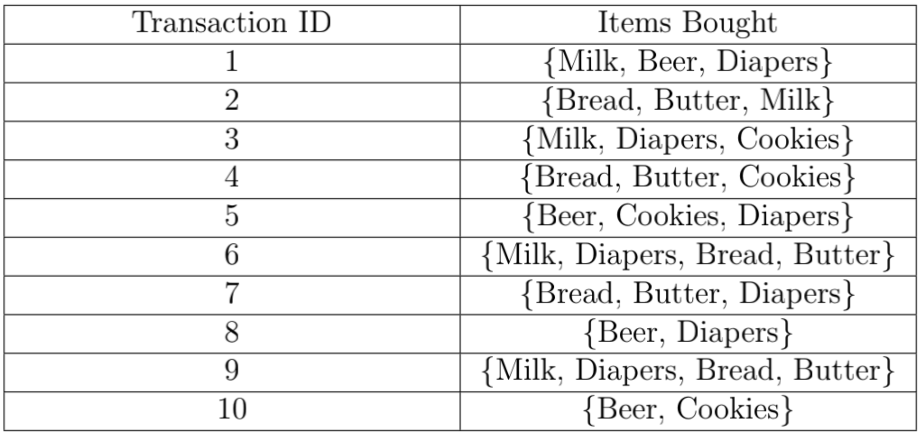
COSI 126A HOMEWORK 3

Jie Tang

APRIL 13, 2020

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)?
Answer: Since there are 6 unique items in the transactions. They are {Milk, Beer, Diapers, Bread, Butter, Cookies}. So the maximum number of association rules are 3^6 – 2^7 +1 = 601. Here we exclude the empty itemset.

(b) What is the maximum size of frequent itemsets that can be extracted (assuming minsup > 0)?
Answer: From the transactions shown above, the maximum length transaction is only 4, so the maximum size of frequent itemsets that can be extracted is also 4.

(c) Write an expression for the maximum number of size-3 itemsets that can be derived from this data set.
Answer: Since there are 6 unique items in the dataset, and we want to choose 3 out of them. The expression would be $\binom{6}{3}$ = 20

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

Itemset	Support
Beer	4
Bread	5
Butter	5
Cookies	4
Milk	5
Diaper	7

Two Itemsets:

Itemset	Support
{Bread, Beer}	0
{Butter, Beer}	0
{Milk, Beer}	1
{Milk, Cookies}	1
{Bread, Cookies}	1
{Butter, Cookies}	1
{Diapers, Cookies}	2
{Milk, Butter}	2
{Beer, Cookies}	2
{Diapers, Bread}	3
{Diapers, Beer}	3
{Diapers, Butter}	3
{Milk, Bread}	3
{Diapers, Milk}	4
{Bread, Butter}	5

Since the supersets of a itemset will have less or equal support compare to itself. As a result, the largest support should be S{Bread, Butter} = 5

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

Answer: Since conf(A->B) = sup({A,B})/sup({A}), conf(B->A) = sup({A,B})/sup({B}).

conf(Bread → Butter) = 5/5 =1

conf(Bread → Butter) =5/5 =1

As a result, {Bread, Butter} can be a pair of itemset.

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 Fk−1 × F1 merging strategy.

Answer:

One Itemset:

Itemset	Support
1	5
2	5
3	6
4	4
5	4

And then we can get:

{1, 2, 3} → {1, 2, 3, 4} & {1, 2, 3, 5}
{1, 2, 4} → {1, 2, 4, 5 }
{1, 3, 4} → {1, 3, 4, 5}
{2, 3, 4} → {2, 3, 4, 5}

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

Answer:

{1, 2, 3, 4}, {1, 2, 3, 5}, {1, 2, 4, 5}, {2, 3, 4, 5}, {1, 3, 4, 5}

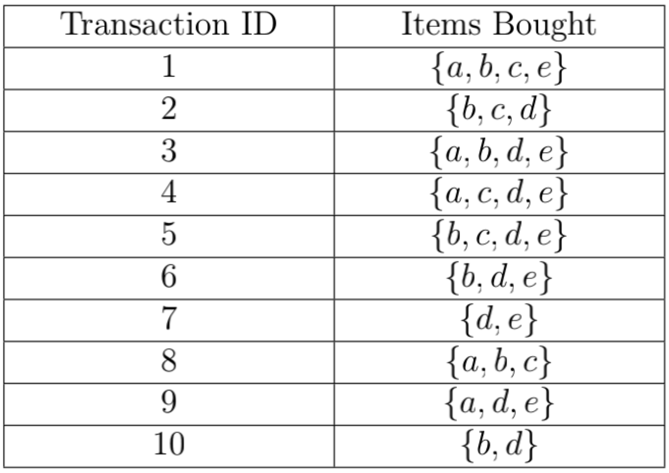
(c) List all candidate 4-itemsets that survive the candidate pruning step of the Apriori algorithm.

Answer: Only {1, 2, 3, 4} survived, because all {1, 2, 3}, {1, 2, 4}, {1, 3, 4}, {2, 3, 4} are frequent 3-itemsets.

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} → {c}, {a} → {d}, {b} → {d}, {e} → {c}, {c} → {a}.

Answer:

contingency table for {b} →{c}			
	c	c*	
b	4	3	7
b*	1	2	3
	5	5	10

contingency table for {a} →{d}			
	d	d*	
a	3	2	5
a*	5	0	5
	8	2	10

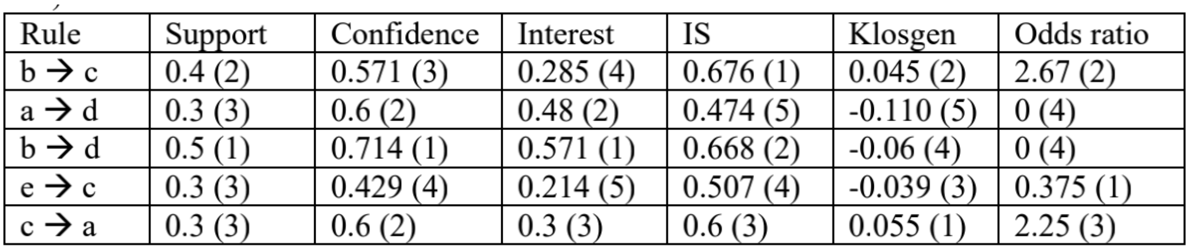
contingency table for {b} →{d}			
	d	d*	
b	5	2	7
b*	3	0	3
	8	2	10

contingency table for {e} →{c}			
	c	c*	
e	3	4	7
e*	2	1	3
	5	5	10

contingency table for {c} → {a}			
	a	a*	
c	3	2	5
c*	2	3	5
	5	5	10

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

Answer:



Rule	Support	Confidence	Interest	IS	Klosgen	Odds ratio
b → c	0.4 (2)	0.571 (3)	0.285 (4)	0.676 (1)	0.045 (2)	2.67 (2)
a → d	0.3 (3)	0.6 (2)	0.48 (2)	0.474 (5)	-0.110 (5)	0 (4)
b → d	0.5 (1)	0.714 (1)	0.571 (1)	0.668 (2)	-0.06 (4)	0 (4)
e → c	0.3 (3)	0.429 (4)	0.214 (5)	0.507 (4)	-0.039 (3)	0.375 (1)
c → a	0.3 (3)	0.6 (2)	0.3 (3)	0.6 (3)	0.055 (1)	2.25 (3)

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?

Answer:

Correlation between the rankings of confidence and others					
	Support	Interest	IS	Klosgen	Odds Ratio
Confidence	0.54	0.97	0.14	-0.28	-0.94

From the results above, we can see that Interest has the highest positive correlation with confidence, while Odds Ratio has the highest negative correlation with confidence. Meanwhile, IS has the least correlation 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} → {b}. Is the rule interesting according to the confidence measure?
Answer: From the above information, we can compute that conf({a} → {b}) = 0.17/0.22 = 77%. The rule is interesting because it is more than the confidence threshold.

(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.

Answer: We can know that the formula of calculating ***Interest*** = $\frac{P(X,Y)}{P(x)P(Y)}$
As a result, *Interest({a, b}) = 0.17/(0.22*0.91) =84.9%. So they are negatively correlated.*

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

Answer: From the results of parts (a) and parts (b), I can see that high confidence rule may not be interesting. Sometimes, we need to find other measurements.