Data mining, Spring 2010. Exercises 2, April 20, 2010

1. Given the transactions given in the table 1, draw the itemset lattice induced by the items, and label each node of the lattice with the following letters:

   - M, if the node is a maximal frequent itemset
   - C, if it is a closed frequent itemset
   - N, if it is frequent but neither maximal nor closed
   - I, if it is infrequent.

   Assume a $minsup$ threshold of 30%.

2. Using the transaction data in table 1, draw the FP-tree data structure used by the FP-growth algorithm, using two different ordering for the items:

   (a) Alphabetical order

   (b) Descending order of support

3. Using the FP-tree structure for transaction data in table 1 using the alphabetical order for items, draw the

   (a) conditional FP-tree for itemsets ending with {d}

   (b) conditional FP-tree for itemsets ending with {c,d}

4. Using the transaction data of table 1, draw a contingency table for each of the following rules

Transaction ID	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}
8	{a,b,c}
9	{a,d,e}
10	{b,d}

Table 1: Marker basket transactions

4	8	2	5
6	1	6	9
2	10	8	4
10	4	7	2
3	2	10	3
5	9	3	8
7	6	9	7
8	5	4	10
9	3	5	6
1	7	1	1

Table 2: Matrix with randomly permuted columns

(a) $b \rightarrow c$

(b) $a \rightarrow d$

(c) $b \rightarrow d$

(d) $e \rightarrow c$

(e) $c \rightarrow a$

5. Using the contingency tables computed above, rank the rules $b \rightarrow c$, $a \rightarrow d$, $e \rightarrow c$ into decreasing order using the following measures

   (a) Confidence

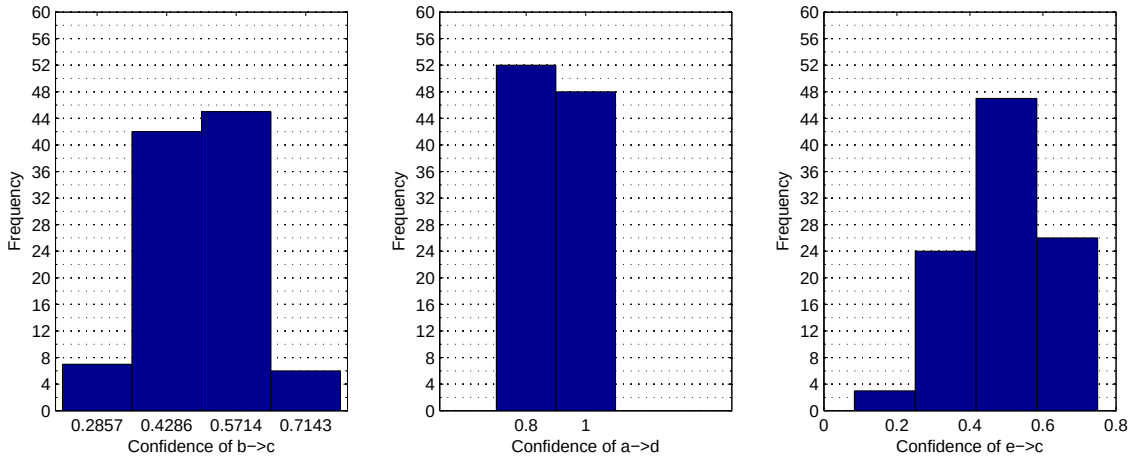
   (b) Interest factor

   (c) IS

6. Simulate a few steps of the 'simple' randomization method for assessing the statistical significace of the confidence of the association rule $b \rightarrow c$. Use the values in the matrix of Table 2 as the source of random numbers. Record the confidence for $b \rightarrow c$ after each step.

7. In table below the distribution of confidence values of the association rules $b \rightarrow c$, $a \rightarrow d$, $e \rightarrow c$ computed from 100 randomly permuted datasets are shown.

   Use these distributions to determine the statistical significance (p-value) for the confidence values for $b \rightarrow c$, $a \rightarrow d$, $e \rightarrow c$ obtained from the original data. Rank the rules based on the p-value.



8. Simulate a few iterations of the swap randomization algorithm on the data in table 1. Use the following sequence of $(row, column)$ indices as the random source: (8,2), (4,5), (7,2), (1,5), (2,4), (8,1), (4,2), (3,4), (9,5), (2,2).