

MACHINE LEARNING – WORKSHEET (CLUSTERING)

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 steps is not required for K-means clustering?

- a. A distance metric
- b. Initial number of clusters
- c. Initial guess as to cluster centroids
- d. all of the mentioned

Ans: d. all of the mentioned

6. Which of 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 neighbor is same as k-means
- d. None

Ans: c. k-nearest neighbor 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

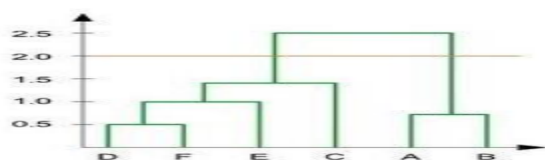
Options:

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

Ans: a. 1 only

9. In the figure 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?
- Given sales data from a large number of products in a supermarket, estimate future sales for each of these products.
 - Given a database of information about your users, automatically group them into different market segments.
 - Predicting whether the stock price of a company will increase tomorrow.
 - 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 |

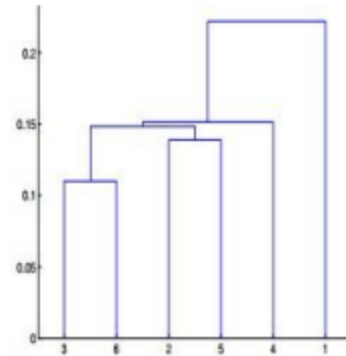
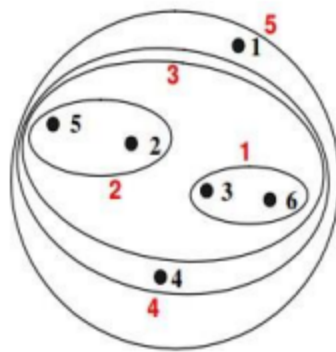
Table : X-Y coordinates of six points.

| | 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 |

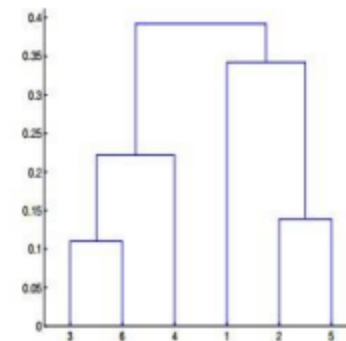
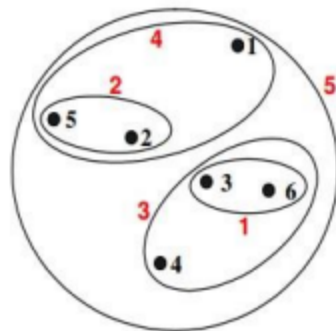
Table : Distance Matrix for Six Points

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

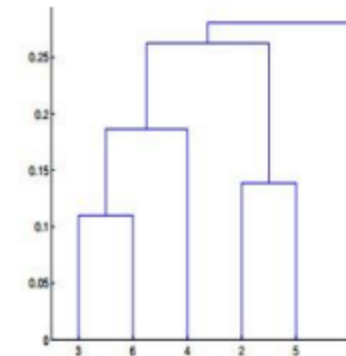
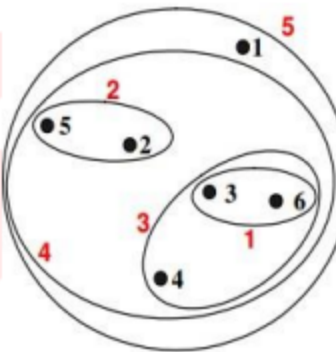
a.



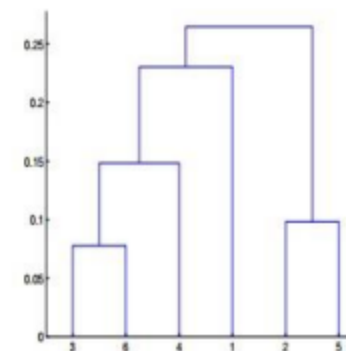
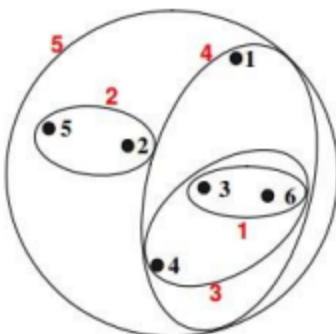
b.

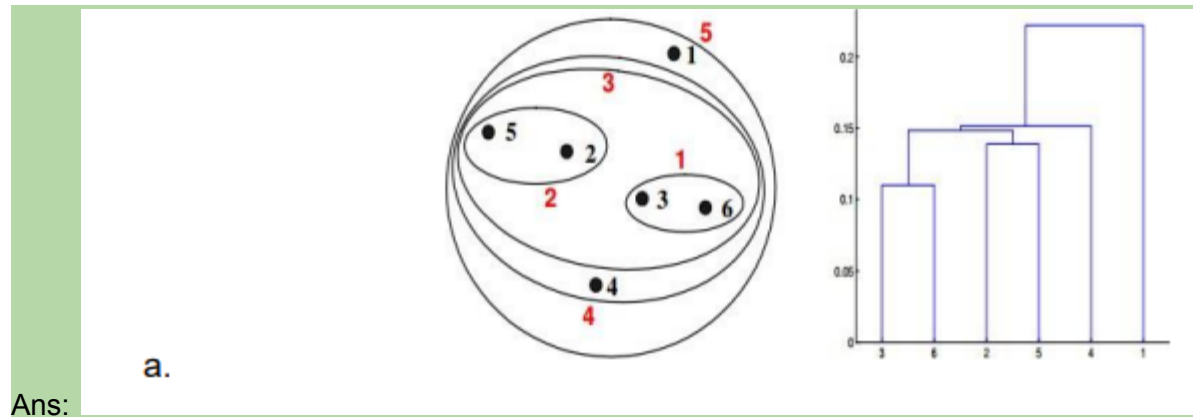


c.



d.





Ans:

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 |

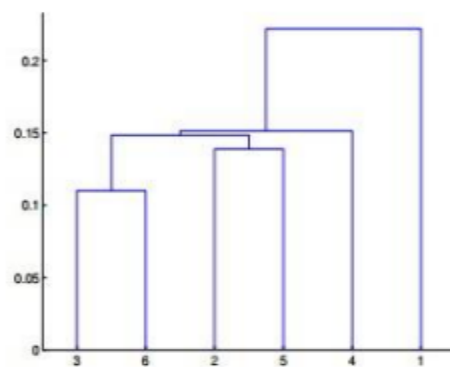
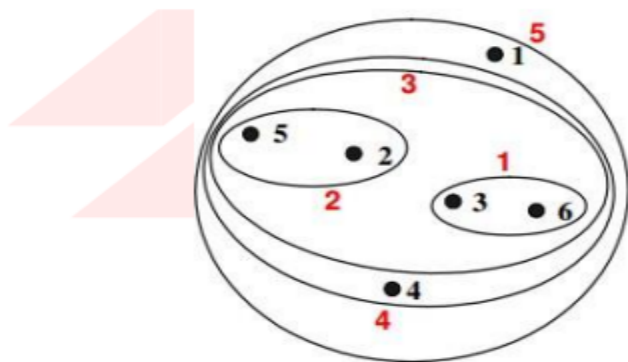
Table : X-Y coordinates of six points.

| | 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 |

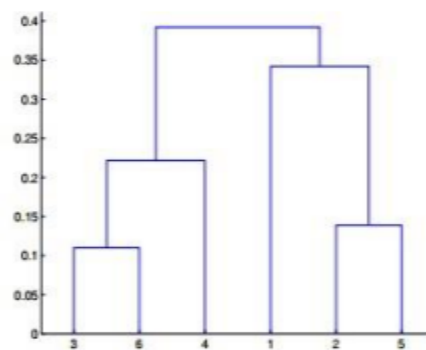
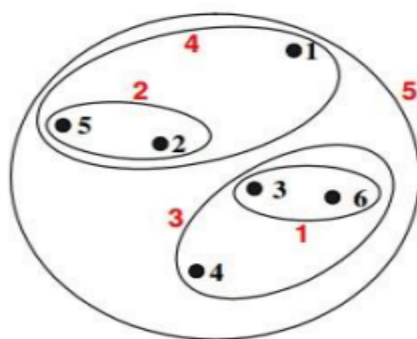
Table : Distance Matrix for Six Points

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

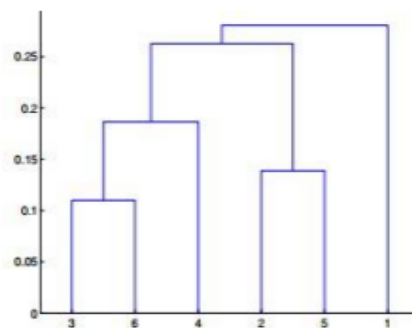
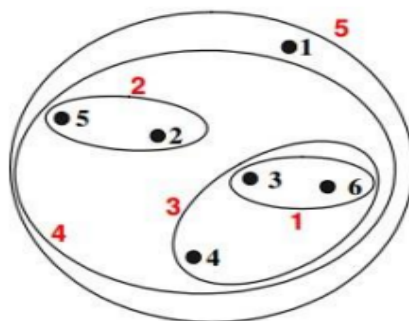
a.



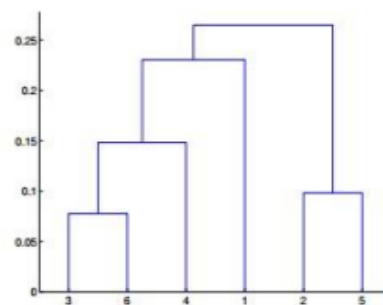
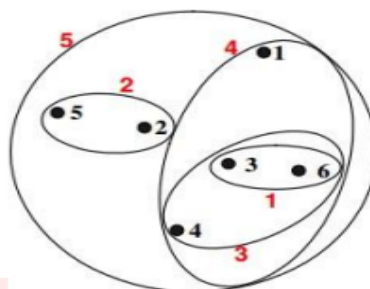
b.



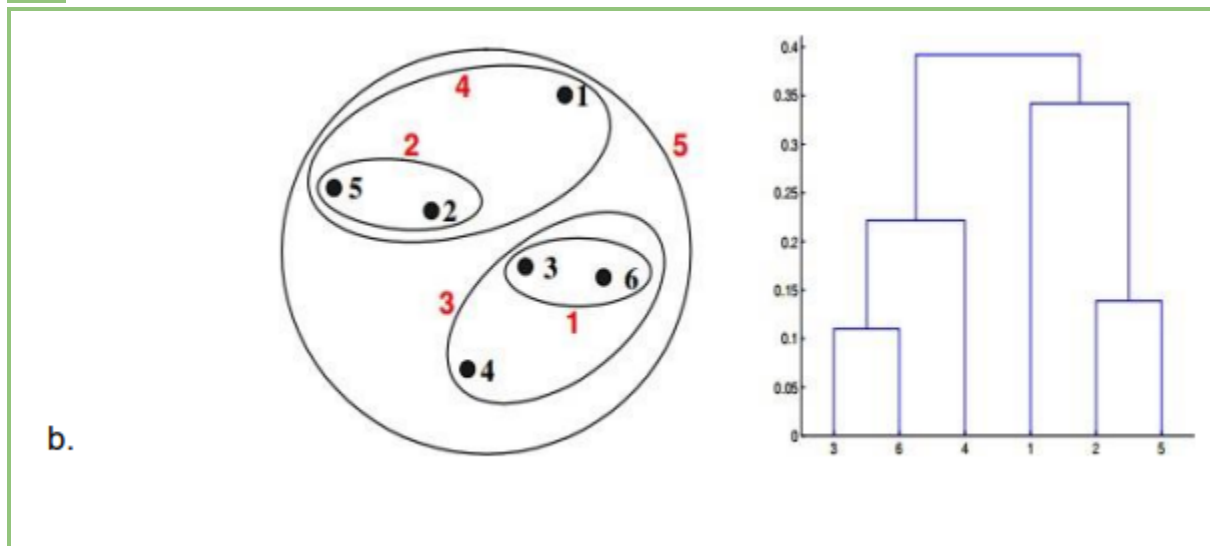
c.



d.



Ans:



13. What is the importance of clustering?

Ans: Clustering is important as, as a data scientist you will not get labeled data in your datasets. Most of the problems are not segregated into classes. Clustering helps to find out the patterns in the data without taking any particular variable into consideration. This could very well help us to find the relations between the features of the data. After clustering we could treat the records based on its group. We can also find similar patterns in the records which could be used for recommendation system applications.

14. How can I improve my clustering performance?

Ans: Improving the clustering performing rather a subject and domain specific tasks. But there are some steps that we could perform to ensure betterment of the performance. First of all the more the data the better the results. Also this goes without saying that garbage in garbage out, so clean the data as much as possible before using it for analysis. Using an appropriate clustering algorithm is also very important during cluster analysis. Choosing the optimal number of clusters is also a very important step during clustering. We should choose the algorithms for finding the optimal number of clusters very carefully. Such as elbow method, gap stats method, silhouette method, etc.). Reducing dimensions could be one of the optimization method

