ASSIGNMENT – 3

MACHINE LEARNING

Q1 to Q12 have only one correct answer. Choose the correct option to answer your question.

Q1. 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:- All of the above.

Q2. On which data type, we cannot perform cluster analysis?

a. Time series data

b. Text data

c. Multimedia data

d. None

ANS:- None

Q3. Netflix’s movie recommendation system uses

a. Supervised learning

b. Unsupervised learning

c. Reinforcement learning and Unsupervised learning d. All of the above

ANS:- Reinforcement learning and Unsupervised learning

Q4. 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:- The tree representing how close the data points are to each other

Q5. 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:-None

Q6. 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:- k-nearest neighbour is same as k-means

Q7. 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:- 1, 2 and 3

Q8. 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

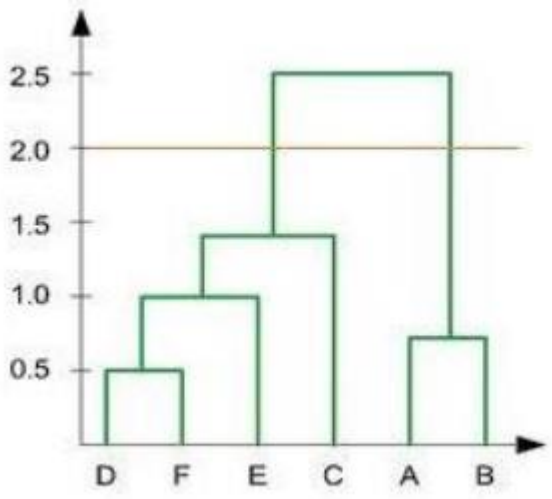
b. 2 only

c. 1 and 2

d. None of them

ANS:- 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:- 4

Q10. 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:- a.Given sales data from a large number of products in a supermarket, estimate future sales for each of these products.

Q11. 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.

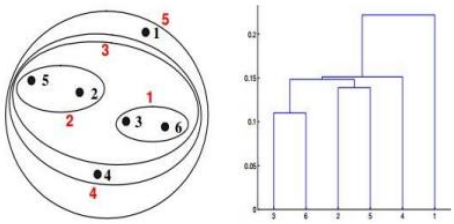
	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

Table : Distance Matrix for Six Points

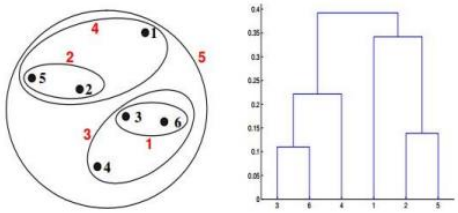
Which of the following clustering representations and dendrogram depicts the use of MIN or Single link

proximity function in hierarchical clustering:

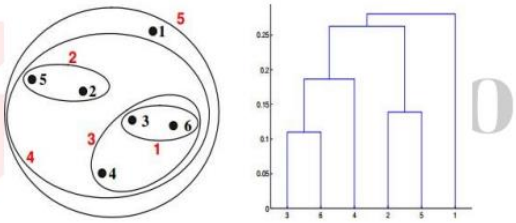
a.



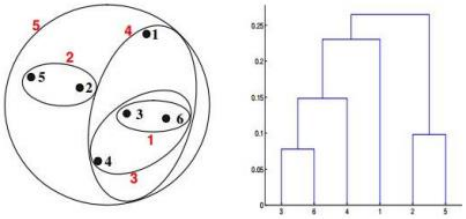
b.

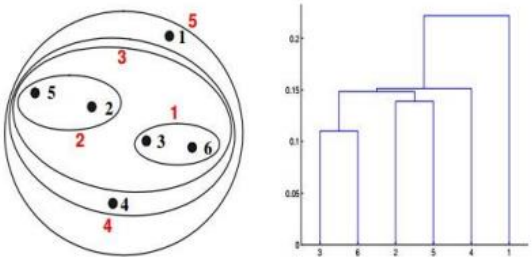


c.



d.





a.

ANS:-A

For the single link or MIN version of hierarchical clustering, the proximity of two clusters is defined to be the minimum of the distance between any two points in the different clusters. For instance, from the table, we see that the distance between points 3 and 6 is 0.11, and that is the height at which they are joined into one cluster in the dendrogram. As another example, the distance between clusters {3, 6} and {2, 5} is given by dist({3, 6}, {2, 5}) = min(dist(3, 2), dist(6, 2), dist(3, 5), dist(6, 5)) = min(0.1483, 0.2540, 0.2843, 0.3921) = 0.1483.

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

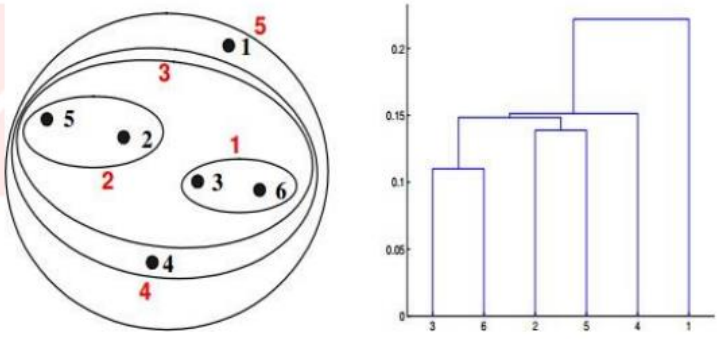
Table : X-Y coordinates of six points.

	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

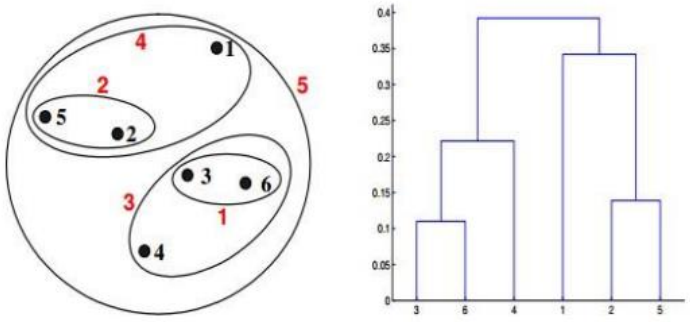
Table : Distance Matrix for Six Points

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

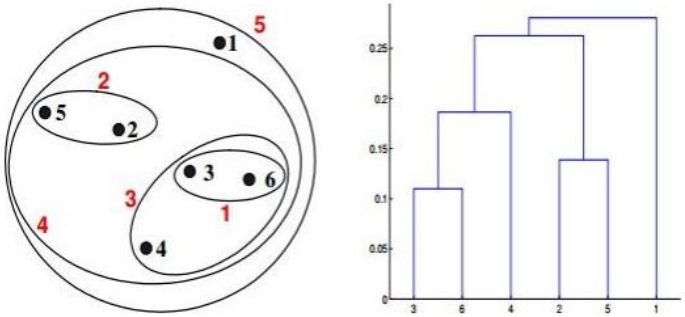
a.



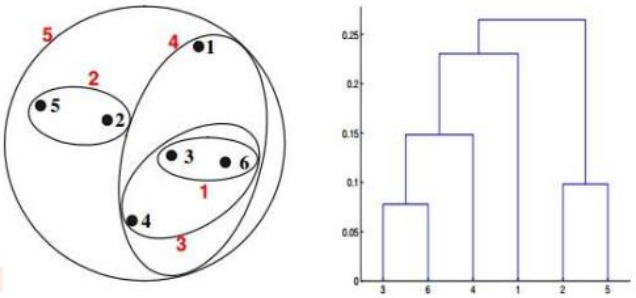
b.



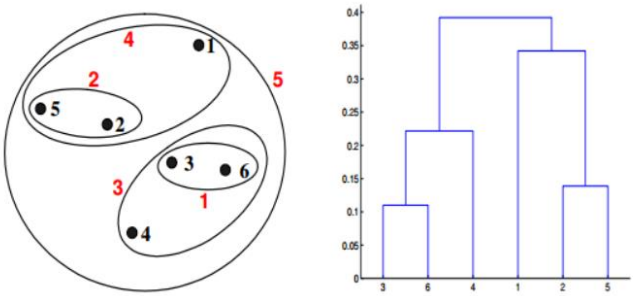
c.



d.



B.



ANS:- B

For the single link or MAX version of hierarchical clustering, the proximity of two clusters is defined to be the maximum of the distance between any two points in the different clusters. Similarly, here points 3 and 6 are merged first. However, {3, 6} is merged with {4}, instead of {2, 5}. This is because the dist({3, 6}, {4}) = max(dist(3, 4), dist(6, 4)) = max(0.1513, 0.2216) = 0.2216, which is smaller than dist({3, 6}, {2, 5}) = max(dist(3, 2), dist(6, 2), dist(3, 5), dist(6, 5)) = max(0.1483, 0.2540, 0.2843, 0.3921) = 0.3921 and dist({3, 6}, {1}) = max(dist(3, 1), dist(6, 1)) = max(0.2218, 0.2347) = 0.2347.

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

13. What is the importance of clustering?

ANS:- 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. Cluster analysis groups objects (observations, events) based on the information found in the data describing the objects or their relationships. The goal is that the objects in a group will be similar (or related) to one other and different from (or unrelated to) the objects in other groups. The greater the similarity (or homogeneity) within a group, and the greater the difference between groups, the —better‖ or more distinct the clustering. Data mining is the process of analysing data from different viewpoints and summarising it into useful information. Data mining is one of the top research areas in recent days. Cluster analysis in data mining is an important research field it has its own unique position in a large number of data analysis and processing. Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. Data mining is the process of analysing data from different perspectives and summarizing it into useful information. Data mining involves the anomaly detection, association rule learning, classification, regression, summarization and clustering. In this paper, clustering analysis is done. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. Clustering is important in data analysis and data mining . It is the task of grouping a set of objects so that objects in the same group are more similar to each other than to those in other groups. A good clustering algorithm is able to identity clusters irrespective of their shapes.

Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings (including values such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It will often be necessary to modify data pre-processing and model parameters until the result achieves the

desired properties….. Besides the term clustering, there are a number of terms with similar meanings, including automatic classification, numerical taxonomy, The subtle differences are often in the usage of the results: while in data mining, the resulting groups are the matter of interest, in automatic classification the resulting discriminative power is of interest. This often leads to misunderstandings between researchers coming from the fields of data mining and machine learning, since they use the same terms and often the same algorithms, but have different goal.

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. Clustering was performed on six benchmark datasets, consisting of five image datasets used in object, face, digit recognition tasks (COIL20, COIL100, CMU-PIE, USPS, and MNIST) and one text document dataset (REUTERS-10K) used in topic recognition. K-means, spectral clustering, Graph Regularized Non-negative Matrix Factorization, and K-means with principal components analysis algorithms were used for clustering. For each clustering algorithm, blind source separation (BSS) using Independent Component Analysis (ICA) was applied. Unsupervised feature learning (UFL) using reconstruction cost ICA (RICA) and sparse filtering (SFT) was also performed for feature extraction prior to the cluster algorithms. Clustering performance was assessed using the normalized mutual information and unsupervised clustering accuracy metrics. Grouping observed data into cohesive clusters without any prior label information is an important task. Especially, in the era of big-data, in which very large and complex amounts of data from various platforms are collected, such as image content from Facebook or vital signs and genomic sequences measured from patients in hospitals. Often, these data are not labelled and a significant undertaking is typically required (usually by individuals with domain knowledge). Even in simple tasks, such as labelling images or video data can require thousands of hours. Therefore, using the unsupervised learning technique of cluster analysis can aide in the process of providing labels to observed data .