Machine Learning
Exercise Sheet 2

Prof. Dr. Dr. Lars Schmidt-Thieme, Martin Wistuba
Information Systems and Machine Learning Lab
University of Hildesheim

November 5th, 2013
Submission until November 12th, 13.00 to wistuba@ismll.de

Exercise 3: Linear Regression (6 Points)

a)

Given are the data instances of the example from the lecture (gas consumption):

$$\mathcal{D} = \{(2, 6), (6, 5), (8, 4.5)\}$$

Estimate the target $\hat{y}$ for $x = 10$ using the method of least squares. The true value is $y = 2$. Estimate the error. Interprete the result. Create a plot of all distances and show for each data point the least square error.

b)

In the lecture was proven for the simple linear regression that

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

minimizes the residual sums of squares (RSS).
Reconsider the proof and provide intermediate steps for the partial derivative.
Setting the derivative to zero is a necessary criterion for the existence of an extremum. Justify that the given solution is a global minimum.

Exercise 4: R (3 Points)

a)

Read Chapter 2 and 3 of „An Introduction to R". In R there are different possibilites to create vectors. Give an example for three of them. Explain in three sentences something about objects and classes in R.

b)

Create a linear regression model for the in R integrated data set `cars`. We want to find a predictor for the variable `cars$dist` given `cars$speed`. Get the coefficients, plot the data and add a regression line. Add the used R code to your solution.
Hint: The needed commands can be found in Appendix A.

c)

Are you satisfied with the linear model? How could you model the relationship between speed and breaking distance instead?

Exercise 5: Weka (1 Point)

Open the data set `lymph.arff` with Weka. Transform the nominal attributes of the data set using the `unsupervised/attributes/NominalToBinary` filter to binary variables and save the data as `lymph-bin.arff`. Compare both ARFF files. What happened?