Assignment: Worksheet 3 Machine Learning
Author: Yashshree Bawiskar
Internship Batch: 25

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?**
a. Biological network analysis
b. Market trend prediction
c. Topic modelling
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

Assignment: Worksheet 3 Machine Learning
Author: Yashshree Bawiskar
Internship Batch: 25

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. None

8. **Which of the following are true?**

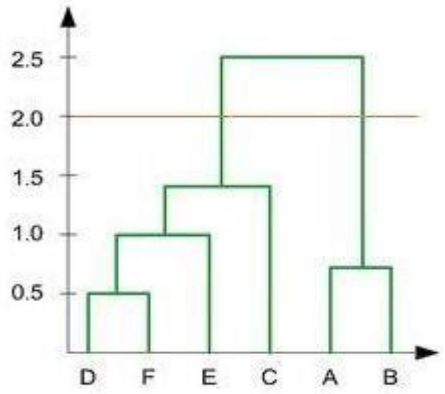
i. Clustering analysis is negatively affected by multicollinearity of features
ii. Clustering analysis is negatively affected by heteroscedasticity Options:

   a. 1 only
   b. 2 only
   c. 1 and 2
   d. None of them

   Ans: a. 1 only

Assignment: Worksheet 3 Machine Learning
Author: Yashshree Bawiskar
Internship Batch: 25

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.
    e.

    Ans: a. Given sales data from a large number of products in a supermarket, estimate future sales for each of these products.

Assignment: Worksheet 3 Machine Learning
Author: Yashshree Bawiskar
Internship Batch: 25

11. **Given, six points with the following attributes:**

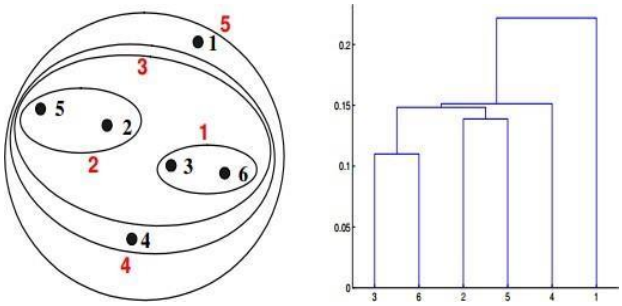
point	x coordinate	y coordinate
p1	0.4005	0.5306
p2	0.2148	0.3854
p3	0.3457	0.3156
p4	0.2652	0.1875
p5	0.0789	0.4139
p6	0.4548	0.3022

Table : X-Y coordinates of six points.

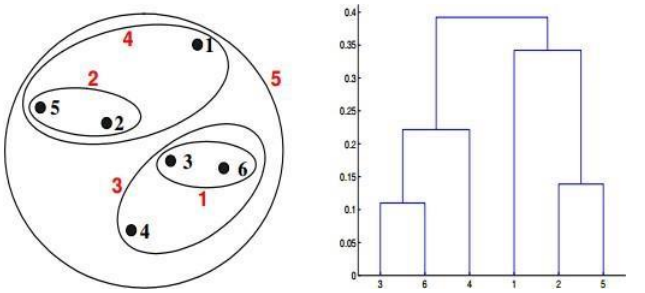
	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

Table : Distance Matrix for Six Points

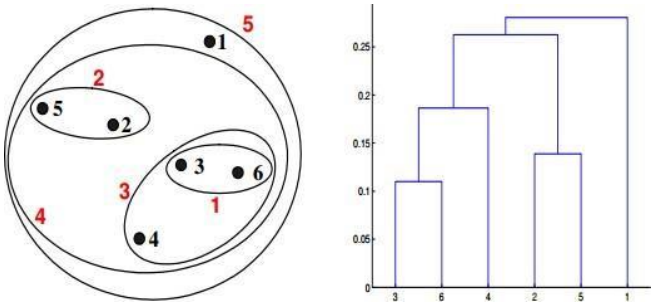
Which of the following clustering representations and dendrogram depicts the use of MIN or Single link proximity function in hierarchical clustering:



a.

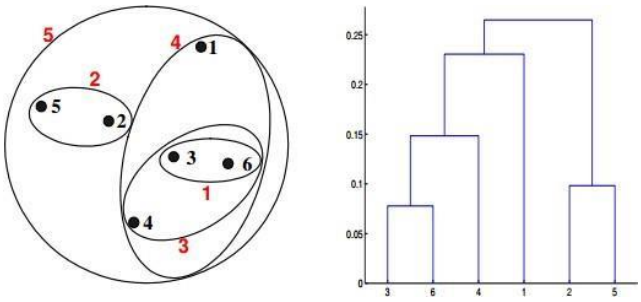


b.



Assignment: Worksheet 3 Machine Learning
Author: Yashshree Bawiskar
Internship Batch: 25

c.



d.

Ans: a

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.

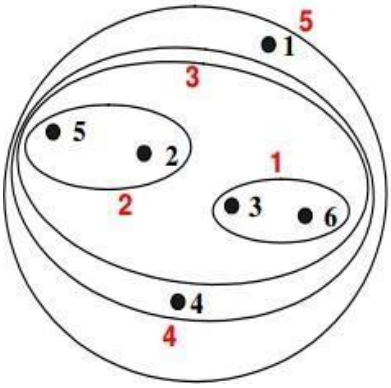
	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

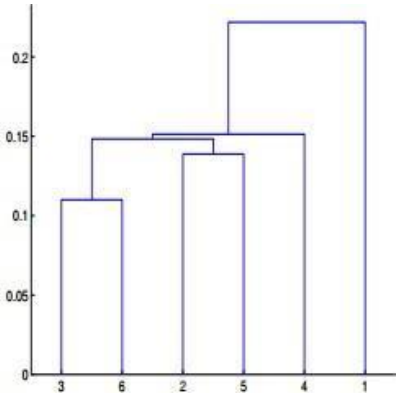
Table : Distance Matrix for Six Points

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

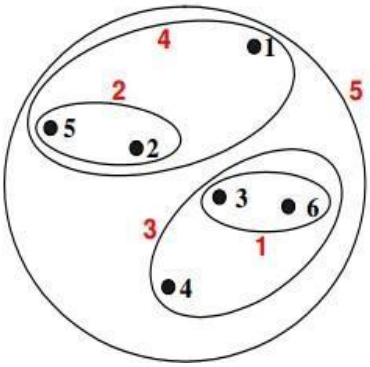
Assignment: Worksheet 3 Machine Learning
Author: Yashshree Bawiskar
Internship Batch: 25

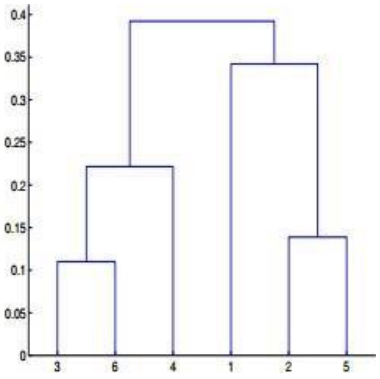
a.



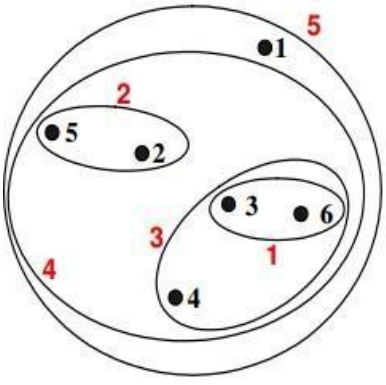


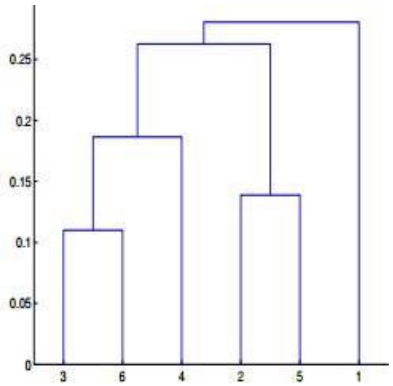
b.





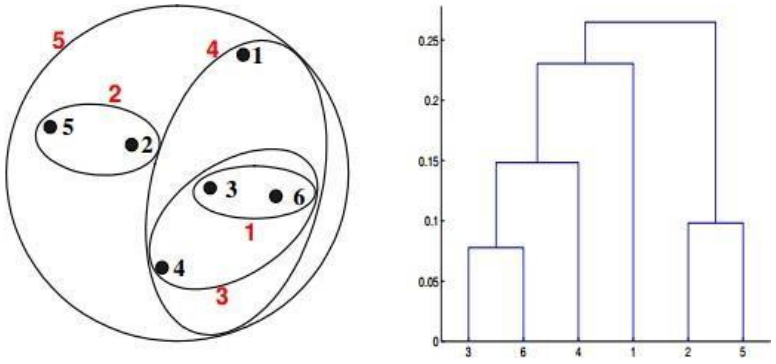
c.





Assignment: Worksheet 3 Machine Learning
Author: Yashshree Bawiskar
Internship Batch: 25

d.



Ans: b

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

13. **What is the importance of clustering?**

- Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters.
- Clustering helps in understanding the natural grouping in a dataset. Their purpose is to make sense to partition the data into some group of logical groupings. Clustering quality depends on the methods and the identification of hidden patterns.
- The biggest advantage of clustering over-classification is it can adapt to the changes made and helps single out useful features that differentiate different groups.
- It can be used to characterize & discover customer segments for marketing purposes. And for classification among different species of plants and animals. By learning the earthquake- affected areas we can determine the dangerous zones.

14. How can I improve my clustering performance?
By performing the visualization and instead of random initialization, we choose only the first center randomly. with a probability that is proportional to their squared distance from all current centres. Points further away from current centres get a higher probability to become a center in the next iteration of initialization.