Jun.-Prof. Dr. Thomas Schultz
Michael Ankele (ankele@cs.uni-bonn.de)
Shekoufeh Gorgi Zadeh (gorgi@cs.uni-bonn.de)

University of Bonn
Institute of Computer Science II
November 2, 2016

Winter term 2016/17

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 6.

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

In the first project, we will work with a Chronic Kidney Disease (CKD) Dataset, which contains an experiment on 400 people, 250 of them having CKD and 150 are healthy. In the experiment, on each instance, 24 variables are measured such as age, blood pressure, sugar etc. Note that there are some instances with missing data. For more details on attributes and the experiment read the `chronic_kidney_disease.info.txt` file.

Please download the `chronic_kidney_disease_full.xls` file and 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) Write a piece of code to replace nominal attribute values with 0 and 1. For instance, for the "rbc" and "pc" columns, map "normal" to 1 and "abnormal" to 0. For the rest, map "present" to 1 and "notpresent" to 0, "yes" to 1 and "no" to 0, "good" to 1 and "poor" to 0, "ckd" to 1 and "notckd" to 0 (3P).

c) Interpolate the missing values in a sensible way (2P).

d) Extract subgroups with CKD and without CKD and print the number of instances for each subgroup (3P).

e) 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 attribute and for the two CKD and not CKD class labels (8P). Use it to identify the five attributes that best separate the classes CKD vs. not CKD (3P).

f) Write a reduced dataset to disk. It should only contain the five most relevant attributes from d), and the interpolated replacements of missing values from c) (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!