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

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:

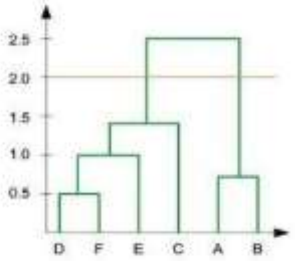
a. 1 only

b. 2 only

c. 1 and 2

d. None of them

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

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

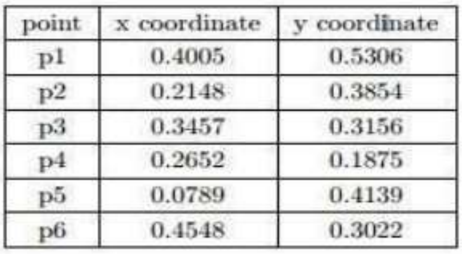
a. Given sales data from a large number of products in a supermarket, estimate future sales for each of these products.

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.

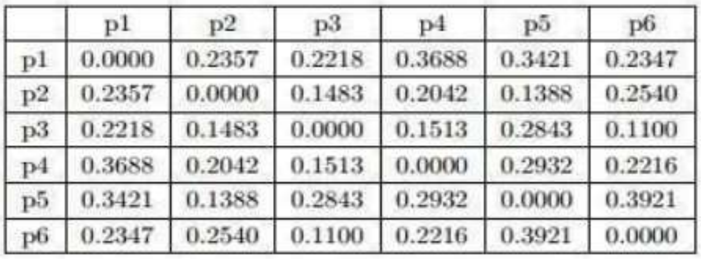
d. Given historical weather records, predict if tomorrow's weather will be sunny or rainy.

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

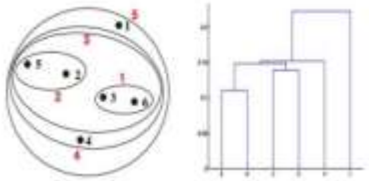
Table : X-Y coordinates of six points.



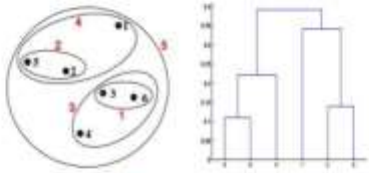
	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

Table : Distance Matrix for Six Points

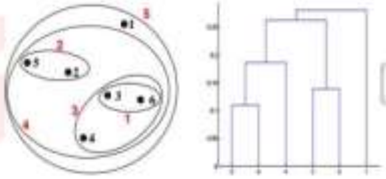
Which of the following clustering representations and dendrogram depicts the use of MIN or Single link proximity function in hierarchical clustering:



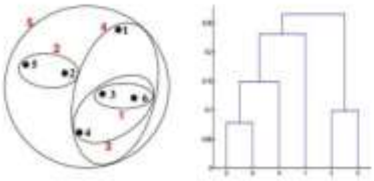
a.



b.



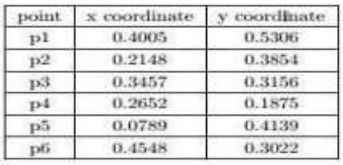
c.



d.

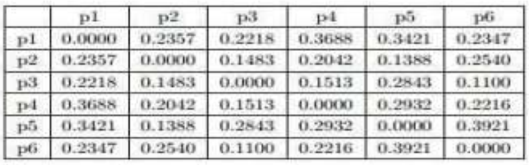
ANSWER: a

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.

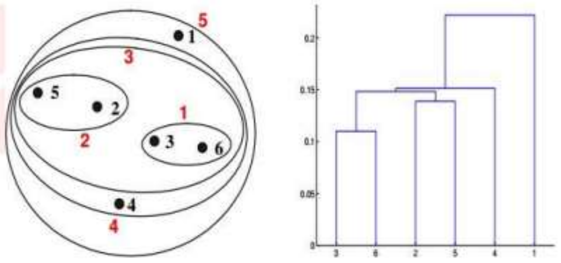


	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

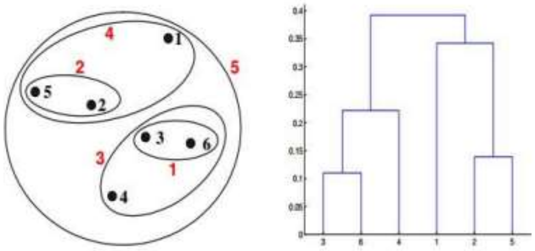
Table : Distance Matrix for Six Points

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

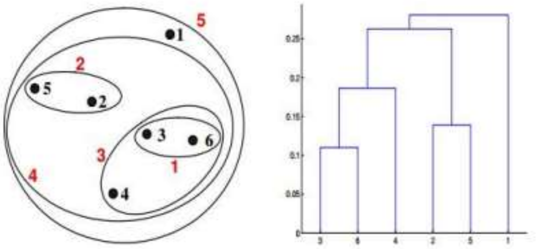
a.



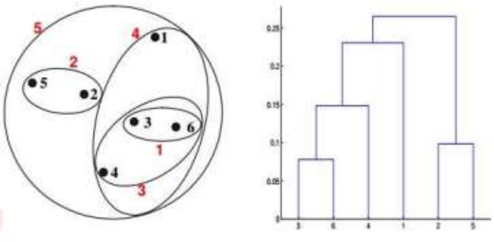
b.



c.



d.



ANSWER: a

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

13. What is the importance of clustering?

ANSWER:

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 then to those in other groups.

14. How can I improve my clustering performance?

ANSWER:

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 RTCA or SFT, improves clustering performance.

14. How can I improve my clustering performance?

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

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

3. How many outcomes are possible with Bernoulli trial?

a) 2

b) 3

c) 4

d) None of the mentioned

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

5. Level of significance is also called:

a) Power of the test

b) Size of the test

c) Level of confidence

d) Confidence coefficient

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

a) Decrease

b) Increase

c) Both of them

d) None

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

a) Probability

b) Hypothesis

c) Causal

d) None of the mentioned

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

WORKSHEET

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

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

10. What Is Bayes' Theorem?

ANSWER:

Bayes’ Theorem states that the conditional probability of an event, based on the occurrence of another event, is equal to the likelihood of the second event given the first event multiplied by the probability of the first event.

11. What is z-score?

ANSWER:

Z-score indicates how much a given value differs from the standard deviation. The Z-Score, or standard score, is the number of standard deviation is a given data point lies above or below mean. Standard deviation is essentially a reflection of the amount of variability within a given data set.

12. What is t-test?

ANSWER:

A t-test is an inferential statistic used to determine if there is a significant difference between the means of two group and how they are related. T-tests are used when the data sets follow a normal distribution and have unknow variances, like data set recorded from flipping a coin 100 times.

13. What is percentile?

ANSWER:

In statistics, a percentile is a term describes how a score compares to other scores from the same set.

14. What is ANOVA?

ANSWER:

Analysis of variance (ANOVA) is a statistical formula used to compare variances across the mean(or average) of different groups.

15. How can ANOVA help?

ANSWER:

ANOVA is helpful for testing three or more variable. It is similar to multiple two-sample t-tests. However, it results in fewer type I errors and is appropriate for a range of issues. ANOVA group differences by comparing the means of each group and includes spreading out the variance into diverse sources.