DATA MINING

COURSE: B.Sc.(H)-VI Semester

TEACHER: MS. SONAL LINDA

Solved Examples and Exercises

Chapter 6. Association Analysis

Solved Examples

1. Consider the dataset shown in Table 1.

Table 1: Example of market basket transaction

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 $\{e\}$, $\{b, d\}$ and $\{b, d, e\}$ by treating each transaction ID as a market basket.

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

(c) Repeat part(a) by treating each CustomerID 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).

(d) Use the results in part(c) to compute the confidence for the association rules $\{b, d\} \rightarrow \{e\}$ and $\{e\} \rightarrow \{b, d\}$.

Solution:

- Support determines how often a rule is applicable to a given dataset.

$$\text{Support, } s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N}$$

- Confidence determines how frequently items in Y appear in transactions that contain X.

$$\text{Confidence, } c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

Where, N is total number of transactions and $\sigma(X)$ is the support count of X (support count refers to the number of transactions that contain a particular itemset X).

(a) The support for itemsets:

$$s(\{e\}) = \frac{8}{10} = 0.8$$

$$s(\{b,d\}) = \frac{2}{10} = 0.2$$

$$s(\{b,d,e\}) = \frac{2}{10} = 0.2$$

(b) The confidence for the association rules:

$$c(\{b,d\} \rightarrow \{e\}) = \frac{\sigma(\{b,d,e\})}{\sigma(\{b,d\})} = \frac{2}{2} = 1$$

$$c(\{e\} \rightarrow \{b,d\}) = \frac{\sigma(\{b,d,e\})}{\sigma(\{e\})} = \frac{2}{8} = 0.25$$

(c) Repeat part(a) by treating each CustomerID as a market basket

Customer ID	Items Bought
1	$\{a,d,e\}, \{a,b,c,e\}$
2	$\{a,b,d,e\}, \{a,c,d,e\}$
3	$\{b,c,e\}, \{b,d,e\}$
4	$\{c,d\}, \{a,b,c\}$
5	$\{a,d,e\}, \{a,b,e\}$

The support for itemsets:

$$s(\{e\}) = \frac{4}{5} = 0.8$$

$$s(\{b,d\}) = \frac{5}{5} = 1$$

$$s(\{b,d,e\}) = \frac{4}{5} = 0.8$$

(d) The results in part(c) is used to compute the confidence for the association rules $\{b, d\} \rightarrow \{e\}$ and $\{e\} \rightarrow \{b, d\}$

$$c(\{b, d\} \rightarrow \{e\}) = \frac{\sigma(\{b, d, e\})}{\sigma(\{b, d\})} = \frac{4}{5} = 0.8$$

$$c(\{e\} \rightarrow \{b, d\}) = \frac{\sigma(\{b, d, e\})}{\sigma(\{e\})} = \frac{4}{4} = 1$$

2. 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 dataset.

(a) List all candidate 4-itemsets obtained by a candidate generation procedure using $F_{k-1} \times F_1$ 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 pruning step of the Apriori algorithm.

Solution:

(a) All candidate 4-itemsets obtained by a candidate generation procedure using $F_{k-1} \times F_1$ 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:

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

Merging frequent 3-itemsets and frequent 1-itemsets to obtain frequent 4-itemsets:

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

(b) All candidate 4-itemsets obtained by the candidate generation procedure using Apriori.

Frequent 3-itemsets:	Frequent 4-itemsets:

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

Itemset	Count
$\{1,2,3,4\}$	4
$\{1,2,3,5\}$	3
$\{1,2,4,5\}$	1
$\{1,3,4,5\}$	2
$\{2,3,4,5\}$	1

(c) Let minimum support count be 3, all candidate 4-itemsets that survive the pruning step of the Apriori algorithm:
$\{1,2,3,5\}, \{1,2,3,4\}$

Exercises

1. Consider the market basket transactions shown in Table 2.

Table 2: Market Basket Transactions.

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}

(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 dataset.

(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 $\{a\} \rightarrow \{b\}$ and $\{b\} \rightarrow \{a\}$ have the same confidence.

2. 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 dataset shown in Table 3 with $minsup = 30\%$, i.e., any itemset occurring in less than 3 transactions is considered to be infrequent.

Table 3: Market Basket Transaction

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	$\{b, d\}$

(a) Draw an itemset lattice representing the data set given in Table 3. 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.

(a) What is the confidence for the rules $\emptyset \longrightarrow A$ and $A \longrightarrow \emptyset$?

(b) Let c_1, c_2, and c_3 be the confidence values of the rules $\{p\} \longrightarrow \{q\}$, $\{p\} \longrightarrow \{q, r\}$, and $\{p, r\} \longrightarrow \{q\}$, respectively. If we assume that c_1, c_2, and c_3 have different values, what are the possible relationships that may exist among c_1, c_2, and c_3? Which rule has the lowest confidence?

(c) Repeat the analysis in part (b) assuming that the rules have identical support. Which rule has the highest confidence?

(d) Transitivity: Suppose the confidence of the rules $A \longrightarrow B$ and $B \longrightarrow C$ are larger than some threshold, $minconf$. Is it possible that $A \longrightarrow C$ has a confidence less than $minconf$?

4.

Suppose we have market basket data consisting of 100 transactions and 20 items. If the support for item a is 25%, the support for item b is 90% and the support for itemset $\{a, b\}$ is 20%. Let the support and confidence thresholds be 10% and 60%, respectively.

(a) Compute the confidence of the association rule $\{a\} \rightarrow \{b\}$. 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)?

5.

For each of the following questions, provide an example of an association rule from the market basket domain that satisfies the following conditions. Also, describe whether such rules are subjectively interesting.

(a) A rule that has high support and high confidence.

(b) A rule that has reasonably high support but low confidence.

(c) A rule that has low support and low confidence.

(d) A rule that has low support and high confidence.