APRIL 14, 2020

COSI126A _HW3

CHUYUE WU

INSTRUCTOR: DR. HONGFU LIU

SCHOOL: BRANDEIS UNIVERSITY

Course Name: Introduction to Data Mining

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)?

There are 6 unique items: Milk, Beer, Diapers, Bread, Butter, and Cookies.

R = $3^d - 2^{d+1} + 1$ = $3^6 - 2^{6+1} + 1$ = 602

The maximum number of association rules that can be extracted from this data (including rules that have zero support) is 602.

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

{Milk, Diapers, Bread, Butter} has the maximum size with 4 items, and its support count is 2 so it is frequent.

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

$$\binom{6}{3} = \frac{6*5*4}{3*2*1} = 20$$

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

Transaction ID	Milk	Beer	Diapers	Bread	Butter	Cookies
1	√	√	√			
2	√			√	√	
3	√		√			√
4				√	√	√
5		√	√			√
6	√		√	√	√	
7			√	√	√	
8		√	√			
9	√		√	√	√	
10		√				√

From the table above, I found the itemset that has the largest support is {Bread, Butter}, which has support count of 5 and support of 0.5.

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

$$\text{Confidence}(\{\text{Bread}\} \dashrightarrow \{\text{Butter}\}) = \frac{support\text{count}(\text{Bread, Butter})}{support\text{count}(\text{Bread})} = \frac{5}{5} = 1$$

$$\text{Confidence}(\{\text{Butter}\} \dashrightarrow \{\text{Bread}\}) = \frac{support\text{count}(\text{Bread, Butter})}{support\text{count}(\text{Butter})} = \frac{5}{5} = 1$$

So we can see if {a} and {b} has same support count, then the rules {a} −→ {b} and {b} −→ {a} would have the same confidence. Such pattern will also be applied to {Milk} −→ {Diapers} and { Diapers } −→ { Milk }, {Milk} −→ {Bread} and { Bread } −→ { Milk }, and so on.

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.

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}, Frequent 1-itemsets: {1}, {2}, {3}, {4}, {5}

Candidate Generation: {1,2,3,4}, {1,2,3,5}, {1,2,4,5}, {1,3,4,5}, {2,3,4,5}

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

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

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

For {1,2,3,4}, {1,2,3}, {1,2,4}, {1,3,4}, {2,3,4} are all frequent, survive
For {1,2,3,5}, {1,2,3}, {1,2,5}, {1,3,5}, {2,3,5} are all frequent, survive
For {1,2,4,5}, {1,2,4}, {1,2,5}, {3,4,5} are frequent, but {1,4,5} are not, remove
For {1,3,4,5}, {1,3,4}, {1,3,5}, {3,4,5} are frequent, but {1,4,5} are not, remove
For {2,3,4,5}, {2,3,4}, {2,3,5}, {3,4,5} are frequent, but {2,4,5} are not, remove
Above all, only {1,2,3,4} and {1,2,3,5} survive.

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}.

{b} −→ {c}:

	c	$\bar{c}$	
b	4	3	7
$\bar{b}$	1	2	3
	5	5	10

{a} −→ {d}:

	d	$\bar{d}$	
a	3	2	5
$\bar{a}$	5	0	5
	8	2	10

{b} −→ {d}:

	d	$\bar{d}$	
b	5	2	7
$\bar{b}$	3	0	3
	8	2	10

{e} −→ {c}:

	c	$\bar{c}$	
e	3	4	7

$\bar{e}$	2	1	3
	5	5	10

{c} −→ {a}:

	a	$\bar{a}$	
c	3	2	5
$\bar{c}$	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.

a. Support.

	Support	Rank
{b} −→ {c}	0.40	2
{a} −→ {d}	0.30	3
{b} −→ {d}	0.50	1
{e} −→ {c}	0.30	3
{c} −→ {a}	0.30	3

(b) Confidence.

	Confidence	Rank
{b} −→ {c}	0.57	4
{a} −→ {d}	0.60	2
{b} −→ {d}	0.71	1
{e} −→ {c}	0.43	5
{c} −→ {a}	0.60	2

(c) Interest (X −→ Y) = $\frac{P(X,Y)}{P(X)P(Y)}$

	Interest	Rank
{b} −→ {c}	1.14	2

{a} −→ {d}	0.75	5
{b} −→ {d}	0.89	3
{e} −→ {c}	0.86	4
{c} −→ {a}	1.20	1

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

	IS	Rank
{b} −→ {c}	0.68	1
{a} −→ {d}	0.47	5
{b} −→ {d}	0.67	2
{e} −→ {c}	0.51	4
{c} −→ {a}	0.60	3

(e) $Klosgen(X \rightarrow Y) = \sqrt{P(X, Y)} * (P(Y|X) - P(Y))$, where $P(Y|X) = \frac{P(X,Y)}{P(X)}$

	Klosgen	Rank
{b} −→ {c}	0.04	2
{a} −→ {d}	-0.11	5
{b} −→ {d}	-0.06	4
{e} −→ {c}	-0.04	3
{c} −→ {a}	0.05	1

(f) Odds ratio$(X \rightarrow Y) = \frac{P(X,Y)P(\bar{X},\bar{Y})}{P(X,\bar{Y})P(\bar{X},Y)}$

	Odds Ratio	Rank
{b} −→ {c}	2.67	1
{a} −→ {d}	0.00	4
{b} −→ {d}	0.00	4
{e} −→ {c}	0.38	3
{c} −→ {a}	2.25	2

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?

np.corrcoef(confidence, support) = 0.408

np.corrcoef(confidence, interest) = 0.096

np.corrcoef(confidence, is) = 0

np.corrcoef(confidence, klosgen) = -0.289

np.corrcoef(confidence, oddsratio) = -0.490

Odds Ratio is most highly correlated with confidence, and IS 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} −→ {b}. Is the rule interesting according to the confidence measure?

$$\text{Confidence}(\{a\} \rightarrow \{b\}) = \frac{support(\text{a, b})}{support(\text{a})} = \frac{17\%}{22\%} = 77.3\%$$

The rule is interesting because it exceeds 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.

$$\text{Interest}(\{a, b\}) = \frac{\text{P(a,b)}}{\text{P(a)P(b)}} = \frac{17\%}{22\% * 91\%} = 0.849$$

In terms of the interest measure, because P(a, b) < P(a) x P(b), item a and item b are negatively correlated.

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

If the confidence of a rule is high, it doesn’t mean the items needs to be interesting.