**Birla Institute of Technology & Science, Pilani**

**Work-Integrated Learning Programmes Division**

**Second Semester 2013-2014**

**Mid-Semester Test(EC-2 Regular)**

Course No. : IS ZC415

Course Title : DATA MINING

Nature of Exam : Closed Book

Weightage : 35%

No. of Pages = 1

# No. of Questions = 5

Duration : 2 Hours

Date of Exam : 15/02/2014 (AN)

Note:

1. Please follow all the *Instructions to Candidates* given on the cover page of the answer book.
2. Assumptions made if any, should be stated clearly at the beginning of your answer.
3. Mention one application where cosine similarity is more useful than supremum distance measure and state the reason for justification of your answer. [2]
4. Mention one application where supremum distance measure can be more useful and state the reason for justification of your answer. [2]

* 1. Suppose a group of persons with the sorted medical store credit points listed as follows: 0, 400, 1200, 1600, 1600, 1800, 2400, 2600, 2800

1. Partition them by Equi-width binning – for bin width of e.g., 500
2. Partition them by Equi-frequency binning – for bin density of e.g., 3: [3 x 2 = 6]
3. Suppose you are the new administrator at XYZ Ltd, an aggressive web advertising company. The company has been collecting web statistics but has not been able to make use of them because of their large volume. Give a specific example in which you might evaluate the performance of a given web ad using data mining. [2]
4. Define Overfitting and Underfitting. Mention the application/criteria where Overfitting is not at all a problem. Mention the application/criteria where Underfitting is not at all a problem. [4]
5. State the advantage of the RIPPER in terms of Instance Elimination approach. [2]
   1. Consider the following data set **DS2**.  [8 + 4 = 12]

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| TID | Attribute1 | Attribute2 | Attribute3 | Attribute4 | Class |
| 1 | C1 | 1 | 3 | no | YES |
| 2 | C1 | 2 | 2 | yes | NO |
| 3 | C2 | 0 | 2 | yes | NO |
| 4 | C1 | 0 | 2 | no | YES |
| 5 | C3 | 1 | 1 | no | YES |
| 6 | C2 | 2 | 1 | no | NO |
| 7 | C2 | 1 | 1 | no | NO |
| 8 | C1 | 0 | 3 | yes | NO |

1. Construct a decision tree for predicting the attribute **Class** using ID3 algorithm [**Note:** Use entropy calculations for feature/attribute selection].
2. Is the ID3 decision tree induction algorithm guaranteed to find an optimal tree always (that is, a tree that best classifies the training tuples over all possible trees)? Why or why not?
   1. Suppose if you have 10 students EC2 (mid semester examination) component marks and EC3 (Comprehensive examination) component marks. Which measure will you select, if you have to compare students’ performance in EC2 and EC3 component? State the reason for the justification of your answer and mention the significance of the values of the measure you have chosen. [5]

**\*\*\*\*\*\*\*\*\*\*\***