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?

Answer: d) All of the above

2. on which data type, we cannot perform cluster analysis?

Answer: d) None

3. Netflix's movie recommendation system uses-

Answer: c) Reinforcement learning and unsupervised learning

4. The final output of Hierarchical clustering is-

Answer: 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?

Answer: d) None

6. Which is the following is wrong?

Answer: 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?

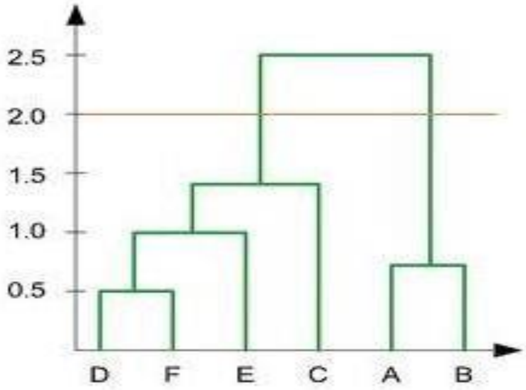
Answer: d) 1, 2 and 3 (Single-link, Complete-link and Average-link)

8. Which of the following are true?

Answer: a) 1 only (Clustering analysis is negatively affected by multicollinearity of Features)

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?

MACHINE LEARNING ASSIGNMENT-3

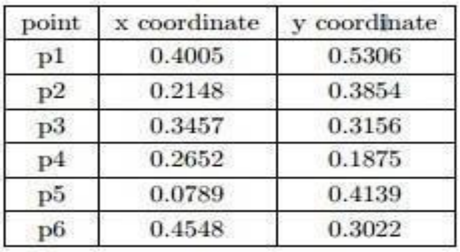


Answer: a) 2

10. For which of the following tasks might clustering be a suitable approach?

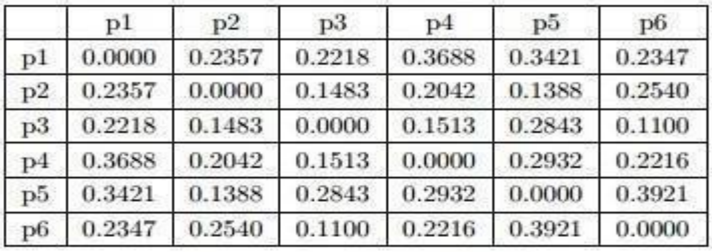
Answer: b) Given a database of information about your users, automatically group them into different market segments.

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.



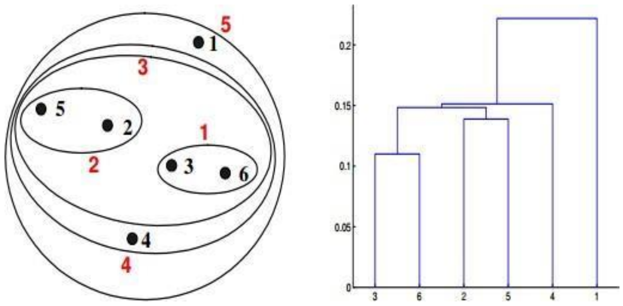
	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

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:

MACHINE LEARNING ASSIGNMENT-3

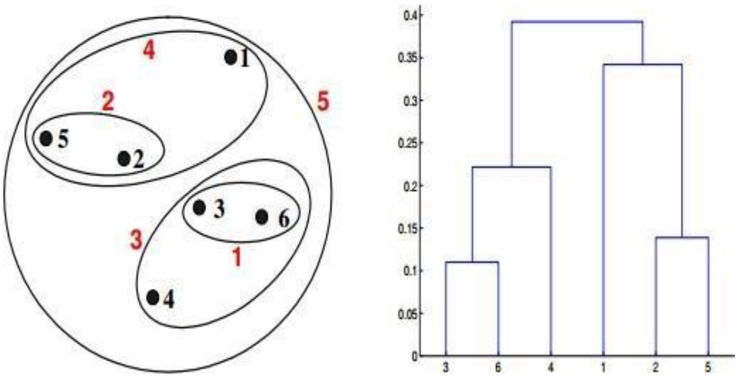
Answer: a)



12. Given, six points with the following attributes:

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

Answer: b)



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

13. What is the importance of clustering?

<u>Answer:</u>

Clustering is important in data analysis and data mining applications. 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.

There are six main methods of data clustering:-

1. The partitioning method.
2. Hierarchical method.
3. Density based method.
4. Grid based method.
5. The model based method.
6. The constraint-based method.

Each method groups the data in a different way. In the density based method, for instance, the data is clustered together according to its density, as the name suggests. In the grid based method, the objects are organized to create a grid structure.
It mainly divides many unstructured data sets into clusters and, according to the common attributes present in them, it helps create more and more clusters.
Clustering is an unsupervised method that works on datasets in which there is no outcome (target) variable, the relationship between the observations, that is, unlabeled data.
Marketers commonly use cluster analysis to develop market segments, which allow for better positioning of products and messaging.
Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing.
Clustering algorithms are the largest group of data mining algorithms used for unsupervised learning.
They are often used as a preprocessing step for supervised algorithms, given a set of n objects, clustering algorithms find k groups based on a similarity measure.

14. How can I improve my clustering performance?

Answer:
Clustering is an unsupervised machine learning methodology that aims to partition data into distinct groups, or clusters. There are a few different forms including hierarchical, density, and similarity based. Each have a few different algorithms associated with it as well. One of the hardest parts of any machine learning algorithm is feature engineering, which can especially be difficult with clustering as there is no easy way to figure out what best segments your data into separate but similar groups.

The guiding principle of similarity based clustering is that similar objects are within the same cluster and dissimilar objects are in different clusters. This is not different than the goal of most conventional clustering algorithms. With similarity based clustering, a measure must be given to determine how similar two objects are. This similarity measure is based off distance, and different distance metrics can be employed, but the similarity measure usually results in a value in [0,1] with 0 having no similarity and 1 being identical.

Measuring Improvement:
However, when the data has well separated clusters, the performance of k-means depends completely on the goodness of the initialization. Therefore, if high clustering accuracy is needed, a better algorithm should be used instead.
Clustering algorithms K-means Initialization Clustering accuracy Prototype selection.

MACHINE LEARNING ASSIGNMENT-3

K-means algorithm, groups N data points into k clusters by minimizing the sum of squared distances between every point and its nearest cluster mean (centroid). This objective function is called sum-of-squared errors (SSE). Although k-means was originally designed for minimizing SSE of numerical data, it has also been applied for other objective functions.
Wrong choice of the function can easily reverse the benefit of a good algorithm so that a proper objective function with a worse algorithm can provide better clustering than good algorithm with wrong objective function.

There are other algorithms that are known, in many situations, to provide better clustering results than k-means. However, k-means is popular for good reasons. First, it is simple to implement. Second, people often prefer to use an extensively studied algorithm whose limitations are known rather than a potentially better, but less studied, algorithm that might have unknown or hidden limitations. Third, the local fine-tuning capability of k-means is very effective.
K-means starts by selecting k random data points as the initial set of centroids.
The centroid of every cluster is recalculated as the mean of all data points assigned to the cluster. Together, these two steps constitute one iteration of k-means. These steps fine-tune both the cluster borders and the centroid locations. The algorithm is iterated a fixed number of times.

K-means has excellent fine-tuning capabilities. Given a rough allocation of the initial cluster centroids, it can usually optimize their locations locally. However, the main limitation of k-means is that it rarely succeeds in optimizing the centroid locations globally. The reason is that the centroids cannot move between the clusters if their distance is big, or if there are other stable clusters in between preventing the movements.
The k-means result therefore depends a lot on the initialization. Poor initialization can cause the iterations to get stuck into an inferior local minimum.

To compensate for the mentioned weaknesses of k-means, two main approaches have been considered:

- ➢ Using a better initialization.
- ➢ Repeating k-means several times by different initial solution.

Numerous initialization techniques have been presented in the literature, including the following:

1) Random points.
2) Furthest point heuristic
3) Sorting heuristic
4) Density-based

5) Projection-based
6) Splitting technique

K-means initialization techniques:

- Better initialization
- Repeating k-means

K-means is a good algorithm for local fine-tuning but it has serious limitation to relocate the centroids when the clusters do not overlap. It is therefore unrealistic to expect the clustering problem to be solved simply by inventing a better initialization for k-means. The question is merely, how much a better initialization can compensate for the weakness of k-means.

Any clustering algorithm could be used as an initialization technique for k-means. However, solving the location of initial centroids is not significantly easier than the original clustering problem itself. Therefore, for an algorithm to be considered as initialization technique for k-means, in contrast to being a standalone algorithm, we set the following requirements:

- Simple to implement
- Lower (or equal) time complexity than k-means
- No additional parameters