Processing of Biological Data

Prof. Dr. Volkhard Helms
Summer Semester 2017

Saarland University
Chair for Computational Biology

Exercise Sheet 5

Due: July 11, 2017 10:15

Submit your solutions on paper, hand-written or printed at the *beginning* of the lecture or in building E2 1, Room 3.03. Alternatively you may send an email with a single PDF attachment. Additionally hand in all source code via mail to maryam.nazarieh@bioinformatik.uni-saarland.de.

Gene Expression Prediction

Exercise 5.1: Read data and Normalization (25 points)

The data given in the supplementary comprises of two already data sets for gene expression and histone modification of a mouse cell. The data is divided into two sets of training data and test data. In this assignment, we aim to predict gene expression based on histone modification.

- Read the data into a data matrix where the rows correspond to the set of genes in each sample and columns correspond to the different samples.
- Filter the data, for both expression and methylation data, by removing entries with empty/NA expression and methylation values. If there are several entries with the same gene name, substitute the rows by taking the average mean expression and methylation value for each gene in every sample.
- Submit the final matrices as your solution.

Exercise 5.2: Model Prediction (50 points)

Linear regression is a method for modelling the relationship between a dependent or response variable y and one or more explanatory variables denoted by X. The model comprises a linear combination of the parameters.

$$Y = \alpha + \beta X \tag{1}$$

The above formula describes a line with slope β and y-intercept α. Linear regression models are often fitted using the least square approach, where the least square error is computed by the sum of the squares of the differences between the response variable y in the traing data and the best-fit regression line as follow:

$$SSE = \sum_{i=1}^{n} (Y_i - Y_i)^2 \tag{2}$$

Your task in this assignment is to build a linear regression model from training data (gene expression and histne modification) to predict the gene expression from histone modification in the test data. (You can write a program that calculates best fitted α and β or alternatively use one of the packages in python or R to do the task.)

- Quantify the strength of the relationship between Y and each of the explanatory variables.
- Determine which variables have no relationship with Y at all.
- Identify which subsets of the X contain redundant information about Y.

Exercise 5.3: Performance Measurement (25 points)

ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. In ROC curve the sensitivity is plotted in function of the $(1 - specificity)$ for different cut-off points of a parameter. Your task is to evaluate the classifier in the second exercise with an ROC curve.

- Plot the ROC curve for the classifier.
- Provide the area under the curve (AUC) for the classifier.
- Provide the sensitivity and specificity for a chosen cutoff of probability=0.5.

Have fun!