Machine Learning 2
Exercise Sheet 1

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

April 12th, 2016

Submission until April 19th, 18.00 to schilling@ismll.de

Exercise 1: Exponential Family (5 Points)

a) Show that the binomial $Y \sim \text{Bin}(n, p)$ distribution, where the number of repitions n is known, belongs to the exponential family, i.e. its probability density f has a representation:

$$f(y \mid p) = \binom{n}{y} p^y (1-p)^{(n-y)} = \exp(y \cdot b(p) + c(p) + d(y))$$

b) Then, compute the expected value of Y using:

$$\text{E}(Y) = \frac{-c'(p)}{b'(p)}$$

Exercise 2: Count Data (5 Points)

a) Read Chapter 11.6 in „An Introduction to R“ `http://cran.r-project.org/doc/manuals/R-intro.pdf` to understand how to use GLMs in R.

b) Given is the following data:

x_i	1	2	3	4	5	6	7	8	9	10	11	12	13	14
y_i	0	1	2	3	1	4	9	18	23	31	20	25	37	45

with y_i number of deaths and x_i is a time point.
Compare the Linear regression model with the Poisson regression model for this data, plot both models and the data.

c) Compare the linear regression model on the small real world data set „ceb.txt“. The last column in the data set is the target.

Estimate the RMSE in a 2-fold cross-validation. Do not forget to shuffle the data. Which of the models performs better? More information about the data set can be found on `http://data.princeton.edu/wws509/datasets/#ceb`.