Data Mining:
Exercise Session 5

(1) Association Rule Mining: Basics

What is an association rule?

Given a set of items $I = \{i_1, i_2, \ldots, i_m\}$ and defined an Itemset A as a collection of items, $A \subseteq I$, an association rule is an implication $A \implies B$, where:

- $A, B \subseteq I$;
- $A \cap B = \emptyset$.

The implication $A \implies B$ can be read as “if a basket contains the items in A, then it is likely to contain the items in B”.

(1) Association Rule Mining: Basics

Three important definitions:

- *Support* of an itemset A:
  - *Support Count*: number of transactions containing all items in the itemset A;
  - *Support Frequency*: frequency of transactions containing all items in A, i.e. $P(A) =$ Support Count / Tot Transactions;
- *Confidence* of a rule $A \implies B$: $P(B|A) = \frac{sup(\{A,B\})}{sup(\{A\})}$; it measures the accuracy of the rule;
- *Interest* of a rule $A \implies B$: $conf(B|A) - sup(B)$; it measures the influence of A on B;

(1.1) Confidence, Support and Interest

TID	Items
1	1 2 4 9
2	3 4 5 9 10
3	1 2 4 9
4	3 4 5 9 10
5	1 3 4 5
6	1 4 5 6
7	1 2 4 9
8	1 3 4 5 6 9 10
9	3 4 5 9 10
10	3 4 5 9 10

- What is the support of $\{5, 9\}$?

(1.1) Confidence, Support and Interest

TID	Items
1	1 2 4 9
2	3 4 5 9 10
3	1 2 4 9
4	3 4 5 9 10
5	1 3 4 5
6	1 4 5 6
7	1 2 4 9
8	1 3 4 5 6 9 10
9	3 4 5 9 10
10	3 4 5 9 10

- What is the support of $\{5, 9\}$?
  Count $= 5$, Frequency $= \frac{1}{2}$

(1.1) Confidence, Support and Interest

TID	Items
1	1 2 4 9
2	3 4 5 9 10
3	1 2 4 9
4	3 4 5 9 10
5	1 3 4 5
6	1 4 5 6
7	1 2 4 9
8	1 3 4 5 6 9 10
9	3 4 5 9 10
10	3 4 5 9 10

- What is the support of $\{5, 9\}$?
  Count $= 5$, Frequency $= \frac{1}{2}$
- What is the support of $\{1, 3, 4, 5\}$?

(1.1) Confidence, Support and Interest

TID	Items
1	1 2 4 9
2	3 4 5 9 10
3	1 2 4 9
4	3 4 5 9 10
5	1 3 4 5
6	1 4 5 6
7	1 2 4 9
8	1 3 4 5 6 9 10
9	3 4 5 9 10
10	3 4 5 9 10

- What is the support of $\{5, 9\}$?
  Count $= 5$, Frequency $= \frac{1}{2}$
- What is the support of $\{1, 3, 4, 5\}$?
  Count $= 2$, Frequency $= \frac{1}{5}$

(1.1) Confidence, Support and Interest

TID	Items
1	1 2 4 9
2	3 4 5 9 10
3	1 2 4 9
4	3 4 5 9 10
5	1 3 4 5
6	1 4 5 6
7	1 2 4 9
8	1 3 4 5 6 9 10
9	3 4 5 9 10
10	3 4 5 9 10

- What is the support of $\{5, 9\}$?
  Count $= 5$, Frequency $= \frac{1}{2}$
- What is the support of $\{1, 3, 4, 5\}$?
  Count $= 2$, Frequency $= \frac{1}{5}$
- What is the confidence of $\{5\} \implies \{9\}$?

(1.1) Confidence, Support and Interest

TID	Items
1	1 2 4 9
2	3 4 5 9 10
3	1 2 4 9
4	3 4 5 9 10
5	1 3 4 5
6	1 4 5 6
7	1 2 4 9
8	1 3 4 5 6 9 10
9	3 4 5 9 10
10	3 4 5 9 10

- What is the support of $\{5, 9\}$?
  Count $= 5$, Frequency $= \frac{1}{2}$
- What is the support of $\{1, 3, 4, 5\}$?
  Count $= 2$, Frequency $= \frac{1}{5}$
- What is the confidence of $\{5\} \implies \{9\}$?
  Confidence $= \frac{sup(\{5,9\})}{sup(\{5\})} = \frac{5}{7}$

(1.1) Confidence, Support and Interest

TID	Items
1	1 2 4 9
2	3 4 5 9 10
3	1 2 4 9
4	3 4 5 9 10
5	1 3 4 5
6	1 4 5 6
7	1 2 4 9
8	1 3 4 5 6 9 10
9	3 4 5 9 10
10	3 4 5 9 10

- What is the support of $\{5, 9\}$?
  Count $= 5$, Frequency $= \frac{1}{2}$
- What is the support of $\{1, 3, 4, 5\}$?
  Count $= 2$, Frequency $= \frac{1}{5}$
- What is the confidence of $\{5\} \implies \{9\}$?
  Confidence $= \frac{sup(\{5{,}9\})}{sup(\{5\})} = \frac{5}{7}$
- What is the confidence of $\{3, 4, 5\} \implies \{1\}$?

(1.1) Confidence, Support and Interest

TID	Items
1	1 2 4 9
2	3 4 5 9 10
3	1 2 4 9
4	3 4 5 9 10
5	1 3 4 5
6	1 4 5 6
7	1 2 4 9
8	1 3 4 5 6 9 10
9	3 4 5 9 10
10	3 4 5 9 10

- What is the support of $\{5, 9\}$?
  Count $= 5$, Frequency $= \frac{1}{2}$
- What is the support of $\{1, 3, 4, 5\}$?
  Count $= 2$, Frequency $= \frac{1}{5}$
- What is the confidence of $\{5\} \implies \{9\}$?
  Confidence $= \frac{sup(\{5,9\})}{sup(\{5\})} = \frac{5}{7}$
- What is the confidence of $\{3, 4, 5\} \implies \{1\}$?
  Confidence $= \frac{sup(\{1,3,4,5\})}{sup(\{3,4,5\})} = \frac{1}{3}$

(1.1) Confidence, Support and Interest

TID	Items
1	1 2 4 9
2	3 4 5 9 10
3	1 2 4 9
4	3 4 5 9 10
5	1 3 4 5
6	1 4 5 6
7	1 2 4 9
8	1 3 4 5 6 9 10
9	3 4 5 9 10
10	3 4 5 9 10

- What is the support of $\{5, 9\}$?
  Count $= 5$, Frequency $= \frac{1}{2}$
- What is the support of $\{1, 3, 4, 5\}$?
  Count $= 2$, Frequency $= \frac{1}{5}$
- What is the confidence of $\{5\} \implies \{9\}$?
  Confidence $= \frac{sup(\{5,9\})}{sup(\{5\})} = \frac{5}{7}$
- What is the confidence of $\{3, 4, 5\} \implies \{1\}$?
  Confidence $= \frac{sup(\{1,3,4,5\})}{sup(\{3,4,5\})} = \frac{1}{3}$
- What is the interest of $\{5\} \implies \{9\}$?

(1.1) Confidence, Support and Interest

TID	Items
1	1 2 4 9
2	3 4 5 9 10
3	1 2 4 9
4	3 4 5 9 10
5	1 3 4 5
6	1 4 5 6
7	1 2 4 9
8	1 3 4 5 6 9 10
9	3 4 5 9 10
10	3 4 5 9 10

- What is the support of $\{5, 9\}$?
  Count $= 5$, Frequency $= \frac{1}{2}$
- What is the support of $\{1, 3, 4, 5\}$?
  Count $= 2$, Frequency $= \frac{1}{5}$
- What is the confidence of $\{5\} \implies \{9\}$?
  Confidence $= \frac{sup(\{5,9\})}{sup(\{5\})} = \frac{5}{7}$
- What is the confidence of $\{3, 4, 5\} \implies \{1\}$?
  Confidence $= \frac{sup(\{1,3,4,5\})}{sup(\{3,4,5\})} = \frac{1}{3}$
- What is the interest of $\{5\} \implies \{9\}$?
  Interest $=$ conf($\{5\} \Rightarrow \{9\}$) - sup($\{9\}$) $= -\frac{3}{35}$

(1.1) Confidence, Support and Interest

TID	Items
1	1 2 4 9
2	3 4 5 9 10
3	1 2 4 9
4	3 4 5 9 10
5	1 3 4 5
6	1 4 5 6
7	1 2 4 9
8	1 3 4 5 6 9 10
9	3 4 5 9 10
10	3 4 5 9 10

- What is the support of $\{5, 9\}$?
  Count $= 5$, Frequency $= \frac{1}{2}$
- What is the support of $\{1, 3, 4, 5\}$?
  Count $= 2$, Frequency $= \frac{1}{5}$
- What is the confidence of $\{5\} \implies \{9\}$?
  Confidence $= \frac{sup(\{5,9\})}{sup(\{5\})} = \frac{5}{7}$
- What is the confidence of $\{3, 4, 5\} \implies \{1\}$?
  Confidence $= \frac{sup(\{1,3,4,5\})}{sup(\{3,4,5\})} = \frac{1}{3}$
- What is the interest of $\{5\} \implies \{9\}$?
  Interest $=$ conf($\{5\} \Rightarrow \{9\}$) - sup($\{9\}$) $= -\frac{3}{35}$
- What is the interest of $\{3, 4, 5\} \implies \{1\}$?

(1.1) Confidence, Support and Interest

TID	Items
1	1 2 4 9
2	3 4 5 9 10
3	1 2 4 9
4	3 4 5 9 10
5	1 3 4 5
6	1 4 5 6
7	1 2 4 9
8	1 3 4 5 6 9 10
9	3 4 5 9 10
10	3 4 5 9 10

- What is the support of $\{5, 9\}$?
  Count $= 5$, Frequency $= \frac{1}{2}$
- What is the support of $\{1, 3, 4, 5\}$?
  Count $= 2$, Frequency $= \frac{1}{5}$
- What is the confidence of $\{5\} \implies \{9\}$?
  Confidence $= \frac{sup(\{5,9\})}{sup(\{5\})} = \frac{5}{7}$
- What is the confidence of $\{3, 4, 5\} \implies \{1\}$?
  Confidence $= \frac{sup(\{1,3,4,5\})}{sup(\{3,4,5\})} = \frac{1}{3}$
- What is the interest of $\{5\} \implies \{9\}$?
  Interest $=$ conf($\{5\} \Rightarrow \{9\}$) - sup($\{9\}$) $= -\frac{3}{35}$
- What is the interest of $\{3, 4, 5\} \implies \{1\}$?
  Interest $= \frac{1}{3} - \frac{3}{5} = -\frac{4}{15}$

(1.1) Confidence, Support and Interest

TID	Items
1	1 2 4 9
2	3 4 5 9 10
3	1 2 4 9
4	3 4 5 9 10
5	1 3 4 5
6	1 4 5 6
7	1 2 4 9
8	1 3 4 5 6 9 10
9	3 4 5 9 10
10	3 4 5 9 10

For the itemset $\{3, 4, 5, 9, 10\}$**, do:**

- Write one association rule based on this itemset.

(1.1) Confidence, Support and Interest

TID	Items
1	1 2 4 9
2	3 4 5 9 10
3	1 2 4 9
4	3 4 5 9 10
5	1 3 4 5
6	1 4 5 6
7	1 2 4 9
8	1 3 4 5 6 9 10
9	3 4 5 9 10
10	3 4 5 9 10

For the itemset $\{3, 4, 5, 9, 10\}$, do:

- Write one association rule based on this itemset.
  $\{3, 4, 5, 9\} \implies \{10\}$

(1.1) Confidence, Support and Interest

TID	Items
1	1 2 4 9
2	3 4 5 9 10
3	1 2 4 9
4	3 4 5 9 10
5	1 3 4 5
6	1 4 5 6
7	1 2 4 9
8	1 3 4 5 6 9 10
9	3 4 5 9 10
10	3 4 5 9 10

For the itemset $\{3, 4, 5, 9, 10\}$**, do:**

- Write one association rule based on this itemset.
  $\{3, 4, 5, 9\} \implies \{10\}$
- What is its support??

(1.1) Confidence, Support and Interest

TID	Items
1	1 2 4 9
2	3 4 5 9 10
3	1 2 4 9
4	3 4 5 9 10
5	1 3 4 5
6	1 4 5 6
7	1 2 4 9
8	1 3 4 5 6 9 10
9	3 4 5 9 10
10	3 4 5 9 10

For the itemset $\{3, 4, 5, 9, 10\}$, do:

- Write one association rule based on this itemset.
  $\{3, 4, 5, 9\} \implies \{10\}$
- What is its support??
  Count $= 5$, Frequency $= \frac{1}{2}$

(1.1) Confidence, Support and Interest

TID	Items
1	1 2 4 9
2	3 4 5 9 10
3	1 2 4 9
4	3 4 5 9 10
5	1 3 4 5
6	1 4 5 6
7	1 2 4 9
8	1 3 4 5 6 9 10
9	3 4 5 9 10
10	3 4 5 9 10

For the itemset $\{3, 4, 5, 9, 10\}$, do:

- Write one association rule based on this itemset.
  $\{3, 4, 5, 9\} \implies \{10\}$
- What is its support??
  Count $= 5$, Frequency $= \frac{1}{2}$
- What is its confidence??

(1.1) Confidence, Support and Interest

TID	Items
1	1 2 4 9
2	3 4 5 9 10
3	1 2 4 9
4	3 4 5 9 10
5	1 3 4 5
6	1 4 5 6
7	1 2 4 9
8	1 3 4 5 6 9 10
9	3 4 5 9 10
10	3 4 5 9 10

For the itemset $\{3, 4, 5, 9, 10\}$, do:

- Write one association rule based on this itemset.
  $\{3, 4, 5, 9\} \implies \{10\}$
- What is its support??
  Count $= 5$, Frequency $= \frac{1}{2}$
- What is its confidence??
  Confidence $= \frac{sup(\{3,4,5,9,10\})}{sup(\{3,4,5,9\})} = 1$

(1.1) Confidence, Support and Interest

TID	Items
1	1 2 4 9
2	3 4 5 9 10
3	1 2 4 9
4	3 4 5 9 10
5	1 3 4 5
6	1 4 5 6
7	1 2 4 9
8	1 3 4 5 6 9 10
9	3 4 5 9 10
10	3 4 5 9 10

For the itemset $\{3, 4, 5, 9, 10\}$, do:

- Write one association rule based on this itemset.
  $\{3, 4, 5, 9\} \implies \{10\}$
- What is its support??
  Count $= 5$, Frequency $= \frac{1}{2}$
- What is its confidence??
  Confidence $= \frac{sup(\{3,4,5,9,10\})}{sup(\{3,4,5,9\})} = 1$
- Write another association rule based on this itemset.

(1.1) Confidence, Support and Interest

TID	Items
1	1 2 4 9
2	3 4 5 9 10
3	1 2 4 9
4	3 4 5 9 10
5	1 3 4 5
6	1 4 5 6
7	1 2 4 9
8	1 3 4 5 6 9 10
9	3 4 5 9 10
10	3 4 5 9 10

For the itemset $\{3, 4, 5, 9, 10\}$**, do:**

- Write one association rule based on this itemset.
  $\{3, 4, 5, 9\} \implies \{10\}$
- What is its support??
  Count $= 5$, Frequency $= \frac{1}{2}$
- What is its confidence??
  Confidence $= \frac{sup(\{3,4,5,9,10\})}{sup(\{3,4,5,9\})} = 1$
- Write another association rule based on this itemset.
  $\{3\} \implies \{4, 5, 9, 10\}$

(1.1) Confidence, Support and Interest

TID	Items
1	1 2 4 9
2	3 4 5 9 10
3	1 2 4 9
4	3 4 5 9 10
5	1 3 4 5
6	1 4 5 6
7	1 2 4 9
8	1 3 4 5 6 9 10
9	3 4 5 9 10
10	3 4 5 9 10

For the itemset $\{3, 4, 5, 9, 10\}$, do:

- Write one association rule based on this itemset.
  $\{3, 4, 5, 9\} \implies \{10\}$
- What is its support??
  Count $= 5$, Frequency $= \frac{1}{2}$
- What is its confidence??
  Confidence $= \frac{sup(\{3,4,5,9,10\})}{sup(\{3,4,5,9\})} = 1$
- Write another association rule based on this itemset.
  $\{3\} \implies \{4, 5, 9, 10\}$
- What is its confidence?

(1.1) Confidence, Support and Interest

TID	Items
1	1 2 4 9
2	3 4 5 9 10
3	1 2 4 9
4	3 4 5 9 10
5	1 3 4 5
6	1 4 5 6
7	1 2 4 9
8	1 3 4 5 6 9 10
9	3 4 5 9 10
10	3 4 5 9 10

For the itemset $\{3, 4, 5, 9, 10\}$, do:

- Write one association rule based on this itemset.
  $\{3, 4, 5, 9\} \implies \{10\}$
- What is its support??
  Count $= 5$, Frequency $= \frac{1}{2}$
- What is its confidence??
  Confidence $= \frac{sup(\{3,4,5,9,10\})}{sup(\{3,4,5,9\})} = 1$
- Write another association rule based on this itemset.
  $\{3\} \implies \{4, 5, 9, 10\}$
- What is its confidence?
  Confidence $= \frac{sup(\{3,4,5,9,10\})}{sup(\{3\})} = \frac{5}{6}$

(1.2) Lift

The lift of an association rule $A \implies B$ is defined as follows:

$$Lift(A \implies B) = \frac{Conf(A \implies B)}{P(B)} = \frac{sup(\{A, B\})}{sup(\{A\}) \cdot P(B)}$$

(1.2) Lift

The lift of an association rule $A \implies B$ is defined as follows:

$$Lift(A \implies B) = \frac{Conf(A \implies B)}{P(B)} = \frac{sup(\{A, B\})}{sup(\{A\}) \cdot P(B)}$$

Situation 1
A school has 500 students in it. Out of these students, 300 take Machine Learning (ML), 200 take Data Mining (DM) and 50 take both classes. Calculate the Lift of the rule $ML \implies DM$.

(1.2) Lift

The lift of an association rule $A \implies B$ is defined as follows:

$$Lift(A \implies B) = \frac{Conf(A \implies B)}{P(B)} = \frac{sup(\{A, B\})}{sup(\{A\}) \cdot P(B)}$$

Situation 1

A school has 500 students in it. Out of these students, 300 take Machine Learning (ML), 200 take Data Mining (DM) and 50 take both classes. Calculate the Lift of the rule $ML \implies DM$.

$$Lift(ML \implies DM) = \frac{sup(\{ML, DM\})}{sup(\{ML\}) \cdot P(DM)} = \frac{50}{300 \cdot \frac{200}{500}} = \frac{5}{12}$$

(1.2) Lift

The lift of an association rule $A \implies B$ is defined as follows:

$$Lift(A \implies B) = \frac{Conf(A \implies B)}{P(B)} = \frac{sup(\{A, B\})}{sup(\{A\}) \cdot P(B)}$$

Situation 2

A party has 1000 confirmed guests. Out of the guests, 600 drink Hoegaarden (H), 300 drink Kriek (K) and 200 drink both. Calculate the Lift of the rule $H \implies K$.

(1.2) Lift

The lift of an association rule $A \implies B$ is defined as follows:

$$Lift(A \implies B) = \frac{Conf(A \implies B)}{P(B)} = \frac{sup(\{A, B\})}{sup(\{A\}) \cdot P(B)}$$

Situation 2

A party has 1000 confirmed guests. Out of the guests, 600 drink Hoegaarden (H), 300 drink Kriek (K) and 200 drink both.
Calculate the Lift of the rule $H \implies K$.

$$Lift(H \implies K) = \frac{sup(\{H, K\})}{sup(\{H\}) \cdot P(K)} = \frac{200}{600 \cdot \frac{300}{1000}} = \frac{10}{9}$$

(2) Apriori: Join and Prune

Iterative: generate and prune candidates

Based on 4 steps:

- Compute the support of each candidate;
- Prune based on support threshold;
- Generate new candidates through the join function;
- Prune based on subset infrequency.

Key insight: if count of an itemset i is less than the threshold s, then the count of any extension of i has a count less than s too.

(2) Apriori: Join and Prune

Given the following set of frequent 3-itemsets

$$F_3 = \{\{1,3,5\},\{1,5,8\},\{1,3,10\},\{2,3,4\},\\ \{2,3,5\},\{3,8,9\},\{3,8,10\}\}$$

(2) Apriori: Join and Prune

Given the following set of frequent 3-itemsets

$$F_3 = \{\{1,3,5\},\{1,5,8\},\{1,3,10\},\{2,3,4\},\\ \{2,3,5\},\{3,8,9\},\{3,8,10\}\}$$

1. Generate all legal candidates of the next level.

(2) Apriori: Join and Prune

Given the following set of frequent 3-itemsets

$$F_3 = \{\{1,3,5\},\{1,5,8\},\{1,3,10\},\{2,3,4\},\\ \{2,3,5\},\{3,8,9\},\{3,8,10\}\}$$

1. Generate all legal candidates of the next level.
   $\{1,3,5,10\}$

(2) Apriori: Join and Prune

Given the following set of frequent 3-itemsets

$$F_3 = \{\{1,3,5\},\{1,5,8\},\{1,3,10\},\{2,3,4\},\\ \{2,3,5\},\{3,8,9\},\{3,8,10\}\}$$

1. Generate all legal candidates of the next level.
   $\{1,3,5,10\}$, $\{2,3,4,5\}$

(2) Apriori: Join and Prune

Given the following set of frequent 3-itemsets

$$F_3 = \{\{1,3,5\},\{1,5,8\},\{1,3,10\},\{2,3,4\},\\ \{2,3,5\},\{3,8,9\},\{3,8,10\}\}$$

1. Generate all legal candidates of the next level.
   $\{1,3,5,10\}$, $\{2,3,4,5\}$, $\{3,8,9,10\}$

(2) Apriori: Join and Prune

Given the following set of frequent 3-itemsets

$$F_3 = \{\{1,3,5\},\{1,5,8\},\{1,3,10\},\{2,3,4\},\\ \{2,3,5\},\{3,8,9\},\{3,8,10\}\}$$

1. Generate all legal candidates of the next level.
   $\{1,3,5,10\}$, $\{2,3,4,5\}$, $\{3,8,9,10\}$
2. Perform pruning on this candidate set.

(2) Apriori: Join and Prune

Given the following set of frequent 3-itemsets

$$F_3 = \{\{1,3,5\},\{1,5,8\},\{1,3,10\},\{2,3,4\},\\ \{2,3,5\},\{3,8,9\},\{3,8,10\}\}$$

1. Generate all legal candidates of the next level.
   $\{1,3,5,10\}$, $\{2,3,4,5\}$, $\{3,8,9,10\}$
2. Perform pruning on this candidate set.
   All three are pruned.
   Missing subsets: $\{1,\underline{3,5,10}\}$,$\{2,\underline{3,4,5}\}$,$\{3,\underline{8,9,10}\}$

(3) Closed and Maximal Itemsets

For example: Frequent patterns AB, AC, BC and ABC
$\Rightarrow$ sometimes only interested in ABC

Return only either *closed* or *maximal* itemsets, where

- *Closed*: no immediate superset has the same count;
- *Maximal*: **frequent** + no immediate superset is frequent.

(3) Closed and Maximal Itemsets

Itemset	Support	Frequent	Closed	Maximal
A	15			
B	20			
C	33			
D	25			
AB	15			
AC	12			
AD	15			
BC	18			
BD	5			
CD	25			
ABC	10			
ABD	2			
ACD	12			
BCD	3			
ABCD	1			

min. sup
$s = 10$

(3) Closed and Maximal Itemsets

Itemset	Support	Frequent	Closed	Maximal
A	15	x		
B	20			
C	33			
D	25			
AB	15			
AC	12			
AD	15			
BC	18			
BD	5			
CD	25			
ABC	10			
ABD	2			
ACD	12			
BCD	3			
ABCD	1			

min. sup

$s = 10$

(3) Closed and Maximal Itemsets

Itemset	Support	Frequent	Closed	Maximal
A	15	x		
B	20	x	x	
C	33			
D	25			
AB	15			
AC	12			
AD	15			
BC	18			
BD	5			
CD	25			
ABC	10			
ABD	2			
ACD	12			
BCD	3			
ABCD	1			

min. sup

$s = 10$

(3) Closed and Maximal Itemsets

Itemset	Support	Frequent	Closed	Maximal
A	15	x		
B	20	x	x	
C	33	x	x	
D	25			
AB	15			
AC	12			
AD	15			
BC	18			
BD	5			
CD	25			
ABC	10			
ABD	2			
ACD	12			
BCD	3			
ABCD	1			

min. sup

$s = 10$

(3) Closed and Maximal Itemsets

Itemset	Support	Frequent	Closed	Maximal
A	15	x		
B	20	x	x	
C	33	x	x	
D	25	x		
AB	15			
AC	12			
AD	15			
BC	18			
BD	5			
CD	25			
ABC	10			
ABD	2			
ACD	12			
BCD	3			
ABCD	1			

min. sup

$s = 10$

(3) Closed and Maximal Itemsets

Itemset	Support	Frequent	Closed	Maximal
A	15	x		
B	20	x	x	
C	33	x	x	
D	25	x		
AB	15	x	x	
AC	12			
AD	15			
BC	18			
BD	5			
CD	25			
ABC	10			
ABD	2			
ACD	12			
BCD	3			
ABCD	1			

min. sup

$s = 10$

(3) Closed and Maximal Itemsets

Itemset	Support	Frequent	Closed	Maximal
A	15	x		
B	20	x	x	
C	33	x	x	
D	25	x		
AB	15	x	x	
AC	12	x		
AD	15			
BC	18			
BD	5			
CD	25			
ABC	10			
ABD	2			
ACD	12			
BCD	3			
ABCD	1			

min. sup
$s = 10$

(3) Closed and Maximal Itemsets

Itemset	Support	Frequent	Closed	Maximal
A	15	x		
B	20	x	x	
C	33	x	x	
D	25	x		
AB	15	x	x	
AC	12	x		
AD	15	x	x	
BC	18			
BD	5			
CD	25			
ABC	10			
ABD	2			
ACD	12			
BCD	3			
ABCD	1			

min. sup
$s = 10$

(3) Closed and Maximal Itemsets

Itemset	Support	Frequent	Closed	Maximal
A	15	x		
B	20	x	x	
C	33	x	x	
D	25	x		
AB	15	x	x	
AC	12	x		
AD	15	x	x	
BC	18	x	x	
BD	5			
CD	25			
ABC	10			
ABD	2			
ACD	12			
BCD	3			
ABCD	1			

min. sup
$s = 10$

(3) Closed and Maximal Itemsets

Itemset	Support	Frequent	Closed	Maximal
A	15	x		
B	20	x	x	
C	33	x	x	
D	25	x		
AB	15	x	x	
AC	12	x		
AD	15	x	x	
BC	18	x	x	
BD	5		x	
CD	25			
ABC	10			
ABD	2			
ACD	12			
BCD	3			
ABCD	1			

min. sup

$s = 10$

(3) Closed and Maximal Itemsets

Itemset	Support	Frequent	Closed	Maximal
A	15	x		
B	20	x	x	
C	33	x	x	
D	25	x		
AB	15	x	x	
AC	12	x		
AD	15	x	x	
BC	18	x	x	
BD	5		x	
CD	25	x	x	
ABC	10			
ABD	2			
ACD	12			
BCD	3			
ABCD	1			

min. sup

$s = 10$

(3) Closed and Maximal Itemsets

Itemset	Support	Frequent	Closed	Maximal
A	15	x		
B	20	x	x	
C	33	x	x	
D	25	x		
AB	15	x	x	
AC	12	x		
AD	15	x	x	
BC	18	x	x	
BD	5		x	
CD	25	x	x	
ABC	10	x	x	x
ABD	2			
ACD	12			
BCD	3			
ABCD	1			

min. sup

$s = 10$

(3) Closed and Maximal Itemsets

Itemset	Support	Frequent	Closed	Maximal
A	15	x		
B	20	x	x	
C	33	x	x	
D	25	x		
AB	15	x	x	
AC	12	x		
AD	15	x	x	
BC	18	x	x	
BD	5		x	
CD	25	x	x	
ABC	10	x	x	x
ABD	2		x	
ACD	12			
BCD	3			
ABCD	1			

min. sup
$s = 10$

(3) Closed and Maximal Itemsets

Itemset	Support	Frequent	Closed	Maximal
A	15	x		
B	20	x	x	
C	33	x	x	
D	25	x		
AB	15	x	x	
AC	12	x		
AD	15	x	x	
BC	18	x	x	
BD	5		x	
CD	25	x	x	
ABC	10	x	x	x
ABD	2		x	
ACD	12	x	x	x
BCD	3			
ABCD	1			

min. sup

$s = 10$

(3) Closed and Maximal Itemsets

Itemset	Support	Frequent	Closed	Maximal
A	15	x		
B	20	x	x	
C	33	x	x	
D	25	x		
AB	15	x	x	
AC	12	x		
AD	15	x	x	
BC	18	x	x	
BD	5		x	
CD	25	x	x	
ABC	10	x	x	x
ABD	2		x	
ACD	12	x	x	x
BCD	3		x	
ABCD	1			

min. sup

$s = 10$

(3) Closed and Maximal Itemsets

Itemset	Support	Frequent	Closed	Maximal
A	15	x		
B	20	x	x	
C	33	x	x	
D	25	x		
AB	15	x	x	
AC	12	x		
AD	15	x	x	
BC	18	x	x	
BD	5		x	
CD	25	x	x	
ABC	10	x	x	x
ABD	2		x	
ACD	12	x	x	x
BCD	3		x	
ABCD	1		x	

min. sup
$s = 10$

Data Mining:
Esercise Session 5
Second part

(1) Apriori

min. support threshold $s = 2$

TID	Items
1	1 4 10
2	3 5 6
3	3 5 6 8
4	3 4 6
5	3 5 6 8
6	2 6 7 8
7	2 6 7 8
8	1 4 9
9	3 4
10	3 5 6 7

(1) Apriori

min. support threshold $s = 2$

TID	Items
1	1 4 10
2	3 5 6
3	3 5 6 8
4	3 4 6
5	3 5 6 8
6	2 6 7 8
7	2 6 7 8
8	1 4 9
9	3 4
10	3 5 6 7

Level 1

Count frequencies of individual items

(1) Apriori

min. support threshold $s = 2$

TID	Items
1	1 4 10
2	3 5 6
3	3 5 6 8
4	3 4 6
5	3 5 6 8
6	2 6 7 8
7	2 6 7 8
8	1 4 9
9	3 4
10	3 5 6 7

Level 1

Count frequencies of individual items

1:2, 2:2, 3:6, 4:4, 5:4, 6:7, 7:3, 8:4, 9:1, 10:1

(1) Apriori

min. support threshold $s = 2$

TID	Items
1	1 4 10
2	3 5 6
3	3 5 6 8
4	3 4 6
5	3 5 6 8
6	2 6 7 8
7	2 6 7 8
8	1 4 9
9	3 4
10	3 5 6 7

Level 1

Count frequencies of individual items

1:2, 2:2, 3:6, 4:4, 5:4, 6:7, 7:3, 8:4, ~~9:1~~, ~~10:1~~

(1) Apriori

min. support threshold $s = 2$

TID	Items
1	1 4 10
2	3 5 6
3	3 5 6 8
4	3 4 6
5	3 5 6 8
6	2 6 7 8
7	2 6 7 8
8	1 4 9
9	3 4
10	3 5 6 7

Level 2

Count frequencies of pairs of frequent items

(1) Apriori

min. support threshold $s = 2$

TID	Items
1	1 4 10
2	3 5 6
3	3 5 6 8
4	3 4 6
5	3 5 6 8
6	2 6 7 8
7	2 6 7 8
8	1 4 9
9	3 4
10	3 5 6 7

Level 2

Count frequencies of pairs of frequent items

~~1 2:0~~, ~~1 3:0~~, 1 4:2, ~~1 5:0~~, ~~1 6:0~~, ~~1 7:0~~, ~~1 8:0~~,
~~2 3:0~~, ~~2 4:0~~, ~~2 5:0~~, 2 6:2, 2 7:2, 2 8:2,
3 4:2, 3 5:4, 3 6:5, ~~3 7:1~~, 3 8:2,
~~4 5:0~~, ~~4 6:1~~, ~~4 7:0~~, ~~4 8:0~~,
5 6:4, ~~5 7:1~~, 5 8:2,
6 7:3, 6 8:4
7 8:2

(1) Apriori

min. support threshold $s = 2$

TID	Items
1	1 4 10
2	3 5 6
3	3 5 6 8
4	3 4 6
5	3 5 6 8
6	2 6 7 8
7	2 6 7 8
8	1 4 9
9	3 4
10	3 5 6 7

Level 3

Join step (candidate 3-itemsets)

Prune step (subset pruning)

Count frequencies of remaining itemsets

(1) Apriori

min. support threshold $s = 2$

TID	Items
1	1 4 10
2	3 5 6
3	3 5 6 8
4	3 4 6
5	3 5 6 8
6	2 6 7 8
7	2 6 7 8
8	1 4 9
9	3 4
10	3 5 6 7

Level 3

Join step (candidate 3-itemsets)
2 6 7, 2 6 8, 2 7 8,
3 4 5, 3 4 6, 3 4 8, 3 5 6, 3 5 8, 3 6 8,
5 6 8,
6 7 8
Prune step (subset pruning)

Count frequencies of remaining itemsets

(1) Apriori

min. support threshold $s = 2$

TID	Items
1	1 4 10
2	3 5 6
3	3 5 6 8
4	3 4 6
5	3 5 6 8
6	2 6 7 8
7	2 6 7 8
8	1 4 9
9	3 4
10	3 5 6 7

Level 3

Join step (candidate 3-itemsets)
2 6 7, 2 6 8, 2 7 8,
3 4 5, 3 4 6, 3 4 8, 3 5 6, 3 5 8, 3 6 8,
5 6 8,
6 7 8

Prune step (subset pruning)
3 4 5 (‘4 5’ is infrequent), 3 4 6, 3 4 8

Count frequencies of remaining itemsets

(1) Apriori

min. support threshold $s = 2$

TID	Items
1	1 4 10
2	3 5 6
3	3 5 6 8
4	3 4 6
5	3 5 6 8
6	2 6 7 8
7	2 6 7 8
8	1 4 9
9	3 4
10	3 5 6 7

Level 3

Join step (candidate 3-itemsets)
2 6 7, 2 6 8, 2 7 8,
3 4 5, 3 4 6, 3 4 8, 3 5 6, 3 5 8, 3 6 8,
5 6 8,
6 7 8

Prune step (subset pruning)
3 $\underline{4\ 5}$ ('4 5' is infrequent), 3 $\underline{4\ 6}$, 3 $\underline{4\ 8}$

Count frequencies of remaining itemsets
2 6 7:2, 2 6 8:2, 2 7 8:2,
3 5 6:4, 3 5 8:2, 3 6 8:2,
5 6 8:2,
6 7 8:2 (all are frequent)

(1) Apriori

min. support threshold $s = 2$

TID	Items
1	1 4 10
2	3 5 6
3	3 5 6 8
4	3 4 6
5	3 5 6 8
6	2 6 7 8
7	2 6 7 8
8	1 4 9
9	3 4
10	3 5 6 7

Level 4

Join step

Prune step

Count

(1) Apriori

min. support threshold $s = 2$

TID	Items
1	1 4 10
2	3 5 6
3	3 5 6 8
4	3 4 6
5	3 5 6 8
6	2 6 7 8
7	2 6 7 8
8	1 4 9
9	3 4
10	3 5 6 7

Level 4

Join step
2 6 7 8, 3 5 6 8
Prune step

Count

(1) Apriori

min. support threshold $s = 2$

TID	Items
1	1 4 10
2	3 5 6
3	3 5 6 8
4	3 4 6
5	3 5 6 8
6	2 6 7 8
7	2 6 7 8
8	1 4 9
9	3 4
10	3 5 6 7

Level 4

Join step
2 6 7 8, 3 5 6 8
Prune step
All 3-subsets are frequent.
Count

(1) Apriori

min. support threshold $s = 2$

TID	Items
1	1 4 10
2	3 5 6
3	3 5 6 8
4	3 4 6
5	3 5 6 8
6	2 6 7 8
7	2 6 7 8
8	1 4 9
9	3 4
10	3 5 6 7

Level 4

Join step
2 6 7 8, 3 5 6 8
Prune step
All 3-subsets are frequent.
Count
2 6 7 8:2, 3 5 6 8:2

(1) Apriori

min. support threshold $s = 2$

TID	Items
1	1 4 10
2	3 5 6
3	3 5 6 8
4	3 4 6
5	3 5 6 8
6	2 6 7 8
7	2 6 7 8
8	1 4 9
9	3 4
10	3 5 6 7

Level 5

Cannot generate any 5-itemsets
$\Rightarrow$ the algorithm terminates.

(2) PCY

- Construct a bit array B of size b;
- Initialize each position to be 0;
- Select a hash function h with range $[0, b-1]$;
- Hash each element and increase the bucket count by 1;
- If a bucket has a count greater than the minimum support threshold, then we can not eliminate any member of it;
- if the count is less than the minimum, then we can eliminate all the pairs hashed to the bucket (best case).

Pay attention!
We do not save the pairs but the count inside the buckets.

(2) PCY

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

(2) PCY

Baskets:

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

Support for each individual item

(2) PCY

Baskets:

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

Support for each individual item

$supp(\{1\}) = 4$
$supp(\{2\}) = 6$
$supp(\{3\}) = 8$
$supp(\{4\}) = 8$
$supp(\{5\}) = 6$
$supp(\{6\}) = 4$

(2) PCY

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

Hash $\{i, j\}$
to bucket
$i \times j$ *mod* 11

Buckets

(2) PCY

Baskets:

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

Hash $\{i,j\}$
to bucket
$i \times j$ *mod* 11

Buckets

$\{1,2\} \mapsto 2$

(2) PCY

Baskets:

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

Hash $\{i,j\}$
to bucket
$i \times j \ mod \ 11$

Buckets

$\{2,6\}$, $\{3,4\}$	$\mapsto$	1
$\{1,2\}$, $\{4,6\}$	$\mapsto$	2
$\{1,3\}$	$\mapsto$	3
$\{1,4\}$, $\{3,5\}$	$\mapsto$	4
$\{1,5\}$	$\mapsto$	5
$\{1,6\}$, $\{2,3\}$	$\mapsto$	6
$\{3,6\}$	$\mapsto$	7
$\{2,4\}$, $\{5,6\}$	$\mapsto$	8
$\{4,5\}$	$\mapsto$	9
$\{2,5\}$	$\mapsto$	10

(2) PCY

Baskets:

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

Hash $\{i,j\}$
to bucket
$i \times j \ mod \ 11$

Buckets

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

(2) PCY

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

Hash $\{i,j\}$
to bucket
$i \times j$ *mod* 11

Buckets

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

Which itemsets are counted on the second pass?
The support threshold is 4.

(2) PCY

Baskets:

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

Hash $\{i,j\}$
to bucket
$i \times j \ mod \ 11$

Buckets

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

Which itemsets are counted on the second pass?
The support threshold is 4.

(3) FP Growth

Avoid candidate generation by using a FP tree which can compactly and completely represent the frequent itemsets

Steps:

- Find the frequent singletons;
- Build FP tree: remove infrequent items, sort the remaining items and insert them into a tree;
- Construct conditional pattern-base for each node in the FP tree;
- Construct conditional FP tree from each conditional pattern-base by summing the counts for each item;
- Recursively mine conditional FP trees.

(3) FP Growth

TID	Items
1	A, B, C, E, F
2	A, C, D, E, F
3	A, B, C, G, I
4	A, B, C, G
5	B, E, F, H, I, J

(3) FP Growth

TID	Items
1	A, B, C, E, F
2	A, C, D, E, F
3	A, B, C, G, I
4	A, B, C, G
5	B, E, F, H, I, J

Item	Count
A	4
B	4
C	4
D	1
E	3
F	3
G	2
H	1
I	2
J	1

(3) FP Growth

TID	Items
1	A, B, C, E, F
2	A, C, D, E, F
3	A, B, C, G, I
4	A, B, C, G
5	B, E, F, H, I, J

Item	Count
A	4
B	4
C	4
~~D~~	~~1~~
E	3
F	3
~~G~~	~~2~~
~~H~~	~~1~~
~~I~~	~~2~~
~~J~~	~~1~~

min. sup
$s = 3$

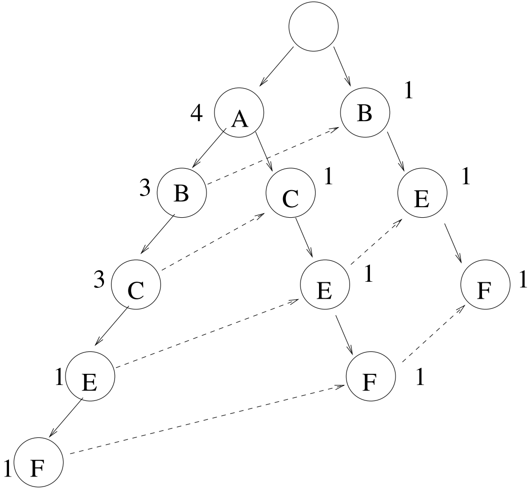
(3) FP Growth

TID	Items
1	A, B, C, E, F
2	A, C, D, E, F
3	A, B, C, G, I
4	A, B, C, G
5	B, E, F, H, I, J

TID	Items
1	A, B, C, E, F
2	A, C, E, F
3	A, B, C
4	A, B, C
5	B, E, F

(3) FP Growth

TID	Items
1	A, B, C, E, F
2	A, C, E, F
3	A, B, C
4	A, B, C
5	B, E, F



(4) Thought Question

What are the causes of pattern explosion?

Primary causes of pattern explosions include:

- The nature of the problem. If $\{A, B, C\}$ is frequent, then all its 7 subsets are necessarily returned;
- Functional or statistical relations between attributes. Dependencies between multiple variables might result in a large number of itemsets;
- Locality of the support constraint. A frequent itemset is always returned, independent of already returned itemsets.

(4) Thought Question

Can you think of a way to solve or alleviate this issue?

Possible solutions are:

- Top-*k* mining, i.e. only returning *k* most frequent itemsets. Furthermore, frequency can be replaced with another interestingness measure;
- Condensed representations, lossless (closed itemsets) or lossy (maximal itemsets);
- Pattern set mining, i.e. introducing global constraints on the result sets to eliminate redundancy. For example, ensuring that covers of returned itemsets do not overlap. More advanced methods rely on heuristics rooted in probability theory, information theory, compression, etc.