**Birla Institute of Technology & Science - Pilani, Hyderabad Campus**

**Second Semester 2015-2016 CS F415 / CS C415 : Data Mining**

**Test 1**

**Type: Closed Time: 60 mins Max Marks: 40 Date: 29.02.2016**

**All parts of the same question should be answered together.**

**1.** Suppose a group of 12 *sales price* records has been sorted as follows: 5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215. Partition them into three bins by each of the following methods. [6 Marks]
(a) equal-frequency partitioning (b) equal-width partitioning (c) clustering

2. In real-world data, tuples with *missing values* for some attributes are a common occurrence. Describe all possible methods for handling this problem. [5 Marks]

3.a. What is the confidence for the rules $\emptyset \rightarrow A$ and $A \rightarrow \emptyset$? [7 Marks]
3.b. Let $c1$, $c2$, and $c3$ be the confidence values of the rules $\{p\} \rightarrow \{q\}$, $\{p\} \rightarrow \{q, r\}$, and $\{p, r\} \rightarrow \{q\}$, respectively. If we assume that $c1$, $c2$, and $c3$ have different values, what are the possible relationships that may exist among $c1$, $c2$, and $c3$? Which rule has the lowest confidence?
3.c. Repeat the analysis in part (b) assuming that the rules have identical support. Which rule has the highest confidence?
3.d. Transitivity: Suppose the confidence of the rules $A \rightarrow B$ and $B \rightarrow C$ are larger than some threshold, *minconf*. Is it possible that $A \rightarrow C$ has a confidence less than *minconf*?

4.a. Use the transactional database of the below example with support threshold of 33.34%, confidence threshold of 60% and build a frequent pattern tree (FP-Tree). Show for each transaction how the tree evolves.

| Transaction ID | Items |
|---|---|
| T1 | HotDogs, Buns, Ketchup |
| T2 | HotDogs, Buns |
| T3 | HotDogs, Coke, Chips |
| T4 | Chips, Coke |
| T5 | Chips, Ketchup |
| T6 | HotDogs, Coke, Chips |

4.b. Use Fp-Growth to discover the frequent itemsets from this FP-tree. [11 Marks]

5. Let $I = \{i_1, i_2, \ldots, i_d\}$ be the item set of a rule generation data mining problem. Prove that the total number of possible rules extracted from a data set that contain d items is $3^d - 2^{d+1} + 1$. [5 Marks]

6. Consider the problem of finding the $K$ nearest neighbors of a data object. A programmer designs the below mentioned algorithm for this task. [6 Marks]

**Algorithm** Algorithm for finding $K$ nearest neighbors.

1: **for** $i$ = 1 to *number of data objects* **do**
2: Find the distances of the *ith* object to all other objects.
3: Sort these distances in decreasing order.
(Keep track of which object is associated with each distance.)
4: **return** the objects associated with the first $K$ distances of the sorted list
5: **end for**

(a) Describe the potential problems with this algorithm if there are duplicate objects in the data set. Assume the distance function will only return a distance of 0 for objects that are the same.
(b) How would you fix this problem?