CS 363D/ SSC 358 Statistical Learning and Data Mining Spring 2014

Homework 5

Lecturer: Prof. Pradeep Ravikumar Date Due: Apr. 28, 2014

Keywords: *Association Rules*

1. (a) (1 point) Can a superset of an *infrequent* itemset be *frequent*? Why or why not?

   (b) (1 point) Let x denote an item that occurs in *every* transaction of a dataset. What can you say about the confidence of a rule of the form $\{y\} \rightarrow \{x\}$, where y is some item in the dataset that appears in at least one transaction?

   (c) (2 points) Let Y denote a frequent itemset. We are interested in generating rules from Y that have a confidence of at least c. Let $X \rightarrow Y - X$ be a rule that *does not* satisfy the confidence threshold. Let $X' \subseteq X$. Can $X' \rightarrow Y - X'$ satisfy the confidence threshold? Give reasons.

2. (6 points)

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: Transaction Data set

Recall the *Apriori* algorithm that uses a generate-and-count strategy for deriving frequent itemsets. Candidate itemsets of size $k + 1$ are created from frequent itemsets of size k (candidate generation step). A candidate is discarded if any one of its subsets is found to be infrequent (candidate pruning step). Finally, if the support of any candidate is less than the minimum support threshold, mark it as infrequent; otherwise mark it as frequent (support counting step). Suppose the *Apriori* algorithm is applied to the dataset of transactions shown in Table 1 with minimum support threshold 30%, i.e. any itemset occurring in less than 3 transactions is considered to be infrequent.

(a) Draw an itemset lattice representing the dataset given in Table 1. Label each node in the lattice with the following letter(s).

- **N**: If the itemset is not considered to be a candidate itemset by the algorithm (Note that an itemset is not considered either if it is not generated at all or it is generated but subsequently removed because one of its subsets is found to be infrequent).
- **F**: If the candidate itemset is found to be frequent by the algorithm.

- **I**: If the candidate itemset is found to be infrequent after support counting.

(b) What is the percentage of frequent itemsets (with respect to all itemsets in the lattice)?

(c) What is the percentage of candidate itemsets that are found to be infrequent after support counting?