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: May 6, 10:30 (by the end of the lecture)

Exercise Series 2

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. Continue investigating the iris data set, introduced in exercise series 1.

a) Generate boxplots visualizing the distribution of values for each of the variables sepal length, sepal width, petal length and petal height. The boxplots should be plotted separately for each of the 3 species classes. (4 points + 1 bonus)

b) As a possible measure of distribution skewness (so-called *non-parametric skew*) one may consider the ratio:

$$S = \frac{\mu - \theta}{\sigma}$$

where μ is the arithmetic mean, θ the median and σ the standard deviation.

Compute the distribution skewness for each of the variables sepal length, sepal width, petal length and petal height. Explain the results. (4 points + 1 bonus)

c) Calculate Pearson and Spearman rank correlations between each pair of variables, yielding a so-called correlation matrix. Do this once using the whole data and once using each of the 3 species classes separately. Which differences do you observe? What could be a possible explanation? (6 points + 1 bonus)

d) We now want to visualize the correlation matrices calculated in c) via heatmaps. The color code indicates the strength of the correlation. R-package `lattice` provides a function `levelplot` for that purpose.

Draw heatmaps for each of the correlation matrices using the `heat.colors` color scheme. Learn, how to save the obtained figures into a common pdf-file and interpret the results. (6 points + 1 bonus)

2. Consider the following four datasets.
(available for download as `exercise2-2.csv`):

I		II		III		IV	
x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

a) For each dataset:
i. Calculate the mean and variance for x and y
ii. Calculate the correlation between x and y
iii. Linear regression: The linear regression line is a linear function that 'explains' the dependent variable y in terms of x by a linear function $f(x)$ that fits the real data values y as good as possible, i.e. it looks for a line so that the $f(x_i)$ are a "good" approximation for the given data values y_i . (Linear regression will be covered in the lecture in more detail at a later stage). It is given by the formula $f(x) = a + bx$, where $b = \rho_{xy}\frac{s_x}{s_y}$ and $a = \bar{y} - b\bar{x}$. Here, ρ_{xy} is the Pearson correlation between x and y . s_x and s_y are the square roots of the variances of x and y. Determine b and a for each data set.
(3 points + 1 bonus)

b) i. Now might be a good idea to look at the data: For each data set plot the data in a scatter plot and add the regression line to the scatter plot.
ii. In which cases does the linear regression line give a good fit to the data? On the basis of the scatter plots, describe the relationship/dependence between x and y for each data set. (5 points + 1 bonus)