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

**Work-Integrated Learning Programmes Division**

**Second Semester 2013-2014**

**Comprehensive Examination**

**(EC-3 Regular)**

Course No. : IS ZC415

Course Title : DATA MINING

Nature of Exam : Open Book

No. of Pages = 2

# No. of Questions **=** 5

Weightage : 50%

Duration : 3 Hours

Date of Exam : 05/04/2014 (AN)

Note:

1. Please follow all the *Instructions to Candidates* given on the cover page of the answer book.
2. All parts of a question should be answered consecutively. Each answer should start from a fresh page.
3. Assumptions made if any, should be stated clearly at the beginning of your answer.
4. What are the major differences between K-NN classifiers and the Decision Tree classifiers? Explain how the K-NN classifier can be used for making prediction. [4]
5. Consider a training set ***TS*** that contains 100 positive examples (i.e. +) and 400 negative examples(i.e. -). For each of the following candidate rules,

R1: A → + (covers 4 positive and 1 negative examples),

R2: B → + (covers 30 positive and 10 negative examples),

R3: C → + (covers 100 positive and 90 negative examples),

Determine which rule is the best and worst candidate rule according to:

1. Rule accuracy.
2. Laplace measure. [2 + 2 = 4]
   1. Consider the following dataset ***DS3*** with the set of attributes (items) i.e. R = {A, B, C, D, E, F}. [6 + 2 + 2 + 2 + 2 + 6 + 3 = 23]

|  |  |
| --- | --- |
| **TID** | **Items** |
| 1 | A, C, D |
| 2 | A, C, D,F |
| 3 | A, D,E, F |
| 4 | B, D,E, F |
| 5 | A,B,F |
| 6 | A,B,D |

1. Apply the Apriori algorithm on the given dataset DS3 to find the frequent itemsets which satisfy minimum support threshold (min\_sup) of 35%.
2. List all frequent closed item sets.
3. List all frequent maximal item sets.
4. List all item sets that are infrequent but would be considered as candidate item sets by the Apriori algorithm.
5. What is the most significant advantage of FP-Tree? Why FP-Tree is complete in relevance to frequent pattern mining?
6. Construct FP-Tree for the data set ***DS3*** with minimum support threshold 35%.
7. Generate association rules from the frequent itemsets identified in Q3.a. , Which satisfy the minimum confidence (min\_conf ) of 80%.

***IS ZC415 (EC-3 Regular) Second Semester 2013-2014 Page 2***

* 1. Consider the following frequent 3-sequences:

< {i1, i2, i3} >, < {i2, i3}{i4} >, < {i2}{i3}{i3} >, < {i1, i2}{i3} >, < {i1, i3}{i4} >, < {i1}{i2, i3} >, < {i1, i2}{i4} >, < {i1, i2, i4} >, < {i2, i3}{i3} >, and < {i2}{i3}{i4} >.

1. List all the candidate 4-sequences produced by the candidate generation step of the Generalized Sequential Patterns (GSP) mining algorithm.
2. List all the candidate 4-sequences pruned during the candidate pruning step of the GSP algorithm (assuming no timing constraints).
3. List all the candidate 4-sequences pruned during the candidate pruning step of the GSP algorithm (assuming maxgap = 1). [2 + 2 + 2 = 6]
   1. Consider the following dataset ***DS5*** with eight points.

A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9).

1. Use single-link and complete-link hierarchical agglomerative clustering algorithm to cluster the dataset ***DS5***(use Euclidean distance) and show the Dendrogram.
2. Suppose that the initial seeds (centers of each cluster) are A1, A4 and A7. Run the k-means algorithm for one iteration only and show the new clusters at the end of this first iteration and the centers of the new clusters. (**Note:** State the assumption, if any). [8 + 5 = 13]

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