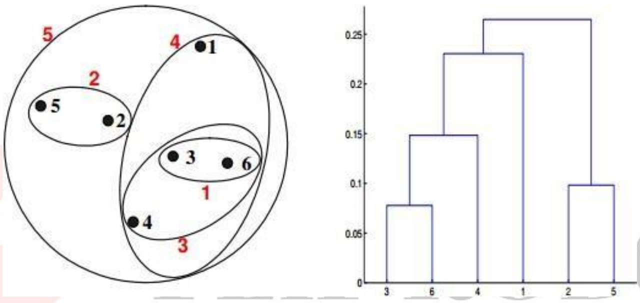
b.


4
1
5
2
5
2
3
6
3
1
4
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
3 6 4 1 2 5

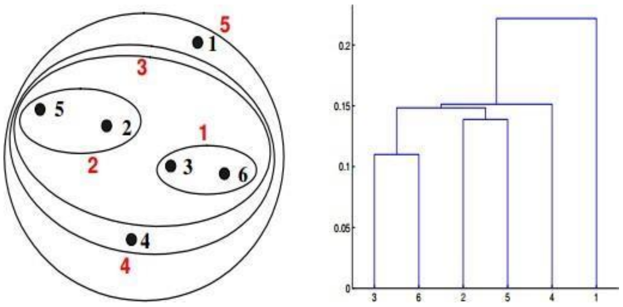
c.


5
1
2
5
2
3
6
3
1
4
4
0.25
0.2
0.15
0.1
0.05
0
3 6 4 2 5 1

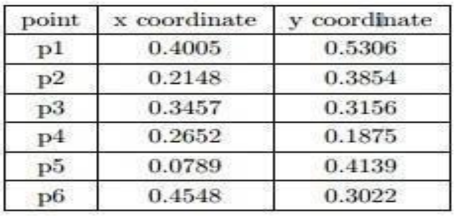
d.


5
4
1
2
5
2
3
6
1
4
3
0.25
0.2
0.15
0.1
0.05
0
3 6 4 1 2 5

Answer: a)

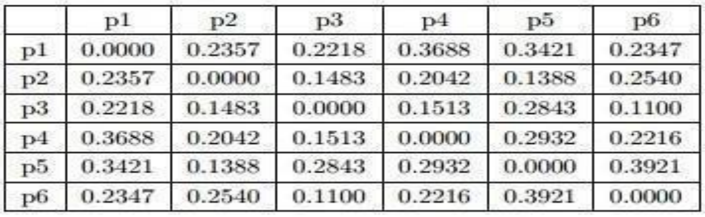

5
1
3
5
2
1
2
3
6
4
4
0.2
0.15
0.1
0.05
0
3 6 2 5 4 1

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.

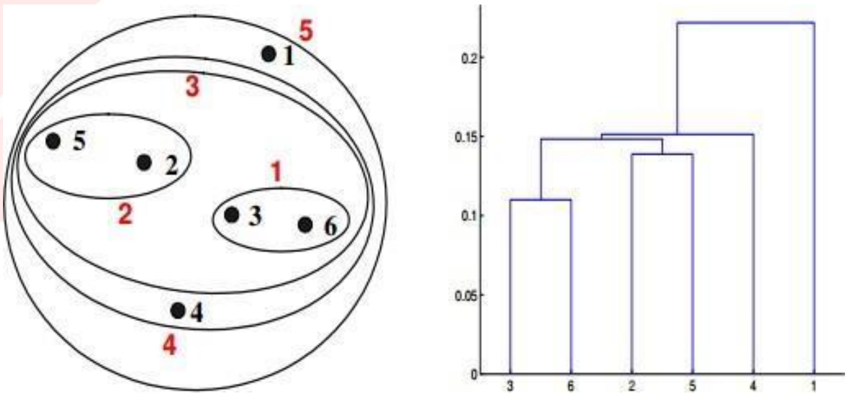


	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

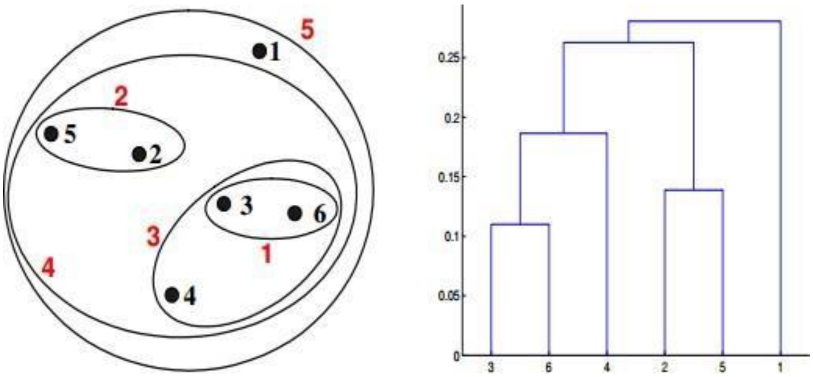
Table : Distance Matrix for Six Points

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

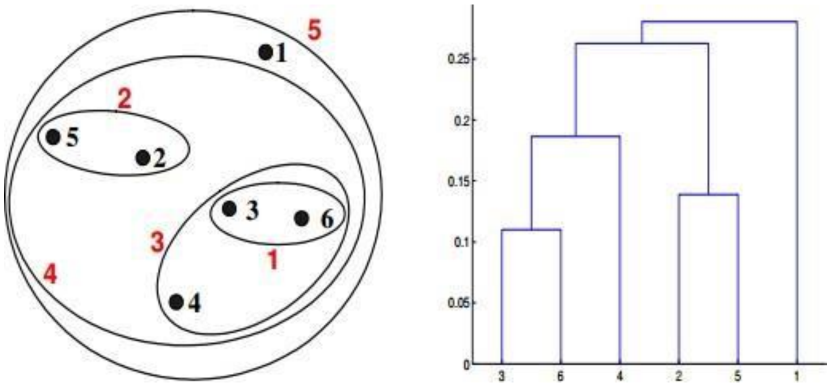
a.



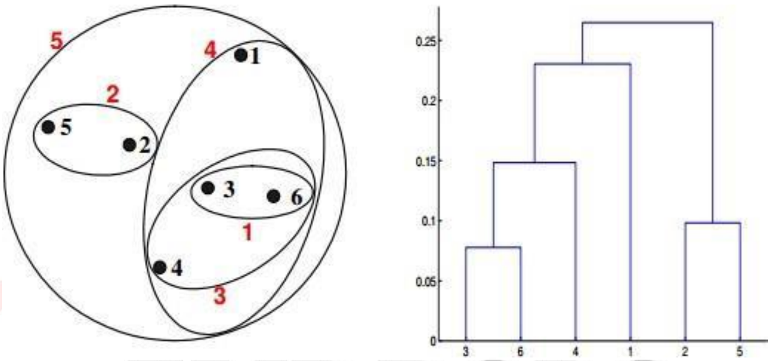
b.



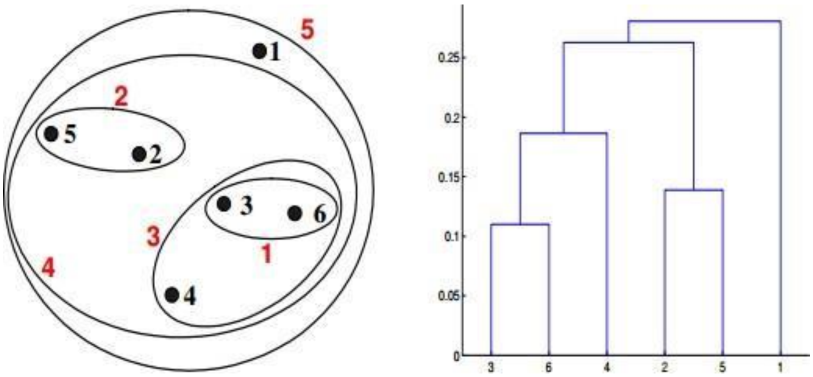
c.


5
1
2
5
2
3
6
3
1
4
4
0.25
0.2
0.15
0.1
0.05
0
3
6
4
2
5
1

d.


5
4
1
2
5
2
3
6
1
4
3
0.25
0.2
0.15
0.1
0.05
0
3
6
4
1
2
5

Answer: b)


5
1
2
5
2
3
6
3
1
4
4
0.25
0.2
0.15
0.1
0.05
0
3
6
4
2
5
1

MACHINE LEARNING

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

13. What is the importance of clustering?

Answer:

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) to each other than to those in other groups (clusters).

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

Cluster analysis itself is not one specific algorithm, but the general task to be

Solved.

14. How can I improve my clustering performance?

Answer:

K-means is a numerical, unsupervised, non-deterministic, iterative method. It is simple and very fast, so in many practical applications, the method is proved to be a very effective way that can produce good clustering results. The standard k-means algorithm is effective in producing clusters for many practical applications. But the computational complexity of the original k-means algorithm is very high in high-dimensional data.

The method for improving the accuracy and efficiency by reducing dimension and initializing the cluster for modified k-means using **PCA** (PRINCIPAL COMPONENT ANALYSIS).