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 15, 2016

Winter term 2016/17

Bioinformatics II

Assignment Sheet 3

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

We strongly encourage you to continuously work on the assignments 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 (Producing a Scatterplot Matrix, *25 Points*)

In the previous assignment, you wrote a reduced dataset to disk that is limited to the CKD and notCKD classes and five variables that distinguish most strongly between subjects with and without chronic kidney disease. This week, you will create and interpret a basic visualization of that data.

Your final visualization should be a 5×5 matrix whose rows and columns are the measurements of the variables you selected last week. Diagonal cells visualize how the variables are distributed; off-diagonal cells visualize the relationship between the values of pairs of variables.

Please proceed in the following steps and submit your final script, the final image, and answers to the questions:

a) Each diagonal cell should contain two overlaid density plots, one for the CKD and one for the not CKD class. In the density plot, variable values should be on the x axis, the frequency of observing that value in each class should be on the y axis. Use different colors to distinguish between the classes, and add a legend. Your visual design should make it easy to answer the following questions (5P for implementation, 1P for justifying choice of colors, 3P for answering questions):
   - For which variable you could devide the range of the values to three subranges where each indicate either healthy, with CKD or uncertain? Roughly list the subranges.
   - Which variable(s) is(are) almost always constant for the healthy subjects?

b) In each non-diagonal cell, display a scatter plot that visualizes the values of the corresponding pair of variables. Use different colors, opacities and markers so that it is simple to relate these scatter plots to the density plots on the diagonal, and the size of the marker reflects the number of overlapping points. (5P for implementation, 2P for answering questions):
   - Point out a pair of variables whose values have a clear positive correlation overall.
   - Can you identify a pair of variables for which the values are highly correlated in one group of subjects (e.g. healthy), but less so in the other group?

c) Compute the distance consistency of all scatter plots. Which pair of variables leads to the highest distance consistency? (6P)

d) Imagine that, given only the values of two variables, you will be asked to decide whether they are from a healthy subject, or a person with chronic kidney disease. Which pair of variables would you choose to make that decision? Modify the visualization to best answer the question, and justify both your answer and your modification. (3P)

Hint: You can use the Python toolkit matplotlib to create plots. More information on it is available from http://matplotlib.org/.

Good Luck!