Northeastern University, Khoury College of Computer Science

CS 6220 Data Mining — Assignment 2

Due: Jan 25, 2023(100 points)

Yu Wang
Git User Name: titojojo
Email: wang.yu25@northeastern.edu
https://github.com/Titojojo/CS6220-Data-Mining

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.

   ```
   {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 $F_{k-1} \times F_{k-1}$.

   ```
   {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, 4}
   ```

Association Rules

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

   There are {Beer, Diapers, Milk, Bread, Butter, Cookies, Eggs}.
   In total 7 items.
   So we have $2^7 - 1 = 127$ possible association rules.

   b) What is the confidence of the rule $\{Milk, Diapers\} \Rightarrow \{Butter\}$?

$$Confidence = \frac{\sigma(\{\text{Milk, Diapers, Butter}\})}{\sigma(\{\text{Milk, Diapers}\})} = \frac{2}{4} = 0.5$$

   c) What is the support for the rule $\{Milk, Diapers\} \Rightarrow \{Butter\}$?

$$Support = \frac{\sigma(\{\text{Milk, Diapers, Butter}\})}{|T|} = \frac{2}{10} = 0.2$$

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

   True.
   According to Apriori principle, if an itemset is frequent, then all of its subsets must also be frequent:

$$\forall X, Y : (X \subseteq Y) \Rightarrow s(Y) \leq s(X)$$

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.

   False.
   {a, b}, {b, c}, and {a, c} are frequent does not guarantee that their union {a, b, c} is also frequent.
   For example, suppose we have the following set: {a, b} with support 5, {b, c} with support 5, and {a, c} with support 6, {a, b, c} with support 1. If minsup is 2 in this case, then {a, b}, {b, c}, and {a, c} are frequent itemsets, but {a, b, c} is not 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.

   False.
   Based on the Anti-monotone property of support, the support of an itemset never exceeds that of its subsets:

$$\forall X, Y : (X \subseteq Y) \Rightarrow s(Y) \leq s(X)$$

   {b} is a subset of both {a, b} and {b, c}, so we have:

$$20 = s(a, b) \leq s(b)$$

$$30 = s(b, c) \leq s(b)$$

   So the support of b is larger or equal to 30.

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.

   False.

$$\binom{5}{2} = \frac{5 \times 4}{2 \times 1} = 10$$

   The maximum number of size-2 frequent itemsets should be 10.

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