1.Which of the following is an application of clustering?

a. Bio-logical network analysis

b. Market trend prediction

c. Topic Modelling

d. All of the above

Ans is: 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

Ans is: 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

Ans is: c. 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

Ans is: b. 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

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

Ans is: c. 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

Ans is: 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 the

Ans is: a. 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

Ans is: a. 2

Since the number of vertical lines intersecting the red horizontal line at y=2 in the dendrogram are 2, therefore, two clusters will be formed.

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.

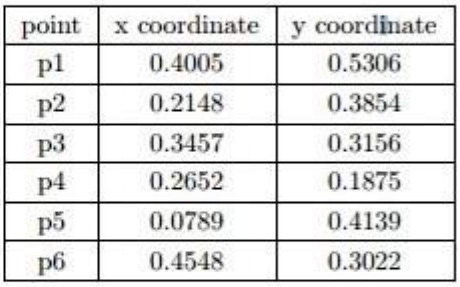
b. Given a database of information about your users, automatically group them into different market segments.

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.

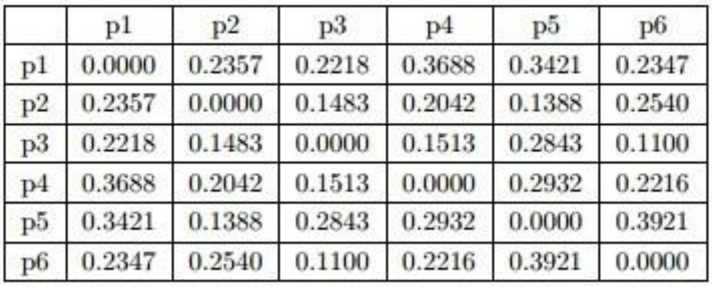
Ans is: b. Given a database of information about your users, automatically group them into different market segments.

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.

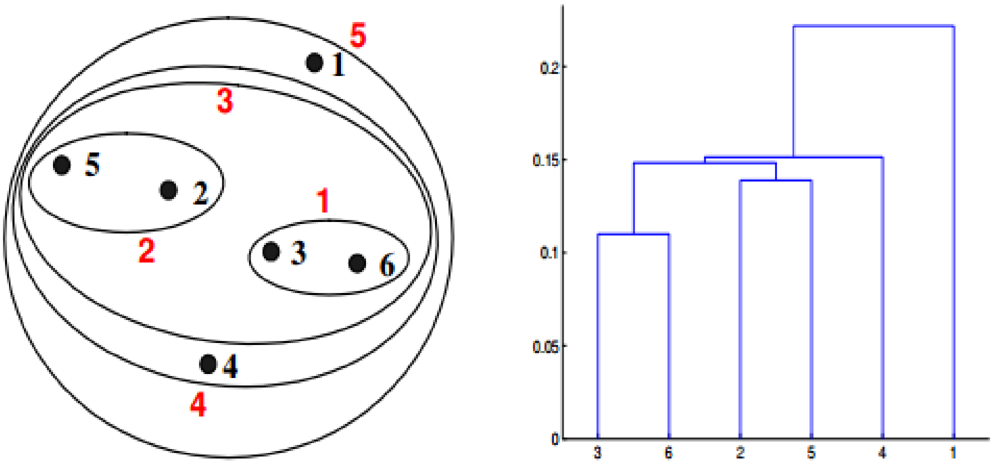


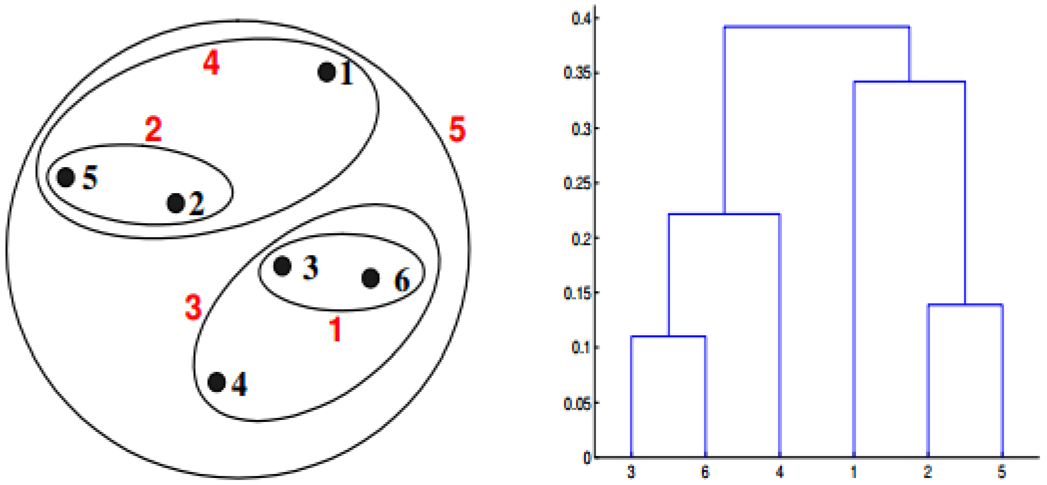
	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

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：

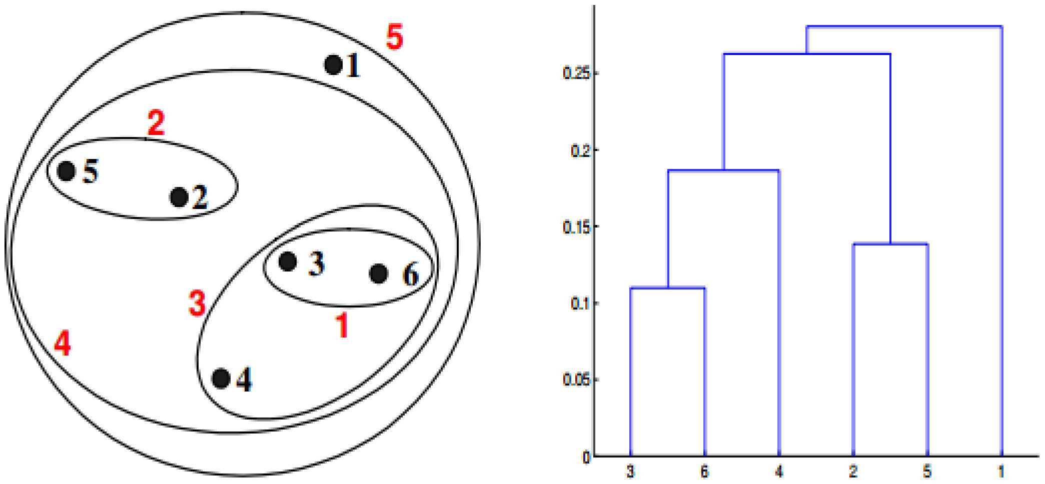
a.

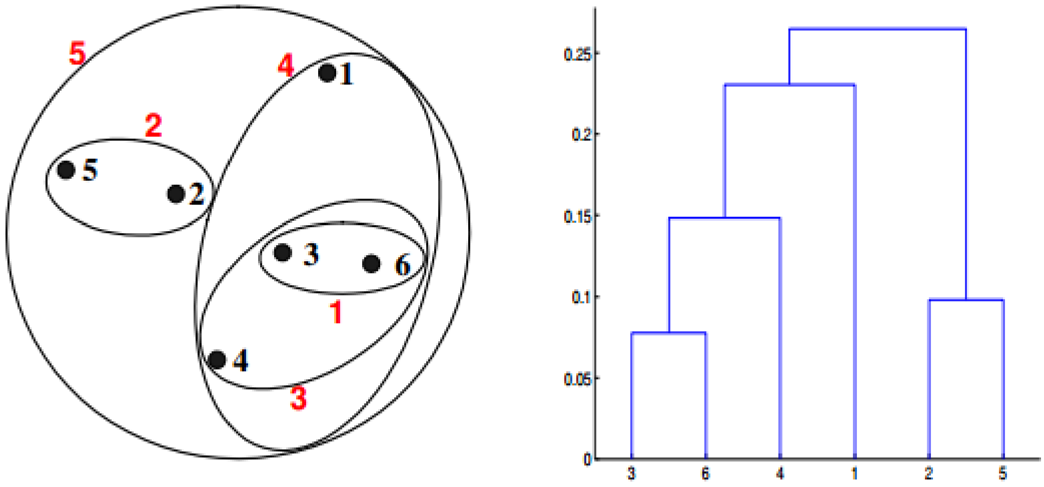



4
1
5
2
5
2
3
6
3
1
4
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
3
6
4
1
2
5

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

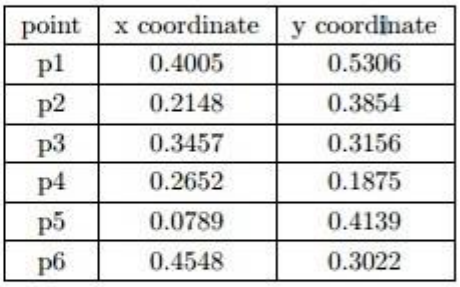

5
4
1
2
5
2
3
6
1
4
3
0.25
0.2
0.15
0.1
0.05
0
3
6
4
1
2
5

d.

Ans is: a

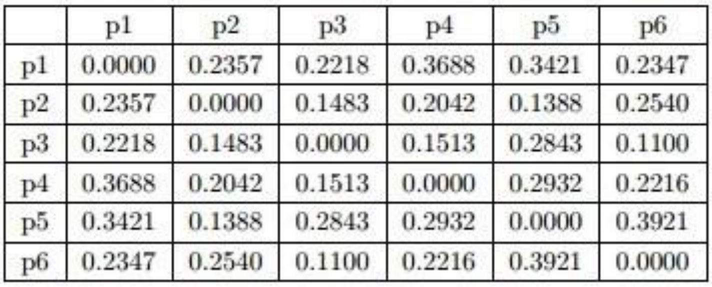
For the single link or MIN version of hierarchical clustering, the proximity of two clusters is defined to be the minimum of the distance between any two points in the different clusters. For instance, from the table, we see that the distance between points 3 and 6 is 0.11, and that is the height at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

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.

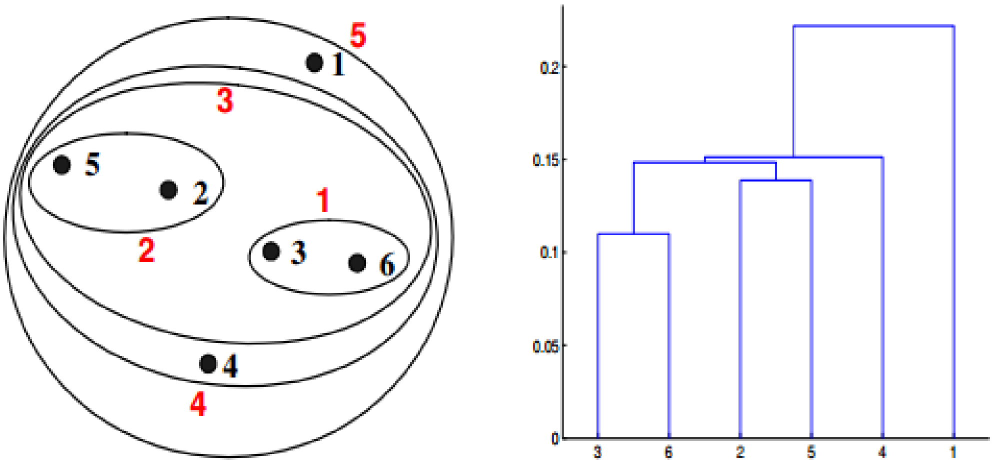


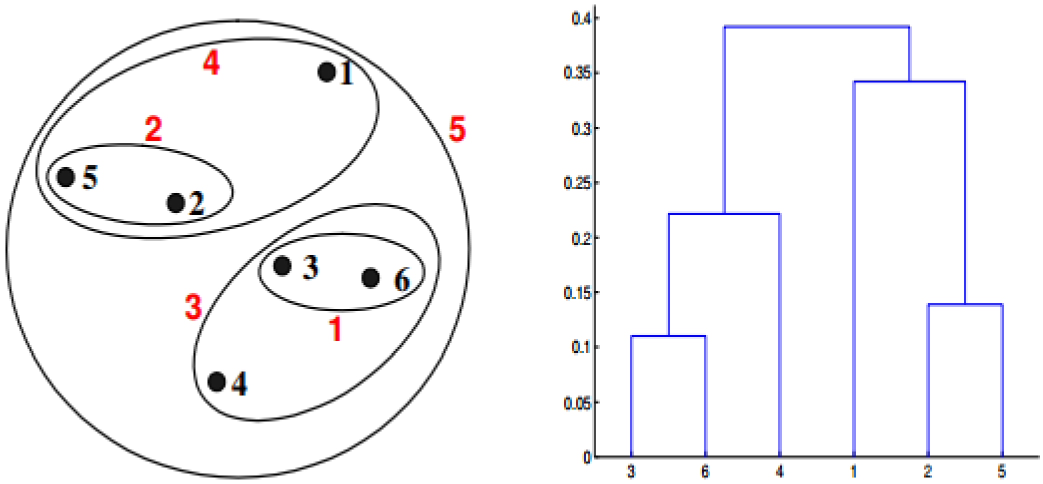
	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

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:

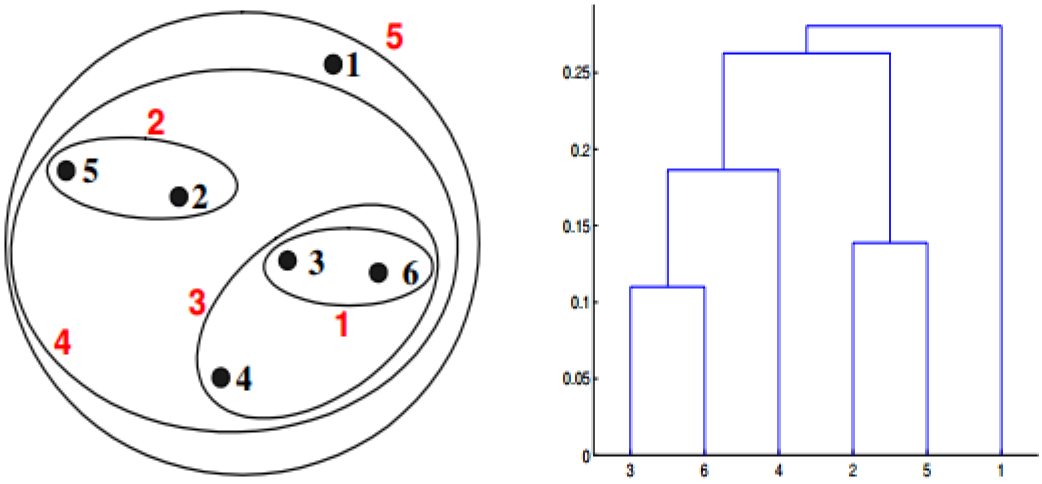
a.


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

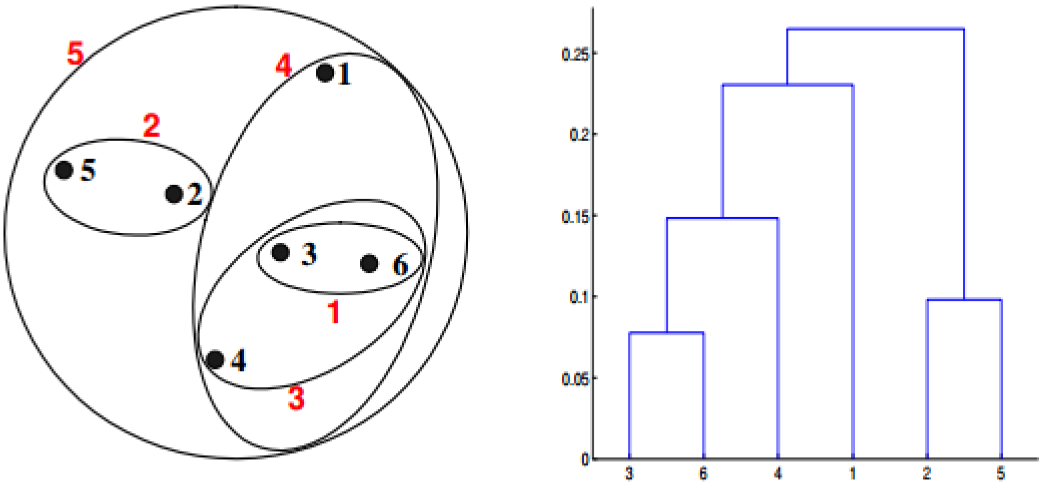

4
1
5
2
5
2
3
6
3
1
4
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
3
6
4
1
2
5

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.



Ans is :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. 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.

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

13. What is the importance of clustering?

- Ans. *Network traffic classification*. Organizations seek various ways of understanding the different types of traffic entering their websites, particularly what is spam and what traffic is coming from bots. Clustering is used to group together common characteristics of traffic sources, then create clusters to classify and differentiate the traffic types. This allows more reliable traffic blocking while enabling better insights into driving traffic growth from desired sources.
- *Marketing and sales*. Marketing success means targeting the right people or prospects in the right way. Clustering algorithms group together people with similar traits, perhaps based on

their likelihood to purchase. With these groups or clusters defined, test marketing across them becomes more effective, helping to refine messaging to reach them.

- *Document analysis*. Any organization dealing with high volumes of documents will benefit by being able to organize them effectively and quickly as they’re generated. That means being able to understand underlying themes in the documents, and then being able to compare that to other documents. Clustering algorithms examine text in documents, then group them into clusters of different themes. That way they can be speedily organized according to actual content.

14. How can I improve my clustering performance?

Ans.

- 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.
- Surprisingly for some cases, high clustering performance can be achieved by simply performing K-means clustering on the ICA components after PCA dimension reduction on the input data. However, the number of PCA and ICA signals/components needs to be limited to the number of unique classes.