JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

TEST-3 EXAMINATION-2022

B.Tech-V Semester (CS&IT)

COURSE CODE (CREDITS): 20B1WCI531 (2) MAX. MARKS: 35

COURSE NAME: FOUNDATION FOR DATA SCIENCE AND VISUALIZATION

COURSE INSTRUCTORS: Ravindara Bhatt and Prateek Thakral MAX. TIME: 2 Hours

*Note: All questions are compulsory. Marks are indicated against each question in square brackets.*

**Q1. [2 +3]**

a) How would you build a data-driven recommendation system? What are the limitations of this approach? [**CO 1**]

b) Write Python program to build a nearest neighbor model that can predict the class from the IRIS dataset. [**CO 2**]

**Q.2 [2 + 3]**

a) Human eyes are fast and effective at judging the quality of clustering methods for two-dimensional data. Can you design a data visualization method that may help humans visualize data clusters and judge the clustering quality for three-dimensional data? What about for even higher-dimensional data? [CO 4]

b) What types of outliers might you expect to occur in the following data sets:

(i) Student grades.

(ii) Salary data.

(iii) Lifespans in Wikipedia [**CO 3**]

**Q 3. [2 + 3]**

a) How can the data be normalized? [**CO 4**]

b) Write Python program to solve the below Fizz Buzz programming challenge. Print the numbers 1 to 100, except that if the number is divisible by 3, print "fizz"; if the number is divisible by 5, print "buzz"; and if the number is divisible by 15, print "fizzbuzz". [**CO 2**]

**Q.4 [2 + 3]**

a) Explain overfitting and underfitting in detail with an example. [**CO 6**]

b) Explain null and alternative hypothesis by considering the example for a flipping coin. [**CO 5**]

Q 5. **[2 + 3]**

a) How many different 7-place license plates are possible if the first 3 places are to be occupied by letters and the final 4 by numbers? **[CO 5]**

b) Suppose you build a classifier that answers yes on every possible input. What precision and recall will this classifier achieve? **[CO6]**

c) Suppose that the data mining task is to cluster the following eight points (with (x, y) representing location) into three clusters: A1(2, 10), A2(2, 5), A3(8, 4), B1(5, 8), B2(7, 5), B3(6, 4), C1(1, 2), C2(4, 9). The distance function is Euclidean distance. Suppose initially we assign A1, B1, and C1 as the center of each cluster, respectively. Use the k-means algorithm to show only

   i. The three cluster centers after the first round execution

   ii. The final three clusters

**[CO 6]**

**Q 6. [2 + 2 + 1]**

a) Show that the eigenvalues of $MM^T$ are the same as that of $M^TM$. Are their eigenvectors also the same? **[CO 6]**

b) Explain what precision and recall are. How do they relate to the ROC curve? Is it better to have too many false positives, or too many false negatives? Explain. **[CO6]**

c) Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8): Compute the Minkowski distance between the two objects, using q = 3. **[CO4]**

**Q** 7. **[2 + 2+ 1]**

a) What is cross-validation? How might we pick the right value of k for k-fold cross validation? **[CO 6]**

b) Suppose a bank would like to develop a classifier that guards against fraudulent credit card transactions. Illustrate how you can induce a quality classifier based on a large set of non-fraudulent examples and a very small set of fraudulent cases. **[CO 6]**

c) Show that accuracy is a function of sensitivity and specificity. **[CO 6]**