MACHINE LEARNING

Question) 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

Question) On which data type, we cannot perform cluster analysis ?

a. Time series data

b. Text data

c. Multimedia data

d. None

Answer) None

Question) 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

Question) 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

Question) 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

Question) 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

Question) 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

Answer) Single-Link, Complete-Link, Average-link

Question) Which of the following are true ?

i.	Clustering analysis is negatively affected by multicollinearity of features
ii.	Clustering analysis is negatively affected by heteroscedasticity

Answer) Clustering analysis is negatively affected by multicollinearity of features.

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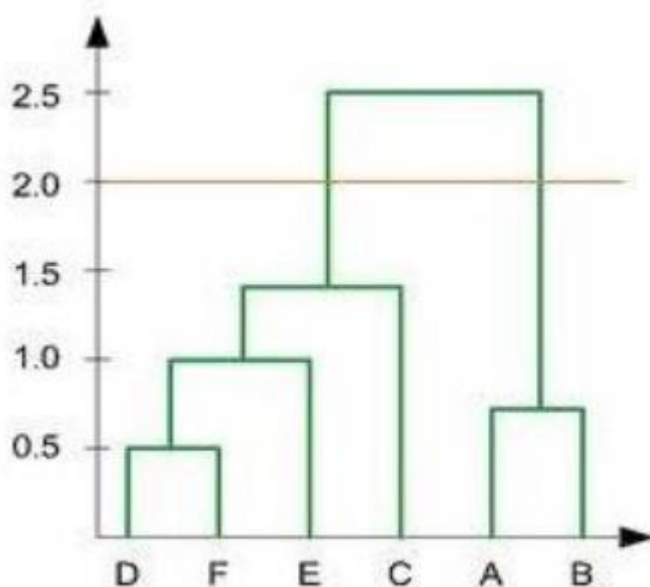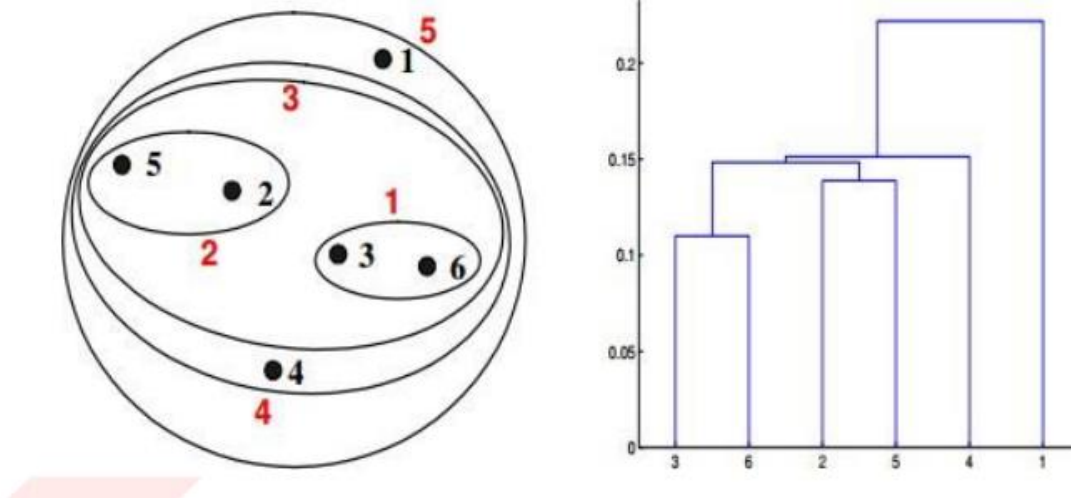

a) 2
b) 4
c) 3
d) 5

Answer) 2

Question) 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) a) Assign each point to its nearest cluster

c) Update the cluster centroids based the current assignment


Question) 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.

c.

d.
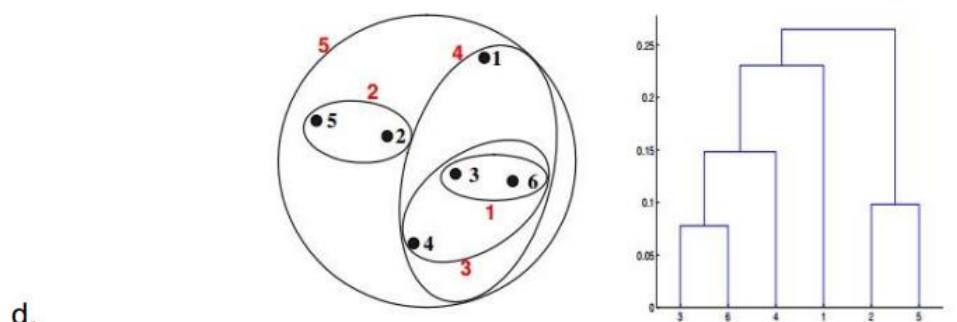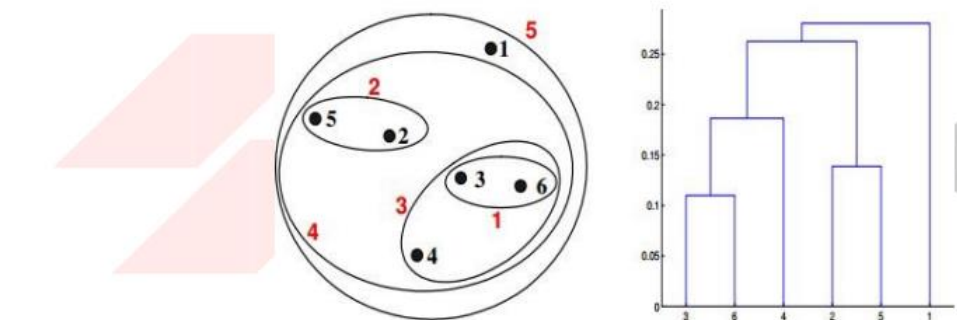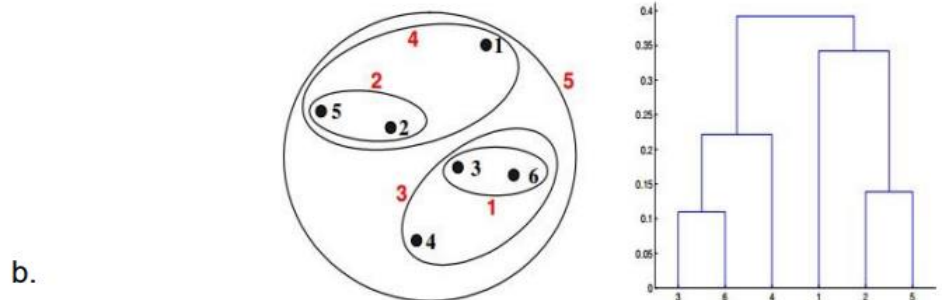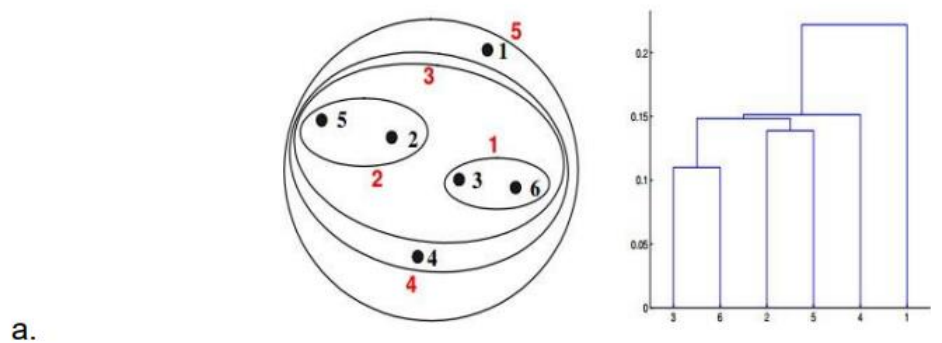
Answer) Explanation: For the single link or MIN version of hierarchical clustering, the proximity of two clusters is defined to be the minimum of the distance between any two points in the different clusters. For instance, from the table, we see that the distance between points 3 and 6 is 0.11, and that is the height at which they are joined into one cluster in the dendrogram. As another example, the distance between clusters {3, 6} and {2, 5} is given by dist({3, 6}, {2,5}) = min(dist(3, 2), dist(6, 2), dist(3, 5), dist(6, 5)) = min(0.1483, 0.2540, 0.2843, 0.3921) = 0.1483

Question) 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

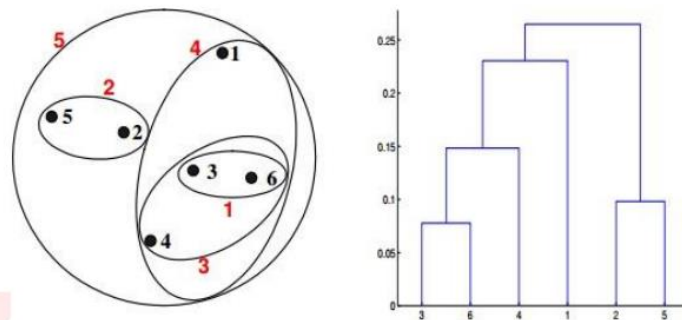**Table :** Distance Matrix for Six Points

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


a.

b.

c.

d.
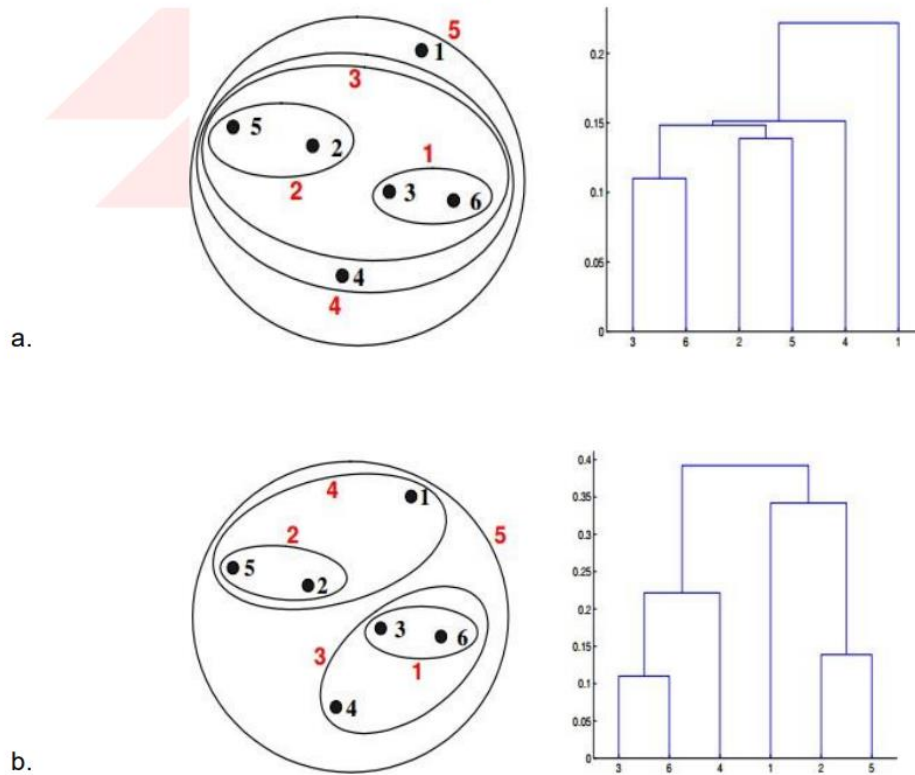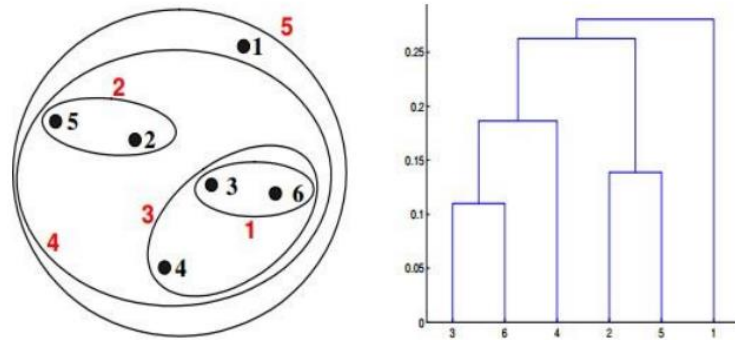
Answer) Explanation: For the single link or MAX version of hierarchical clustering, the proximity of two clusters is defined to be the maximum of the distance between any two points in the different clusters. Similarly, here points 3 and 6 are merged first. However, {3, 6} is merged with {4}, instead of {2, 5}. This is because the dist({3, 6}, {4}) = max(dist(3, 4), dist(6, 4)) = max(0.1513, 0.2216) = 0.2216, which is smaller than dist({3, 6}, {2, 5}) = max(dist(3, 2),dist(6, 2), dist(3, 5), dist(6, 5)) = max(0.1483, 0.2540, 0.2843, 0.3921) = 0.3921 and dist({3, 6}, {1}) = max(dist(3, 1),dist(6, 1)) = max(0.2218, 0.2347) = 0.2347.
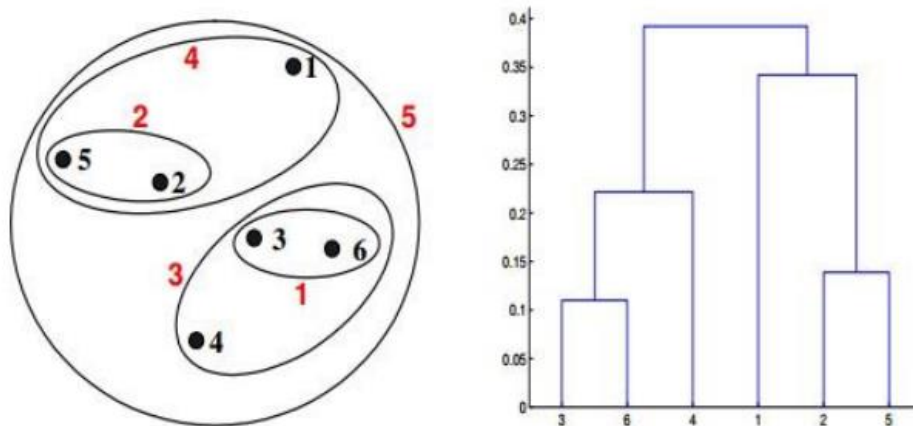
Question) What is the importance of clustering ?

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). Clustering can be done by the different no.

Question) How can I improve my clustering performance ?

Answer) Graph-based clustering performance can easily be improved by applying ICA blind source separation during the graph Laplacian embedding step. Applying unsupervised feature learning to input data using either RICA or SFT, improves clustering performance.