CMPT459 Fall 2017
Data Mining
Martin Ester
TA: Zhilin Zhang

Programming Assignment 4

Total marks: 100
Due date: November 22, 2017

Data
The Groceries data set contains 1 month (30 days) of real-world point-of-sale transaction data from a typical local grocery outlet. The data set contains 9835 transactions with items that are aggregated to 169 categories, stored in sparse format. The dataset is part of the "arules" package and can be loaded by calling data(Groceries).

Tasks
In this assignment, you will gain practical experience with mining frequent itemsets and association rules. Solve the tasks using R (package "arules") and answer the questions.

1. Plot a histogram of the number of items (categories) per transaction. What do you observe? How can you explain this observation?

2. How many frequent itemsets, closed frequent itemsets, and maximal frequent itemsets do you obtain with minimum support = 0.001?

3. How many frequent itemsets, closed frequent itemsets, and maximal frequent itemsets do you obtain with minimum support = 0.01?

4. What are the 10 itemsets with the highest support, and what is their support?

5. How do you explain the relatively small number of frequent itemsets for the already low minimum support of 0.01? How do you explain the observation that the numbers of frequent itemsets, closed frequent itemsets, and maximal frequent itemsets are so similar?

6. At minimum support = 0.01, how many association rules do you obtain with minimum confidence = 0.9? How far do you need to lower the minimum confidence to obtain more than 10 rules?

7. For minimum support = 0.01 and minimum confidence = 0.5, plot only the rules that have "whole milk" in their right hand side.

8. Among the rules produced in task 7, which ones have the highest lift? Can you explain these rules? How interesting are they?

Submission
Submit your R code in pa4.r and a report report4.pdf answering the questions in CourSys.

Note: we do not accept handwritten submissions, but only typed reports!