FIB UPC UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

KMLMM course.
Project 1: Leukemia classification

Course: 2018-2019
Prof. Tomàs Aluja

Leukemia data

Problem: Although cancer classification has improved over the past 30 years, there has been no general approach for identifying new cancer classes (class discovery) or for assigning tumors to known classes (class prediction). The problem focuses in finding a classifier using PLS1 regression approach on gene expression monitoring by DNA microarrays, to automatically differentiate between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) (Golub et al "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring").

Available data: There are two datasets containing the initial training dataset with 38 samples ("data_set_ALL_AML_train.csv"), and a test dataset with 34 samples ("data_set_ALL_AML_independent.csv").

In addition we have the responses in "table_ALL_AML_predic.doc", altogether with the prediction obtained in the aforementioned paper.

These datasets contain measurements corresponding to ALL and AML samples from Bone Marrow and Peripheral Blood. Details about the experimental method and protocol can be found in Experimental_protocol.html. Intensity values have been re-scaled such that overall intensities for each chip (sample) are equivalent. This is done by fitting a linear regression model using the intensities of all genes with "P" (present) calls in both the first sample (baseline) and each of the other samples. The inverse of the "slope" of the linear regression line becomes the (multiplicative) re-scaling factor for the current sample. This is done for every chip in the dataset except the baseline, which gets a re-scaling factor of one.

Steps to guide the practical work:

1. Read the data files "data_set_ALL_AML_train.csv" and "data_set_ALL_AML_independent.csv". Enter the response manually into R

from the “table_ALL_AML_predic.doc” document (the response corresponds to the “actual” column of such word document).

2. Form the data matrices X and Xt, containing the gene expression for the 38 training samples and 34 test samples. Be aware that data is presented in its transposed form, with instances as columns and they are not in order. Only numeric information is pertinent to solve the problem.

3. Perform the PLS1 regression of the training data. Select the number of PLS1 components.

4. Project the test data as supplementary individuals onto the selected PLS1 components (be aware of centering the test data respect to the mean of the training data).

5. Perform a joint plot of the train and test individuals in the plane of the two first PLS1 components, differentiating those individuals with ALL from those with AML.

6. Obtain the logistic regression model (or any other model of your choice) to predict the response in the training data, using as input the selected PLS1 components.

7. Predict the probability of AML leukemia in the test sample.