Exercise 2: Data Mining II – Regression

In this exercise we will mainly focus on regression. We will start with a simple example to get handy with the different operators and then apply the methods to a larger data set.

1) Load the auction dataset (you can find it within ILIAS) into your RapidMiner repository. There are three attributes: the price (integer), the age of the clock (integer) and the number of bidders which are/were interested in buying the clock (integer).
Inspect your data and have a look at the plot view (data visualization – Data Mining I – first lecture) – can you make an assumption about the function to predict the price based on one of the variables?

2) Apply a standard linear regression operator to the data set. The dataset describes auction prices for clocks. Interpret the results within the result view of the linear regression model.

3) Next Step is to measure the performance of the regression. As this is a basic classification problem, apply cross-validation to the process. Try to interpret the result (RMSE) for the current case.

4) In the lecture we had, in addition to “simple” linear regression, also a pure interpolation regression (using k-NN), poly-nominal (local and non-local) regressions and model trees (e.g. M5 (weka extension)). Play around with these regression operators and compare the results. (Keep in mind that you learned additional possibilities in the lecture. If you have time, you can try them as well.)

5) Have a look at this dataset: http://archive.ics.uci.edu/ml/datasets/Automobile and import it into your RapidMiner. We now want to predict the price of a car based on the attributes we have here. Apply the methods learned before and also keep in mind the pre-processing what you have learned in the exercise before. In addition you could also try to add attributes – if you have a good idea.

6) Let us get serious: Load the trainings dataset of the Data Mining Cup 2006 into RapidMiner (https://www.data-mining-cup.com/reviews/dmc-2006/). This dataset is originally not designed for a regression analysis, but we will try to predict the GMS (price) for each article. Again, read the description of the data set to import the data correctly. Apply the operators from part 4 and try to find a model. Think about different strategies for attribute selection (M5 Prime, Greedy, etc.) which can potentially improve your results.