Overview

In the first part of the session, you will gain familiarity with association rule mining by solving some solve pen and paper exercises. The goal of second part of the session is to deepen your understanding of ensemble methods. Have fun!

1 Association Rule Mining: Basics

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

Answer the following questions:

1. What is the support of $\{5, 9\}$?
2. What is the support of $\{1, 3, 4, 5\}$?
3. What is the confidence of $5 \Rightarrow 9$?
4. What is the confidence of $\{3, 4, 5\} \Rightarrow \{1\}$?
5. What is the interest of $5 \Rightarrow 9$?
6. What is the interest of $\{3, 4, 5\} \Rightarrow \{1\}$?

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

1. Write one association rule based on this itemset.
2. What is its support?
3. What is its confidence?
4. Write another association rule based on this itemset.
5. What is its confidence?

1. $supp(\{5, 9\}) = 5$
2. $supp(\{1, 3, 4, 5\}) = 2$
3. $conf(\{5\} \Rightarrow \{9\}) = \dfrac{5}{supp(\{5\}) = 7} \approx 0.7$
4. $conf(\{3, 4, 5\} \Rightarrow \{1\}) = \dfrac{2}{supp(\{3, 4, 5\}) = 6} = \dfrac{1}{3}$

5. $\mathit{interest}(\{5\} \Rightarrow \{9\}) = \mathit{conf}(5 \Rightarrow 9) - \mathit{supp}(9)/|D| = 5/7 - 8/10 \approx -0.1$
6. $\mathit{interest}(\{3,4,5\} \Rightarrow \{1\}) = \mathit{conf}(3,4,5 \Rightarrow 1) - \mathit{supp}(1)/|D| = 1/3 - 6/10 \approx -0.27$

1. $\{3,4,5,9\} \Rightarrow \{10\}$
2. $\mathit{supp}(\{3,4,5,9,10\}) = 5$
3. $\mathit{conf}(\{3,4,5,9\} \Rightarrow \{10\}) = \dfrac{5}{\mathit{supp}(\{3,4,5,9\}) = 5} = 1.0$
4. $\{3\} \Rightarrow \{4,5,9,10\}$
5. $\mathit{conf}(\{3\} \Rightarrow \{4,5,9,10\}) = \dfrac{5}{\mathit{supp}(\{3\}) = 6} = \dfrac{5}{6} \approx 0.83$

1.2 Lift

The lift of an an association rule $A \Rightarrow B$ is defined as follows: $Lift(A \Rightarrow B) = \dfrac{Conf(A \Rightarrow B)}{Supp(B)}$
Use this definition to calculate the lift for the following problems.

Situation 1 The school has 500 students in it. Out of these students, 300 take machine learning (ML) and 200 take data mining (DM) and 50 take both classes. Calculate the lift of the rule $ML \Rightarrow DM$.

Situation 2 A party has 1000 confirmed guests. Out of the guests, 600 drink Hoegaarden (H) and 300 drink Kriek (K) and 200 drink both. Calculate the lift of the rule $H \Rightarrow K$.

Situation 1 $lift(ML \Rightarrow DM) = \dfrac{50/300}{200/500} = 0.417$

Situation 2 $lift(H \Rightarrow K) = \dfrac{200/600}{300/1000} = 1.11$

2 Apriori: Join and Prune

Given the following set of frequent 3-itemsets, $F_3 = \{1,3,5\}, \{1,5,8\}, \{1,3,10\}, \{2,3,4\}, \{2,3,5\}, \{3,8,9\}, \{3,8,10\}$, do

1. Generate all legal candidates of the next level of Apriori's search.
2. Perform pruning of this candidate set.

1. Generate all legal candidates of the next level of Apriori's search:
   $\{1, \underline{3,5,10}\}$,$\{2, \underline{3,4,5}\}$,$\{3, \underline{8,9,10}\}$
2. All pruned: Missing subsets underlined above.

3 Apriori

For the transactional database below, iterate through the Apriori algorithm using a minimum 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, 4:6, 5:4, 6:7, 7:3, 8:4, ~~9:1~~, ~~10:1~~

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

Level 3 *Join* (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 (subset pruning): 3 4 5 (‘4 5’ is infrequent), 3 4 6, 3 4 8
Count (count frequencies of the remaining itemsets):
2 6 7:2, 2 6 8:2, 2 7 8:2,
3 5 6:4, 3 5 8:2, 3 6 8:3,
5 6 8:2,
6 7 8:2 (all are frequent)

Level 4 *Join*: 2 6 7 8, 3 5 6 8
Prune: no pruning, all 3-subsets are frequent, e.g. one should check ‘2 7 8’ and ‘6 7 8’ for ‘2 6 7 8’
Count: 2 6 7 8:2, 3 5 6 8:2

Level 5 *Join*: Cannot generate any 5-itemsets ⇒ the algorithm terminates.

All frequent itemsets, ordered by frequency descending:
$Freq = 7$: 6
$Freq = 6$: 3
$Freq = 5$: 3 6
$Freq = 4$: 3 5 6, 3 5, 5 6, 6 8, 4, 5, 8
$Freq = 3$: 6 7, 7
$Freq = 2$: 2 6 7 8, 3 5 6 8, 2 6 7, 2 6 8, 2 7 8, 3 5 8, 3 6 8, 5 6 8, 6 7 8, 1 4, 2 6, 2 7, 2 8, 3 4, 3 8, 5 8, 7 8, 1, 2

The dataset corresponds to the first 10 rows of `eda.csv`, where items are as follows:
1 - $hospital = sports$, 2 - $hospital = general$, 3 - $hospital = prenatal$, 4 - $gender = f$, 5 - $gender = m$, 6 - $blood\_test = t$, 7 - $ecg = t$, 8 -$ultrasound = t$, 9 - $mri = t$, 10 - $xray = t$

4 PCY

Here is a collection of twelve baskets. Each contains three of the six items 1 through 6.

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

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 $i \times j \ mod \ 11$, where mod is the *modulo* operation, i.e. finding the remainder after division by 11. For example, $Hash\left(\{5, 6\}\right) = (5 \times 6) \ mod \ 11 = 30 \ mod \ 11 = 8$, because $30 = 11 \times 2 + \mathbf{8}$.

The support threshold is 4.

1. Perform the first pass of PCY: compute the support for each individual item and frequencies of buckets.

2. Which pairs hash to which buckets?

3. Which buckets are frequent?

4. Generate all candidate pairs (2-itemsets): which ones are counted on the second pass of PCY?

1. Perform the first pass of PCY: compute the support for each individual item and frequencies of buckets.
   $sup\left(\{1\}\right) = 4$, $sup\left(\{2\}\right) = 6$, $sup\left(\{3\}\right) = 8$, $sup\left(\{4\}\right) = 8$, $sup\left(\{5\}\right) = 6$, $sup\left(\{6\}\right) = 4$.
   See below for bucket frequencies.

2. Which pairs hash to which buckets?
   $\{2, 6\}$, $\{3, 4\}$ $\mapsto$ 1 (bucket frequency = 5).
   $\{1, 2\}$, $\{4, 6\}$ $\mapsto$ 2 (4).
   $\{1, 3\}$ $\mapsto$ 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).
   Note that only the information on the right-hand side of arrows needs to be stored.

3. Which buckets are frequent?
   1, 2, 4, 8.

4. Generate all candidate pairs (2-itemsets): which ones are counted on the second pass of PCY?
   All individual items are frequent, hence all item pairs are valid candidates.
   It is only necessary to count frequencies for the ones that hash into frequent buckets:
   $\{2, 6\}$, $\{3, 4\}$, $\{1, 2\}$, $\{4, 6\}$, $\{1, 4\}$, $\{3, 5\}$, $\{2, 4\}$, $\{5, 6\}$.
   Note that neither $\{1, 2\}$ nor $\{4, 6\}$ is frequent, even though they hash into the frequent bucket 2.