Paper-21A

DATA AND TEXT MINING

DOCUMENTATION RESEARCH UNIT, ISI-BC

(FINAL EXAMINATION)

Full Marks: 70 **Dt: 26/04/18** **Time: 3 hrs.**

1. What is ROC curve? Describe the interpretation of this curve. [2+4]

2. What are generalization and over-fitting aspects in pattern recognition? Discuss the effects of these factors on the said task. [4+4]

3. Discuss and differentiate between the classification and clustering problems with examples. [3+3]

4. Write the K-means and DBSCAN clustering algorithms with their advantages and disadvantages over each other. [4+4]

5. What are principal components? Describe the procedure to get the principal components. [2+4]

6. Describe Bayes decision rule for classification task. List out the merits and demerits of this rule. [5+6]

7. 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 X. Explain the reasons for the different distance values. [10]

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

8. Given the following confusion matrix of a classifier, find [15]
    a. Accuracy
    b. Precision
    c. Recall
    d. Sensitivity
    e. Selectivity, of the classes

		PREDICTED CLASS	
		Class=Yes	**Class=No**
ACTUAL CLASS	**Class=Yes**	6954	46
	Class=No	412	2588