CSCI 4502/5502: Data Mining
Homework 3

Due at **12:30pm on Thursday, Feb 12, 2015**. Submit one file electronically at moodle: "**LastName_FirstName_Homework3.pdf**". Make sure to include your name, student id, and the Honor Code Pledge (`http://honorcode.colorado.edu/student-information/honor-code-pledge`).

1. The following contingency table summarizes the survey data of a student population, where ski refers to students who ski, $\overline{ski}$ refers to students who do not ski, $football$ refers to students who play football, and $\overline{football}$ refers to students who do not play football.

	$football$	$\overline{football}$	$\sum_{row}$
ski	1500	1000	2500
$\overline{ski}$	500	1000	1500
$\sum_{col}$	2000	2000	4000

   (a) Based on the given data, determine the correlation relationship between ski and playing football using the $lift$ measure.

   (b) Suppose that the association rule "$ski \Rightarrow football$" is mined. Given a minimum support threshold of 25% and a minimum confidence threshold of 50%, is this association rule strong (i.e., meet the thresholds)?

2. Given a data set with five transactions, each containing five items, as shown in the table. Let $min\_support = 60\%$.

TID	$items\_bought$
T1	{E, G, S, F, Z}
T2	{B, E, D, I, N}
T3	{B, E, I, N, O}
T4	{B, G, I, N, Z}
T5	{B, G, N, T, Z}

   (a) What is the maximum number of possible frequent itemsets?

   (b) Find all frequent itemsets using the Apriori algorithm. Your answer should include the key steps of the computation process.

   (c) In the computation above, how many rounds of database scan are needed? What is the total number of candidates?

(d) Let n be the total number of transactions, b be the number of items in each transaction, m be the number of k-itemset candidates. Consider the following two different approaches for counting the support values of the candidates. For each transaction, the first approach checks if a candidate occurred in the transaction or not; the second approach enumerates all the possible k-itemsets of the transaction and checks if the itemset is one of the candidates. What is the computation complexity for each approach? Is one always better than the other?

3. Given a data set with four transactions. Let $min\_support = 60\%$, and $min\_confidence = 80\%$.

cust_ID	*TID*	*items_bought* (in the form of *brand – item_category*)
01	T100	{Sunny-Cherry, Dairyland-Milk, Wonder-Bread, Sweet-Pie}
02	T200	{Best-Cheese, Dairyland-Milk, Goldenfarm-Cherry, Sweet-Pie, Wonder-Bread}
01	T300	{King's-Cereal, Sunset-Milk, Dairyland-Cheese, Best-Bread}
03	T400	{Wonder-Bread, Sunset-Milk, Best-Cereal, Sweet-Pie, Dairyland-Cheese}

(a) At the granularity of *item_category* (e.g., $item_i$ could be "$Milk$" and ignore brand name), for the following rule template,

$$\forall X \in \textbf{transaction}, \quad buys(X, item_1) \wedge buys(X, item_2) \Rightarrow buys(X, item_3) \quad [s, c]$$

list the frequent k-itemset for the largest k, and all of the strong association rules (with their support s and confidence c) containing the frequent k-itemset for the largest k.

(b) (**Required for CSCI 5502 students, optional and 5-point extra credit for CSCI 4502 students**) At the granularity of $brand - item\_category$ (e.g., $item_i$ could be "$Sunset - Milk$"), for the following rule template,

$$\forall X \in \textbf{customer}, \quad buys(X, item_1) \wedge buys(X, item_2) \Rightarrow buys(X, item_3)$$

list the frequent k-itemset for the largest k (but do not print any rules).