CS5228: Knowledge Discovery and Data Mining

Tutorial 4 — Association Rule Mining

Question 1

$supp(\{A\} \to \{B\}) = supp(\{A, B\})$

$supp(\{B\} \to \{A\}) =$

tid	transactions
1	B, C, D, E, F
2	E, C
3	D, B, D, A
4	G, E, H, C
5	H, A, G, D, B
6	B, E, G
7	B, A, D
8	A, D, B, C

(a) **Calculate the following values:**

- $support(\{A\})$, $support(\{B\})$ $support(\{A, B\})$
- $support(\{A\} \to \{B\})$, $support(\{B\} \to \{A\})$
- $confidence(\{A\} \to \{B\})$, $confidence(\{B\} \to \{A\})$
- $lift(\{A\} \to \{B\})$, $lift(\{B\} \to \{A\})$

support({A}) = $1/2$

support({A}→{B}) = $1/2$

confidence({A}→{B}) = $1 = 4/4$

lift({A}→{B}) = $4/3$

support({B}) = $3/4$

support({B}→{A}) = $1/2$

confidence({B}→{A}) = $2/3$

lift({B}→{A}) = $4/3$

support({A, B}) = $1/2$

$$\frac{supp(\{A,B\})}{supp(\{B\})} = \frac{1/2}{3/4} = \frac{2}{3}$$

Question 1

tid	transactions
1	B, C, D, E, F
2	E, C
3	D, B, D, A
4	G, E, H, C
5	H, A, G, D, B
6	B, E, G
7	B, A, D
8	A, D, B, C

(a) **Calculate the following values:**

- $support(\{A\}), support(\{B\})\ support(\{A, B\})$
- $support(\{A\} \rightarrow \{B\}), support(\{B\} \rightarrow \{A\})$
- $confidence(\{A\} \rightarrow \{B\}), confidence(\{B\} \rightarrow \{A\})$
- $lift(\{A\} \rightarrow \{B\}), lift(\{B\} \rightarrow \{A\})$

support({A}) = **1/2**

support({B}) = **3/4**

support({A, B}) = **1/2**

support({A}→{B}) = **1/2**

support({B}→{A}) = **1/2**

confidence({A}→{B}) = **1.0**

confidence({B}→{A}) = **2/3**

lift({A}→{B}) = **4/3**

lift({B}→{A}) = **4/3**

Question 1

(b) **Connections between metrics.** In a) we calculated the *support*, *confidence*, and *lift* for different itemsets and association rules. However, we do not need to calculate all values individually. Based on the definitions of the metrics, which calculations can we skip?

Question 1

(b) **Connections between metrics.** In a) we calculated the *support*, *confidence*, and *lift* for different itemsets and association rules. However, we do not need to calculate all values individually. Based on the definitions of the metrics, which calculations can we skip?

Solution

- support(X→Y) = support(X∪Y)
- support(X→Y) = support(Y→X)

No need to calculate support({A}→{B}) and support({B}→{A}) if we already have support({A, B}) (or vice versa)

- lift(X→Y) = lift(Y→X)

Question 1

(c) **"Usefulness" of different metrics.** What makes *support* and *confidence* more useful compared to other metrics such as *lift*, *conviction*, *collective strength*, *leverage*?

Question 1

(c) **"Usefulness" of different metrics.** What makes *support* and *confidence* more useful compared to other metrics such as *lift*, *conviction*, *collective strength*, *leverage*?

Solution

- Only support and confidence are anti-monotone which facilitate the Apriori algorithm(s)
- lift, conviction, collective strength, leverage etc. are all very useful metric
  but can only be calculated for the association rules after running the Apriori algorithm

Question 1

(d) **"Importance" of different metrics.** What makes an association rule particularly interesting? A high *support*, high *confidence*, high *lift*, high *conviction*, etc.?

Question 1

(d) **"Importance" of different metrics.** What makes an association rule particularly interesting? A high *support*, high *confidence*, high *lift*, high *conviction*, etc.?

Solution

- No metric is intrinsically better in describing what makes a rule interesting
- Different metrics look at different aspects of a rule
- Which aspect is more/most relevant depends on the data and the task

Question 2

tid	transactions
1	cough, fatigue, COVID-19-negative
2	anosmia, cough, fatigue, COVID-19-positive
3	anosmia, fatigue, headache, heart palpitations, COVID-19-positive
4	cough, fatigue, headache, COVID-19-negative
5	headache, stomach pain, COVID-19-negative
6	cough, heart palpitations, COVID-19-negative
7	anosmia, headache, stomach pain, COVID-19-positive
...	...

(a) **Choice of *minsup* and *minconf*.** How might the setup and the task above affect which values for *minsup* and *minconf* are meaningful? Hint: Assume that large majority of the COVID-19 test results in our dataset are negative.

$\{\dots\} \rightarrow \{cov+\}$ should be low

~~$\{cough\} \rightarrow \{fever\}$~~

Question 2

tid	transactions
1	cough, fatigue, COVID-19-negative
2	anosmia, cough, fatigue, COVID-19-positive
3	anosmia, fatigue, headache, heart palpitations, COVID-19-positive
4	cough, fatigue, headache, COVID-19-negative
5	headache, stomach pain, COVID-19-negative
6	cough, heart palpitations, COVID-19-negative
7	anosmia, headache, stomach pain, COVID-19-positive
...	...

(a) **Choice of *minsup* and *minconf*.** How might the setup and the task above affect which values for *minsup* and *minconf* are meaningful? Hint: Assume that large majority of the COVID-19 test results in our dataset are negative.

Solution

- Any rule of the form X→{COVID-19-positive} won't have high support
- We cannot afford to set *minsup* to high, or we won't get relevant rules

Question 2

tid	transactions
1	~~cough, fatigue, COVID-19-negative~~
2	anosmia, cough, fatigue, COVID-19-positive
3	anosmia, fatigue, headache, heart palpitations, COVID-19-positive
4	~~cough, fatigue, headache, COVID-19-negative~~
5	~~headache, stomach pain, COVID-19-negative~~
6	~~cough, heart palpitations, COVID-19-negative~~
7	anosmia, headache, stomach pain, COVID-19-positive
...	...

(b) **Tweaking the dataset.** Can we simplify this task by only considering those transactions that contain COVID-19-positive, and remove all transactions that contain COVID-19-negative?

Question 2

tid	transactions
1	cough, fatigue, COVID-19-negative
2	anosmia, cough, fatigue, COVID-19-positive
3	anosmia, fatigue, headache, heart palpitations, COVID-19-positive
4	cough, fatigue, headache, COVID-19-negative
5	headache, stomach pain, COVID-19-negative
6	cough, heart palpitations, COVID-19-negative
7	anosmia, headache, stomach pain, COVID-19-positive
...	...

(b) **Tweaking the dataset.** Can we simplify this task by only considering those transactions that contain COVID-19-positive, and remove all transactions that contain COVID-19-negative?

Solution: Not a good idea

- Example: {cough}→{COVID-19-positive} might have a high support, but "cough" is not truly a good indicator if "cough" is also very common with negative test results
- Sidenote: confidence(X→{COVID-19-positive}) = 1.0 for all itemsets X

Question 3

3. **Complexity Analysis.** The most naive approach for mining association rules would be to generate all possible rules and check if their support and confidence exceeds the specified thresholds $minsup$ and $minconf$. In the lecture, you have learned that, given d unique items in a dataset of transactions, there are $3^d - 2^{d+1} + 1$ possible rules.

$\{\} \Rightarrow \{\}$ $\emptyset$

Proof that d unique items result in $3^d - 2^{d+1} + 1$ possible rules! (Hint: Write out all possible rules for $d = 2, 3, 4, \ldots$ items; you should quickly spot the pattern that will allow you to validate the formula).

$\underbrace{\{A, B, C, D\}}_{d=4}$

$A \to B$

$A \to C$

$A \to BC$

$C \to A$

$\vdots$

$AB \to \{\}$

$ABC \to \{\}$

AC

D

$ABC \to D$

$ACD \Rightarrow B$

$\vdots$

Question 3

Solution:

- Each item has 3 possibilities to appear in a rule: on the left side of the rule, on the right side of the rule, or not at all. That reflects the 3^d possibilities.
- However, these 3^d rules include invalid ones where the left and/or right side of the rule is empty. Of the 3^d rules, there are 2^d where the left side is empty, and 2^d where the right side is empty. We have to subtract these invalid combinations, and $2^d + 2^d = 2^{d+1}$.
- Note that we now have subtracted the rule $\{\} \rightarrow \{\}$ twice. So we need '+1' to correct for this.