KATHMANDU UNIVERSITY
End Semester Examination
February, 2025

Marks Scored:

Level : B.E./B.Sc.
Year : IV

Course : COMP 482
Semester : I

Exam Roll No. :

Time: 30 mins.

F. M. : 10

Registration No.:

Date 04 FEB 2025

SECTION "A"
[20Q. × 0.5 = 10 marks]

Choose the most appropriate answer and **encircle**.

1. Which of the following is NOT a data mining task?
   a. Entering data into the system manually from printed documents.
   b. Detecting and removing noisy data automatically.
   c. Identifying new items in a stream of data items.
   d. Finding groups of similar items.

2. Which of the following visualizations is used for representing distributions?
   a. Scatter Plot b. Line Chart c. Heatmap d. Box Plot

3. What does numerosity reduction aim to achieve?
   a. Reduce the dimensionality of the dataset.
   b. Replace the original data with a smaller representation that preserves trends and patterns.
   c. Cluster data points into fewer groups.
   d. Increase the number of features.

4. A confusion matrix shows 90 true positives, 30 false positives, 10 false negatives, and 70 true negatives. What is the accuracy of the model?
   a. 0.5 b. 0.2 c. 0.6 d. 0.8

5. In k-Nearest Neighbors (k-NN) outlier detection, a data point is considered an outlier if _____
   a. Its distance from the k-nearest neighbors is small.
   b. It has many neighboring data points within a small radius.
   c. It has few neighboring data points or is far from the neighbors.
   d. Its z-score is high.

6. In the Apriori algorithm, what is the significance of the minimum support threshold?
   a. It determines the confidence level of a rule.
   b. It ensures that only frequent itemsets are considered.
   c. It decides the number of clusters.
   d. It reduces dimensionality.

7. Which of the following is a disadvantage of decision trees?
   a. Tendency to overfit the data. b. Lack of interpretability.
   c. Inability to handle categorical data. d. Poor performance on small datasets.

8. In a Bayesian Network, what does each node represent?
   a. A variable
   b. A data point
   c. A conditional probability table
   d. A dependency

9. A population consists of 3 strata with the following sizes and average values:
   - Stratum A: 50 individuals, average age = 30 years
   - Stratum B: 100 individuals, average age = 40 years
   - Stratum C: 150 individuals, average age = 50 years

   If a sample of 60 individuals are drawn using proportional stratified sampling, how many individuals should you sample from each stratum?
   a. Stratum A: 15, Stratum B: 15, Stratum C: 30.
   b. Stratum A: 10, Stratum B: 30, Stratum C: 20
   c. Stratum A: 20, Stratum B: 20, Stratum C: 20
   d. Stratum A: 10, Stratum B: 20, Stratum C: 30

10. What happens in a 5-fold cross validation?
   a. The train and test subsets contain respectively 80% and 20% of the data.
   b. The train and test subsets contain respectively 95% and 5% of the data.
   c. The train and test subsets contain exactly the same number of instances.
   d. The train and test subsets contain 70% and 30% of the data.

11. A dataset contains 100 transactions with 40 transactions containing milk, 30 containing bread, and 20 containing milk and bread. What is the confidence of the rule milk → bread?
   a. 20%
   b. 50%
   c. 66.67%
   d. 100%

12. What is the primary goal of data science methodologies?
   a. To develop complex algorithms.
   b. To optimize database performance.
   c. To focus only on data visualization.
   d. To create a systematic approach for solving data-driven problems.

13. Which performance metric is typically used for evaluating logistic regression models?
   a. Mean squared error
   b. R-squared
   c. Confusion Matrix
   d. Silhouette score

14. What is the main purpose of backpropagation in neural networks?
   a. To initialize the weights
   b. To calculate the output
   c. To determine the activation function
   d. To update the weights by minimizing the error

15. Which of the following steps is NOT part of the k-means algorithm?
   a. Initialize k centroids
   b. Use dimensionality reduction to optimize cluster assignments.
   c. Update the centroids by taking the mean of assigned points.
   d. Assign each data point to the nearest centroid.

04 FEB 2025

16. Which of the following is a key characteristic of ROLAP?
    a. It stores data in a multidimensional array.
    b. It cannot handle large volumes of data.
    c. It requires precomputed aggregates for analysis.
    d. It relies on relational databases to perform OLAP operations.

17. Consider an OLAP cube with three dimensions: Product (P1, P2), Region (North, South), and Time (2023, 2024). The cube stores sales revenue for each combination. If the cube contains 2 x 2 x 2 = 8 data cells, what would happen to the number of cells if a new dimension Customer (C1, C2, C3) is added?
    a. 12 cells b. 16 cells c. 24 cells d. 32 cells

18. What is the main difference between classification and regression in data mining?
    a. Classification predicts continuous values, while regression predicts categories
    b. Classification predicts categories, while regression predicts continuous values.
    c. Classification groups data into clusters, while regression finds patterns in data.
    d. Classification is unsupervised, while regression is supervised.

19. Suppose you are given the following data points in 1D space: {2, 4, 10, 12, 18, 20, 25} and you want to divide them into k = 2 clusters. Initially, the centroids are C1 = 4, and C2 = 1. After one iteration of the k-means algorithm, which of the following will be the new position of the centroids?
    a. C1 = 7, C2 = 21 b. C1 = 5.33, C2 = 21.67.
    c. C1 = 5.33, C2 = 18.67 d. C1 = 6, C2 = 19

20. What is the key assumption in the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) method for outlier detection?
    a. Data points are uniformly distributed.
    b. Clusters are circular in shape.
    c. Data points are normally distributed.
    d. Noise points are sparse in regions of low density.

KATHMANDU UNIVERSITY
End Semester Examination
February, 2025

Level : B.E./B.Sc.
Year : IV
Time : 2 hrs. 30mins.

04 FEB 2025

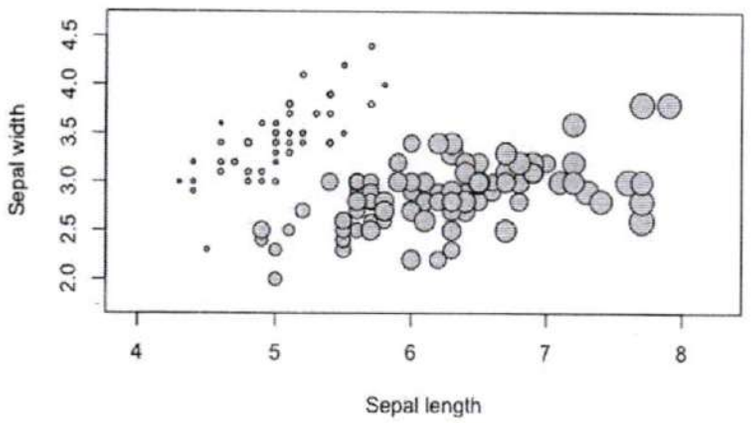
Course : COMP 482
Semester : I
F. M. : 40

SECTION "B"
[6Q. × 4 = 24 marks]

Attempt *ANY SIX* questions.

1. Discuss the Data Science Hierarchy of Needs. Why is it important to follow data mining methodologies? Explain, in brief, any one data mining methodology. [1 + 1 + 2]

2. Discuss the importance of exploratory data analysis. Suppose you are given the iris dataset with 3 attributes - sepal length, sepal width, and petal length. What can you say



about the dataset from the following plot? [2 + 2]

3. Define dimensions, and measures. Give an example of a constellation schema with at least 2 measures, and 3 dimensions. [2 + 2]

4. Explain the role of lift and correlation analysis in measuring the interestingness of association rules. Give an example of a situation where lift can better measure rule interestingness than confidence. [2 + 2]

5. Compare and contrast classification and regression tasks. Design a naive Bayes classifier from Table 1. [2 + 2]

6. What kind of data can be used with a single-layer neural network for binary classification? Design a perceptron for implementing OR gate. [1 + 3]

7. Differentiate between (*ANY TWO)*: [2 + 2]
   a. Numerosity reduction and dimensionality reduction
   b. Divisive and agglomerative clustering
   c. SEMMA and CRISP-DM

P.T.O.

SECTION "C"

[2Q. × 8 = 16 marks]

Attempt *ANY TWO* questions.

8. Consider the following dataset. Simulate a 3-fold cross validation for evaluating two classification algorithms, clearly explaining how a 3-fold cross-validation works. How would you choose the value of k for k-fold cross-validation? Explain the scenario when you would choose k=n, where n is the number of instances in the dataset. [5 + 2 + 1]

Table 1 Dataset for Question 5 and 8

Instance	Feature1	Feature2	Class
1	S	A	+
2	S	B	-
3	S	A	+
4	M	A	+
5	S	B	+
6	M	A	-

9. Construct a dendrogram that shows the hierarchical relationship between the following data points: (3, 1, 1), (2, 2, 2), (1, 2, 5), (3, 2, 4), (4, 5, 5). Use the single link method to compute distance between clusters. Divide the data points into two clusters using the constructed dendogram, and compute Silhouette coefficient of any one data point. [4 + 1 + 3]

10. What is the main idea behind statistical methods of outlier detection? In the Interquartile range (IQR) method of detecting outliers, a data point is an outlier if it is more than 1.5 times IQR above the third quartile or below the first quartile. Note that IQR is the difference between the third quartile (Q3) and the first quartile (Q1). Using this method, determine if the highest and the lowest scores are outliers in the set of 19 test scores: 5, 7, 7, 8, 9, 10, 11, 11, 11, 11, 13, 14, 14, 14, 16, 16, 16, 17, 20. How would you use this method to find outliers in a multivariate data? [2 + 4 + 2]