COSI 126A: Homework 3

Due by Nov. 19th

Section I: Association Problems (50 points)

Problem 1 (10 points)

Transaction ID	Items Bought
1	{Milk, Beer, Diapers}
2	{Bread, Butter, Milk}
3	{Milk, Diapers, Cookies}
4	{Bread, Butter, Cookies}
5	{Beer, Cookies, Diapers}
6	{Milk, Diapers, Bread, Butter}
7	{Bread, Butter, Diapers}
8	{Beer, Diapers}
9	{Milk, Diapers, Bread, Butter}
10	{Beer, Cookies}

Consider the market basket transactions shown above.

(a) What is the maximum number of association rules that can be extracted from this data (including rules that have zero support)?

(b) What is the maximum size of frequent itemsets that can be extracted (assuming $minsup > 0$)?

(c) Write an expression for the maximum number of size-3 itemsets that can be derived from this data set.

(d) Find an itemset (of size 2 or larger) that has the largest support.

(e) Find a pair of items, a and b, such that the rules $\{a\} \longrightarrow \{b\}$ and $\{b\} \longrightarrow \{a\}$ have the same confidence.

Problem 2 (10 points)

Consider the following set of frequent 3-itemsets:

$$\{1,2,3\}, \{1,2,4\}, \{1,2,5\}, \{1,3,4\}, \{1,3,5\}, \{2,3,4\}, \{2,3,5\}, \{3,4,5\}.$$

Assume that there are only five items in the data set.

(a) List all candidate 4-itemsets obtained by a candidate generation procedure using the $F_{k-1} \times F_1$ merging strategy.

(b) List all candidate 4-itemsets obtained by the candidate generation procedure in *Apriori*.

(c) List all candidate 4-itemsets that survive the candidate pruning step of the *Apriori* algorithm.

Problem 3 (10 points)

The original association rule mining formulation uses the support and confidence measures to prune uninteresting rules.

(a) Draw a contingency table for each of the following rules using the transactions shown in the table below.

Transaction ID	Items Bought
1	$\{a, b, c, e\}$
2	$\{b, c, d\}$
3	$\{a, b, d, e\}$
4	$\{a, c, d, e\}$
5	$\{b, c, d, e\}$
6	$\{b, d, e\}$
7	$\{d, e\}$
8	$\{a, b, c\}$
9	$\{a, d, e\}$
10	$\{b, d\}$

Rules: $\{b\} \longrightarrow \{c\}$, $\{a\} \longrightarrow \{d\}$, $\{b\} \longrightarrow \{d\}$, $\{e\} \longrightarrow \{c\}$, $\{c\} \longrightarrow \{a\}$.

(b) Use the contingency tables in part (a) to compute and rank the rules in decreasing order according to the following measures.

(a) Support.

(b) Confidence.

(c) Interest$(X \longrightarrow Y) = \frac{P(X,Y)}{P(X)}P(Y)$.

(d) IS$(X \longrightarrow Y) = \frac{P(X,Y)}{\sqrt{P(X)P(Y)}}$.

(e) Klosgen$(X \longrightarrow Y) = \sqrt{P(X,Y)} \times (P(Y \mid X) - P(Y))$, where $P(Y \mid X) = \frac{P(X,Y)}{P(X)}$.

(f) Odds ratio$(X \longrightarrow Y) = \frac{P(X,Y)P(\overline{X},\overline{Y})}{P(X,\overline{Y})P(\overline{X},Y)}$.

Problem 4 (10 points)

Given the rankings you had obtained in Exercise 12, compute the correlation between the rankings of confidence and the other five measures. Which measure is most highly correlated with confidence? Which measure is least correlated with confidence?

Problem 5 (10 points)

Suppose we have market basket data consisting of 100 transactions and 20 items. If the support for item a is 22%, the support for item b is 91% and the support for itemset $\{a, b\}$ is 17%. Let the support and confidence thresholds be 10% and 60%, respectively.

(a) Compute the confidence of the association rule $\{a\} \longrightarrow \{b\}$. Is the rule interesting according to the confidence measure?

(b) Compute the interest measure for the association pattern $\{a, b\}$. Describe the nature of the relationship between item a and item b in terms of the interest measure.

(c) What conclusions can you draw from the results of parts (a) and (b)?

Section II: Programming (50 points)

Provided is a transformation of the Online Retail dataset. The transformed dataset contains 541,909 transactions and 2603 items. The meaning of each item is given in the file *OnlineRetailAttributes.xlsx*.

You must find the 100 item pairs with the most support, as well as the confidence of the implied association rules.

Please consult *HW3_Association_Rules.ipynb* for more information. Implement your assignment in this file and include it with your submission. As usual, the problem set should be submitted as a PDF, with LaTeX strongly encouraged.