Machine Learning
Exercise Sheet 2

Prof. Dr. Dr. Lars Schmidt-Thieme, Osman Akcatepe
Information Systems and Machine Learning Lab (ISMLL)
University of Hildesheim

08. November 2011
Deadline: 15. November

Exercise 1: Linear Regression (6 Points)

a)

Suppose that the following data instances from the example in lecture (gas consumption) are given: $D = \{(1, 6.25), (2, 6), (4, 5.5)\}$. Calculate the target variable $\hat{y}$ for $x = 10$ with the Least Squares method. Let the actual value be $y = 2$. Calculate the error. Interpret the result. Build a graphic with all data instances and plot the squared error for each data point.

b)

What can be done if the linear model does not reflect the distribution of the data? Which problems can occur?

c)

In the lecture it was proven that, for the simple linear regression,

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

minimizes the Residual Sum of Squares (RSS).

Have a look at this proof of that and provide the intermediate steps of the partial derivation. Setting the derivatives to zero is a necessary criterion for the existence of an extremum. Give reasons to why the previous solution is a global minimum.

Exercise 2: R (3 Points)

a)

Read the chapters 2 and 3 of "An Introduction to R". Write 3 examples in R code how vectors can be produced in different ways. Write 3 sentences about objects/classes in R.

b)

Write a linear regression model for the dataset cars, which is integrated in R. Given `cars$speed`, find the estimator for the variable `cars$dist`. Read out the coefficients, plot the data and add the regression line to the plot. Write down the used R code. Remark: You can find the code needed here in Appendix A, which you worked out in the last exercise sheet.

c)

Is this linear model satisfying? How else could the ratio of velocity and breaking distance be modeled?

Exercise 3: Weka (1 Point

Load the dataset `lymph.arff` in Weka. Using the filter `unsupervised/attributes/NominalToBinary`, convert the nominal attributes of the dataset to binary variables and save the data as `lymph-bin.arff`. Compare now the both ARFF files. Which differences do you see there?