DATA AND TEXT MINING

DOCUMENTATION RESEARCH UNIT, ISI-BC

FINAL EXAMINATION

Full Marks: 50 (Buffer marks:10) **Dt: 02/05/17** **Time: 3 hrs.**

1. Explain [2+2]
   a. Discuss the significance of cross validation?
   b. Enlist two methods of cross validation.

2. What is ROC curve? Describe the interpretation of this curve. [2+3]

3. Differentiate between real and artificial data set for pattern classification task. With the help of Bayes decision rule, demonstrate the behaviour of misclassification error with the increase of training and test samples. [3+6]

4. What are generalization and over-fitting aspects in pattern recognition? Discuss the effect of these factors on the said task. [3+2]

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

6.
   a. Give the mathematical formulation of defining clusters. [3]
   b. Write the K-means and DBSCAN clustering algorithms with their advantages and disadvantages over each other. [4+4]

7. What are principal components? Describe the procedure to get principal components. [3+3]

8. Describe any four statistical descriptions of a data set. [4]

9. Describe the Branch and Bound Feature selection method with an example of selecting **THREE** optimum features out of **SEVEN** features. [10]