CSE597 Fall 2017 Assignment 3

Assigned: Friday, Oct 20, 2017

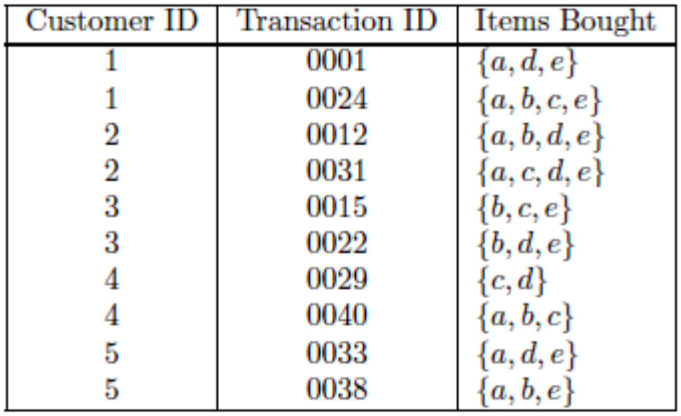
Due: Monday, October 30, 2017 (Typed hardcopy at the beginning of class)

Maximum: 100 point

Note: This assignment is to be done by an individual student, no team work allowed.

1. (20%) Consider the data set shown in Table 1.

Table 1



Customer ID	Transaction ID	Items Bought
1	0001	$\{a, d, e\}$
1	0024	$\{a, b, c, e\}$
2	0012	$\{a, b, d, e\}$
2	0031	$\{a, c, d, e\}$
3	0015	$\{b, c, e\}$
3	0022	$\{b, d, e\}$
4	0029	$\{c, d\}$
4	0040	$\{a, b, c\}$
5	0033	$\{a, d, e\}$
5	0038	$\{a, b, e\}$

   a. Compute the support for itemsets {a} , {a, b} , and {a, c, e} by treating each transaction ID as a market basket.

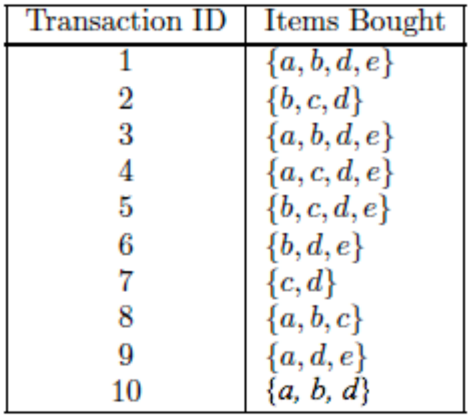
   b. Use the results in part (a) to compute the confidence for the association rules {a, c} → {e} and {e} → {a, c} . Is confidence a symmetric measure?

   c. Repeat part (a) by treating each customer ID as a market basket. Each item should be treated as a binary variable (1 if an item appears in at least one transaction bought by the customer, and 0 otherwise.)

2. (20%) The Apriori algorithm uses a generate-and-count strategy for deriving frequent itemsets. Candidate itemsets of size k +1 are created by joining a pair of

frequent itemsets of size k (this is known as the *candidate generation* step). A candidate is discarded if any one of its subsets is found to be infrequent during the candidate pruning step. Suppose the Apriori algorithm is applied to the data set shown in Table 2 with minsup = 40%, i.e., any itemset occurring in less than 4 transactions is considered to be infrequent.

Table 2



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	$\{a, b, d\}$

a. Draw an itemset lattice representing the data set given in Table 2. 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 Apriori algorithm. There are two reasons for an itemset not to be considered as a candidate itemset: (1) it is not generated at all during the candidate generation step, or (2) it is generated during the candidate generation step but is subsequently removed during the candidate pruning step because one of its subsets is found to be infrequent.
    - F : If the candidate itemset is found to be frequent by the Apriori algorithm.
    - I : If the candidate itemset is found to be infrequent after support counting.

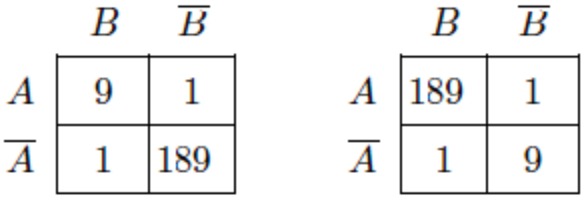
b. What is the percentage of frequent itemsets (with respect to all itemsets in the lattice)?

c. What is the pruning ratio of the Apriori algorithm on this data set? (Pruning ratio is defined as the percentage of itemsets not considered to be a candidate because (1) they are not generated during candidate generation or (2) they are pruned during the candidate pruning step.)

d. What is the false alarm rate (i.e, percentage of candidate itemsets that are found to be infrequent after performing support counting)?

3. (20%) Consider the contingency tables shown in Table 3

Table 3



	B	$\overline{B}$
A	9	1
$\overline{A}$	1	189

(a) Table I.

	B	$\overline{B}$
A	189	1
$\overline{A}$	1	9

(b) Table II.

a. For table I, compute support, the interest measure, and the φ correlation coefficient for the association pattern {A, B} . Also, compute the confidence of rules A → B and B → A.

b. For table II, compute support, the interest measure, and the φ correlation coefficient for the association pattern { A, B} . Also, compute the confidence of rules A → B and B → A .

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

4. (20%) For each of the sequence w $= <e_1, \ldots, e_{last}>$ below, determine whether they are subsequences of the following data sequence:

<{A,B}{C,D}{A,B}{C,D}{A,B}{C,D}>

subjected to the following timing constraints:

mingap = 0 (interval between last event in e_i and first event in e_{i+1} is $>$ 0)
maxgap = 2 (interval between first event in e_i and last event in e_{i+1} is $\leq$ 2)
maxspan = 6 (interval between first event in e_1 and last event in e_{last} is $\leq$ 6)
ws = 1 (time between first and last events in e_i is $\leq$ 1)

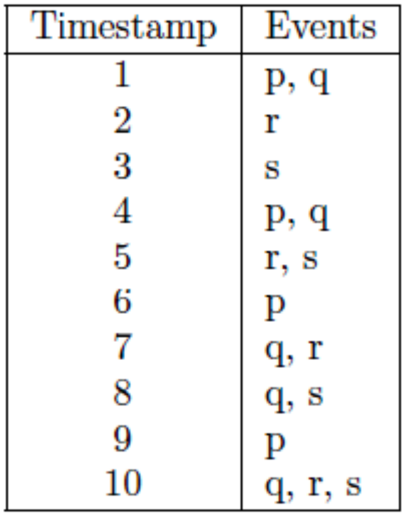
a. w = <{A}{B,C}{ A}>

b. w = < {A}{B}{C}>

5. (20%) Consider the data sequence shown in Table 4 for a given object. Count the number of occurrences for the sequence {p}{s}{r} according to the following counting methods:

Assume that ws = 0, mingap = 0, maxgap = 3, maxspan = 5

Table 4



Timestamp	Events
1	p, q
2	r
3	s
4	p, q
5	r, s
6	p
7	q, r
8	q, s
9	p
10	q, r, s

a. CMINWIN (number of minimal windows of occurrence).

b. CDIST_O (distinct occurrences with possibility of event-timestamp overlap).