Overview

In this exercise, we briefly review concepts from itemset mining and association rule mining and then proceed to sequence mining. This part of the exercise is pen-and-paper. Then we move on to the bike rental prediction with ensemble methods. Have fun!

1 Closed and Maximal Itemsets

In the following table, which itemsets are frequent, frequent closed and frequent maximal for the min.support threshold 10.

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			

2 Sequence mining

For the transactional database below:

1. Convert it to a customer-sequence database.
2. Mine the a-projected database and the c-projected database using FREESPAN with $minsup = 2$.
3. Perform the same task using PREFIXSPAN.

Date	CID	Items
01/12	1	a b
01/12	2	bh
01/12	3	i
01/12	4	b c
02/12	1	c
02/12	2	c
02/12	3	b g f
02/12	4	d
03/12	1	a d
03/12	2	a
03/12	3	i
03/12	4	c
04/12	1	b
04/12	2	g
04/12	3	j
04/12	4	f
05/12	1	g
05/12	2	f
05/12	3	c
05/12	4	a
06/12	1	f
06/12	2	a c
06/12	3	b g
06/12	4	b c
07/12	2	h
07/12	4	a
08/12	2	c d
08/12	4	d
09/12	4	c d

3 FP Growth

Using the transactional database below, construct its frequent pattern tree (FP tree) using a minimum support threshold, $s = 3$.

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

4 Thought Question

Pattern explosion is a well-known problem in frequent itemset mining. High support thresholds typically result only in few well-known patterns; but for low support thresholds, the number of frequent itemsets can easily be orders of magnitude larger than the number of transactions. Knowledge discovery in such humongous itemset collections is virtually impossible.

What are the causes of pattern explosion? Can you think of a way to solve or at least alleviate this issue?

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 completing the skeleton code.