REI502M - Introduction to Data Mining

Solutions to homework 4

Elías Snorrason   October 8, 2019

Problem 5.1

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.

Part	Rule	Interesting
A	Cereal $\rightarrow$ Milk	Nothing new to learn from this rule, as these items are far too common. Not interestin
B	Milk $\rightarrow$ Cereal	Same support as above, but milk has other uses. Not subjectively interesting.
C	Batteries $\rightarrow$ Toothpaste	Not as frequent. No common use cases. Not subjectively interesting.
D	Strawberries $\rightarrow$ Chocolate	Occur relatively frequently together. More interesting.

Problem 5.2

Consider the training examples shown in the following table for a binary classification problem.

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.

10 distinct baskets/transactions.

- $\{e\}$: $s = \frac{8}{10} = 0.8$
- $\{b, d\}$: $s = \frac{2}{10} = 0.2$
- $\{b, d, e\}$: $s = \frac{2}{10} = 0.2$

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

Both rules have support 0.2, (support count is 2):

- $\{b, d\} \to \{e\}$: $c = \frac{0.2}{0.2} = 1$
- $\{e\} \to \{b, d\}$: $c = \frac{0.2}{0.8} = 0.25$

Support is a symmetric measure, but **confidence is not symmetric**!

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).

Now we have 5 baskets in total.

- $\{e\}$: $s = \frac{4}{5} = 0.8$
- $\{b, d\}$: $s = \frac{5}{5} = 1$
- $\{b, d, e\}$: $s = \frac{4}{5} = 0.8$

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

- $\{b, d\} \to \{e\}$: $c = \frac{0.8}{1} = 0.8$
- $\{e\} \to \{b, d\}$: $c = \frac{0.8}{0.8} = 1$

E. Suppose s_1 and c_1 are the support and confidence values of an association rule r when treating each transaction ID as a market basket. Also, let s_2 and c_2 be the support and confidence values of r when treating each customer ID as a market basket. Discuss whether there are any relationships between s_1 and s_2 or c_1 and c_2.

Although support for $\{e\}$ remained the same, nothing can be said about support for $\{b, d\}$ and $\{b, d, e\}$ (except that it increased significantly by using Customer ID). The increase in support is not reflected in the changes in confidence of the rules. This means that in general, <u>**no clear difference**</u> **in treating transaction IDs or customer IDs as market baskets**.

Problem 5.3

A. What is the confidence for the rules $\emptyset \to \{A\}$ and $\{A\} \to \emptyset$?

Confidence of $X \to Y$, where $X \cap Y = \emptyset$, can be written as: $c(X \to Y) = \frac{s(X \cup Y)}{s(X)}$.

- $c(\emptyset \to A) = \frac{s(\emptyset \cup A)}{s(\emptyset)} = s(A)$
- $c(A \to \emptyset) = \frac{s(\emptyset \cup A)}{s(A)} = 1$

The former rule has the same support and confidence, while the latter always has confidence at unity.

B. Let c_1, c_2, and c_3 be the confidence values of the rules $\{p\} \to \{q\}$, $\{p\} \to \{q, r\}$, and $\{p, r\} \to \{q\}$, 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?

Denoting the support of a union of two sets, we will omit the use of $\cup$ (for convenience).

$$c_1 = \frac{s(pq)}{s(p)}$$
$$c_2 = \frac{s(pqr)}{s(p)}$$
$$c_3 = \frac{s(pqr)}{s(pr)}$$

Since $s(pq) \geq s(pqr)$, we can say that $c_1 \geq c_2$ by looking at the denominators. Similarily, since $s(p) \geq s(pr)$, we can say that $c_3 \geq c_2$. Thus, the rule $\{p\} \to \{q, r\}$ has the lowest confidence (c_2).

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

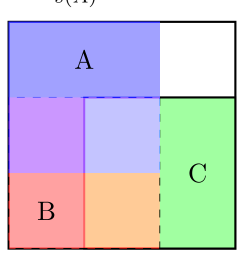
In this case: $s(pq) = s(pqr)$, which leads to $c_1 = c_2$. As we still have $s(p) \geq s(pr)$, we can say that $c_3 \geq c_1$ and $c_3 \geq c_2$.

D. Transitivity: Suppose the confidence of the rules $\{A\} \to \{B\}$ and $\{B\} \to \{C\}$ are larger than some threshold, $minconf$. Is it possible that $\{A\} \to \{C\}$ has a confidence less than $minconf$?

Denote the confidence of rules $\{A\} \to \{B\}$, $\{B\} \to \{C\}$ and $\{A\} \to \{C\}$ as c_1, c_2 and c_3 respectively.
We have that:

$$c_1 = \frac{s(AB)}{s(A)} > minconf$$
$$c_2 = \frac{s(BC)}{s(B)} > minconf$$
$$c_3 = \frac{s(AC)}{s(A)}$$

The lower limit of $s(AC)$ is not well defined, but one can find examples where $c_3 < minconf$. If $s(A) \gtrsim s(B) \geq s(C)$. we can potentially have $s(AC) < s(AB)$ such that $\frac{s(AC)}{s(A)} < minconf$.



All transactions

In this example $s(A) = s(B) = s(C) = \frac{4}{9}$, $c_1 = c_2 = \frac{1}{3}$, yet $c_3 = \frac{1}{4}$. We could easily have set $\frac{1}{4} < minconf < \frac{1}{3}$.

Problem 5.6

Consider the market basket transactions shown in the following table.

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)?

The total number of possible rules, R, extracted from a data set that contains d items is:

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

There are $d = 6$ items in the table(Beer, Bread, Butter, Cookies, Diapers and Milk). Thus:

$$R = 3^6 - 2^7 + 1 = 602$$

602 association rules can be extracted from this data.

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

With $minsup > 0$, we only need to look for the larget itemset in the data set. Itemsets corresponding to ID 6 and 9 have the **maximum size of 4** in the data set.

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

Disregarding the support threshold, there are $\frac{6!}{3!}$ possible 3-itemsets (with duplicates). The number of distinct 3-itemsets is therefore:

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

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

Itemset	Support
cookies | milk	1
bread | cookies	1
milk	5
beer | cookies	2
beer | diapers	3
bread | butter | milk	3
bread | butter | cookies	1
beer | milk	1
butter | cookies	1
butter | milk	3
butter	5
bread | butter | diapers | milk	2
bread | butter	**5**
bread	5
butter | diapers | milk	2
bread | diapers	3
cookies	4
beer	4
butter | diapers	3
diapers	7
diapers | milk	4
beer | cookies | diapers	1
beer | diapers | milk	1
bread | diapers | milk	2
bread | butter | diapers	3
bread | milk	3
cookies | diapers | milk	1
cookies | diapers	2
$\emptyset$	10

Table 1: All itemsets with non-zero support count

Ignoring the 1-itemsets (and $\emptyset$), the itemset with the largest support is $\{\textbf{bread}, \textbf{butter}\}$.

E. Find a pair of items, a and b, such that the rules $\{a\} \to \{b\}$ and $\{b\} \to \{a\}$ have the same confidence.

Bread and butter have the same support ($s = 5$). This means that the rules $\{\textbf{bread}\} \to \{\textbf{butter}\}$ and $\{\textbf{butter}\} \to \{\textbf{bread}\}$ have the same confidence ($c = \frac{5}{5} = 1$). The same can be said with beer and cookies ($s = 4$, $c = \frac{2}{4} = 0.5$).

Problem 5.8

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.

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

Other combinations were duplicates or not extendible from 3-itemsets to 4-itemsets.

B. List all candidate 4-itemsets obtained by the candidate generation procedure in Apriori.

From the frequent 3-itemsets, we can assume that $minsup = 4$.
All 4-itemsets from the previous part were generated from frequent 3-itemsets, so we get the same candidates as before:
$\{1,2,3,4\}$, $\{1,2,3,5\}$, $\{1,2,4,5\}$, $\{1,3,4,5\}$, $\{2,3,4,5\}$.

C. List all candidate 4-itemsets that survive the candidate pruning step of the Apriori algorithm.

$\{1,2,3,4\}$ survives as all of it's subsets ($\{1,2,3\}$, $\{1,2,4\}$, $\{1,3,4\}$, $\{2,3,4\}$) are frequent.
$\{1,2,3,5\}$ survives as all of it's subsets ($\{1,2,3\}$, $\{1,2,5\}$, $\{1,3,5\}$, $\{2,3,5\}$) are frequent.
Other 4-itemsets are pruned.