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 21, 2016

Winter term 2016/17

# Bioinformatics II

**Assignment Sheet 4**

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 (Principal Component Analysis, *25 Points*)

It is difficult to fully visualize a very high-dimensional space. In the previous two assignments, we therefore focused on a few variables that we found to be particularly discriminative. This week, we will instead employ dimensionality reduction on the values of all variables.

a) Perform a Principal Component Analysis (PCA) on the values.
Please download the new `chronic_kidney_disease_full.xls` file, which is uploaded with this exercise sheet. Interpolate missing values, and keep all variables this time. Make a plot that, for any number $n$, shows what fraction of the overall variance in the data is contained in the first $n$ principal components. How many components do we need to cover $\geq 99.99\%$ of the variance? (5P)
*Hint:* You may use the implementation of PCA that is provided in the Python package scikit-learn.

b) Each sample is now characterized by a point in PCA space. Create a scatter plot matrix (in the same manner as in the previous sheet) that shows the first five principal components. In which PCA modes do you see a clear difference between the healthy subjects and the people with CKD, in which modes the difference is less? (4P)

c) In the second PCA mode, you should see a clear cluster of outliers, a group of points that belong together, but are quite far away from the rest of the data. Provide a list of all subject-IDs that form that cluster and then remove the outliers. Use the index of subject's row as the subject-ID. (5P)

d) The results of our Principal Component Analysis are more strongly affected by changes in the values of variables that are very large overall than by others with lower overall values. Account for this by computing the mean value for each variable. Then, replace each value with a factor that describes its deviation from the respective average. To do so, normalize the values of each variable between 0 and 1, then subtract the average from all the values of that variable. Observe how this affects the PCA. How does the number of components needed to cover $\geq 99.99\%$ of the variance change? (4P)

e) See what happens when we re-weight the variables to emphasize those that discriminate well between classes CKD and not CKD. To do so, first normalize as in d), then compute $F$ scores (cf. sheet 2, task 1 e)) and multiply each data value by its corresponding $F$ score. Create two scatter plots to compare PCA results with and without the re-weighting. (4P)

f) By using the PCA plot from task e), could you say whether using the average value of each variable for interpolating the missing values of that variable is a good way of interpolation? (3P)

**Good Luck!**