INFO-F-528

Machine learning for bioinformatics

Project 2014-15

Gianluca Bontempi, Claudio Reggiani

Computer Science Department, ULB

1 Introduction

The project counts for roughly 50% of your grade. This project is an individual homework. It shall be completed independently. and it shall represent the sole efforts of the individual submitting the assignment. The result of another student's efforts, or the copy of another student's efforts (current, or past, semester(s)), is considered academic dishonesty.

2 Goal

The goals of the project are

- to implement and assess different classification algorithms and different methods of feature selection in a microarray classification task,
- to select among the learning and feature selection techniques the ones which appears to be the most accurate and use them for predicting the classes of a test set.

3 Dataset

The dataset comes from a collection of gene expression data from thousands of tumor samples. The goal is the classification of *Kidney* against *Other* tumor types, on the basis of gene expression.

The original dataset has 200 samples, one for each patient of the study, and each sample is described by 50 genes. For the purpose of this project the original dataset has been randomly split into two parts: each one made of 100 samples.

The resulting datasets are contained in *Project1415.Rdata*. The file contains two R data frame objects

- *Kidney.train* with 100 samples described using 50 gene expressions and 1 class variables. This should be used for the model building.
- *Kidney.test* with 100 samples described using 50 gene expressions. It containing the inputs for which a class has to be predicted.

4 Specifications

The student has to choose a learning method and a feature selection method among at least three alternatives. The project report has to specify and justify (with tables, figures) the selection procedure which led to the final choice. The student has to return, together with the code and the report, the set of class predictions for the test set. The accuracy test prediction will be assessed in terms of balanced error rate and AUC.

Your code could include and use only the following packages: ROCR, rpart, tree, nnet, e1071, class, MASS, randomForest. At the same time, keep in mind that is possible to reuse the code explained in the TPs.

5 Report and Deadline

The project must include

- the entire **commented R code source**.
- a **readme** file with the instructons about how to execute the code.
- a pdf file **report.pdf** containing a detailed description of the analysis procedure which led to the to choice of the learning algorithm.
- the list of predictions for the 100 patients of the test set contained into a file **predictions.txt** with a prediction per line.
- the list of the indices of the 20 most relevant features (a file **feature.txt** with an index per line).

The project, compressed in a zip file, must be sent before Sunday the 4th of January 2015 at 1pm by an email with the Subject : INFOF258 project to Gianluca Bontempi (email: gbonte@ulb.ac.be) and cc to Claudio Reggiani (email : claudio.reggiani@ulb.ac.be).