Northeastern University, Khoury College of Computer Science

CS 6220 Data Mining — Assignment 2

Due: September 26, 2023(100 points)

Chun-Wei Tseng
https://github.com/Chun-Wei-Tseng/CS6220-Assignment-2
tseng.chun@northeastern.edu

Frequent Itemsets

Consider the following set of frequent 3-itemsets:

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

Assume that there are only five items in the data set. This question was taken from Tan et al., which may help in reviewing Candidate Generation.

1. List all candidate 4-itemsets obtained by a candidate generation procedure using the $F_{k-1} \times F_1$ merging strategy.

→ Ans:
→ F_1:

Itemset	Support
{1}	4
{2}	5
{3}	5
{4}	4
{5}	3

→ All candidate 4-itemsets: {1, 2, 3, 4}, {1, 2, 3, 5}, {1, 2, 4, 5}, {1, 3, 4, 5}, {2, 3, 4, 5}

2. List all candidate 4-itemsets obtained by the candidate generation procedure in A Priori, using $F_{k-1} \times F_{k-1}$.

→ Ans:
→ k = 4, looking for sets to merge where their first 2 items are identical
→ All candidate 4-itemsets: {1, 2, 3, 4}, {1, 2, 3, 5}, {1, 2, 4, 5}, {2, 3, 4, 5}

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

→ {1, 2, 3, 5} was pruned because {1, 3, 5} not frequent. {1, 2, 4, 5} was pruned because {2, 4, 5} not frequent. {2, 3, 4, 5} was pruned because {2, 4, 5} is not frequent.
→ All candidate 4-itemsets that survive: {1, 2, 3, 4} because {1, 3, 4} and all other subsets are also in the frequent itemsets.

Association Rules

Consider the following table for question 4:

Transaction ID	Items
1	{Beer, Diapers}
2	{Milk, Diapers, Bread, Butter}
3	{Milk, Diapers, Cookies}
4	{Bread, Butter, Cookies}
5	{Milk, Beer, Diapers, Eggs}
6	{Beer, Cookies, Diapers}
7	{Milk, Diapers, Bread, Butter}
8	{Bread, Butter, Diapers}
9	{Bread, Butter, Milk}
10	{Beer, Butter, Cookies}

4. a) What is the maximum number of association rules that can be extracted from this data (including rules that have zero support)?

   → Ans:

   - $R = 3^d - 2^{d+1} + 1$, where R is the total number of possible rules and d is the number of items in the dataset
   - Dataset = {Beer, Bread, Butter, Cookies, Diapers, Eggs, Milk} → d = 7
   - $R = 3^d - 2^{d+1} + 1 = 3^7 - 2^8 + 1 = 2187 - 256 + 1 = 1932$

   b) What is the confidence of the rule {Milk, Diapers} ⇒ {Butter}?

   → Ans:

   - Transactions that have {Milk, Diapers}: 2, 3, 5, 7
   - Transactions that have {Butter} when {Milk, Diapers} are also in the transaction: 2, 7
   - Confidence of {Milk, Diapers} ⇒ {Butter} = $\frac{2}{4} = \frac{1}{2}$

   c) What is the support for the rule {Milk, Diapers} ⇒ {Butter}?

   → Ans:

   - Transactions contain {Milk, Diapers, Butter}: 2, 7
   - Total number of transactions: 10
   - Support for {Milk, Diapers} ⇒ {Butter}: $\frac{2}{10} = \frac{1}{5}$

5. `True` or `False` with an explanation: Given that {a,b,c,d} is a frequent itemset, {a,b} is always a frequent itemset.

→ Ans: True. Based on Apriori Principle, if an itemset is frequent, then all of its subsets must also be frequent.

6. `True` or `False` with an explanation: Given that {a,b}, {b,c} and {a,c} are frequent itemsets, {a,b,c} is always frequent.

→ Ans: False. Reverse of the Apriori Principle is not necessarily true. Just because all smaller subsets of an itemset are frequent does not guarantee that the itemset will be frequent.

7. `True` or `False` with an explanation: Given that the support of {a,b} is 20 and the support of {b,c} is 30, the support of {b} is larger than 20 but smaller than 30.

→ Ans: False. Support of {a,b} is 20 means that {a, b} appears in 20 of the transactions. Support of {b,c} is 30 means that {b, c} appears in 30 of the transactions. Support of {b} is at least larger than 30 because {b} is in all the transactions that have {b, c} in them. But {b} can appear in many more transactions that do not have a or c. Therefore, the upper bound for the support of {b} can be much larger.

8. `True` or `False` with an explanation: In a dataset that has 5 items, the maximum number of size-2 frequent itemsets that can be extracted (assuming minsup > 0) is 20.

→ Ans: False. In dataset that has 5 iteams {a, b, c, d, e}, total number of different item set is 10, including: {a, b}, {a, c}, {a, d}, {a, e}, {b, c}, {b, d}, {b, e}, {c, d}, {c, e}, {d, e}. The maximum number of size-2 frequent itemsets cannot exceed this number.

9. Draw the itemset lattice for the set of unique items $I = \{a, b, c\}$.