BRAINWARE UNIVERSITY

Term End Examination 2024-2025
Programme – BCA-2022
Course Name – Introduction to Data Sciences
Course Code - BCAE602B
(Semester VI)

Full Marks : 60 **Time : 2:30 Hours**

[The figure in the margin indicates full marks. Candidates are required to give their answers in their own words as far as practicable.]

Group-A

(Multiple Choice Type Question) 1 x 15=15

1. *Choose the correct alternative from the following :*

(i) In which of the following cases will K-means clustering fail to give good results? 1) Data points with outliers 2) Data points with different densities 3) Data points with non convex shapes

a) 1 and 2 b) 2 and 3
c) 1, 2, and 3 d) 1 and 3

(ii) Suppose you have trained a logistic regression classifier and it outputs a new example x with a prediction ho(x) = 0.2. This means

a) Our estimate for P(y=1 | x) b) Our estimate for P(y=0 | x)
c) Our estimate for P(y=1 | x) d) none

(iii) Five numbers are given: (5, 10, 15, 5, 15). Now, what would be the sum of deviations of individual data points from their mean?

a) 10 b) 25
c) 50 d) 5

(iv) A test is administered annually. The test has a mean score of 150 and a standard deviation of 20. If Ravi's z-score is 1.50, what was his score on the test?

a) 180 b) 130
c) 50 d) 150

(v) Branch of statistics which deals with development of particular statistical methods is classified as

a) industry statistics b) economic statistics
c) applied statistics d) mathematical statistics

(vi) Which of the following is the probability calculus of beliefs, given that beliefs follow certain rules?

a) Bayesian probability b) Frequency probability
c) Frequency inference d) Bayesian inference

(vii) Choose the correct SQL command to retrieve all columns from a table named `employees`.

a) SELECT * FROM employees; b) GET ALL FROM employees;
c) SHOW employees; d) FETCH * employees;

(viii) Identify a common statistical measure used to describe data spread.

a) Variance b) Mean
c) Mode d) Frequency

(ix) Identify a key characteristic of NoSQL databases.

a) They allow flexible data models b) They only store relational data
c) They require predefined schemas d) They are slower than SQL databases

(x) A cricketer's scores in five matches are: 45, 38, 50, 42, 55. What is the range?

a) 16 b) 17
c) 22 d) 25

(xi) Calculate the median of the data: 3, 7, 9, 12, 15, 18, 21

a) 9 b) 10
c) 11 d) 12

(xii) Which of the following is not the correct syntax for creating a set in Python?

a) set([[1,2],[3,4]]) b) set([1,2,2,3,4])
c) set((1,2,3,4)) d) {1,2,3,4}

(xiii) Which of the following statements do not align with best practices in data processing and dimensionality reduction?

a) Training and testing data must be processed differently. b) Test transformation is often imperfect.
c) The primary goal of PCA is statistical analysis, while the secondary goal is data compression. d) All of the mentioned

(xiv) Which of the following statements does not align with standard data science and statistical practices?

a) ROC curve stands for Receiver Operating Characteristic. b) For time series analysis, data must be processed in chunks.
c) Random sampling must always be done with replacement. d) None of the mentioned.

(xv) Which of the following statements does not correctly describe clustering or classification techniques?

a) k-means clustering is a method of vector quantization. b) k-means clustering aims to partition n observations into k clusters.
c) k-nearest neighbor is the same as k-means. d) k-nearest neighbor is the same as k-means.

Group-B

(Short Answer Type Questions) 3 x 5=15

2. A dataset follows a normal distribution with a mean of 50 and a standard deviation of 5. Calculate the probability of getting a value greater than 60 using the empirical rule. (3)
3. A company wants to minimize the cost function C(x,y)=x^2+y^2, subject to the constraint that the total resource allocation satisfies x^2 + y^2 = 4. Use the Lagrange multiplier method to find the optimal values of x and y. (3)
4. Calculate the Euclidean distance between the two points (2,3) and (5,7) in a k-NN model. (3)
5. Illustrate how to implement multiple linear regression in Python using the scikit-learn library with a sample dataset. (3)
6. A company is using k-means clustering for customer segmentation but is unsure about the optimal number of clusters. Critique whether the Elbow Method or Silhouette Score is a better approach to determine the optimum number of clusters and justify your reasoning. (3)

OR

A company is using Akaike Information Criterion (AIC) and Adjusted R-squared to select the best multiple regression model. Critique which metric is more effective for model selection and justify your reasoning. (3)

Group-C

(Long Answer Type Questions) 5 x 6=30

7. Evaluate the roles of descriptive and inferential statistics in data analysis. Justify the significance of each through practical examples. (5)
8. Draw and explain the bias-variance tradeoff in machine learning. Interpret its impact on model performance. (5)
9. Use Bayes's theorem to solve a probability-based problem and explain the steps involved with a real-life example. (5)
10. Evaluate the performance of different classification models (e.g., Logistic Regression, Random Forest, and SVM) on a given dataset using precision, recall, and F1-score. (5)
11. A regression analysis between apples (y) and oranges (x) resulted in the following least-squares line: $y = 100 + 2x$. Predict the implication if oranges are increased by 1 (5)
12. Critically evaluate the different types of biases encountered during the sampling process by analyzing their impact on research validity. Support your evaluation with real-life examples for each type of bias. (5)

OR

Critically evaluate the concept of degrees of freedom (DF) in statistics by explaining its significance and impact on the validity of statistical tests. Support your evaluation with a practical example that illustrates its role in hypothesis testing. (5)

****