Jun.-Prof. Dr. Thomas Schultz
Michael Ankele (ankele@cs.uni-bonn.de)
Amin Abbasloo (abbasloo@informatik.uni-bonn.de)

University of Bonn
Institute of Computer Science II
November 2, 2015

Winter term 2015/16

Bioinformatics II
Assignment Sheet 2

If you have questions concerning the exercises, please write to our mailing list: vl-bioinf@lists.iai.uni-bonn.de.

There will be two practical projects, one on visualization of multi-dimensional data, the other one on image processing. Subtasks will be given out every week, and we strongly encourage you to continuously work on them and contact us with questions. However, you will only have to hand in your results (for all sheets of the first project) on December 1.

Exercise 1 (Read, Write, and Filter Data, *25 Points*)

In the first project, we will work with a Mice Protein Expression Dataset, which contains expression levels of 77 proteins, measured in the cerebral cortex of 8 classes of mice. The classes result from two genotypes (Ts65Dn, which serves as a mouse model of human down syndrome, vs. normal controls), two treatments (injection of the drug memantine vs. a saline solution as a control), and two experimental conditions related to context fear conditioning (context-shock, which should lead to learning, vs. shock-context, in which no learning takes place). Counting all repeated measurements, there are 1080 instances overall, some with missing data. You can find more information on the data in the corresponding scientific publication.

Please proceed in the following steps and submit your final script. You will also need its results for the next stage of the project.

a) Read the dataset and print the number of instances and columns, as well as the column names, to the terminal (3P).

b) Interpolate the missing values in a sensible way (3P).

c) Extract subgroups t-CS-sal, c-CS-sal, t-CS-mem, c-CS-mem and print the number of instances for each subgroup (5P).

d) The F score is one way to determine how well a given variable distinguishes between two groups. F is large if the differences of the two groups means $\bar{x}_1$ and $\bar{x}_2$ to the grand mean $\bar{x}$ of all data points is large relative to the variances within the groups. Given groups of size n_1 and n_2 with items $x_{1,i}$ and $x_{2,i}$, respectively, F can be defined as

$$F = \frac{(\bar{x}_1 - \bar{x})^2 + (\bar{x}_2 - \bar{x})^2}{\frac{1}{n_1-1}\sum_{i=1}^{n_1}(x_{1,i} - \bar{x}_1)^2 + \frac{1}{n_2-1}\sum_{i=1}^{n_2}(x_{2,i} - \bar{x}_2)^2}$$

Define a function that calculates F for any given protein and pair of class labels (8P). Use it to identify the five proteins that best separate the classes t-CS-sal vs. c-CS-sal (3P).

e) Write a reduced dataset to disk. It should only contain the four classes from c), only the five most relevant proteins from d), and the interpolated replacements of missing values from b) (3P).

Hint: You can use pandas, a powerful Python data analysis toolkit, for this assignment. It provides fast, flexible, and expressive data structures for working with relational or labeled data. To become familar with it, you can refer to `http://pandas.pydata.org/pandas-docs/stable/10min.html`.

Good Luck!