Data Mining
Homework Assignment #3

Dmytro Fishman, Anna Leontjeva and Jaak Vilo

February 27, 2014

In this homework (and further) you are free to use any programming language you like and feel comfortable with.

Task 1

Watch the presentation "How juries are fooled by statistics" by Peter Donnely. List all the mentioned in the presentation common cases of misinterpretation in statistics (HIV, cot death case etc.), also provide the correct interpretations.

Task 2

You are already familiar with the titanic dataset. Use it to form two contingency tables: Sex vs Survival and Age vs Survival. Hint for R-friends: use function *table()*. Calculate from the tables the following (hint: don't be afraid to look through the lecture slides):

- support and interest (lift) for the association patterns {Male, Survived: Yes} and {Adult, Survived: No},
- confidence of the rules {Male} → {Survived: Yes} and {Adult} → {Survived: No}.
- pick at least 2 (one - symmetric and one - non-symmetric) other interestingness measures from the lecture slides (e.g. table on p.102) and apply it on the mentioned association patterns/rules (or if you are bored, pick different ones).
- interpret the results. What is the difference between the measures?

Task 3

Generate 1000 2x2 contingency tables with 1000 elements in each (distributed over f11, f10, f01, f00) so that the cells are skewed (contain more extreme values than uniform distribution). Calculate the Piatetsky-Shapiro, ϕ correlation and J-measure values. Report 2x2 tables with the highest scores according to each measurement.

Task 4

Plot the above three measures values against each other (3 comparisons) and try to characterize how and why the measures are different from each other.

Task 5

Eliminate from the above 1000 tables those with support less than 1%, 10%, 50% - how the comparisons of measures in task 4 changes?

Task 6 (2pt)

Examine the applicability of association rule mining in the following domains ([1]):

- Text documents (e.g. news articles from an online newspaper archive). Each document consists of a collection of words that appear in the document.

- Stock market data (from NASDAQ or Dow Jones Index e.g. `http://www.bloomberg.com/markets/` or `http://finance.yahoo.com/`). The raw data contains the closing price of each stock for the last 5 years. We are only interested in the fluctuations of stock prices (i.e. whether the stock price goes up or down compared to the previous closing price).

- Census data (such as the US Census data which is available at `http://archive.ics.uci.edu/ml/datasets/Census+Income`). The data contains, for each person, information such as age, level of education, relationship, working hours per week, capital gain etc.

Transaction	Item1	Item2	Item3	Item4	...
Basket1	0	1	1	0	...
Basket2	1	0	1	0	...
...	...	...	...	...	...

Table 1: Example of the transaction matrix

In each of the application domain, answer the following questions:

- Give an example of a hypothetical association rule that can be generated from the domain.

- Describe how you would construct the transaction matrix (table 1) in order to derive the example pattern given above. Specifically, what are the baskets and items?
- What are the limitations of your proposed transaction matrix? Specifically, will there be any interesting rules missed out by your choice of transaction matrix?

References

[1] Vipin Kumar, *CSci 8980:Data Mining*. Homework assignment, http://www-users.cs.umn.edu/ han/dmclass/hw7.pdf, 2000.