Data Mining and Machine Learning in Bioinformatics

Dr. Holger Fröhlich, Dr. Martin Vogt
Sabyasachi Patjoshi
sabyasachi2k13@gmail.com, martin.vogt@bit.uni-bonn.de

Due: Jun 10, 10:30 (by the end of the lecture)

Exercise Series 6

General: Exercises are to be solved and submitted in fixed groups by at most 3 students. Every member of a group should help solving each task and thus be able to answer questions to each task. No late submissions are accepted. Copying solutions will automatically lead to a point reduction of at least 50%. N – 1 homework assignments and N – 2 programming tasks have to be submitted in total.
A group can gain additional bonus points, if it presents its solution for a particular task during the tutorials. Accordingly, each task has a defined number of points as well as bonus points.

1. Install package colonCA from the bioconductor homepage (www.bioconductor.org) containing gene expression measurements from 22 normal and 40 colon cancer patients (Alon et al., 1999). Read the manual pages for R-package colonCA and further documentation on the web to see, how you can work with ExpressionSet objects.
a) A typical first question when looking at such data is to understand the relationship of all 62 samples to each other. One popular method for that purpose is the Principal Component Analysis (PCA). Perform a PCA of these data using method `prcomp` in R, after applying a log-transformation (i.e. A ← log(A), if A is the original expression matrix). Caution: the original expression data is in the format genes x samples! (4 + 1 points)
b) Conduct a 2D PCA plot. In that plot mark in cancer and normal patients with a different color (Tip: `pData(colonCA)$class` indicates the class of each patient). What do you observe? (4 + 1 points)
c) Draw a scree plot of the eigenvalues and interpret the results. (2 + 1 points)
d) Which proportion of the overall variance do the first 2 principal components explain? How many principal components would you need to explain 90% and 95% of the overall variance? (4 + 1 points)

2. Cluster the following 6 samples on paper using average linkage hierarchical clustering (UPGMA): A1 = (0; 10); A2 = (2; 2); A3 = (2; 4); A4 = (8; 8); A5 = (7; 6); A6 = (10; 9). Plot the resulting dendrogram. For a working example of the method see: http://en.wikipedia.org/wiki/UPGMA (6 points + 1 bonus point)

3. Use again the colonCA dataset. **(Postponed to next week)**
a) For each gene perform an unpaired t-test to see whether it is differentially expressed between normal and cancer patients. Consider all genes with p-value <= 0.0001 as significant. (2 points + 1 bonus point)
b) Install R-package EMA and produce a heatmap of the data sub-matrix containing only these differentially expressed genes. For clustering use complete linkage, average linkage and Ward's method using the Pearson correlation distance. Do the results differ? (4 points + 1 bonus point)