Data Mining
Homework Assignment #3

Dmytro Fishman, Anna Leontjeva and Jaak Vilo

February 27, 2014

Task 1

Watch the presentation "How juries are fooled by statistics" by Peter Donnely. Find and extract all the mentioned in the presentation common mistakes humans make in interpreting statistics (HIV, cot death case etc.), also provide the correct interpretations.

Task 2

Here comes something like, Download our favorite titanic data set again, form two contingency tables e.g. counting how many... Calculate confidence of the rules {Male} $\rightarrow$ {Survived: Yes} and {Adult} $\rightarrow$ {Survived: No}.

Task 3

Now generate 1000 "random" 2x2 contingency tables for 1000 elements (distributed into f11, f10, f01, f00). Try to make randomness so that the cells are not too evenly distributed but are also likely to contain some more extreme values. Calculate the Piatetsky-Shapiro, Correlation and J-measure values. Identify best 2x2 tables according to your data.

Task 4

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

Task 5

Eliminate from the above 1000 tables those with support less than 1%, 5%, 10%, 20% , 50% - how the comparisons of measures as done in task 5 changes?

Task 6 (2pt)

Some rules do not provide extra knowledge as other rules already contain the information. For example, if there is the rule {Class='2nd', Age='Child'}→ {'Survived'='Yes'}, then the rule {Class='2nd', Age='Child',Sex='Female'}→ {'Survived'='Yes'} is not so informative. Such rules are called 'redundant'. Come up with the definition of the redundancy of the rules. Using the script and the data from task 5 tune default parameters so that there are redundant rules in the output. Next, add the "filter" that outputs only non-redundant rules.