DATA AND TEXT MINING

DOCUMENTATION RESEARCH UNIT, ISI-BC

Total Marks: 40 **Time: 1.5 hrs.**

ANSWER All

1. If X is a data set with 5 samples and 2 features each. Find the COVARIANCE MATRIX and the Euclidean distances between 2^{nd} and 5^{th} samples of X. [3+4]

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

1. Find and draw the BOX-and-WHISKER plot analysis for the following data set [2]

   X=[4.3, 5.1, 3.9, 4.5, 4.4, 4.9, 5.0, 4.7, 4.1, 4.6, 4.4, 4.3, 4.8, 4.4, 4.2, 4.5, 4.4]

2. List out the types of attributes used in a data mining task with an example. [4]
3. Explain probability density function of a NORMAL distribution with its properties? [4]
4. List the common properties of a distance MATRIC [3]
5. Write short notes on different learning methods [4]
6. Describe with examples and equations; the measures of Location, spread, Shape and dependency. [6]
7. Why data pre-processing is so important? Explain any three types of pre-processing approaches with examples. [2+4]
8. Describe the different challenges in a data mining task. [4]