MACHINE LEARNING

Q1 to Q12 have only one correct answer. Answers is marked with a yellow highlighter and written separately after each question.

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

MACHINE LEARNING

d. 1, 2 and 3

Ans:- D- 1,2 and 3

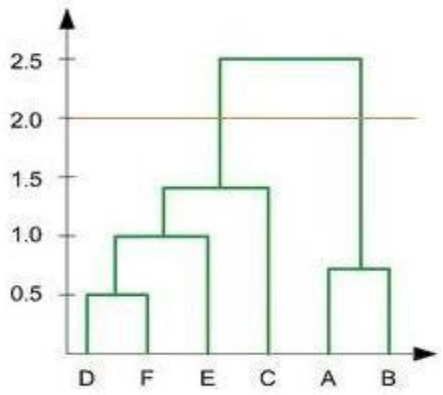
8. Which of the following are true?
   i. Clustering analysis is negatively affected by multicollinearity of features
   ii. Clustering analysis is negatively affected by heteroscedasticity
   Options:
   a. 1 only
   b. 2 only
   c. 1 and 2
   d. None of them

Ans:- A- 1 only

P.T.0

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
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 marketsegments.

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.

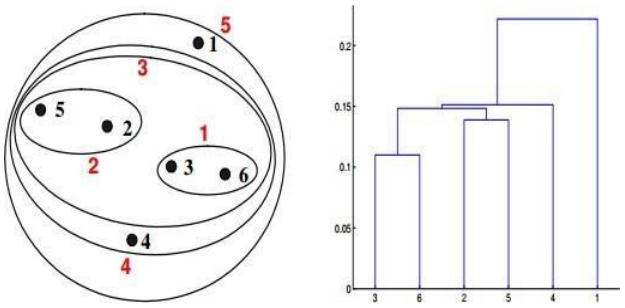
	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

Table : Distance Matrix for Six Points

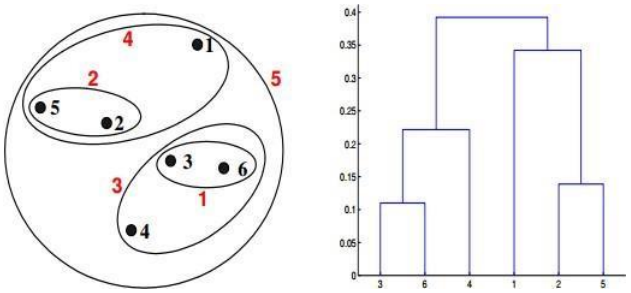
MACHINE LEARNING

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

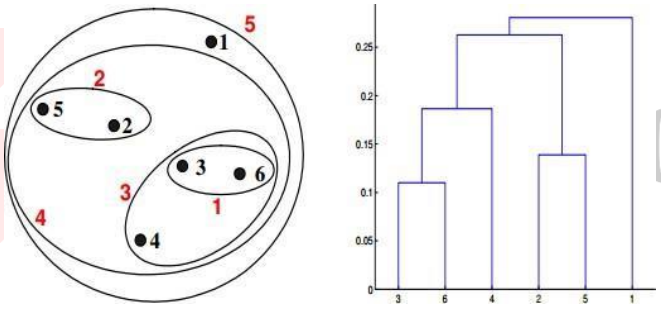
a.



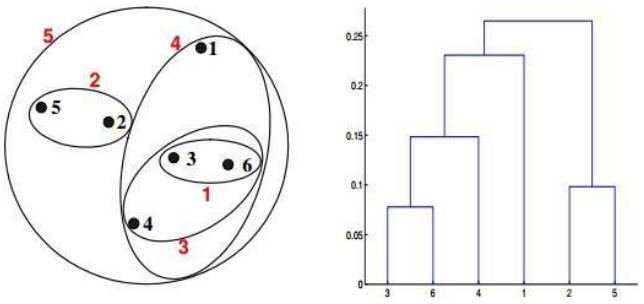
b.



c.



d.



Ans:- A is Correct

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

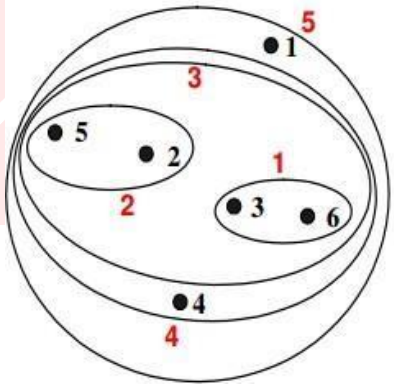
Table : X-Y coordinates of six points.

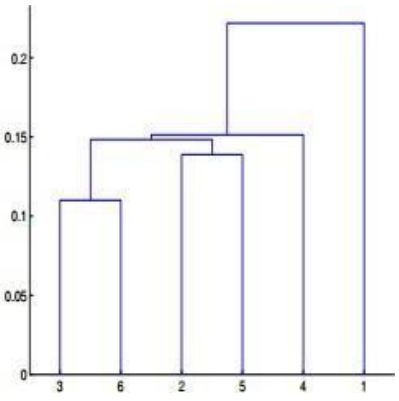
	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

Table : Distance Matrix for Six Points

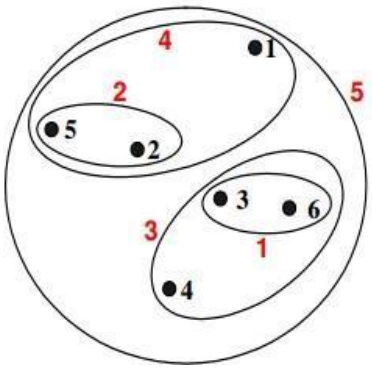
Which of the following clustering representations and dendrogram depicts the use of MAX or Complete link proximity function in hierarchical clustering.

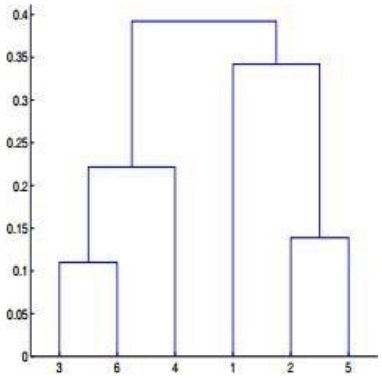
a.





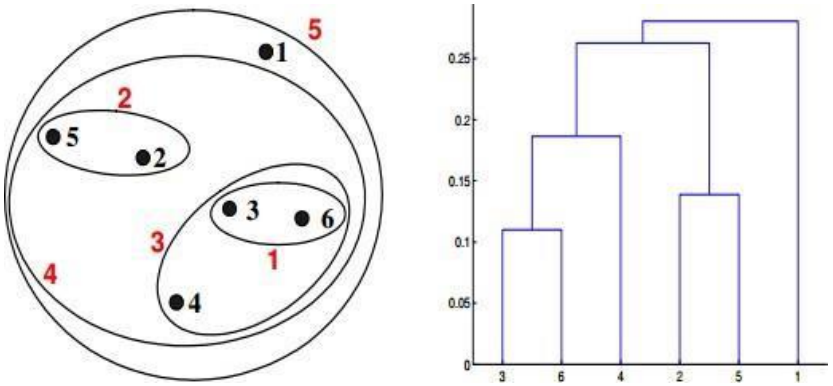
b.



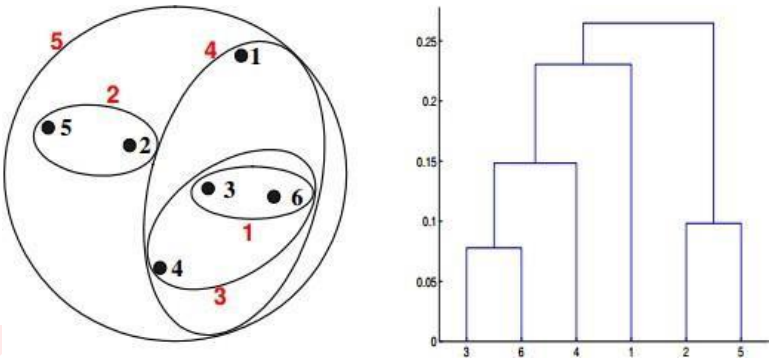


MACHINE LEARNING

c.



d.



Ans:- B Is correct

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

13. What is the importance of clustering?

Answer:-

The process of combining a set of physical or abstract objects into classes of the same objects is known as **clustering**. A cluster is a set of data objects that are the same as one another within the same cluster and are disparate from the objects in other clusters. A cluster of data objects can be considered collectively as one group in several applications. Cluster analysis is an essential human activity.

Cluster analysis is used to form groups or clusters of the same records depending on various measures made on these records. The key design is to define the clusters in ways that can be useful for the objective of the analysis. This data has been **used in several areas, such as astronomy, archaeology, medicine, chemistry, education, psychology, linguistics, and sociology.**

There is one famous use of cluster analysis in **marketing is for market segmentation** − users are segmented based on demographic and transaction history data, and marketing techniques are tailored for each segment.

Another term is for market structure analysis identifying teams of the same products according to competitive measures of similarity. In marketing and political forecasting, clustering of neighborhoods using U.S. postal zip codes has been used strongly to group neighborhoods by lifestyles.

In finance, cluster analysis can be used for making balanced portfolios − Given data on several investment opportunities (e.g., stocks), one can find clusters depending on financial performance variables including return (daily, weekly, or monthly), volatility, beta, and other characteristics, including industry and market

capitalization. Selecting securities from multiple clusters can help make a balanced portfolio.

There is another operation of cluster analysis in finance is **for market analysis**. For a given industry, it is interested in finding teams of the same firms based on measures such as growth rate, profitability, industry size, product range, and presence in several international markets. These teams can then be analyzed to learn the market structure and to decide, for example, who is a competitor.

Cluster analysis can be used for large amounts of data. For example, Internet search engines use clustering methods to cluster queries that users submit. These can then be used for developing search algorithms.

Generally, the basic data used to cluster are a table of measurements on various variables, where each column defines a variable and a row defines a record. The aim is to form groups of data so that the same records are in the same group. The number of clusters can be pre-specified or decided from the data.

14. How can I improve my clustering performance?

Answer:-

By the help of

- Weight training,
- K++ means, etc.

We can improve the clustering performance.