

Seoul National University

M1522.001400 Introduction to Data Mining

Spring 2016, Kang

Homework 5: Frequent Itemsets (Chapter 6)

Due: May 9, 09:30 AM

Reminders

- The points of this homework add up to 100.
- Like all homeworks, this has to be done individually.
- Lead T.A.: Jinhong Jung (montecast9@gmail.com)
- Please type your answers in English. Illegible handwriting may get no points, at the discretion of the graders.
- If you have a question about assignments, please upload your question in eTL.
- If you want to use slipdays or consider late submission with penalties, please note that you are allowed one week to submit your assignment after the due date.

Remember that:

- Whenever you are making an assumption, please state it clearly

Question 1

Suppose there are 20 items, numbered 1 to 20, and also 20 baskets, also numbered 1 to 20. Item i is in basket b if and only if i divides b with no remainder. Thus, item 1 is in all the baskets, item 2 is in all ten of the even-numbered baskets, and so on. Basket 12 consists of items $\{1, 2, 3, 4, 6, 12\}$, since these are all the integers that divide 12. Answer the following questions. [20 points]

(a) If the support threshold is 3 which items are frequent?

Answer: $\{1, 2, 3, 4, 5, 6\}$

(b) If the support threshold is 3, which pairs of items are frequent?

Answer: $(1,2), (1,3), (1,4), (1,5), (1,6), (2,3), (2,4), (2,6), (3,6)$

Question 2

For the data of Question 1, what is the confidence of the following association rules?
[25 points]

(a) $\{3, 5\} \rightarrow 2$.

Answer: 0

(b) $\{1, 2, 4\} \rightarrow 8$.

Answer: 0.4

(c) $\{2, 4, 5\} \rightarrow 5$.

Answer: 1

(d) $\{2, 3\} \rightarrow 6$.

Answer: 1

Question 3

Apply the A-Priori Algorithm with support threshold 2 to the data of Question 1. Answer the following questions. [25 points]

C_k is the set of candidate itemsets of size k – the itemsets that we must count in order to determine whether they are in fact frequent.

L_k is the set of truly frequent itemsets of size k .

(a) Find C_2 and L_2 .

Answer: $C_2 = \{ (x, y) | 1 \leq x < y \leq 10 \}$

$L_2 = \{ (1,2), (1,3), (1,4), (1,5), (1,6), (1,7), (1,8), (1,9), (1,10),$
 $(2,3), (2,4), (2,5), (2,6), (2,8), (2,10), (3,6), (3,9), (4,8), (5,10) \}$

(b) Find the max number of k where L_k is not an empty set.

Answer: $k = 4$

Question 4

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}

Suppose the support threshold is 4. On the first pass of the PCY Algorithm, we use a hash table with 11 buckets, and the set $\{i, j\}$ is hashed to bucket $i \times j \bmod 11$. Answer the following questions. [30 points]

(a) Compute the support for each item and each pair of items.

| Item | | Support | |
|------|--|---------|--|
| 1 | | 4 | |
| 2 | | 6 | |
| 3 | | 8 | |
| 4 | | 8 | |
| 5 | | 6 | |
| 6 | | 4 | |

| Pair | Support | Pair | Support |
|------|---------|------|---------|
| 1,2 | 2 | 2,6 | 1 |
| 1,3 | 3 | 3,4 | 4 |
| 1,4 | 2 | 3,5 | 4 |
| 1,5 | 1 | 3,6 | 2 |
| 1,6 | 0 | 4,5 | 3 |
| 2,3 | 3 | 4,6 | 3 |
| 2,4 | 4 | 5,6 | 2 |
| 2,5 | 2 | | |

(b) Which pairs hash to which buckets?

| Pair | Bucket Number | Pair | Bucket Number |
|------|---------------|------|---------------|
| 1,2 | 2 | 2,6 | 1 |
| 1,3 | 3 | 3,4 | 1 |
| 1,4 | 4 | 3,5 | 4 |
| 1,5 | 5 | 3,6 | 7 |
| 1,6 | 6 | 4,5 | 9 |
| 2,3 | 6 | 4,6 | 2 |
| 2,4 | 8 | 5,6 | 8 |
| 2,5 | 10 | | |

(c) Which buckets are frequent?

Answer: 1, 2, 4, 8

(d) Which pairs are counted on the second pass of the PCY Algorithm?

Answer: (2,6), (3,4), (1,2), (4,6), (1,4), (3,5), (2,4), (5,6)