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 17, 10:30 (by the end of the lecture)

Exercise Series 7

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.

The assignment focuses **clustering** and the **colonCA** data set introduced last week. Be sure to use the log-normalized expression data in the assignments.

1.
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 differentially expressed. (2 points + 1 bonus point)
b) Install R-package EMA and produce heatmaps of the data sub-matrix containing only these differentially expressed genes. For **clustering** of genes use complete linkage, average linkage and Ward's method using the **Pearson correlation "distance"**. Do the results differ for clustering genes differ? (4 points + 1 bonus point)
c) Why is the use of the Pearson correlation more appropriate than the Euclidean distance in this case? (2 points + 1 bonus point)

2. **Gaussian mixture models** (GMM)
a) A GMM can be described as

$$p(x) = \sum_{k=1}^{K} \pi_k N(x \mid \mu_k, \Sigma_k) \text{ with } \sum_k \pi_k = 1, \pi_k \in [0,1]$$

Assuming we like to cluster patients in one of two categories on the basis of expressions from 25 genes. The complexity of the model can be controlled by restricting how the covariance matrices Σ_k are allowed to vary:
i) How many parameters (degrees of freedom) does the most general model possess, i.e., with no restrictions posed on the covariances?
ii) Assuming that all normal distributions are 'diagonal', i.e., there is no correlation between the expression of different genes for each Gaussian (Σ_k is a diagonal matrix), how many parameters does this model possess?

iii) Like ii) assume no correlation but additionally the variation for each gene is the same for each Gaussian, i.e., the Gaussians are spherical. How many parameters does this model possess?
(3 points + 1 bonus point)

Now install the R-package mclust and get familiar with its functionality. The model based clustering Mclust will apply different GMM models depending on the restrictions on the covariance matrices Σ_k. (For the models of a) the abbreviations used in mclust are i) VVV, ii) VVI, and iii) VII.) An optimal model is then selected based on the BIC.

b) Again, using only the significant genes according to 1a), apply GMM based clustering to

i) patients based on their gene expression profiles
ii) differentially expressed genes based on their profiles across patients

What is the optimal number of clusters (according to the BIC criterion) in both cases, if you vary the number the cluster number between 2 and 10? How do mixing proportions look like?
(4 points + 1 bonus point)

c) Now, standardize the gene expressions for each gene separately by unit variance scaling: $x_i' = \frac{x_i - \bar{x}}{\sigma}$, where $\bar{x}$ is the mean of a gene expression across all patients and σ is the standard deviation. Again, perform GMM based clustering using the standardized values for i) patients and ii) genes and report the optimal number of clusters and mixing proportions.
Perform plots of the BIC criterion, and display the classification of data points resulting from the best model (for both, i) and ii)).
(6 points + 2 bonus points)

d) Comparing the clusterings of b) and c). Do you observe large differences in the clusterings for either the patient or the gene clustering? Give reasons why or why not the data should be standardized. (4 points +1 bonus point)

e) Now consider only the patient clustering (of either b) or c)) and the known grouping into normal and cancer patients. Plot the cluster silhouettes and interpret the result. (3 points + 1 bonus point)

3. Consider a **k-means** clustering.
a) Which behavior would you expect, if you plot the **distortion** in dependency of the number of clusters? (2 + 1 points)
b) Consider k relatively well separated clusters, which can be embedded into spheres. Would GMM clustering yield a very different result than k-means? (2 + 1points)

4. Which of the clustering algorithms discussed in the lecture so far can you apply to cluster biological sequences? Give reason for your answer! (2 + 1 points)