Data Mining and Machine Learning in Bioinformatics

Dr. Holger Fröhlich, Dr. Martin Vogt
Sabyasachi Patjoshi
sabyasachi2k13@gmail.com, martin.vogt@bit.uni-bonn.de

Due: Jun 3, 10:30 (by the end of the lecture)

Exercise Series 5

General: Exercises are to be solved and submitted in fixed groups by at most 3 students. Every member of a group should help solving each task and thus be able to answer questions to each task. No late submissions are accepted. Copying solutions will automatically lead to a point reduction of at least 50%. N – 1 homework assignments and N – 2 programming tasks have to be submitted in total.
A group can gain additional bonus points, if it presents its solution for a particular task during the tutorials. Accordingly, each task has a defined number of points as well as bonus points.

1. The following table lists the duration of pregnancy for different species together with their expected life time in years.
   a) Determine the coefficients of the linear regression, which describes the expected life time as a function of the duration of pregnancy **manually**. (3 points + 1 bonus point)
   b) Compute the residual variance. (2 points + 1 bonus point)
   c) Compute the standard error of the slope coefficient β and manually test the hypothesis $H_0: \beta = 0$. (3 points + 1 bonus point)
   d) Compute the 95% confidence interval for the slope coefficient. (3 points + 1 bonus point)

Species	**Pregnancy (weeks)**	**Expected life time**
Lemur	18	18
Macaque	24	26
Gibbon	30	30
Chimpanze	34	40
Human	40	70

2. Consider the iris dataset introduced in the first exercise. Use R to fit a **logistic regression** model to separate the three plant types based on the given four predictor variables. Analyse and discuss the quality of fit and the significance of each of the predictor variables. (4 points + 1 bonus point)

3. We now want to investigate whether one of the predictors can be expressed as a linear combinatioin of the others. Fit **linear regression** models for this purpose and analyse their quality of fit. To which conclusions do you come? (5 points + 1 bonus point)

4. A gene is measured under 3 different stimulation conditions in 2 different cell lines in triplicates. The data looks as shown in the following table:

Cell line	**Stim. 1**	**Stim. 2**	**Stim. 3**
A	3.3	1.2	3.2
A	2.3	0.9	4.0
A	2.5	1.5	2.7
B	1.3	1.5	3
B	2	0.7	3.5
B	1.5	1.8	3.3

a) Fit an appropriate model to these data using two factors, one for the cell line and one for the stimulus. Analyse the model fit, also graphically. (6 points + 1 bonus point)
b) Perform a (two-way) ANOVA and interpret the results. (4 points + 1 bonus point)