## CS565: Data Mining

# Written Assignment 2

Due Date: October 26th, 2007 at 4:30 PM in the drop box.

## Problem 1 (Based on exercises 5.1 and 5.7)

The Apriori algorithm makes use of prior knowledge of subset support properties.

1. Given frequent itemset *L* and a subset *s* of *L*, prove formally that the confidence of the rule “*f* => *L-f* ” cannot be more than the confidence of “ *s* => *L-s* ”, where *f* is a subset of *s*.

2. A *partitioning* version of Apriori divides the transactions of a database *D* into *n* non-overlapping partitions. Prove that any *itemset* that is frequent in *D* must be frequent in at least one partition in *D*.

3. Suppose that all the frequent *itemsets* with minimum support *min_sup* for a large transaction database *D* are saved on a file. At some point, we add a new set of transactions $\delta$ into *D*. Discuss how to *efficiently* mine the new database *D*+ $\delta$ to find frequent itemsets using the same minimum support threshold.

## Problem 2

Consider the following database of transactions.

| **TransID** | **Items Bought** |
|---|---|
| 1 | {a,b,d,e} |
| 2 | {b,c,d} |
| 3 | {a,b,d,e} |
| 4 | {a,c,d,e} |
| 5 | {b,c,d,e} |
| 6 | {b,d,e} |
| 7 | {c,d,f} |
| 8 | {a,b,c,f} |
| 9 | {a,d,e} |
| 10 | {b,d} |

Assuming that *min_sup* = 30% (i.e. an itmeset is frequent if it appears in at least 3 transactions), use the FP-growth algorithm to generate all the frequent itemsets. Show the FP-tree, the conditional FP-trees and the frequent itemsets. Also, give the maximal frequent itemsets.

## Problem 3

Consider a database that stores records with four attributes *A*, *B*, *C* and *Class*. The first three attributes are categorical attributes and the fourth is a class attribute. Build a Naive Bayesian Classifier (NBC) using the following training set:

| **A** | **B** | **C** | **Class** |
|---|---|---|---|
| A1 | B2 | C1 | P |
| A2 | B1 | C2 | N |
| A1 | B1 | C2 | N |
| A1 | B2 | C1 | P |
| A2 | B3 | C2 | N |
| A3 | B1 | C1 | P |
| A1 | B3 | C1 | N |
| A3 | B3 | C1 | P |
| A2 | B2 | C2 | N |
| A1 | B3 | C2 | P |
| A2 | B2 | C1 | P |
| A3 | B1 | C2 | P |
| A3 | B2 | C2 | N |
| A3 | B1 | C1 | P |

Using the NBC that you created, decide the class of the following records: R1=[A2, B1, C1], R2=[A3, B1, C2] and R3=[A1,B1,C1]

## Problem 4

Given the following database below:

| **X** | **Y** | **Z** | **Class** |
|---|---|---|---|
| 15 | 1 | A1 | N |
| 20 | 3 | A2 | P |
| 25 | 2 | A1 | N |
| 30 | 4 | A1 | N |
| 35 | 2 | A2 | P |
| 25 | 4 | A1 | N |
| 15 | 2 | A2 | P |
| 20 | 4 | A1 | P |

build the complete decision tree using binary splits and gini index.

Then, compute the accuracy of the classifier on the following testing dataset.

| **X** | **Y** | **Z** | **Class** |
|---|---|---|---|
| 10 | 2 | A1 | P |
| 20 | 1 | A2 | N |
| 30 | 3 | A1 | P |
| 40 | 2 | A2 | P |