Data Mining
Homework Assignment #2

Dmytro Fishman, Anna Leontjeva and Jaak Vilo

February 20, 2014

Task 1

CustomerID	TransactionID	BasketContent
1	1234	{Aspirin, Panadol}
1	4234	{Aspirin, Sudafed}
2	9373	{Tylenol, Cepacol}
2	9843	{Aspirin, Vitamin C, Sudafed}
3	2941	{Tylenol, Cepacol}
3	2753	{Aspirin, Cepacol}
4	9643	{Aspirin, Vitamin C}
4	9691	{Aspirin, Ibuprofen, Panadol}
5	5313	{Panadol, Vitamin C}
5	1003	{Tylenol, Cepacol, Ibuprofen}
6	5636	{Tylenol, Panadol, Cepacol}
6	3478	{Panadol, Sudafed, Ibuprofen}

a. Compute the support and support count for itemsets {Aspirin}, {Tylenol, Cepacol}, {Aspirin, Ibuprofen, Panadol} by treating each transaction ID as a market basket.

b. Compute the confidence for the following association rules: {Aspirin, Vitamin C → Sudafed}, {Aspirin → Vitamin C}, {Vitamin C → Aspirin}. Why the results for last two rules are different?

c. List all the frequent itemsets under the support count threshold $s_{min} = 3$.

d. What does the anti-monotonicity property of a support imply? Give an example using the above data set.

Task 2

Write down all the steps of Apriori algorithm on the above data set under the support count threshold $s_{min} > 3$. How many steps of Apriori algorithm you needed to perform? Draw a diagram showing all possible combinations of the items (e.g. lecture slide number 68). Mark all maximal, closed and infrequent items on this diagram.

Task 3

Build an FP-tree using data set from exercise 1. Explain all the steps, draw a final tree. How many transactions contain {Aspirin} and {Cepacol}?

Task 4

Read chapter on conditional probability in order to answer the following question:

> At the exam there is 0.8 probability that student has prepared and 0.2 that he has not prepared. Those who are prepared have 0.7 probability of success, those who have not prepared have 0.4 probability of success. What is the probability that randomly selected student will succeed?

Task 5

In this task we will get familiar with the statistical computing language R. Install it. We suggest you to download also the IDE that will make your life much easier: R studio. Once you are set up, take a look at the introduction of R from the CRAN page (Manuals → An Introduction to R) or just google any basic tutorial. R is an open source and has a very powerful community with plenty of tutorials and websites. Once you feel more comfortable with it, go through the following tutorial, runit, check and report the results, describe and interpret them:

Task 6 (2pt)

What is the probability to get 9 or 10 heads when you throw a fair coin 10 times? What is the probability to get 70 or more heads when you throw a fair coin 100 times? Conduct a computational experiment by generating 10,000 times such sequences of 10 coin tosses or 100 coin tosses.