Northeastern University, Khoury College of Computer Science

CS 6220 Data Mining | Assignment 2

Due: September 26, 2023(100 points)

So Man Amanda Au-Yeung
https://github.com/amandaay/CS6220DataMining/tree/main

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

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

   {1, 2, 3, 4}, {1, 2, 4, 5}, {1, 2, 3, 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 $F_{k-1} \times F_{k-1}$.

   Merge pairs of (k – 1) items only if their first k – 2 items are identical, where k = 4

   For the first 2 items to be identical, it results in {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.

Itemset	Support
{1, 2, 3}	1
{1, 2, 4}	1
{1, 2, 5}	1
{1, 3, 4}	1
{1, 3, 5}	0
{2, 3, 4}	1
{2, 3, 5}	1
{2, 4, 5}	0

From #2, we have {1, 2, 3, 4}, {1, 2, 4, 5}, {1, 2, 3, 5}, {1, 3, 4, 5}, {2, 3, 4, 5}.

Since {1, 3, 5}, {2, 4, 5} are pruned out, {1, 2, 3, 4} and {2, 3, 4, 5} are left.

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)?

   $R = 3^7 - 2^{7+1} + 1 = 1932$, where the 7 item sets are = {beer, diapers, milk, bread, butter, cookies, eggs}

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

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

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

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

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. Since {a,b,c,d} is a frequent itemset, {a, b} subset will still always be a frequent itemset.

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. According to Apriori principle, if an itemset is frequent, then all of its subsets must also be frequent. But not the other way around. Thus, the statement does not stand.

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. Since the support of {a, b} is 20, so {b} has at least 20 support. Similarly, {b, c} has a support of 30, meaning the subset of {b} has at least 30 support.

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.

   e.g. {1, 2, 3, 4, 5}, a 5-item dataset

   {1, 2}, {1, 3}, {1, 4}, {1, 5}, {2, 3}, {2, 4}, {2, 5}, {3, 4}, {3, 5}, {4, 5} with a maximum number of 10.

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