

DATA AND TEXT MINING

DOCUMENTATION RESEARCH UNIT, ISI-BC

FINAL EXAMINATION

Handwritten confusion matrix:

		Predicted	
		C	T
Actual	C	8	3
	T	5	4

Total Marks: 70

Time: 3 hrs.

1. Define the relationship between Euclidean, normalised Euclidean and Mahalanobis distance with equations. [6]
2. Give the example of a scenario, where precision and recall overcome the demerits of overall classification accuracy. [4]
3. What are generalization and over-fitting aspects in pattern recognition? [4]
4. In a database of 20 samples, 15 samples belong to CHAIR category and 5 samples belong to TABLE category. The model M classifies 13 samples as CHAIR and 7 as TABLE. Out of 13, 8 are correctly classified as CHAIR. And, Out of 7, 4 are correctly classified as TABLE. Develop the confusion matrix and find the **Precision**, **recall** and **F-measure** of the model. [4+3+3+3]
5. What are properties of Similarity and Dissimilarity measures? Give one example of these measures. [2+2+2]
6. Describe the motivations of standardization and normalization of data sets? Give at least one method of performing these operations. [4+2]
7. Describe Bayes decision model with equation? Discuss the advantages and disadvantages of this model. [5+3+3]
8. Write the KNN data mining algorithm. Describe its merits and demerits. [3+2+2]
9. What is cluster analysis? Give the mathematical formulation/properties of defining clusters [5]
10. Write the K-means and DBSCAN clustering algorithms with their advantage and disadvantages over each other. [4+4]