DATA AND TEXT MINING

DOCUMENTATION RESEARCH AND TRAINING CENTRE, ISI-BC

Final EXAMINATION (2025)

Total Marks: 100 **Time: 180 mins.**

ANSWER Any FIVE

Important Information: PROVIDE SUFFICIENT AND RELEVANT EQUATIONS, DESCRIPTIONS, AND ILLUSTRATIONS TO JUSTIFY YOUR ANSWER.

1. Illustrate the working principle of minimum distance and KNN classification models. Compare their advantages and disadvantages. [20]

2. Discuss measures of Location, Spread, Shape, and Dependence with their applications. [20]

3. If X is a data set with 9 samples and 2 features each. Find the Euclidean and Mahalanobis distances between 2nd and 5th samples of XX. Explain the reasons for the different distance values and their merits and demerits. [20]

XX =[4 5
3 2
1 6
4 0
5 5
6 7
7 7
8 9
9 7];

$$\frac{1}{n-1}\frac{(x-\bar{x})^4}{s^4}$$

4. [10+5+5]

- Provide a complete interpretation of the Box and Whisker Plot Analysis of a data set. How it behaves for symmetric and skewed data.
- What is the interquartile range, and find it for the following vector?
  X=[1, 7, 36, 14, 4, 15, 53, 25, 11]
- Establish the relationship between variance-covariance matrix with correlation matrix. Describe their significances.

5. What is cross-validation in data mining? Why is it so important in the selection of the training and test sets? Discuss different types of cross-validation approaches, their working principles, with examples. [20]

6. What is the distance metric in the decision-making process of data mining? List out the common properties of a distance METRIC. Discuss the various distance metrics starting from Minkowski distance to Mahalanobis distance. [20]

====================END of the Question Paper=====================