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: May 27, 10:30 (by the end of the lecture)

Exercise Series 4

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. The expression of gene EGR1 is measured with sample mean 4 and sample standard deviation of 0.7 using 5 data points. We assume data to be drawn from a normal distribution N(μ, σ), which is unknown.
   a) Compute the **standard error of the mean** (1 point + 1 bonus point).
   b) Based on the sample mean and the standard error we ask, in which range the expectation value μ will lie with high probability, given sample size n=5. In particular, the 1 – α **confidence interval** for μ is given by:

$$\left[\bar{x} \pm se(\bar{x}) t_{1-\frac{\alpha}{2}}(n-1)\right]$$

   Here $t_{1-\frac{\alpha}{2}}(n-1)$ denotes the 1 – $\alpha/2$ quantile of a t-distribution with n – 1 degrees of freedom.
   Compute the 95% confidence interval for the expectation value of EGR1 measurements. **Tip**: use R-function qt to compute t-distribution quantiles (2 points + 1 bonus point)
   c) Assume that in order to get a more confident estimate of the mean we would like to half the size of the confidence interval. Roughly, how many data points would we need? (2 points + 1 bonus point)

2. Here, let us consider the gene expression data of the Golub data set. It is part of the **multtest** package of R and can be loaded using `library(multtest);data(golub)`.
   It contains gene expression data of 3051 genes from 38 tumor mRNA samples. Read the help information of R for more information on this data set. The expression data is organized in a matrix where rows correspond to genes and columns to samples. The tumor class of the columns is given in the vector `golub.cl`. The names of the genes (rows) are given in `golub.gnames`.
   a) Calculate the mean and variance of **all pooled expression data** of the `golub` matrix. Explain the result. (Hint: You may want to check the help information) (2 points + 1 bonus point)
   b) Determine the means and standard deviations of the expression level for every gene for the classes ALL and AML.
      i. `golub.cl` identifies the tumor class by 0 for acute lymphoblastic leukemia (ALL) and 1 for acute myeloid leukemia (AML). Convert `golub.cl` to a vector of factors `golub.fac` with levels ALL and AML. (You can use golub.fac instead of golub.cl in the following to make your code clearer.)
      ii. Determine the 5 genes with the largest mean expression for AML and ALL.
      iii. Determine the 5 genes with the largest mean expression for AML and ALL that are known oncogenes. (Hint: The oncogene information is given as part of the gene name)
      iv. Determine the 5 genes with the largest difference in expression between the two classes.

      and write their names, means, and standard deviations to a csv file.
      (6 points + 2 bonus points)
   c) For the moment, we are interested in the expression data of the single gene in row 1042 (CCND3 Cyclin D3).
      i. Make a boxplot for the expression data of this gene grouped by tumor class.
      ii. Make Q-Q plots for ALL and AML comparing the distributions to a normal distribution. Use `qqline` to add a theoretical normal distribution to the plot.
      iii. We want to investigate whether the means of two sample distributions are different. To this end, assume the distributions for ALL and AML are essentially normal. Which test/test variant should we apply? Use R to perform the test and interpret the result.
      iv. Perform an appropriate non-parametric test as an alternative to the test performed in iv. Do the conclusions differ?

      (6 points + 2 bonus points)
   d) Perform Student t-tests for all genes comparing the distributions for ALL and AML.
      i. How many genes show significant differences of the mean at the p=0.05 level?

ii. Here, we performed more than 3000 tests. If we assume that all genes for both classes were actually drawn independently from identical distributions (i.e., in each case the null hypothesis is true), how often can you expect to falsely reject the null hypothesis, i.e., commit a type I error, given a p-level of 0.05 just by chance in 3000 t-tests?
iii. Use the **Bonferroni correction** (you can read up on it on Wikipedia, for instance) to adjust the p=0.05 level. How many genes show significant differences now?

(6 points + 2 bonus points)

3. Implement the permutation test for testing the significance of the correlation between species richness and lake area. Choose N=1000 permutations. **Tip:** `sample(1:n)` generates a random permutation of n numbers. (4 points + 1 bonus point)

Species Richness	Lake Area
32	2.0
29	0.9
35	3.1
36	3.0
41	3.0