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 16, 2015

Winter term 2015/16

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 1.

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 proteins that we found to be particularly discriminative. This week, we will instead employ dimensionality reduction on the expression levels of all proteins.

a) Perform a Principal Component Analysis (PCA) on the expression levels. To this end, return to the original dataset. Interpolate missing values and keep the same four classes t-CS-sal, c-CS-sal, t-CS-mem, c-CS-mem as previously, but keep all proteins 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. (5P)

   How many components do we need to cover $\geq 95\%$ of the variance? (1P)

   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 effect of the treatment, which modes are less affected? (4P)

c) In the third 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 MouseIDs that form that cluster. (5P)

d) In part c), you should have found that the outliers correspond to samples from a single mouse which exhibited anomalous expression patterns. Exclude all data from this mouse and repeat the PCA analysis. In the resulting scatter plot matrix, identify other clusters that appear to correspond to samples from individual mice. Verify your hypothesis and highlight at least two such clusters. The simplest way to do so is to circle them in an image processing program. (5P)

e) The results of our Principal Component Analysis are more strongly affected by changes in the expression levels of proteins that are strongly expressed overall than by others with weak overall levels. Account for this by computing the baseline expression level for each protein (as the average over the “normal” group c-CS-sal). Then, replace each expression level with a factor that describes its deviation from the respective baseline. Observe how this affects the PCA. How does the number of components needed to cover $\geq 95\%$ of the variance change? (5P)

f) See what happens when we re-weight the proteins to emphasize those that discriminate well between classes t-CS-sal and c-CS-sal. To do so, first normalize as in e), then compute F scores (cf. sheet 2, task 1 d)) and multiply each data value by its corresponding F score. Create two scatter plots to compare PCA results with and without the re-weighting. Each plot should show the first and second mode of the respective PCA space. To avoid distortions, make sure that both plots use the aspect ratio setting "equal". In the re-weighted plot, you should now see that the trisomic mice that were treated with memantine are closer to the control mice than the trisomic mice that were treated with saline. (5P)

Good Luck!