Machine learning

Assignment-3

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

d All 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:- c

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

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

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

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

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

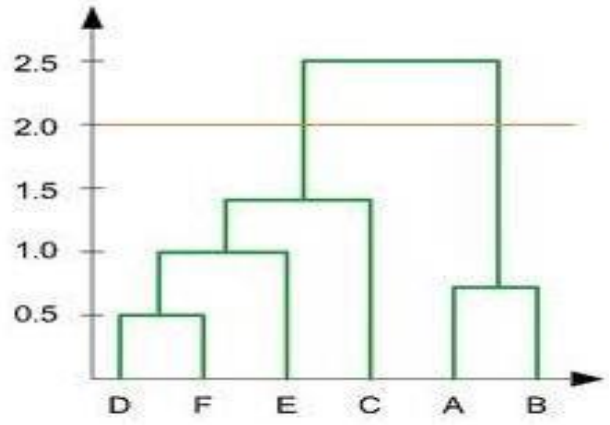
Ans:-d

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:- b

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?



Ans is 2

Explanation :- here in cluster what happens , D and F combined and form a cluster and correspondingly form a dendrogram, the height is decided with Euclidean distance.

Again it is combined with E and new cluster and next dendrogram is formed and again with C and finally it combined with A and B .So, the horizontal line is drawn from y=2 which cut two dendrogram which formed two cluster. We can take any horizontal variable as per our convenient.

10 Given, six points with the following attributes:

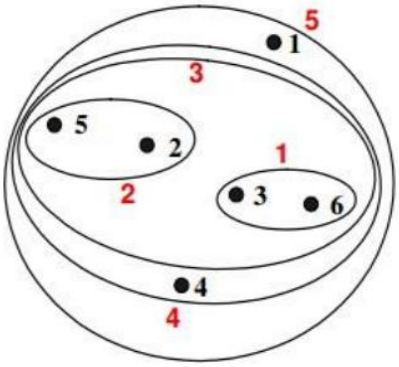
point	x coordinate	y coordinate
p1	0.4005	0.5306
p2	0.2148	0.3854
p3	0.3457	0.3156
p4	0.2652	0.1875
p5	0.0789	0.4139
p6	0.4548	0.3022

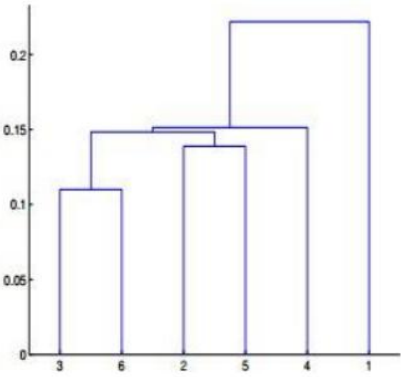
Table : X-Y coordinates of six points.

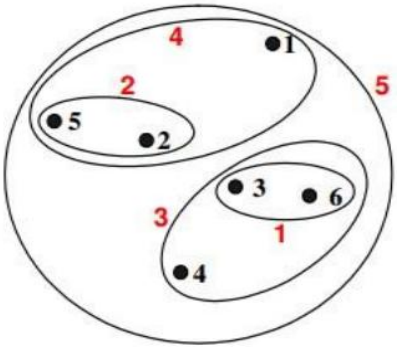
	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

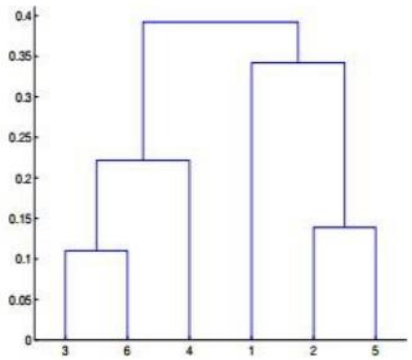
Table : Distance Matrix for Six Points

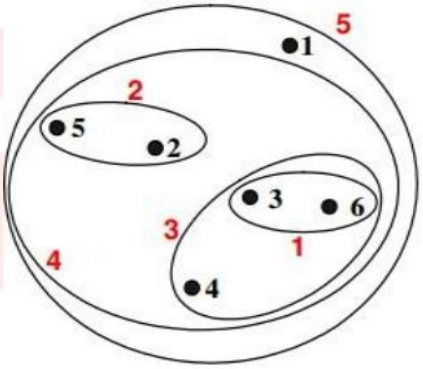
Which of the following clustering representations and dendrogram depicts the use of MIN or Single link proximity function in hierarchical clustering:

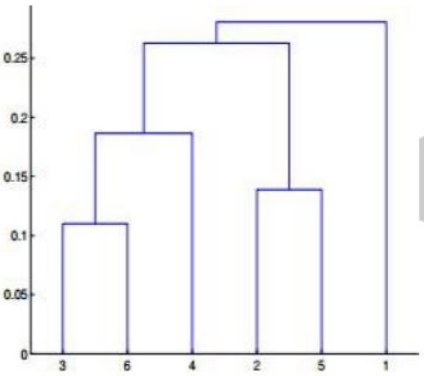

5
1
3
5
2
1
2
3
6
4
4

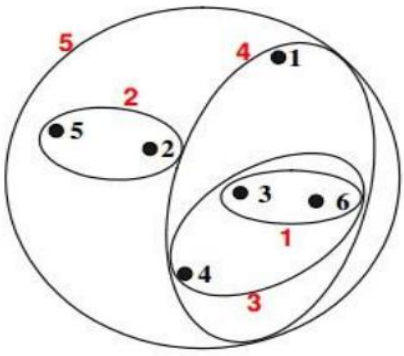

0.2
0.15
0.1
0.05
0
3
6
2
5
4
1

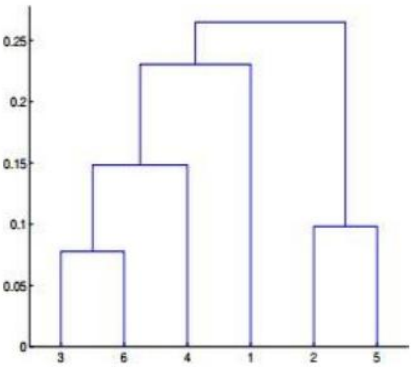

4
1
2
5
5
2
3
6
3
1
4


0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
3
6
4
1
2
5


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


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

Answer: - 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

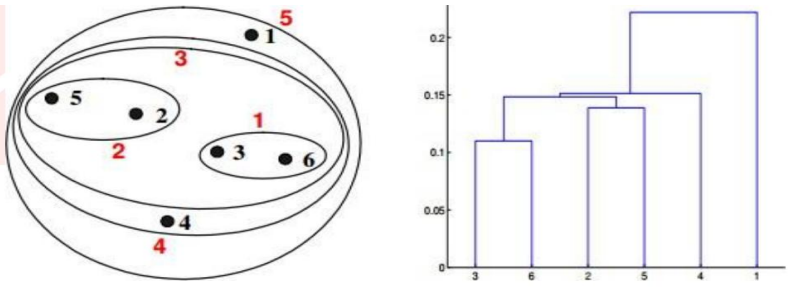
Table : X-Y coordinates of six points.

	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

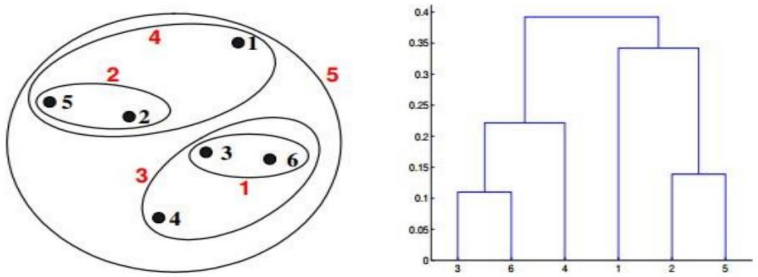
Table : Distance Matrix for Six Points

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

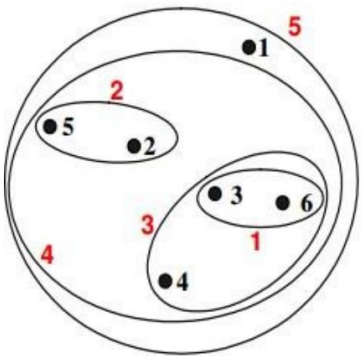
a.

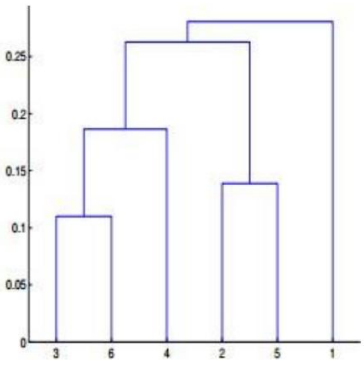


b.

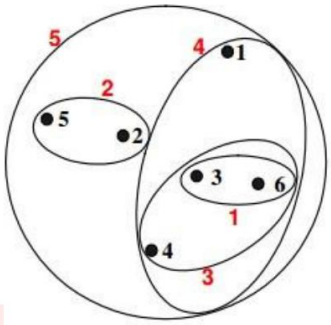


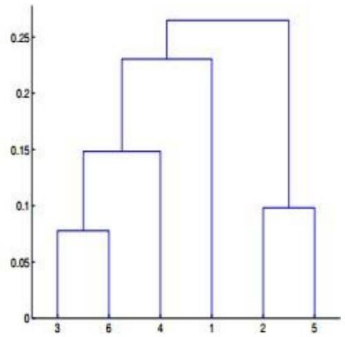
c.





d.





Answer: - b

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

13. What is the importance of clustering?

A cluster is a group of similar things or people occurring closely together.Clustering helps in understanding the natural grouping in datasets involving sets of attributes. Machine Learning has two primary ‘techniques’ for creating a machine learning algorithm which are:

- Supervised learning method
- Un-supervised learning method

Clustering comes in the domain of the unsupervised learning method of machine learning, in which it draws inferences from

the data sets of variables that do not have a labelled output variable.It basically groups data sets with common characteristics. The entire data sets present are many for a particular problem, and it is impossible to analyze them individually; hence, clustering makes it easy to handle and gather insightful data from it. The creation of such clusters mainly depends on its creator, i.e., the programmer writing the code for it and the algorithm which they use. The algorithm depends on the type of data set, the number of data sets, and the type of inferences required.

Cluster Importance in ML

The Primary use of clustering in ML is to extract valuable inferences from many unstructured data sets. Clustering and classification allow you to take a sweeping glance at your data and then form some logical structures based on what we find there. Clustering is a significant component of machine learning and its importance is highly significant in providing better machine learning techniques.

Some use cases of clustering in ML:
Social Network analysis
Image segmentation
Anomaly detection

14. How can I improve my clustering performance?

Clustering analysis is one of the main analytical methods in data mining. K-means is the most popular and partition-based clustering algorithm. But it is computationally expensive and the quality of resulting clusters heavily depends on the selection of the initial centroid and the dimension of the data. Several methods have been proposed in the literature for improving the performance of the k-means clustering algorithm. Lets discuss a method to make the algorithm more effective and efficient by using PCA and modified k-means.

PCA to find initial centroids for k-means and for dimension reduction k-means method is modified by using the heuristics approach to reduce the number of distance calculation to assign the data-point to cluster. A clustering algorithm typically considers all features of the data in an attempt to learn as much as possible about the objects. However, with high-dimensional data, many features are redundant or irrelevant. The redundant features are of no help for clustering, even worse, the irrelevant features may hurt the clustering results by hiding clusters in noises. There are many approaches to address this problem. The simplest approach is the dimension reduction technique including PCA. In these methods, dimension reduction is carried out as a pre-processing step.

K-means is a numerical, unsupervised method. It is simple and very fast, so in many practical applications, this is very effective way that can produce good clustering results. The standard K-means algorithm computational complexity is very high in high dimension. So the accuracy of the k-means clusters heavily depending on the random choice of initial centroids. If the initial partition is not chosen carefully, the computation will run the chance of conversing to a local minimum rather than global minimum solution. TO handle this situation, run the algorithm several times with different initializations. If the results converge to the same partition than it is likely that a global minimum has been reached. Final words: Initial centres determine using PCA and k-means method is modified to assign the data point to cluster.