Name of the Assignment: Machine Learning (Worksheet-3)

Submitted by : Shalini Joshi

Designation : Data Science Intern

Date of Submission : 31st Dec,2022

Objective Type Questions:

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

Ans: 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: 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: 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: 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

Ans: 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: 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: d. 1, 2 and 3

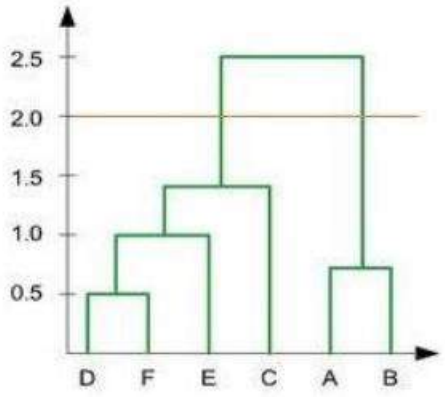
8. Which of the following are true?

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

Ans: a. 1 only

9. In the figure below, 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

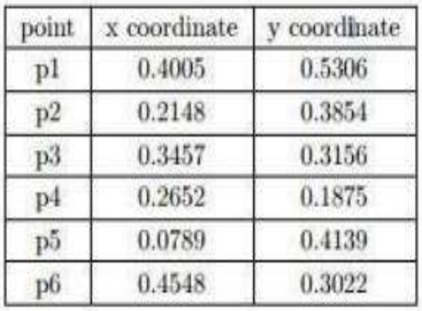
Ans: a. 2

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 stock price of a company will increase tomorrow. d. Given historical weather records, predict if tomorrow's weather will be sunny or rainy.

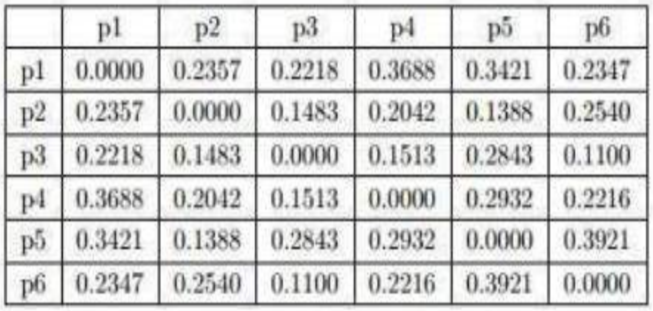
Ans: b. Given a database of information about your users, automatically group them into different market segments

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.

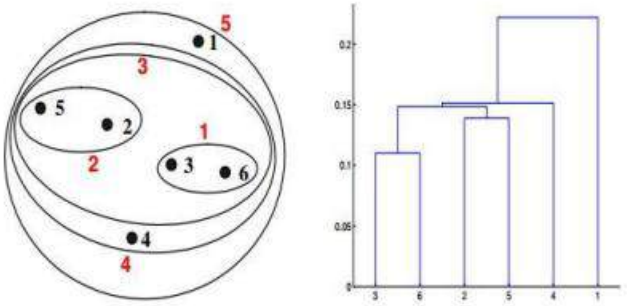


	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

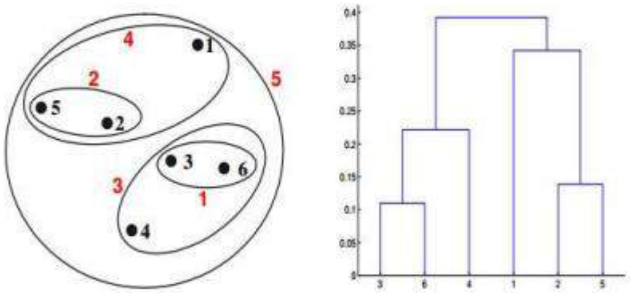
Table : Distance Matrix for Six Points

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

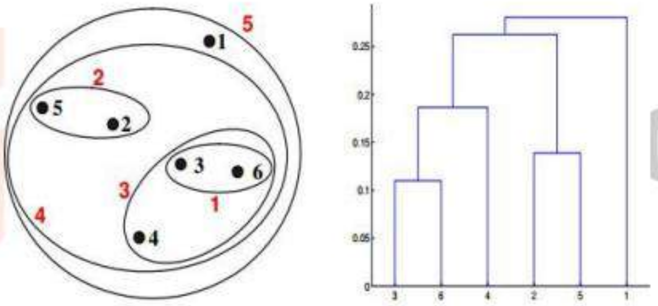
a.



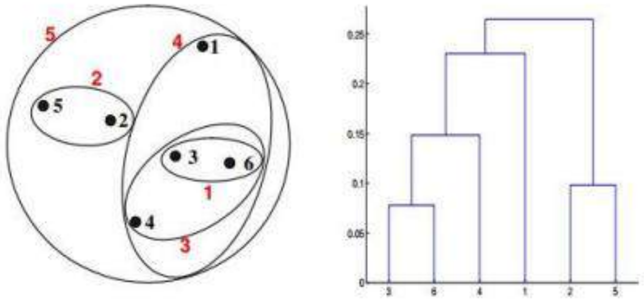
b.



c.



d.



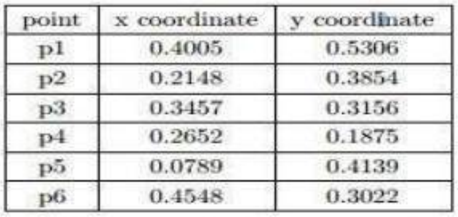
Ans: a

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.

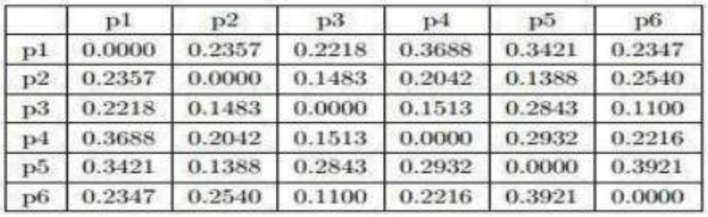
For instance, from the table, we see that the distance between points 3 and 6 is 0.11, and that is the height at which they are joined into one cluster in the dendrogram.

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

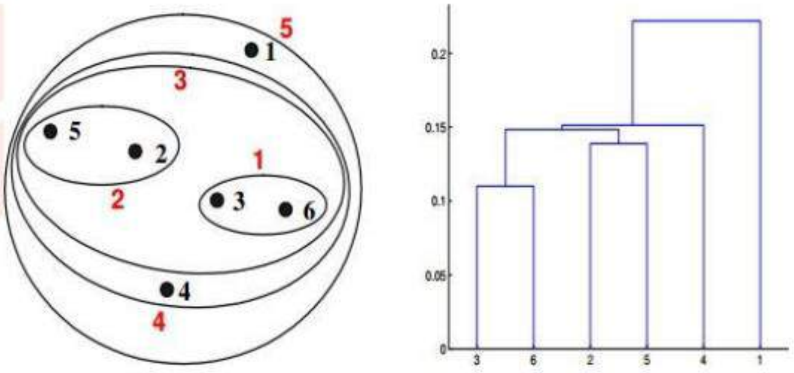
Table : X-Y coordinates of six points.



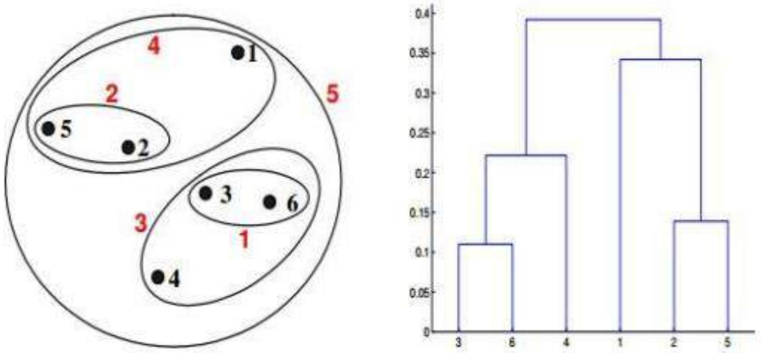
	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

Table : Distance Matrix for Six Points

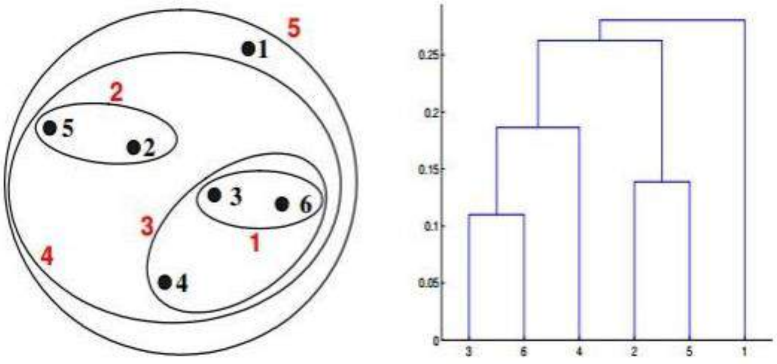
Which of the following clustering representations and dendrogram depicts the use of MAX or Complete link proximity function in hierarchical clustering.


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

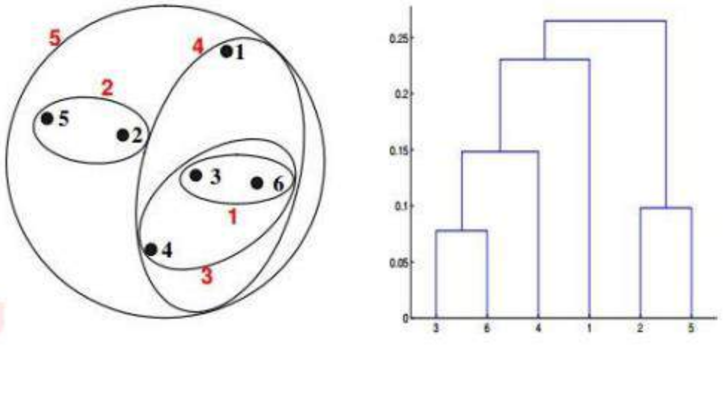
a.


4
1
2
5
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.


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

c.


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: b

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.

Subjective Type Questions:

13. What is the importance of clustering?

Ans: The method of identifying similar groups of data in a dataset is called clustering. It is main task of exploratory data mining, statistical data analysis, used in fields like Machine Learning, Pattern recognition, image analysis and bioinformatics.

Importance of Clustering:

- ✓ Helps in removing inefficiency and determine the internal structure of the data.
- ✓ Helps to understand natural grouping in a dataset with the purpose to partition data logically.
- ✓ They play a wide role in applications like marketing economic research and weblogs to identify similar measures, Image processing and spatial research.
- ✓ Used in outlier detections to detect credit card fraudulence

14. How can I improve my clustering performance?

Ans: Clustering performance can be improved by:

- Use of Supervised learning
- Use of LSTM networks; Long Short-Term Memory is an advanced version of recurrent neural network (RNN) architecture that was designed to model chronological sequences and their long-range dependencies more precisely than conventional RNNs.
- Use of Deep learning-based clustering
- Use of Spectral Clustering

************************ **The End*********************************