BUSINESS INTELLIGENCE AND ANALYTICS

ASSIGNMENT WEEK 9:

Total marks = 1 Marks
15 Qns * 1 marks = 15 marks

1. What is a key challenge associated with unsupervised learning? (1 Mark)

a) Lack of data availability
b) Subjectivity and absence of a clear analysis goal
c) Overfitting issues
d) Limited model interpretability

Answer: B) Subjectivity and absence of a clear analysis goal

2. For clustering, we do not require- (1 Mark)

A. Labeled data
B. Unlabeled data
C. Numerical data
D. Categorical data
ANSWER: A

3. Which of the following is an example of an unsupervised learning algorithm? (1 Mark)
A. Linear regression
B. Logistic regression
C. K-means clustering
D. Support vector machines

Ans: C. K-means clustering

4. What distinguishes K-means clustering from hierarchical clustering? (1 Mark)

A) K-means clustering produces a tree-like representation, while hierarchical clustering uses pre-specified clusters.

B) K-means clustering requires knowing the number of clusters beforehand, while hierarchical clustering does not.

C) K-means clustering creates a dendrogram, while hierarchical clustering creates distinct clusters.

D) There is no difference between K-means and hierarchical clustering techniques.

Answer: B) K-means clustering requires knowing the number of clusters beforehand, while hierarchical clustering does not.

5. What does a dendrogram represent in hierarchical clustering? (1 Mark)

   A) A scatter plot of feature clusters

   B) A visual display of K-means clustering

   C) A tree-like structure showing clustering at various levels

   D) A linear representation of data distributions

   Answer: C) A tree-like structure showing clustering at various levels

6. Which of the following is a method of choosing the optimal number of clusters for k-means? (1 Mark)

   A. Shadow method
   B. the silhouette method
   C. the elbow method
   D. B and C

   ANSWER: D

7. Which of the following statements best describes the goal of SMOTE preprocessing technique? (1 Mark)
   a) Reduce the dimensionality of the data
   b) Balance the class distribution in imbalanced datasets
   c) Improve the interpretability of a machine learning model
   d) Detect outliers in the data

   Ans:b) Balance the class distribution in imbalanced datasets

8. What defines a good clustering according to the K-means approach? (1 Mark)
   A) Maximizing the distance between clusters
   B) Minimizing the total number of observations in each cluster
   C) Minimizing the sum of squared distances within each cluster
   D) Maximizing the number of overlapping observations between clusters

Answer: C) Minimizing the sum of squared distances within each cluster

9. Which of the following is a limitation of K-means clustering? (1 Mark)

   A. Sensitivity to the initial placement of cluster centroids
   B. Inability to handle missing data
   C. Inability to handle categorical data
   D. All of the above
   Explanation: D

10. Which of the following statements about distance between clusters is true? (1 Mark)
    A) Single linkage computes distances between cluster centroids.
    B) Complete linkage uses the average similarity of all objects within clusters.
    C) Single linkage calculates the distance between individual objects in different clusters.
    D) Complete linkage considers the maximum distance between objects in different clusters.
    Answer: D) Complete linkage considers the maximum distance between objects in different clusters.

11. In a 3-dimensional space represented by coordinates (x, y, z), two cluster centroids, A and B, have coordinates A(2, 4, 6) and B(5, 1, 3) respectively. What is the precise Euclidean distance between these centroids, denoting their dissimilarity in the cluster space? (1 Mark)
    A) 5.20 units

    B) 3.00 units

    C) 4.36 units

    D) 6.48 units
    Answer: A) 5.20 units

12. In K-means clustering, what is the purpose of the "elbow method"? (1 Mark)
    A. To determine the optimal number of clusters
    B. To identify the best distance metric
    C. To select the best initialization method
    D. To determine the convergence criteria

    ANSWER:A

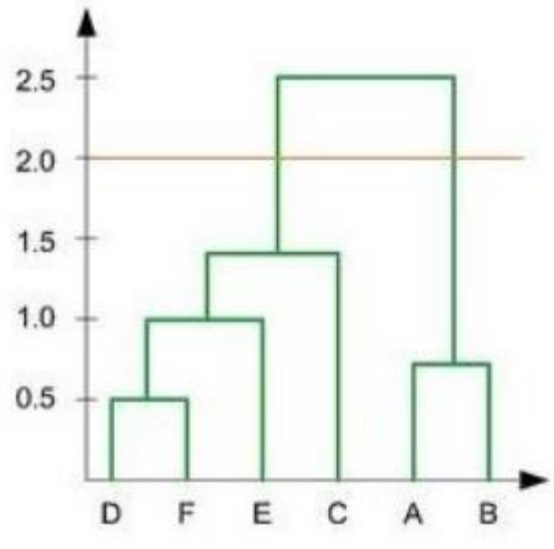
13. Suppose that a customer transaction table contains 9 items and 3 customers. What is the Jaccard coefficient (similarity measure for asymmetric binary variables) for C1 and C2? (1 Mark)

	ITEM 1	ITEM 2	ITEM 3	ITEM 4	ITEM 5	ITEM 6	ITEM 7	ITEM 8	ITEM 9
C1	0	1	0	0	0	1	0	0	1
C2	0	0	1	0	0	0	0	0	1
C3	1	1	0	0	0	1	0	0	0

    a) 0.75
    b) 0.25
    c) 0.35
    d) 0.85

Ans: b. 0.25

14. In the figure below, if you draw a horizontal line on the y-axis for y=2. What will be the number of clusters formed? (1 Mark)



Options:
A. 1
B. 2
C. 3
D. 4

Solution: (B)

15. Assume you want to cluster 7 observations into 3 clusters using the K-Means clustering algorithm. After first iteration, clusters C1, C2, C3 have following observations:

C1: {(2,2), (4,4), (6,6)}
C2: {(0,4), (4,0)}
C3: {(5,5), (9,9)}
What will be the Manhattan distance for observation (9, 9) from cluster centroid C1 in the second iteration? (1 Mark)

Options:
A. 10
B. 5* sqrt (2)
C. 13*sqrt (2)
D. None of these

Solution: (A)