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?

1.2 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 $H \Rightarrow K$.

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.

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

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?

5 Ensemble Methods: Effects of bagging and boosting

The `datasets.zip` on Toledo includes the bike sharing system data (`train.cvs` and `test.csv`).
Today, we will be focusing on regression task. The goal is to predict the total number of bikes in use on hourly basis.
Open `bikes-ensembles.ipynb`. Get familiar with different ensemble methods provided by `scikit-learn` by complete the skeleton code.