Assignment 3
CIS 453/553 Data Mining, Spring 2018
due 11:59 pm, Monday, May 7th

1. Explain why the following algorithm is more efficient than the method for generating association rules from frequent itemsets in section 6.2.2.

```
Algorithm:Rule_Generator. Given a set of frequent itemsets, output all of its strong rules.
Input:
       ls, set of frequent itemsets;
       min_conf, the minimum confidence threshold.
Output: Strong rules of itemsets in ls.
Method:
   1) for each frequent itemset l of ls
   2) rule_generator_helper(l, l, min_conf);

  procedure rule_generator_helper
       (s: current subset of l; l: original frequent itemset; min_conf)
   (1) k = length(s);
   (2) if (k>1) then {
   (3) Generate all the (k-1)-subsets of s;
   (4) for each (k-1)-subset x of s
   (5)    if (support_count(l)/support_count(x) >= min_conf) then {
   (6)        output the rule "x=>(l-x)";
   (7)        rule_generator_helper(x, l, min_conf);
   (8)   }
   (9)   //else do nothing
   (10)}
```

2. A database has five transactions. Let min_sup = 60% and min_conf = 75%.

TID	items_sold
T100	A, B, C, D, E, F
T200	B, D, S, C, F, T
T300	A, U, O, F, W, H
T400	D, A, E, C, F, G
T500	X, A, C, O, E, F

(a) Find all frequent itemsets using the ideas of Apriori and FP-growth, respectively. List the results of each step manually. Compare the efficiency of the two mining processes.

(b)List all of the strong association rules (with support s and confidence c) matching the following metarule, where X is a variable representing customers, and $item_i$ denotes variables representing items (e.g, A, B, C):

$$buys(X, item_1) \wedge buys(X, item_2) \Rightarrow buys(X, item_3)[s, c]$$

(c) Suppose A is "Apple Juice" and O is "Orange Juice". If we consider both of them as "Juice" (J), can we get more rules than (b)? If yes, list them.

3. Are Max patterns also Closed patterns? Are Closed patterns also Max patterns? Prove your conclusions.

4. Why avg(X)<=v is a convertible constraint? Why mining with convertible constraints (e.g., avg(X)<=25) is efficient?

5. Please report interesting association rules you can find from FoodMart.xls: `http://www.cs.uoregon.edu/classes/18S/cis453/data/FoodMart.xls`. It contains purchase information on 100 grocery items for 2127 shopping orders. "1" means the item was purchased, "0" means the item was not purchased, and the rows represent different shopping orders or "transactions." You need to write an Apriori or FP-growth program to mine the data if you are from the CIS major. Or you can download Weka from `http://www.cs.waikato.ac.nz/ml/weka/` and learn how to use it to mine the real life data.

6. (Extra 20% credits): Can you find any rare or negative patterns from FoodMart.xls? If yes, please explain your method/algorithm and justify your results.

To turn in by paper version: Ask Cheri to put your answers to Prof. Dejing Dou's mailbox.

To turn in by emails: We prefer that you send in a pdf file. If you are using Word, you should be able to convert your word file to a pdf file.