Quiz 2

Online (In-class) Quiz –

1. (1.5 pt) Considering the following set {A,B,C,D}. How could you apply Toivonen's algorithm for finding frequent item sets? Give one example of a singleton and one example of a pair in the negative border. You need to explain why your examples are considered as itemsets in the negative border. Explain how and why Toivonen's algorithm uses the itemsets in the negative border (you can use your examples).

Answer –

Example – Suppose the items are {A,B,C} and {A} {C} are frequent itemsets

Negative Border -

I. Singleton – {B}

All the single items which are not in frequent itemsets are in negative border as their immediate subset which is ∅ or empty set is a frequent set (as long as there are at least as many baskets as the support threshold, which is true in our algorithm)

II. Pair – {A,C}

The immediate subsets of (A,C} are {A} , {C} and they are in frequent itemsets while {A,C} is not

Furthermore, {A,B} or {B,C}, cannot be in negative border as their immediate subset {B} is not in frequent itemsets,

How and Why the Toivonen's algorithm uses the itemsets in the negative border?

1. After we get the candidate frequent itemset, For handling false negatives, it creates a negative border which has all the items which are not frequent, but all its immediate subsets are frequent. In our example, {B} and {A,C} are in negative border while {A,B} {B,C} is not as {B} is not frequent.

2. The negative border leads to two cases:

a) If all the items in negative border are not frequent in the entire dataset.

b) If at least one of the items in negative border is a false negative i.e. it is frequent in the whole dataset.

The "case a" gives us the result. But "case b" will lead us to reiterate the entire process. Which means, even if we have one false negative, we will have to redo the process as that one false negative item could be an immediate subset of some other items which could also have been in the negative border.

In our example, had {B} been a false negative, the validity of {A,B} {B,C} should have been rechecked as {B} is a subset of them and it was wrongly identified and so, they had to be placed in the negative border. So, the entire process has to be redone.

2. Why high confidence association rules might not be interesting? give an example of a such a rule. (1.5 pt)

   Answer -

   - Not all high-confidence rules are interesting as common items which have high support can have many association rules which are not that useful

   - The rule X → milk may have high confidence for many itemsets X because milk is just purchased very often (independent of X)

3. (3 pts) Consider the PCY algorithm. Support the minimum support threshold = 3 and the hash function h(i,j)= (i+j)%5. Show the content of count table and the frequent-buckets table produced by PCY in the first pass. For the frequent-buckets table, show the actual counts in the bucket and also the bitmap generated from the table.

1,2,3
2,3
1,2,4
3,4
1,2,3,4

1	3
2	4
3	4
4	3

0	$(1,4)^2$, $(2,3)^3$	5	1
1	$(2,4)^1$	2	0
2	$(3,4)^2$	2	0
3	$(1,2)^3$	3	1
4	$(1,3)^2$	2	0

4. [0.5 point] All high-confidence rules are interesting? True or False?
   False
5. [1.5 points] When should you use the **tables of triples** approach instead of the **triangular matrix** and why?
   Answer -
   If fewer than ⅓ of possible pairs actually occur in the market basket data(pairs with count>0) then we use tables of triples as we just need to store the pairs with count>0 so, it can save space
6. [0.5 points] Finding frequent triplets is easier than frequent pairs as they have 3 items? True or False?
   True
7. [0.5 point] Confidence is an indication of how frequently the items appear in the baskets?

   False

8. [1 point] Circle ALL of the statements that are TRUE about finding frequent item sets:

   A. The true cost of mining disk-resident data is usually the number of disk I/O.

B. In the Triangular-Matrix Method, pair{i,j} is at position (i-1)(n-i/2)+j-i, here n = total number of items.

C. The Triples Method always saves more space than the Triangular-Matrix Method.

D. In the Triples Method, we only keep a table of triples when pairs' count > 0.

Offline (Take-home) Quiz –

9. [1 point] In the PCY algorithm, what should be the conditions for a pair {i, j} for being a candidate pair?
   Both i and j are frequent items [0.5 points]
   The pair {i, j} hashes to a bucket whose bit in the bit vector is 1 [0.5 points]

10. For the A Priori algorithm, consider the following input file of basket data, where each basket lists (i.e., { }) the items it contains. For a support threshold s = 3, answer the following questions.
    *Basket data: {a, b, c, d, e}{d, e, c} {a, b, c, f} {a, b, c, d}

    a) [1 pts] What are the item counts produced in pass 1 and which of these items are frequent?

    b) [2 pts] For pass 2, which are the candidate pairs for each basket? (Only include the pairs that will be counted.)

    c) [1 pts] What is the count for each candidate pair and which of the candidate pairs are frequent?

Item	Count	Frequent
a	3	Yes
b	3	Yes
c	4	Yes
d	3	Yes
e	2	No
f	1	No

Basket	Candidate pairs
1	(a,b), (a,c), (a,d), (b,c), (b,d), (c,d)
2	(c,d)
3	(a,b), (a,c), (b,c)
4	(a,b), (a,c), (a,d), (b,c), (b,d), (c,d)

Candidate pair	Count	Frequent
(a,b)	3	Yes
(a,c)	3	Yes
(a,d)	2	No
(b,c)	3	Yes
(b,d)	2	No
(c,d)	3	Yes

11. [2 points] For the following question, consider the entire set of items contains: A, B, C, D, E,..., J (a total of 10 items). In the A Priori algorithm, how much memory do you need if you use
    A. Triangular-Matrix method
    B. Triples Method to count the occurrence of each possible pair assuming only ¼ of the pairs (doublets) have an occurrence > 0?
    (you can assume that a counter uses 4 bytes)
    (Just write the number for the answer)

a. 10 * 9/2 * 4 = 180 bytes
b. 10 * 9/2 * ¼ = 11.25 = 11 pairs
   11*12 = 132 bytes

If you have assumed there are 11.25 pairs, it is wrong as 11.25 pairs has no physical meaning. I have considered the answer to be correct for both ceil(11.25) = 12 and floor(11.25) = 11.
It is always important to consider the physical meaning of the quantities in consideration and ponder whether these values actually make sense or not.

12. [3 points] Here is a collection of twelve baskets. Each contains three of the six items 1 through 6. {1, 2, 3 } {2, 3, 4 } {3, 4, 5 } {4, 5, 6 }{1, 3, 5 } {2, 4, 6 }

{1, 3, 4 } {2, 4, 5 }{3, 5, 6 } {1, 2, 4 } {2, 3, 5 } {3, 4, 6 } The support threshold is 4. The hash function is i x j mod 11. Using the PCY Algorithm, you need to show 1. frequent singles, 2. frequent buckets, and 3. frequent pairs.

Frequent Singles: 1, 2, 3, 4, 5, 6

Frequent Buckets - 1, 2, 4, 8

Frequent Pairs - (2,4), (3,4), (3,5)

Rubrics

Incorrect Frequent Singles(max 2): Subtract 0.25 Marks

Incorrect Frequent Bucket Numbers (max 2): Subtract 0.25 Marks

Frequent Pairs consist of only the candidate pairs: Subtract 0.25 Marks

Incorrect Frequent Pairs (max 1): Subtract 0.25 Marks