Machine Learning
Exercise Sheet 5

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

November 21st, 2016
Submission until November 28th, 13.00 via learnweb!

Exercise 9: Backward Selection (10 Points)

a) Explain in your words the main difference between the Akaike Information Criterion (AIC) and the Bayes Information Criterion (BIC)!

b) For the following data:

x_1	x_2	y
1	3	6
0	1	6
2	2	3
5	1	-4

a linear regression model given by the parameters

$$\beta = \begin{pmatrix} 4 & -2 & 1 \end{pmatrix}$$

has been learned. Compute its error as well as its AIC by using the negative RSS as logarithm of the likelihood.

$$\log(\mathcal{L}) = -\sum_{i=1}^{N} (y_i - \hat{y}(x_i))^2$$

c) Perform a backward search on the employed variables and compare the three resulting models to the *full* model. Which one do you choose in the end? *Hint:* Use Cramer's rule for inversion of 2×2 matrices:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

Exercise 10: Regularization (10 Points)

a) Explain in your own words why regularization is a key aspect in machine learning!

b) Ridge Regression learns model parameters β by minimizing the following objective function:

$$f(\beta, \lambda) = \sum_{i=1}^{N} (y_i - \beta^\top x_i)^2 + \lambda \|\beta\|_2^2 \qquad \lambda > 0$$

Learn two regression models using the **closed form solution** for $\lambda = 2$ and $\lambda = 5$ for the following data

x	y
2	0
4	-2
-2	4
-1	3

Which model performs better for $x = 5$, where $y = -3$ is the ground truth?

c) Compute the partial derivative $\frac{\partial f}{\partial \lambda}$! Why do we not simply learn λ using gradient descent? Explain what happens if we would.