RWTH Aachen
Lehrstuhl für Informatik 9
Prof. Dr. T. Seidl

November 16th, 2015

Exercise 4 for the lecture

Data Mining Algorithms

WS 2015/2016

Hand in your solutions on November 23rd before the lecture. The tutorial for this exercise will be held on November 27th. Solutions of groups with less than 3 or more than 4 students will not be graded.

Note: All commands for the R-exercises are required to be provided with comments, indicating which task the commands belong to. All R script files should contain a comment-line with the names and matriculation numbers of all group-members. Send all R-files to faerber@informatik.rwth-aachen.de. The subject of the mail must start with "[DMA1]".

Exercise 4.1) FP-growth algorithm — 6 points

Consider the following data set containing 13 transactions:
(b,f,g,m); (d,g,h,l); (b,c,d); (d,g,h,i); (d,g,h,j); (a,b,d,g,k);(a,d,h); (e,f,g);
(b,d); (a,f,g); (a,b,d,g); (b,f,g); (f,g)
In the following, mine all frequent itemsets using the FP-growth algorithm for a minimum support threshold of 10%.

Exercise 4.2) Correlation (Lift) — 7 = 4+3 points

Consider the following dataset from Exercise 3.2:

TID	date	items_bought
T1	10-15-99	{K, A, D, B}
T2	10-15-99	{D, A, C, E, B}
T3	10-19-99	{C, A, B, E}
T4	10-22-99	{B, A, D}

a) From Exercise 3.2, the set of frequent patterns with *min_sup = 60%* are:

A	100%	A B	100%
B	100%	A D	75%
D	75%	B D	75%
A B D	75%		

List all of the strong association rules with *min_sup = 60%* and *min_conf = 80%*.

b) For each strong association rule, compute the correlation of head and body. For each of the strong association rules, state whether its head and body are positively correlated, negatively correlated or independent.

Exercise 4.3) Hierarchical Association Rules **3 = 1+1+1 points**

Let A, A1, A2, B be items. Furthermore let A be a generalization of A1 and A2. Prove or contradict (by giving a counterexample) the following assumptions:

a) *support*("A $\Rightarrow$ B") = *support* ("A1 $\Rightarrow$ B") + *support* ("A2 $\Rightarrow$ B")
b) If *support*("A1 $\Rightarrow$ B") > *min_sup*, then *support*("A $\Rightarrow$ B") > *min_sup*.
c) If *support*("A $\Rightarrow$ B") > *min_sup*, then *support*("A1 $\Rightarrow$ B") > *min_sup*.

Exercise 4.4) Interestingness of Hierarchical Association Rules **5 points**

For the following four rules determine whether they are R-interesting (R=1.95) with respect to each of their ancestors.

Item	count
vessel	**100**
passenger car	**40**
Seat	**15**
Skoda	**20**

Rule #	rule	count
1	**vessel ⇒ aluminium wheels**	**25**
2	**p. car ⇒ aluminium wheels**	**20**
3	**Seat ⇒ aluminium wheels**	**10**
4	**Skoda ⇒ aluminium wheels**	**10**

Hint: For the hierarchy: *Vessel* is an ancestor of *passenger car* and *passenger car* is an ancestor of *Seat* and *Skoda*.

Exercise 4.5) Frequent Pattern & Association Rules Mining with R **3=1+1+1 points**

You need the following packages for this task: "arules" and "arulesViz".
Use the built in dataset "Groceries" in the package arules for mining frequent itemsets using the apriori() algorithm.

a) Mine all frequent itemsets, having a support of 5%, sort them according to their frequency (support).
b) Mine all strong association rules having a support of 0.4%, a confidence of 70%, and at least a length of 2 (i.e. neither head nor body is empty). Sort them according to their frequency (support).
c) Use the "arulesViz" package for a quick visualization of the mined association rules (e.g. scatter plot). What is an interesting rule you found?

Hand in your Rscript together with outputs (itemsets, rules, and plots) in a digital and a paper version.