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 24, 10:30 (by the end of the lecture)

Exercise Series 8

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 the package `ConsensusClusterPlus` from `Bioconductor` and install the `GSVAdata` library from `Bioconductor` and get the `gbm_eset` expression data from `gbm_VerhaakEtAl`.
a) Prioritize the gene expressions based on their median absolute deviation (MAD) and select the 2000 genes showing the highest deviation to be used in the following. (2 points + 1 bonus point)
b) Perform **consensus clustering** of the samples using 80% item resampling, 80% gene resampling, a maximum number of 6 clusters, and a total of 100 resamplings. Generate heatmaps of the **consensus matrices**. Can you deduce something about the quality of the clusterings? (3 points + 1 bonus point).
c) Show and interpret plots for the i) **cluster consensus** and ii) **item consensus**. (3 points + 1 bonus point)
d) Show the **empirical cumulative distribution curves** (empirical CDF) for two to six clusters. How can you use these curves to determine a good number of clusters? Make a sketch how an ideal empirical CDF would look like. (3 points + 1 bonus point).

2. Make sure that the NMF package for **non-negative matrix factorization** is installed. The `GSVAdata` also contains some `leukemia` data. Get the leukemia_eset of gene expressions.
a) Prioritize the gene expressions based on their median absolute deviation (MAD) and select the 2000 genes showing the highest deviation to be used in the following . (1 point)
b) Perform a **non-negative matrix factorizations** (NMF) of rank 2 and 3 using the Brunet method with 50 repetitions. Plot **heatmaps of the coefficients** and **consensus matrices** and interpret the result. (6 points + 2 bonus points)