FLIP ROBO TECHNOLOGIES

Worksheet Set – 3

Batch DS2301

Intern: Vishnukanth

Mail ID: vishnukh25@gmail.com

- ✓ **Machine Learning**
- ✓ **Statistics**

Machine Learning

Worksheet – 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

Answer : 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

Answer : 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

Answer : d) 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

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?

a. A distance metric

b. Initial number of clusters

c. Initial guess as to cluster centroids

d. None

Answer : 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

Answer : 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

Answer : d) 1, 2 and 3

8. Which of the following are true?

i. Clustering analysis is negatively affected by multicollinearity of features

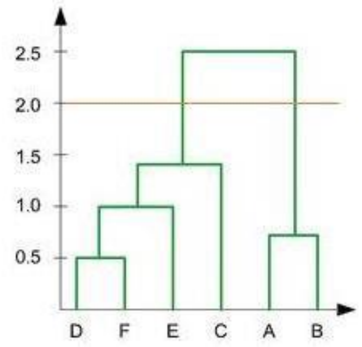
ii. Clustering analysis is negatively affected by heteroscedasticity

Options:

a. 1 only b. 2 only c. 1 and 2 d. None of them

Answer : d) None of them

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

Answer : 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.

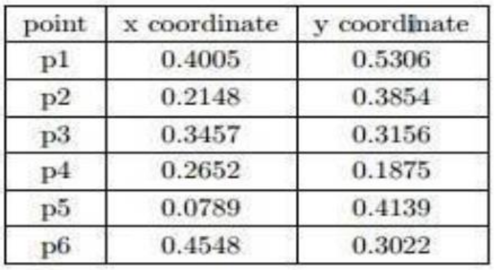
c. Predicting whether stock price of a company will increase tomorrow.

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

Answer : b) and **c)**

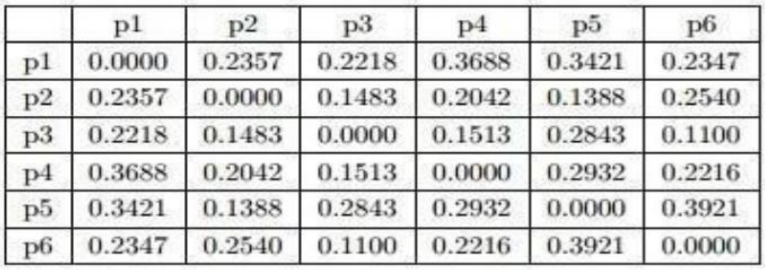
- Given a database of information about your users, automatically group them into different market segments.
- Predicting whether stock price of a company will increase tomorrow.

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.

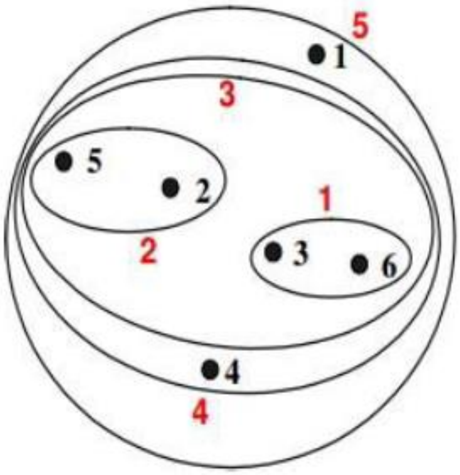


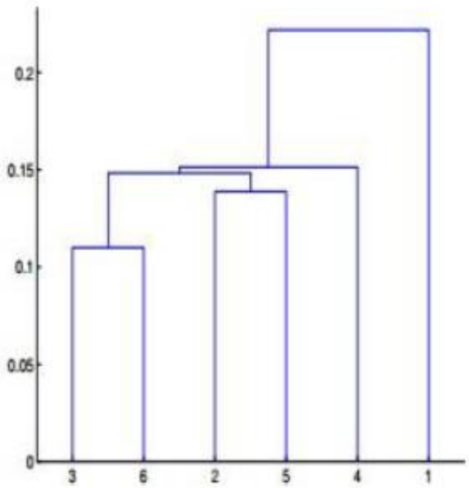
	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

Table : Distance Matrix for Six Points

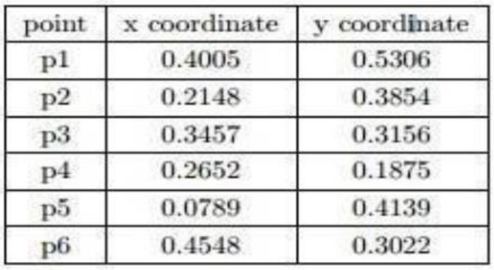
Which of the following clustering representations and dendrogram depicts the use of MIN or Single link proximity function in hierarchical clustering:

Answer : a)



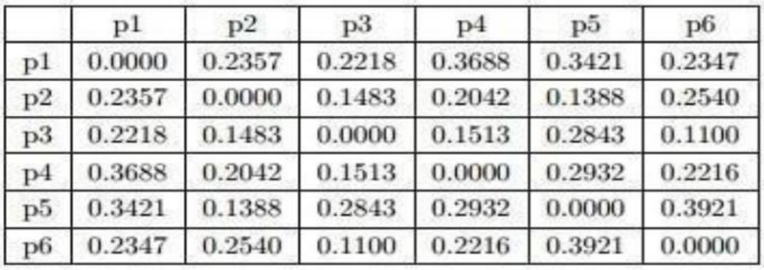


12. Given, six points with the following attributes:



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.

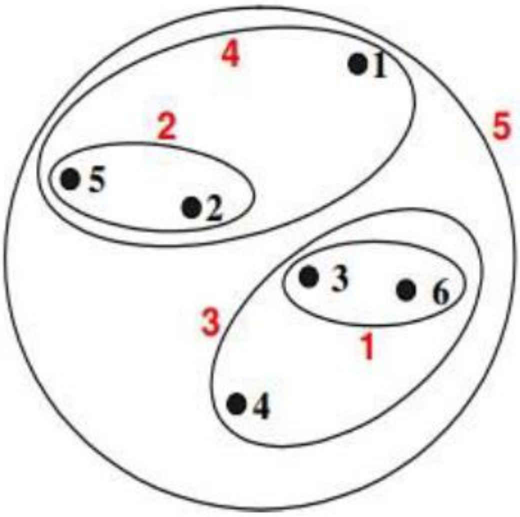


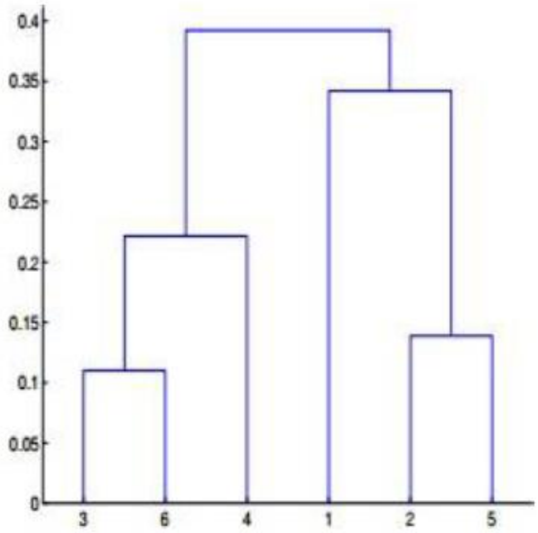
	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

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?

Answer : b)





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

13. What is the importance of clustering?

Clustering is an important technique in data analysis and machine learning as it helps to discover patterns and structure in data. It is used to group similar objects together, which can be useful for a variety of tasks, such as:

Market segmentation: Clustering can be used to group customers with similar characteristics into different market segments. This can be useful for targeted marketing and advertising.

Image and video analysis: Clustering can be used to group similar images or video frames together, which can be useful for image and video compression, content-based retrieval, and indexing.

Anomaly detection: Clustering can be used to identify observations that do not belong to any clusters. These observations can be considered as anomalies or outliers.

Dimensionality reduction: Clustering can be used to reduce the dimensionality of high-dimensional data. This can be useful for visualizing high-dimensional data, as well as for improving the performance of machine learning algorithms.

Data mining: Clustering can be used to uncover hidden patterns in data, such as grouping similar records together or identifying groups of customers with similar purchase behavior.

Clustering is a powerful technique that can be used to gain insights from data and make better decisions.

14. How can I improve my clustering performance?

There are several ways to improve the performance of clustering algorithms:

Feature selection and dimensionality reduction: Selecting the most relevant features and reducing the dimensionality of the data can improve the performance of clustering algorithms. This can be achieved through techniques such as principal component analysis (PCA), linear discriminant analysis (LDA), or independent component analysis (ICA).

Scaling the data: Scaling the data before applying a clustering algorithm can improve the performance. This is particularly important for algorithms such as k-means that rely on distance measures.

Choosing the right number of clusters: Choosing the right number of clusters is important for the performance of clustering algorithms. This can be achieved through techniques such as the elbow method, silhouette scores, or the gap statistic.

Choosing the right distance metric: Choosing the right distance metric can also improve the performance of clustering algorithms. This can depend on the characteristics of the data and the clustering algorithm used.

Choosing the right algorithm: Different clustering algorithms are suitable for different types of data and tasks. It's important to choose the right algorithm for the problem.

Evaluating the clustering results: It's important to evaluate the results of the clustering algorithm in order to measure the performance and identify any potential problems. Metrics such as adjusted Rand index, adjusted Mutual information, Fowlkes-Mallow's index and Davies-Bouldin index are commonly used to evaluate the results of clustering algorithms.

Hyperparameter tuning: Clustering algorithms often have a set of parameters that can be adjusted in order to improve the performance. This can be done through techniques such as grid search or random search.

It's important to note that there is no one-size-fits-all solution for improving clustering performance, as the best approach depends on the characteristics of the data and the specific problem being addressed.

STATISTICS

Worksheet – 3

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

1. Which of the following is the correct formula for total variation?

a) Total Variation = Residual Variation – Regression Variation

b) Total Variation = Residual Variation + Regression Variation

c) Total Variation = Residual Variation * Regression Variation

d) All of the mentioned

Answer : b) Total Variation = Residual Variation + Regression Variation

2. Collection of exchangeable binary outcomes for the same covariate data are called outcomes.

a) random

b) direct

c) binomial

d) none of the mentioned

Answer : c) binomial

3. How many outcomes are possible with Bernoulli trial?

a) 2 b) 3 c) 4 d) None of the mentioned

Answer : a) 2

4. If Ho is true and we reject it is called

a) Type-I error b) Type-II error

c) Standard error d) Sampling error

Answer : a) Type-I error

5. Level of significance is also called:

a) Power of the test

b) Size of the test

c) Level of confidence

d) Confidence coefficient

Answer : c) Level of confidence

6. The chance of rejecting a true hypothesis decreases when sample size is:

a) Decrease

b) Increase

c) Both of them

d) None

Answer : b) Increase

7. Which of the following testing is concerned with making decisions using data?

a) Probability

b) Hypothesis

c) Causal

d) None of the mentioned

Answer : b) Hypothesis

8. What is the purpose of multiple testing in statistical inference?

a) Minimize errors

b) Minimize false positives

c) Minimize false negatives

d) All of the mentioned

Answer : d) All of the mentioned

9. Normalized data are centred at and have units equal to standard deviations of the original data

a) 0

b) 5

c) 1

d) 10

Answer : c) 1

Q10 to Q15 are subjective answer type questions, Answer them in your own words briefly.

10. What Is Bayes' Theorem?

Bayes' theorem is a mathematical formula that describes how the probability of an event changes with new information. It expresses the probability of an event (A) in terms of the probability of another event (B) that has already occurred, and the probability of B given A. It can be expressed mathematically as: P(A|B) = P(B|A) * P(A) / P(B) where P(A|B) is the probability of A given B, P(B|A) is the probability of B given A, P(A) is the prior probability of A, and P(B) is the prior probability of B. The theorem is named after Thomas Bayes, an 18th-century statistician and theologian who first described it in his work on probability and induction.

11. What is z-score?

A z-score is a measure of how many standard deviations an observation or data point is from the mean of a distribution. A z-score of 0 represents the mean, while a z-score of 1 represents a value one standard deviation above the mean and a z-score of -1 represents a value one standard deviation below the mean. Z-scores are useful for comparing data points from different distributions and for identifying outliers.

12. What is t-test?

A t-test is a statistical test used to determine whether there is a significant difference between the means of two groups. It is used to compare the means of two samples to determine if they come from the same population with the same mean. There are two main types of t-tests:

Student's t-test (also known as the independent samples t-test) is used to compare the means of two independent groups.

Paired sample t-test (also known as the dependent samples t-test) is used to compare the means of two related groups, such as before and after measurements on the same group of individuals.

The t-test calculates a t-value, which is a measure of the difference between the means of the two groups and the amount of variation in the data. The t-value is then compared to a critical value from a t-distribution table to determine the probability that the difference between the means is due to chance. If this probability is less than a certain threshold (usually 0.05), then the difference is considered statistically significant and the null hypothesis (that the means are equal) is rejected.

13. What is percentile?

A percentile is a measure used in statistics to indicate the value below which a given percentage of observations in a group of observations falls. For example, the 50th percentile is the value that separates the lowest 50% of observations from the highest 50% of observations. Percentiles can be used to summarize and compare data sets and to understand the distribution of values within a data set.

To calculate a percentile, you first need to arrange the data in increasing order. Then, you take the value of the observation at the point where a certain percentage of the observations fall below it. Percentiles are commonly used in fields such as education, health, and psychology to understand how a given score or measurement compares to others in a population. For example, a student's score on a test might be compared to the scores of other students in the same grade to understand how well the student performed relative to their peers.

There are different ways to calculate percentiles, such as nearest-rank method and linear interpolation method. Percentiles can be used in various forms such as quartiles, deciles and so on.

14. What is ANOVA?

ANOVA stands for Analysis of Variance. It is a statistical method used to test the equality of means across two or more groups. It is used to determine whether there are significant differences between the means of different groups, and it can be used to compare means across multiple groups. It is also used to determine whether there is a significant interaction between different variables. ANOVA is a powerful tool for understanding the relationships between variables and for making inferences about populations based on sample data.

15. How can ANOVA help?

ANOVA can help in several ways:

- Understanding relationships between variables: ANOVA can be used to determine whether there is a significant relationship between two or more variables. For example, it can be used to determine whether there is a significant relationship between a person's income level and their level of education.
- Making inferences about populations: ANOVA allows researchers to make inferences about populations based on sample data. For example, if a study finds a significant difference in means between two groups, the researcher can infer that the same difference would be found in the population from which the sample was drawn.
- Identifying sources of variation: ANOVA can be used to identify which variables are responsible for the differences in means between groups. This can help researchers to identify which variables are most important in explaining the differences in means.
- Comparing multiple groups: ANOVA can be used to compare means across multiple groups. For example, it can be used to compare the means of different treatments in an experiment or to compare the means of different groups in a survey.
- Identifying interactions between variables: ANOVA can also be used to identify whether there is a significant interaction between different variables. For example, it can be used to determine whether there is an interaction between the type of fertilizer used and the type of crop grown on crop yield.