# MESIO

April, 2019

# Statistical Learning

# Task 1. Ensemble Methods

### Task description

Gene expression profiling using cDNA microarrays has become a very popular way of studying diseases. In this task, we analyze data from microarray experiments (Khan et al., 2001) on the small, round, blue-cell tumors (SRBCTs) of childhood, which include the distinct diagnostic categories of neuroblastoma (NB), rhabdomyosarcoma (RMS), non–Hodgkin lymphoma (NML), and the Ewing family of tumors (EWS). SRBCTs are so-named because of their similar appearance on routine histology; they often masquerade as each other, making correct clinical diagnosis difficult. The data initially consisted of 63 cases (23 EWS, 8 BL, 12 NB, and 20 RMS) of both tumor biopsy material and cell lines measured on microarrays containing 6567 genes. Requiring that each gene should have a certain minimal level of intensity reduced the number of genes to 2309. SRBCT data can be loaded from `srbct` data in `mixOmics` package from Bioconductor.

Questions:

1. Use Leave-One-Out cross-validation in order to determine the better combination of the random forest parameters `ntree` and `mtry`. To this end implement a two dimensional grid search approach. Explore LOOCV error for (`ntree`,`mtry`) values ranging in $\{5, 25, 100, 400\} \times \{2, 10, 50\}$. Represent the LOOCV error values attained on the grid using a heat color map.

2. Run random forests in 100 bootstrap samples of SRBCT data and setting `ntree` and `mtry` parameters equal to the optimal values according with the previous point. Save the 30 most-important variables you find at each bootstrap loop.

3. You should see different sets of variables being ranked as the 30 most important for each loop. Implement a procedure for visualizing the overall ranking of the variables and the stability of ranks across the loops.

4. Apply the discrete AdaBoost algorithm (with an exponential loss function, and number of trees equal to 1000) in SRBCT data. Compare the LOOCV error attained by different ensemble classifiers based on trees of sizes: stumps, 4-node trees, 8-node trees, and 16-node trees.

**Important remarks**

- You should deliver an R markdown (or R latex) dynamic report as dynamic as you can (see TreeClassification.Rmd file). Try to use R code into paragraph.
- It is recommended to use caret package but it is not mandatory.
- Use relative paths instead of absolute paths to read / write files, to make it easier to run the code outside of your computer.

**Delivery / Deadline**

A zip file including the data set, the Rmd (or Rsw) file used as template for the report and output reports in pdf and html files.

Deadline: 24th of April, 2018

**Score of each question**

- Question 1 (20%)
- Question 2 (20%)
- Question 3 (30%)
- Question 4 (20%)
- Dynamic report quality (10%)