Homework Assignment #1

Dmytro Fishman, Anna Leontjeva and Jaak Vilo

February 13, 2014

Task 1

Use your own experience and knowledge you have gained during the lecture to come up with the problem that can be solved by the means of Data Mining. Describe it. Make an example of the data. What are the features (variables)? How many observations (size of the data) needed to solve it? Explain what is the type of the data (temporal, spatial, stream, etc.), type of your features? What are the techniques and approaches that can be used to solve it?

Task 2

Read Probability theory sections 1, 2 and 3 from MathWiki web-site and solve all the included problems. Be fair and report number of problems you solved correctly.

Task 3

Choose your favorite visualization tool, play around with it and report few fancy pictures produced by this tool, e.g. ggplot2 (in R), Excel (hope you manage to find it yourself), Google Code Playground, GNU plot, matplotlib (for Python) etc.

Task 4

Take a look at Titanic dataset, and its description. Analyze the data using your favorite tool:

- describe types of features
- characterize them (frequency tables, variance, mean and etc.)
- find interesting patterns and visualize them

Task 5

Consider an example of a recommendation engine e.g. Amazon online shop, make an educated guess on how this kind of system is built. Find several weak points in Amazon recommendation engine that you would like to improve, explain.

Task 6 (1pt)

Use the dataset from Task 4 and find a way to visualize a meaningful (interesting) dependence between any three features and explain it, e.g. place together "age", "class" and "survival indicator" on the same chart.