MACHINE LEARNING ASSIGNMENT-3

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 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

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

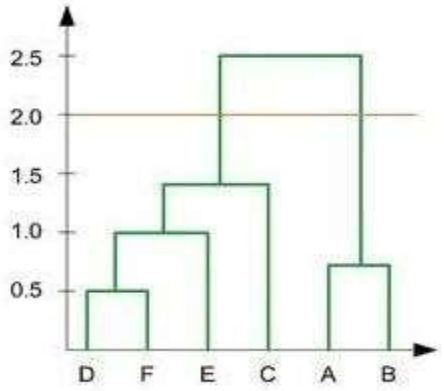
b. 2 only

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?
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,c

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

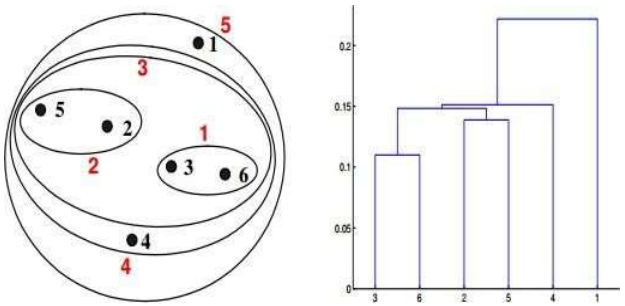
Table : X-Y coordinates of six points.

	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

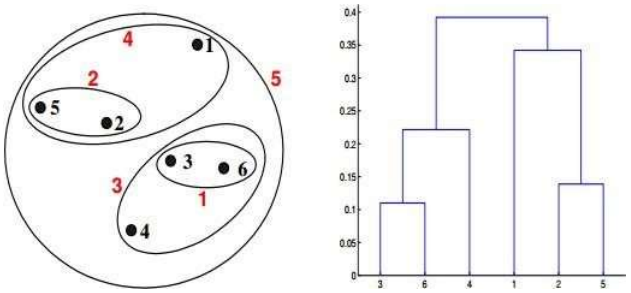
Table : Distance Matrix for Six Points

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

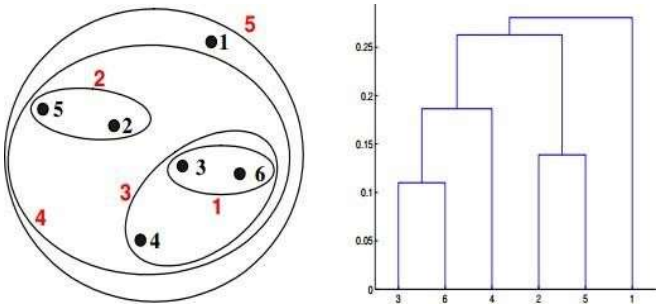
a.



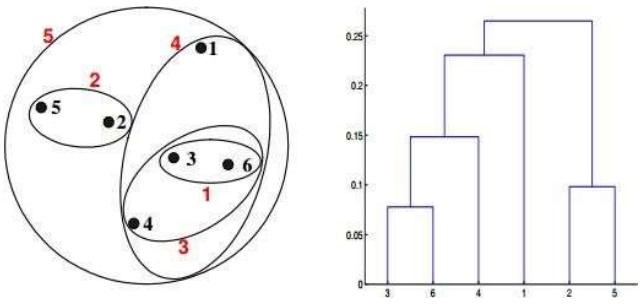
b.



c.



d.



ANS: (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

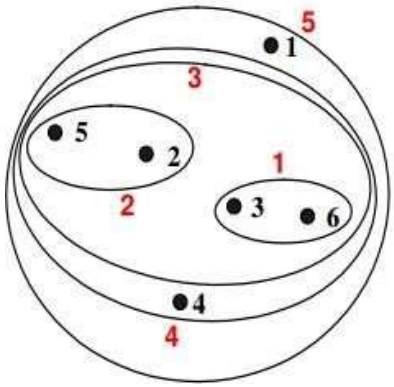
Table : X-Y coordinates of six points.

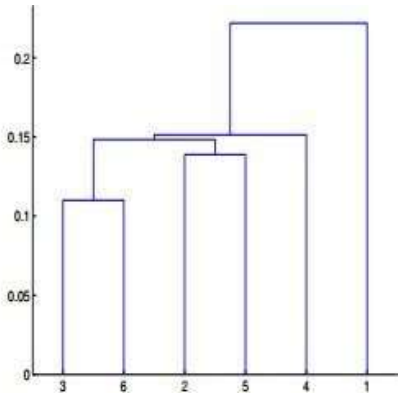
	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

Table : Distance Matrix for Six Points

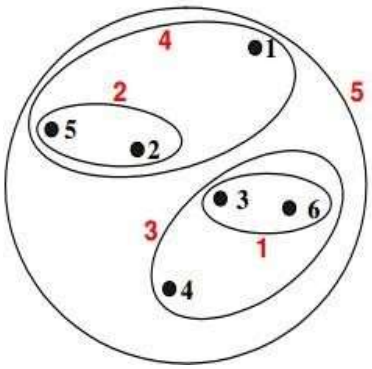
Which of the following clustering representations and dendrogram depicts the use of MAX or Complete link proximity function in hierarchical clustering.

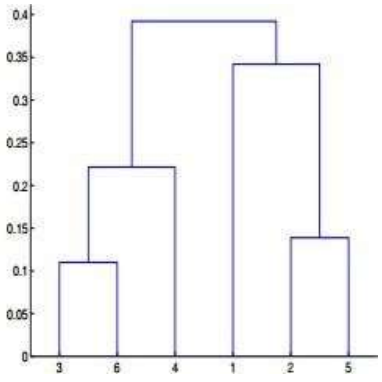
a.



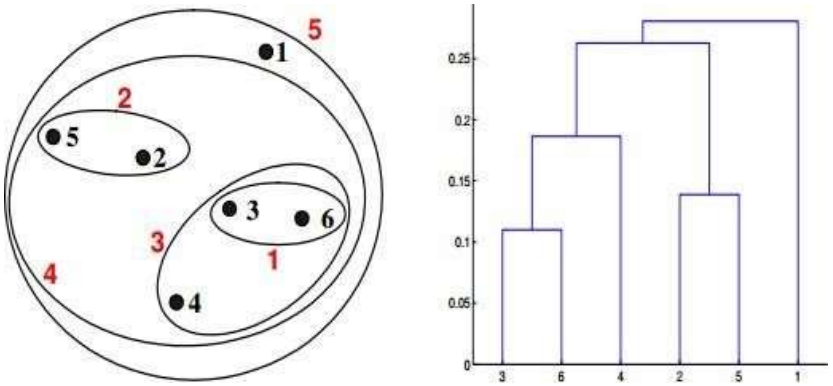


b.

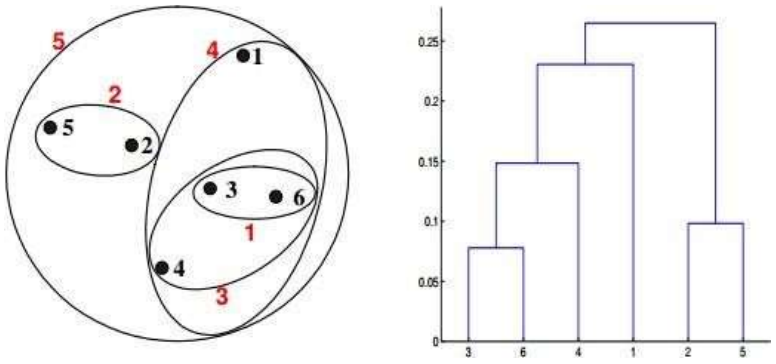




c.



d.



Ans: (d)

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

13. What is the importance of clustering?

It's important to note that analysis of clusters is not the job of a single algorithm. Rather, various algorithms usually undertake the broader task of analysis, each often being significantly different from others. Ideally, a clustering algorithm creates clusters where intra-cluster similarity is very high, meaning the data inside the cluster is very similar to one another. Also, the algorithm should create clusters where the inter-cluster similarity is much less, meaning each cluster contains information that's as dissimilar to other clusters as possible.

There are many clustering algorithms, simply because there are many notions of what a cluster should be or how it should be defined. In fact, there are more than 100 clustering algorithms that have been published to date. They represent a powerful technique for machine learning on unsupervised data. An algorithm built and designed for a specific type of cluster model will usually fail when set to work on a data set containing a very different kind of cluster model.

The common thread in all clustering algorithms is a group of data objects. But data scientists and programmers use differing cluster models, with each model requiring a different algorithm. Clustering or sets of clusters are often distinguished as either hard clustering where each object belongs to a cluster or not, or soft clustering where each object belongs to each cluster to some degree.

This is all apart from so-called server clustering, which generally refers to a group of servers working together to provide users with higher availability and to reduce downtime as one server takes over when another fails temporarily.

Clustering algorithms group together people with similar traits, perhaps based on their likelihood to purchase. With these groups or clusters defined, test marketing across them becomes more effective, helping to refine messaging to reach them. 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.

Clustering is very much important as it determines the intrinsic grouping among the un labelled data present. There are no criteria for good clustering. It depends on the user, what is the criteria they may use which satisfy their need. For instance, we could be interested in finding representatives for homogeneous groups (data reduction), in finding “natural clusters” and describe their unknown properties (“natural” data types), in finding useful and suitable groupings (“useful” data classes) or in finding unusual data objects (outlier detection). This algorithm must make some assumptions that constitute the similarity of points and each assumption make different and equally valid clusters.

Importance of Clustering Methods

1. Having clustering methods helps in restarting the local search procedure and remove the inefficiency. In addition, clustering helps to determine the internal structure of the data.
2. This clustering analysis has been used for model analysis, vector region of attraction.
3. 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.
4. Clustering quality depends on the methods and the identification of hidden patterns.
5. They play a wide role in applications like marketing economic research and weblogs to identify similarity measures, Image processing, and spatial research.
6. They are used in outlier detections to detect credit card fraudulence.

Applications of Cluster Analysis

- Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing.
- Clustering can also help marketers discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns.
- In the field of biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionalities and gain insight into structures inherent to populations.
- Clustering also helps in identification of areas of similar land use in an earth observation database. It also helps in the identification of groups of houses in a city according to house type, value, and geographic location.
- Clustering also helps in classifying documents on the web for information discovery.
- Clustering is also used in outlier detection applications such as detection of credit card fraud.
- As a data mining function, cluster analysis serves as a tool to gain insight into the distribution of data to observe characteristics of each cluster.

14. How can I improve my clustering performance?

Clustering of data is of great interest in many data mining applications. In applying a clustering technique, however, many attributes or features can be irrelevant to the clusters produced. Therefore, reducing the number of dimensions has proven to be a valuable technique for improving the efficiency of a clustering algorithm, especially when the input data vectors contain a large number of features. Feature selection and extraction techniques aim at selecting a subset of the features that is relevant for a given problem. Usually all features do not generate a corresponding increase in performance of the clustering method. Some of these features may be noisy, meaningless, correlated, or irrelevant for the clustering task. In particular, the attributes that have similar data across the majority of components (data vectors) should be deleted because these attributes do not add any useful information for producing different clusters. In this work, we apply feature selection and feature extraction techniques as a pre-processor for our proposed conceptual clustering method to improve clustering performance and to reduce its computational complexity. The results presented from the application of these methods to computer workload characterization in particular indicate that the integration of feature selection and extraction methods with conceptual clustering has potential for producing meaningful categories.

For performance enhancement of our conceptual clustering algorithm, we have considered three feature selection and extraction methods, namely QR, PCA, and ICA. The experimental results achieved show that the removal of linear dependency of features has improved the numerical stability of our algorithm. Reduction of feature space has also considerably improved not only the performance of the algorithm but also the clustering accuracy. The clustering results were evaluated using four different datasets: computer workload dataset, iris dataset, new-thyroid dataset, and image segmentation dataset. The evaluation criterion is based on entropy as a measure of goodness of the cluster produced. In particular, the results of our experiments show that the use of PCA as a pre-processor for our clustering algorithm worked best for these types of datasets. The system presented is a valuable tool for clustering unlabeled points. We compared the results of its application with those produced by the self- organizing map (SOM) to determine the strong and weak aspects of each system especially when they are combined with a feature selection and extraction technique. The results showed when we applying our system we achieved better clustering than when applying SOM on the same datasets.

- 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.
- K-means clustering algorithm can be significantly improved by using a better initialization technique, and by repeating (re-starting) the algorithm.
- When the data has overlapping clusters, k-means can improve the results of the initialization technique.
- When the data has well separated clusters, the performance of k-means depends completely on the goodness of the initialization.
- Initialization using simple furthest point heuristic (Maxmin) reduces the clustering error of k-means from 15% to 6%, on average.