Northeastern University, Khoury College of Computer Science

CS 6220 Data Mining - Assignment 2

Due: January 25, 2024(100 points)

Name: Xinyue Han

Git Username: aiC0ld

Github repo link:
https://github.com/aiC0ld/CS6220-DataMining/tree/main

E-mail: han.xinyue@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} × F_1 merging strategy.

Answer:

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

Answer:

Merge pairs of k - 1 items only if their k - 2 items are identical, where k = 4. So the 4-itemsets obtained by the candidate generation procedure in A Priori are {1, 2, 3, 4}, {1, 2, 3, 5}, {1, 2, 4, 5}, {2, 3, 4, 5}

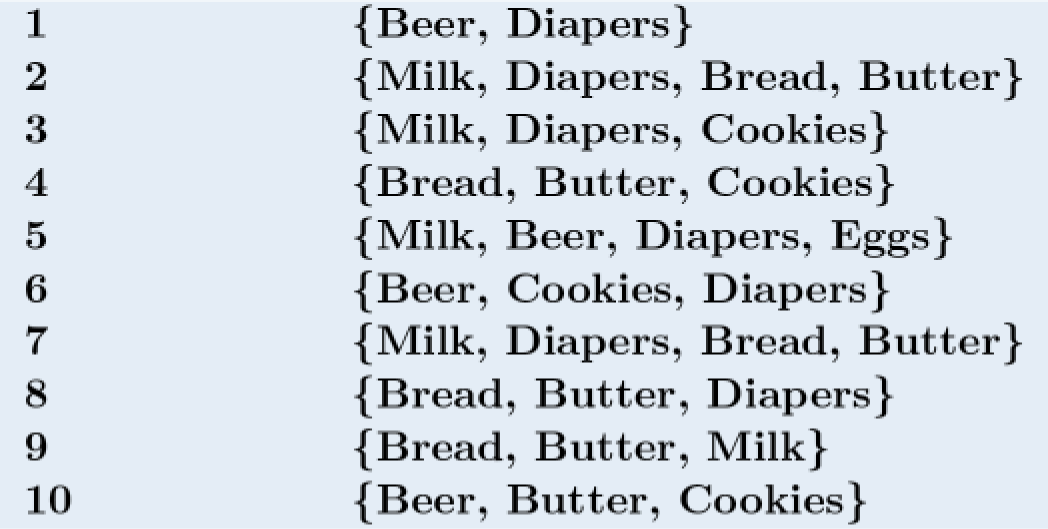
3. List all candidate 4-itemsets that survive the candidate pruning step of the Apriori algorithm.

Answer:

From Q2, {1, 3, 5} and {2, 4, 5} are pruned out. So 4-itemsets that survive the candidate pruning step of the Apriori algorithm: {1, 2, 3, 4}

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)?

Answer:

There are 7 items: beer, diapers, milk, bread, butter, cookies, eggs

$3^n - 2^{n+1} + 1 = 3^7 - 2^8 + 1 = 1932$

So the maximum number of association rules is 1932.

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

Answer:

Confidence = σ(Milk, Diaper, Butter) / σ(Milk, Diaper) = 2 / 4 = 0.5

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

Answer:

Support = σ(Milk, Diaper, Butter) / |T| = 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. If {a,b,c,d} is frequent, then any subset, including {a,b}, must also be frequent, as subsets cannot appear less frequently than their supersets.

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.

Fasle. Even if {a,b}, {b,c}, and {a,c} are all frequent, there is no guarantee that {a,b,c} is frequent. This is because the frequency of a larger itemset may be less than the frequency of any of its subsets. An itemset is frequent only if it meets a minimum support threshold, and the combined frequency of all three items may not meet this requirement.

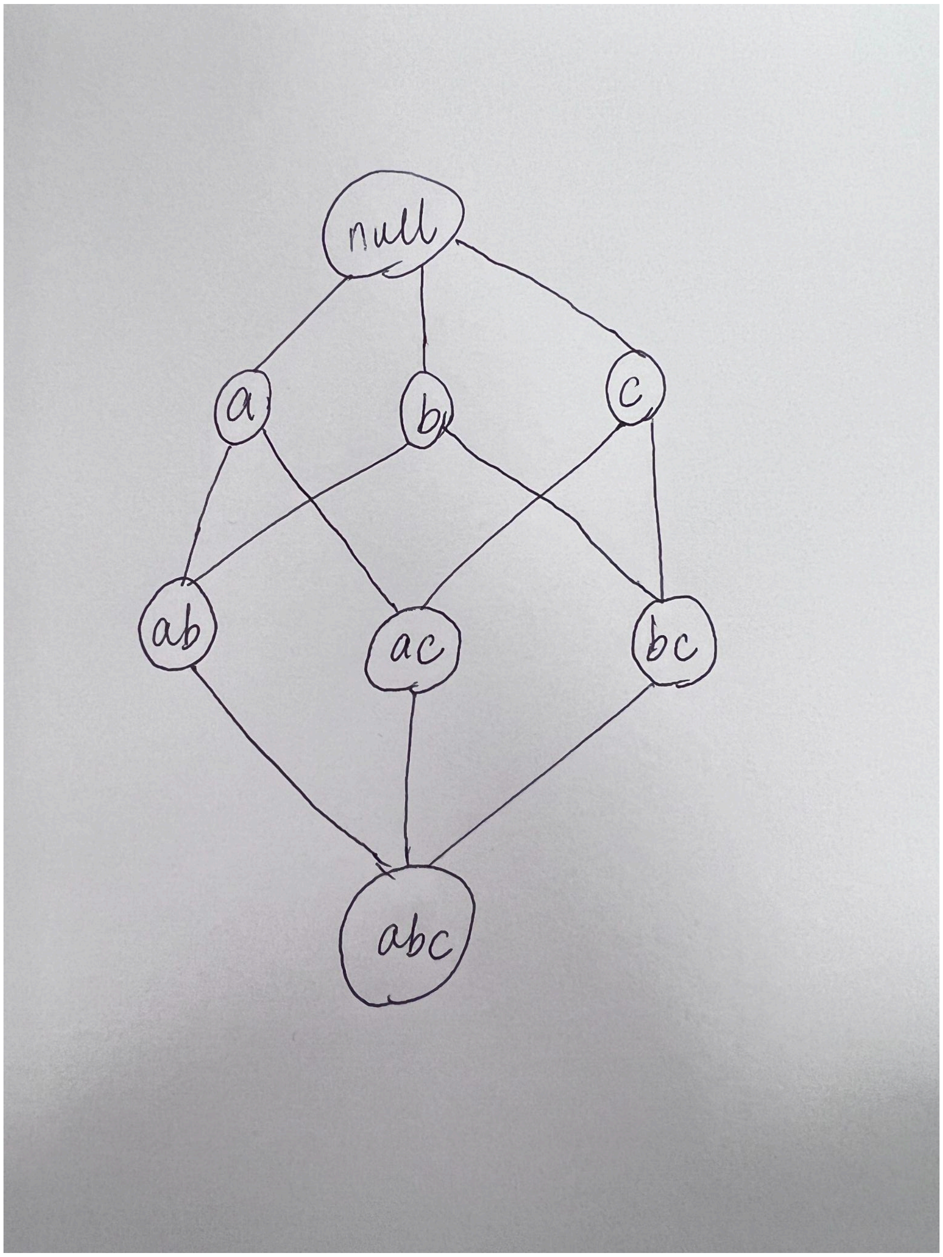
7. True or False with an explanation: Given that the support of {a,b} is 20 and the support of {b,c} is 30, the support of {b} is larger than 20 but smaller than 30.

False.The support of {b} may be greater than 20 and 30, or equal to one of these values. The support of {b} depends on the number of times item bbb appears in combination with other items or alone. Therefore, it is not guaranteed to be strictly between the supports of {a,b} and {b,c}.

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. The maximum number of size-2 itemsets for 5 items can be calculated as C(5, 2) = 10. The assumption of minsup > 0 ensures that all combinations are possible, but there can only be 10 size-2 itemsets.

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


null
a
b
c
ab
ac
bc
abc