DATA AND TEXT MINING

DOCUMENTATION RESEARCH AND TRAINING CENTRE, ISI-BC

FINAL-TERM EXAMINATION (2024)

Total Marks: 60 **Time: 180 mins.**

ANSWER ANY SIX

1. Explain
   - What is overall classification accuracy and cost-sensitive error? [2]
   - Mention the demerits of overall classification accuracy and discuss different cost-sensitive measures. [2+6]

2. State and discuss the merits and demerits of Bayes decision theory. [10]

3. Write the KNN algorithm. Describe its merits and demerits. [5+5]

4.
   - What is a cluster analysis and its challenges? [4]
   - Write the K-means algorithm with its advantages and disadvantages. In what way K-medoid overcome such a situation? [3+3]

5. Given the following confusion matrix of a classifier, find [10]
   a. Accuracy — $\frac{a+d}{a+b+c+d}$
   b. Precision → $\frac{a}{a+c}$
   c. Recall → $\frac{a}{a+b}$
   d. Sensitivity → $\frac{a}{a+b}$
   e. selectivity

   of the classes

		PREDICTED CLASS	
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a (TP) 6954	FN b 46
	Class=No	FP c 412	TN d 2588

6. What are generalization and over-fitting aspects in pattern recognition? Discuss the effect of these factors on the said task. [5+5]

7. Describe the motivations for standardization and normalization of data sets. Give one method for performing these operations. [3+3+4]

8. Differentiate Feature selection and Feature extraction methods of data pre-processing. Describe the challenges involved in these operations. [5+5]

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