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?
   d. All of the above

2. On which data type, we cannot perform cluster analysis?
   d. None

3. Netflix's movie recommendation system uses-
   c. Reinforcement learning and Unsupervised learning

4. The final output of Hierarchical clustering 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?
   d. None

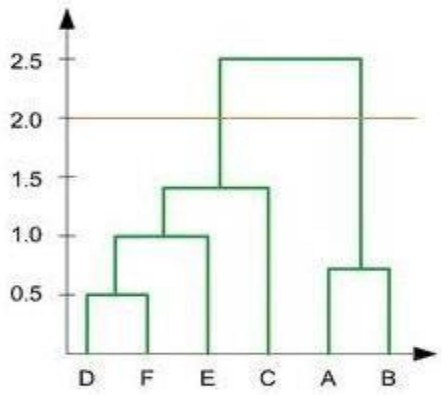
6. Which is the following is wrong?
   c. k-nearest neighbour is same as k-means

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:
   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

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

10. For which of the following tasks might clustering be a suitable approach?
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.

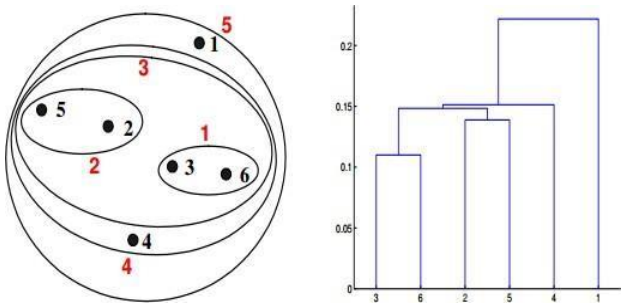
	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

Table : Distance Matrix for Six Points

MACHINE LEARNING

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

a.



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

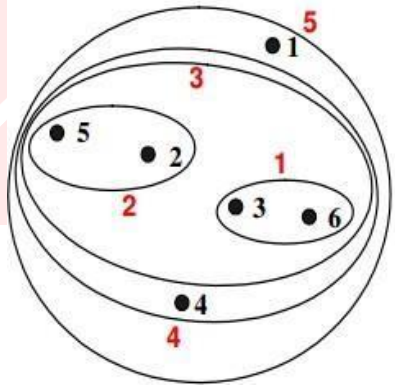
Table : X-Y coordinates of six points.

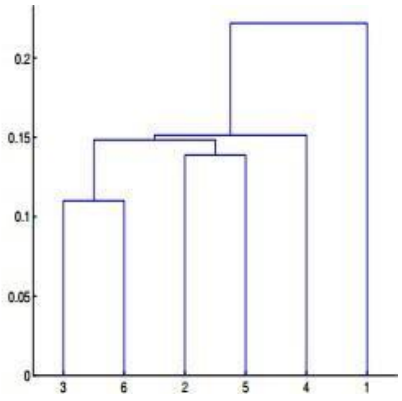
	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

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.





MACHINE LEARNING

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

13. What is the importance of clustering?
14. How can I improve my clustering performance?

13. What is the importance of clustering?

ANS

Clustering refers to making group of data by distinct features in order to make future prediction or to make findings out of data. It's an imp data analysis and data mining applications. It has 3 types of clustering.

1. Centroid based clustering - This clustering technique is simple and efficient among others as it requires K-mean to make group of data.
2. Density based clustering – This technique makes groups based on density of the data. This clustering is useful in case of less dimensional data.
3. Hierarchical Clustering – It creates a tree of clusters and it is suitable for hierarchical data.

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.