CMSC5724: Exercise List 11

Consider the mining of association rules on the transactions:

transaction id	items
1	A, B, E
2	A, B, D, E
3	B, C, D, E
4	B, D, E
5	A, B, D
6	B, E
7	A, E

Problem 1. What is the support of the itemset $\{B, D, E\}$?

Answer.
The support count is 3 because transactions 2, 3 and 4 contain the itemset.

Problem 2. What is the support and confidence of the association rule $BD \rightarrow E$?

Answer.
The support $BD \rightarrow E$ is the support of $\{B, D, E\}$ which is 3. The confidence is

$$conf(BD \rightarrow E) = \frac{support(\{B, D, E\})}{support(\{B, D\})} = \frac{3}{4}.$$

Problem 3. Consider the application of the Apriori algorithm to find all the frequent itemsets whose counts are at least 3. Recall that the algorithm scans the transaction list a number of times, where the i-th scan generates the set F_i of all size-i frequent itemsets from a candidate set C_i. Show C_i and F_i for each possible i.

Answer.
For the first scan, the candidate set C_1 contains all the singleton sets, i.e., C_1 includes $\{A\}$, $\{B\}$, $\{C\}$, $\{D\}$ and $\{E\}$. After the scan, only $\{A\}$, $\{B\}$, $\{D\}$ and $\{E\}$ remain in F_1. In particular, $\{C\}$ is eliminated because its count 1 is smaller than 3.

From F_1, the algorithm generates:

$$C_2 = \{\{A, B\}, \{A, D\}, \{A, E\}, \{B, D\}, \{B, E\}, \{D, E\}\}$$

The second scan produces:

$$F_2 = \{\{A, B\}, \{A, E\}, \{B, D\}, \{B, E\}, \{D, E\}\}$$

$\{A, D\}$ is removed because its count 2 is lower than 3.

From F_2, the algorithm generates:

$$C_3 = \{\{A, B, E\}, \{B, D, E\}\}$$

as follows. For each pair of distinct itemsets $\{a_1, a_2\}$ and $\{b_1, b_2\}$ in F_2, the algorithm adds to C_3 an itemset $\{a_1, a_2, b_2\}$ if and only if $a_1 = b_1$. Hence, $\{A, B\}$ and $\{A, E\}$ give rise to $\{A, B, E\}$, whereas $\{B, D\}$ and $\{B, E\}$ give rise to $\{B, D, E\}$.

Finally, the third scan produces:

$$F_3 = \{\{B, D, E\}\}$$

as you can verify easily by yourself. The algorithm terminates here.

Problem 4. Find all the association rules with support at least 3 and confidence at least 3/4. For your convenience, all the itemsets with support at least 3 are $\{\{A\}, \{B\}, \{D\}, \{E\}, \{A, B\}, \{A, E\}, \{B, D\}, \{B, E\}, \{D, E\}, \{B, D, E\}\}$.

Answer.
The following table lists all the possible association rules and their confidence values. The ones in bold are the final answers.

rule	confidence
$\boldsymbol{A \to B}$	3/4
$B \to A$	1/2
$\boldsymbol{A \to E}$	3/4
$E \to A$	1/2
$B \to D$	2/3
$\boldsymbol{D \to B}$	1
$\boldsymbol{B \to E}$	5/6
$\boldsymbol{E \to B}$	5/6
$\boldsymbol{D \to E}$	3/4
$E \to D$	1/2
$B \to DE$	1/2
$\boldsymbol{BD \to E}$	3/4
$BE \to D$	3/5
$\boldsymbol{D \to BE}$	3/4
$\boldsymbol{DE \to B}$	1
$E \to BD$	1/2

Problem 5. If the universe U (the set of all possible items) has size n, prove:

- the maximum number of distinct association rules is $\sum_{a=1}^{n-1} \sum_{b=1}^{n-a} \binom{n}{a}\binom{n-a}{b}$.
- $\sum_{a=1}^{n-1} \sum_{b=1}^{n-a} \binom{n}{a}\binom{n-a}{b} = \sum_{\ell=2}^{n} \binom{n}{\ell}(2^\ell - 2)$.

Answer. An association rule has the form $I_1 \to I_2$ where I_1 and I_2 are disjoint non-empty subsets of U. Subject to the constraint $|I_1| = a$ and $|I_2| = b$ where a and b are integers satisfying $a \geq 1$, $b \geq 1$, and $a + b \leq n$, we have $\binom{n}{a}$ ways to choose I_1 and then $\binom{n-a}{b}$ ways to choose I_2. Therefore, the total number of possible rules is

$$\sum_{a=1}^{n-1} \sum_{b=1}^{n-a} \binom{n}{a}\binom{n-a}{b}.$$

To prove the second bullet, let us analyze the maximum number of rules in a different way. For each $\ell \in [2, n]$, there are $\binom{n}{\ell}$ itemsets I of size ℓ. Given such an I, there are $2^\ell - 2$ subsets $I_1 \subseteq I$ satisfying $1 \leq |I_1| \leq \ell - 1$. Every such I_1 defines an association rule $I_1 \to I_2$ where $I_2 = I \setminus I_1$. No two association rules thus obtained are the same. Therefore, the total number of possible rules is

$$\sum_{\ell=2}^{n} \binom{n}{\ell}(2^\ell - 2).$$