QPC: RM17001107 AR - 17 Reg. No.

GIET MAIN CAMPUS AUTONOMOUS GUNUPUR – 765022

B. Tech Degree Examinations, May – 2021

(Eighth Semester)

BCSOE 8031 – DATA SCIENCE

(AEI & ECE)

Time: 2 hrs Maximum: 50 Marks

Answer ALL Questions

The figures in the right hand margin indicate marks.

PART – A: (Multiple Choice Questions) (1 x 10 = 10 Marks)

Q.1. Answer ***ALL*** questions

a. Which of the following is not a application for data science?

(i) Recommendation Systems (ii) Image Recognition

(iii) Privacy Checker (iv) Online Price Comparison

b. What is pca.components in Sklearn?

(i) Set of all Eigen vectors for the projection space (ii) Matrix of principal components

(iii) Result of the multiplication matrix (iv) None of the above

c. Which of the following are the Data Sources in data science?

(i) Structured (ii) Unstructured

(iii) Structured and Unstructured (iv) None of the above

d. If X is random variable that follows Normal distribution, then 5X + 10 is a

(i) Chi square distribution (ii) Exponential distribution

(iii) Normal distribution (iv) t-distribution

e. In a Hypothesis test the objective is to

(i) Reject the null hypothesis (ii) Retain the alternate hypothesis

(iii) Decide whether to retain or reject null hypothesis (iv) Decide whether to retain or reject alternate hypothesis

f. The point where the null hypothesis gets rejected is called as?

(i) Significance Value (ii) Acceptance value

(iii) Rejection value (iv) Critical value

g. An urn B1 contains 2 white and 3 black balls and another urn B2 contains 3 white and 4 black balls. One urn is selected at random and a ball is drawn from it. If the ball drawn is found black, Find the probability that the urn chosen was B1.

(i) 21/41 (ii) 7/15

(iii) ½ (iv) 11/20

h. Which of the following is an assumption made by the Naive Bayes classier?

(i) The feature values are conditionally independent given the label (but not necessarily independent) (ii) The feature values are independent

(iii) The feature values are independent, but not conditionally independent given the label (iv) The feature values are independent AND conditionally independent given the label

i. Design a minimum distance classifier with three classes using the following training data:

Class 1: $\begin{bmatrix}-1.0\\0.5\end{bmatrix}$ $\begin{bmatrix}-1.0\\-1.5\end{bmatrix}$ Class 2: $\begin{bmatrix}1.0\\0.5\end{bmatrix}$ $\begin{bmatrix}1.0\\-0.5\end{bmatrix}$ Class 3 : $\begin{bmatrix}-1.0\\0.5\end{bmatrix}$ $\begin{bmatrix}-1.0\\1.5\end{bmatrix}$ Then

classify the test vector $[0.5, -1]^T$ with the trained classifier. Which class does this vector belong to?

(i) Class 1 (ii) Class 2

(iii) Class 3 (v) None of the above

j. What do you mean by a hard margin?

(i) The SVM allows very low error in classification (ii) The SVM allows high amount of error in classification

(iii) The SVM allows no error in classification (iv) None of the above

PART – B: (Short Answer Questions) (2 x 5 = 10 Marks)

Q.2. Answer ***ALL*** questions

a. What is "overfitting" and generalization?

b. Suppose, you are working on a binary classification problem with 3 input features. And you chose to apply a bagging algorithm(X) on this data. You chose max_features = 2 and the n_estimators =3. Now, think that each estimator has 70% accuracy. What will be the maximum accuracy you can get?

c. How to select best hyper parameters in tree based models?

d. What are the different parameters used to validate a simple linear regression model?

e. What do you mean by Web scraping?

PART – C: (Long Answer Questions) (6 x 5 = 30 Marks)

Answer ***ANY FIVE*** questions Marks

3. Build a linear regression model to find a best fit to the following equation. (6)
   $Y = 4 + 2\ x_1 + 3x_2$. Use at least four data samples to find the regression parameters
4. Explain how t-test and F-test is used in Multiple linear regression model building. (6)
5. Discuss the different data visualisation functions in Python. Use examples. (6)
6. Explain different metrics used in nearest neighbour classification and also write the properties of Metrics. (6)
7. Discuss how Multilayer Perceptron can be used to solve nonlinear classification problem. (6)
8. Discuss different types of regularization techniques used in machine learning. (6)
9. Given following 2D data points (6)
   class 1: $x1 = [1; 1]^T$; $x2 = [1; 0]^T$; $x3 = [-1; 0]^T$; $x4 = [-1; -1]^T$
   class 2:; $x5 = [4; 7]^T$; $x6 = [7; 4]^T$; $x7 = [4; 2]^T$; $x8 = [2; 4]^T$. Design an SVM for an optimized hyperplane
10. Derive the discriminant functions for the normal density for different cases of covariance matrix. (6)

--- End of Paper ---