Machine Learning
Exercise Sheet 3

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

November 4th, 2014
Submission until November 11th, 13.00 to wistuba@ismll.de

Exercise 9: Variable Selection (5 Points)

a) What is the main difference between Akaike Information Criterion (AIC) and Bayes Information Criterion (BIC)?

b) Apply a forward and backward for the linear regression model for one step only. Which variables are added or removed? To fit the linear regression model you can use the tool of your choice (e.g. R using the command lm()). The error measure is the RSS.
The data is already split into train and validation:

$\mathcal{D}^{train}$

x_1	x_2	x_3	y
4	7	-7	18
-7	9	-6	5
2	-2	-2	6
-4	10	-2	0

$\mathcal{D}^{valid}$

x_1	x_2	x_3	y
-5	9	-5	5

Exercise 10: R-Tutorial 3 (5 Points)

Read chapters 4 and 5 of „An Introduction to R“.

a) What are „factors“ in R? How are they created and how can they be used?

b) What is the difference between an array and a vector in R? Mention three operations on arrays and matrices in R.

c) In Exercise 1b) you created a linear regression model for the in R integrated data set `cars`. Now write a program in R that fits a quadratic regression model to this dataset.
You are allowed to use the built-in matrix operations and equation solvers. Submit the code.

Exercise 11: Shooting Algorithm (5 Points)

Learn the L1-regularized Linear Regression with the Shooting Algorithm for the following dataset with $\lambda = 0.1$ and $i_{max} = 2$.

x_1	x_2	x_3	y
3.7	3.8	3.9	4
2.4	2.5	2.3	2
2.2	3.0	4.1	5
3.2	3.1	2.9	3

Exercise 12: Nearest-Neighbor and Kernel Regression (5 Points)

Given is following data set:

x	y	x	y
1	2	6	12
2	4	7	14
3	6	8	16
4	8	9	18
5	10	10	20

a) Predict the target for 0, 2.5 and 5.75 using 2-nearest-neighbor regression

b) The nearest-neighbor regression considers only instances in its neighborhood but does not consider that they might be very far away. Kernel regression is similar to nearest-neighbor regression but the neighborhood does not have a fixed size. Instead, all instances that are close enough contribute to the prediction with a weight defined by its distances. A common prediction function for the kernel regression is

$$\hat{y}(x_0) = \frac{\sum_{(x,y)\in\mathcal{D}_{train}} K(x, x_0)\, y}{\sum_{(x,y)\in\mathcal{D}_{train}} K(x, x_0)}$$

where K is a similarity measure.
Use

$$K(x, x_0) = D\left(\frac{|x - x_0|}{\lambda}\right)$$

with

$$D(t) = \begin{cases} \frac{3}{4}\left(1 - t^2\right) & t < 1 \\ 0 & \text{otherwise} \end{cases}$$

and $\lambda = 2$ to predict the target for 0, 2.5 and 5.75.

c) Plot the data and the models of parts a) and b). Compare the models. Where do you see advantages or disadvantages?