MACHINE LEARNING

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

2. On which data type, we cannot perform cluster analysis?
   a. Time series data
   b. Text data
   c. Multimedia data
   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

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

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

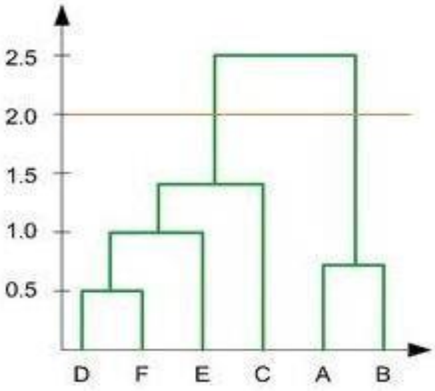
6. Which is the following is wrong?
   a. k-means clustering is a vector quantization method
   b. k-means clustering tries to group n observations into k clusters
   c. k-nearest neighbour is same as k-means
   d. None

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

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

MACHINE LEARNING

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

Solution B

Since the number of vertical lines intersecting the red horizontal line at y=2 in the dendrogram are 2, therefore, two clusters will be formed.

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

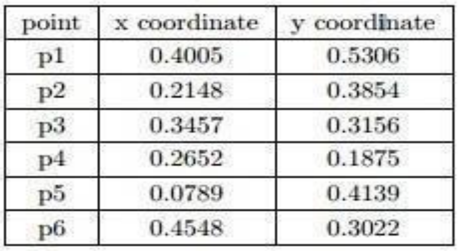
a. **Given sales data from a large number of products in a supermarket, estimate future sales for eachof these products**.

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.

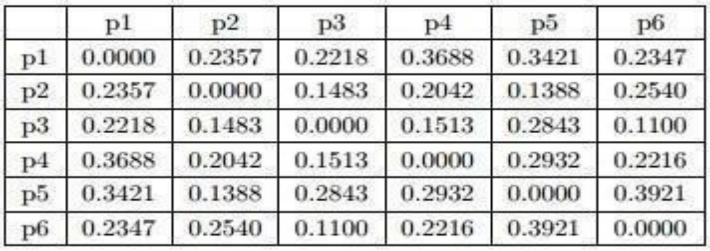
d. Given historical weather records, predict if tomorrow's weather will be sunny or rainy.

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.

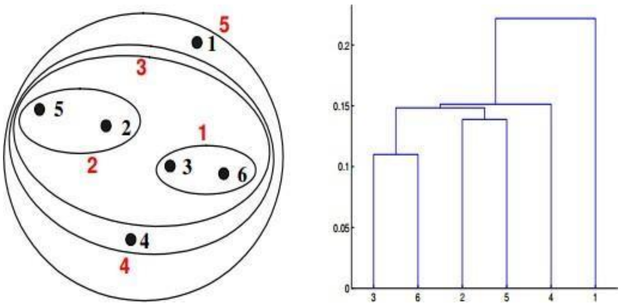


	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

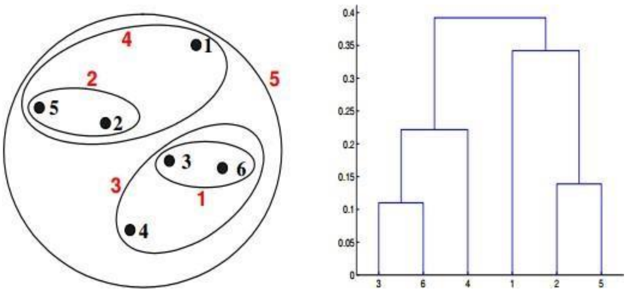
Table : Distance Matrix for Six Points

MACHINE LEARNING

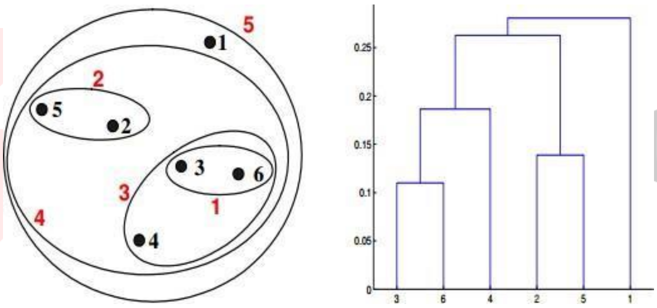
Which of the following clustering representations and dendrogram depicts the use of MIN or Single link proximity function in hierarchical clustering:



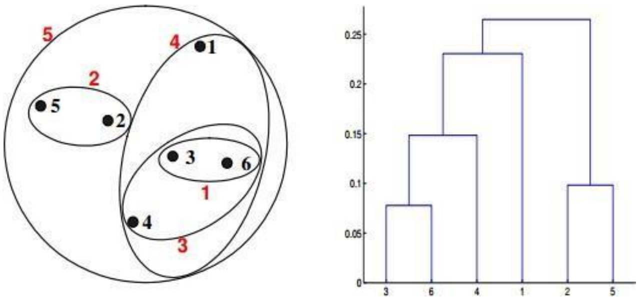
a.



b.



c.

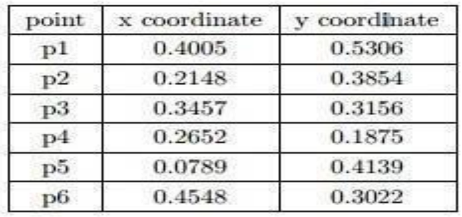


d.

Solution 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.

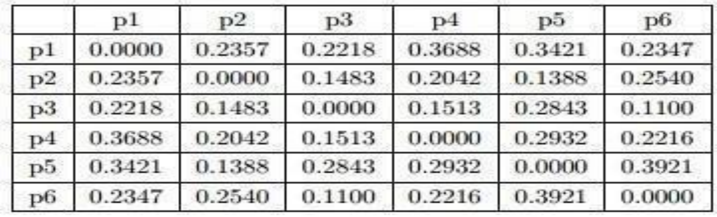
MACHINE LEARNING

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.

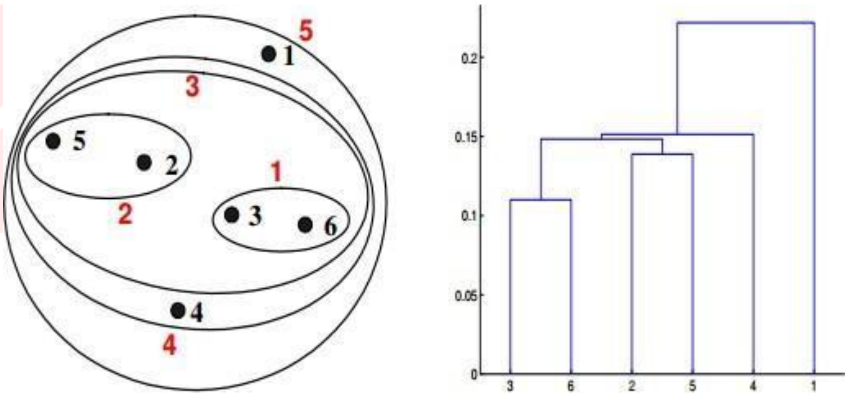


	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

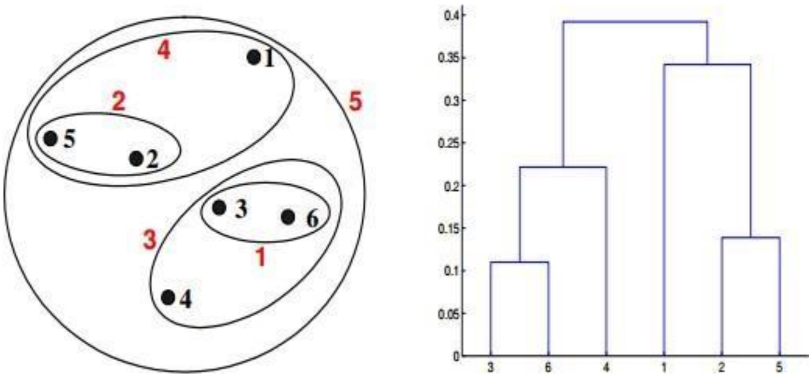
Table : Distance Matrix for Six Points

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

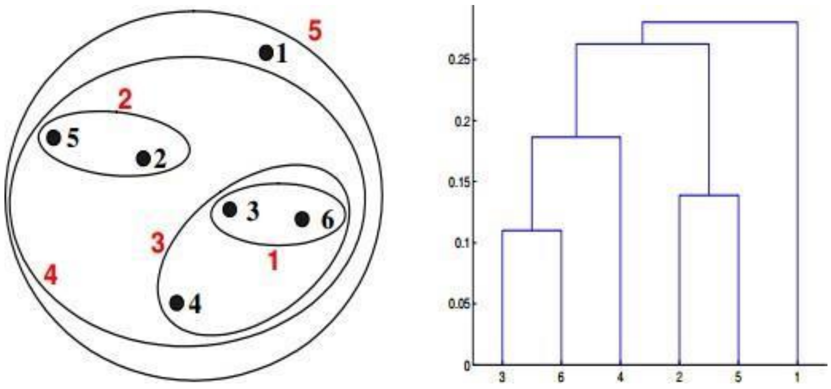
a.



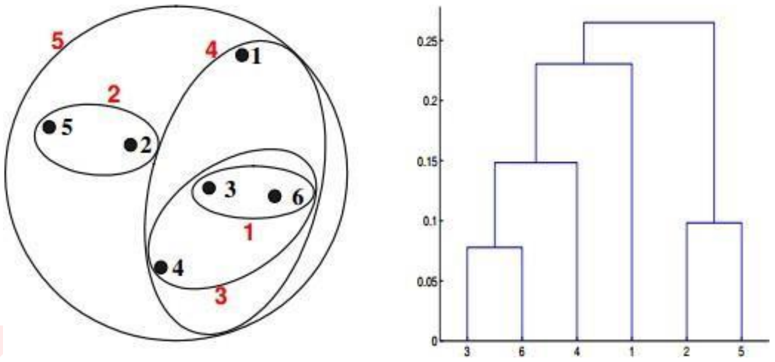
b.



c.



d.



Solution 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?
14. How can I improve my clustering performance?

13. What is the importance of Clustering:

Clustering is the process of arranging a group of objects in such a manner that the objects in the same group are more similar to each other than to the objects in any other group. Data Professionals often use clustering in the exploratory Data Analysis phase to discover new information and patterns in the data. As clustering is unsupervised machine learning, it doesn't require a labelled dataset.

Clustering itself is not one specific algorithm but the general task to be solved. You can achieve this goal using algorithms that differ significantly in their understanding of what constitutes a cluster and how to find them efficiently.

MACHINE LEARNING

Key Success Criteria for Clustering Analysis:

Clustering, unlike supervised learning use-cases such as classification or regression, cannot be completely automated end-to-end. Instead, it is an iterative process of information discovery that requires domain expertise and human judgment used frequently to make adjustments to the data and the model parameters to achieve the desired result.

Most importantly, because clustering is unsupervised learning and doesn't use labeled data, we cannot calculate performance metrics like accuracy, AUC, RMSE, etc., to compare different algorithms or data preprocessing techniques. As a result, this makes it really challenging and subjective to assess the performance of clustering models.

Business Applications of Clustering:

Clustering is a very powerful technique and has broad applications in various industries ranging from media to healthcare, manufacturing to service industries, and anywhere you have large amounts of data. Let's take a look at some practical use-cases:

- Customer Segmentation
- Retail Clustering
- Clustering in Clinical Care / Disease Management
- image segmentation

Customer Segmentation:

Customers are categorized by using clustering algorithms according to their purchasing behavior or interests to develop focused marketing campaigns.

Retail Clustering:

There are many opportunities for clustering in retail businesses. For example, you can gather data on each store and cluster at store level to generate insights that may tell you which locations are similar to each other based on attributes like foot traffic, average store sales, number of SKUs, etc.

Clustering in Clinical Care / Disease Management:

Healthcare and Clinical Science is again one of those areas that are full of opportunities for clustering that are indeed very impactful in the field. One such example is research published by Komaru & Yoshida et al. 2020, where they collected demographics and laboratory data for 101 patients and then segmented them into 3 clusters.

Image Segmentation:

Image segmentation is the classification of an image into different groups. Much research has been done in the area of image segmentation using clustering. This type of clustering is useful if you want to isolate objects in an image to analyze each object individually to check what it is.

5 Essential Clustering Algorithms K-Means:

- User specifies the number of clusters.
- Initialize centroids randomly based on the number of clusters. In the diagram below in Iteration 1, notice three centroids are initialized randomly in blue, red, and green colors.
- Calculate the distance between data points and each centroid and assign each data point to the nearest centroids.

MACHINE LEARNING

- Recalculate the mean of the centroid based on all the assigned data points, and this will change the position of the centroid, as you can see in Iteration 2 - 9, until it finally converges.
- Iteration keeps on going until there is no change to the centroid's mean or a parameter max_iter is reached, which is the maximum number of the iterations as defined by the user during training. In scikit-learn, max_iter by default is set to 300.

15. How can I improve my clustering performance?

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. 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:

A good representation of its effectiveness is fuzzy c-means, a relative of the commonly used k-means algorithm. It works in a very similar fashion to k-means, but rather results in something called the fuzzy partition matrix instead of just a cluster label.

The fuzzy partition matrix is a set of weights that measure how similar a single point is to a given cluster center, close to how our similarity matrix is used previously. It can also be calculated using a weighted distance metric which we can feed our new found optimal weights. This will also then go back into updating the cluster centers. Like K-means, this results in the cluster centers shifting with each iteration, until the maximum number of iterations or a certain improvement threshold has been met.

In fuzzy c-means, you would have a very similar goal as to our original loss function. You would like less “fuzzyness” from points, and you want them all to be as close as possible to their cluster centers, and further away from others. A good measure of the fuzzy clustering algorithm is Dunn’s partition coefficient, a sum of all components of the fuzzy partition matrix.

Let’s try using fuzzy c-means on the Iris data set with and without our learned feature weights. Here the output of fuzzy c-means comparing all variables, assuming 3 clusters.

Classical clustering algorithms:

Classical clustering algorithms used for analysis are K-means, Gaussian Mixture Models, and hierarchical clustering, all of which are based on using a distance measure to assess the similarity of observations. The choice for distance measure is typically data dependent. For instance, in image data, the similarity between pixels can be represented by the Euclidean distance, where as in text documents cosine distance matrix is typically used. Moreover, appropriate feature representation of the observations is even more critical in order to obtain correct clusters of the data, since improved features provide a better representative similarity matrix.

MACHINE LEARNING

Deep learning-based clustering:

Early approaches for learning the appropriate feature space in clustering algorithms implemented deep autoencoders (DAEs). Song et al. used DAEs to directly learn the data representations and cluster centers. Huang et al. employed a DAE with locality and sparsity preserving constraints, which is followed by a K-means to obtain the cluster memberships. A more recent and popular approach by Xie et Learned the feature space and cluster membership directly using a stacked denoising autoencoder. Following Xie et al. there have been many studies proposing deep clustering algorithms to learn the feature space and cluster membership simultaneously using some form of an autoencoder. A departure from the autoencoder framework was demonstrated by Yang et, who used recurrent and convolutional neural networks with agglomerative (hierarchical) clustering.

Spectral clustering:

Spectral clustering usually performs better than K-means and the aforementioned classical algorithms due to its ability to cluster non-spherical data. A key issue in spectral clustering is to solve the multiclass clustering problem. This is accomplished by representing the graph Laplacian in terms of k eigenvectors, k being the number classes.