INFO-F-422 : Statistical foundations of machine learning
Project 2016-17 - 2nd Session

Gianluca Bontempi, Yann-Aël Le Borgne
Computer Science Department, ULB

The project counts for 60% 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.

1 Goal

The goals of the project are

1. to participate to the "Digit Recognizer" Kaggle competition [1] by implementing and assessing different supervised learning algorithms and different methods of feature selection in the related classification task,
2. to select among the learning and feature selection techniques the ones which appear to be the most accurate and use them for submitting to the Kaggle competition.
3. to report your analyses and results as a Jupyter notebook.

2 Kaggle competition

The goal in this competition is to design a classifier able to take an image of a handwritten single digit, and determine what that digit is. The data for this competition were taken from the MNIST dataset. The MNIST ("Modified National Institute of Standards and Technology") dataset is a classic within the Machine Learning community that has been extensively studied.

The student should register with the login "INFOF422Lastname" (e.g. INFOF422Bontempi) and accept the rules of the competition (notably no hidden additional accounts).

1. `http://www.kaggle.com/c/digit-recognizer`

3 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 accuracy of the classifiers during the selection process should be assessed by using the misclassification error.

For the learning method, the only packages that may be used are those seen during the exercise classes : `stats`/`ridge`, `nnet`, `tree`/`rpart`, `lazy`, and `e1071`, for linear/ridge models, neural networks, decision trees, nearest neighbours and SVM, respectively. A list of additional packages that can also be used is provided on the Github page of the project [2] - see note at the end of this document.

The report must be an R Jupyter notebook which has to specify and justify (with tables, figures) the selection procedures which led to the final choice.

The student has to return, together with the report, the set of predictions submitted to the Kaggle competition.

4 Tasks

The student will have to

1. implement in the R language three feature selection procedures. These procedures must be detailed in the notebook. The text must contain the list of final selected variables, and the motivation for their choice. The use of formulas, tables and pseudo-code to describe the feature selection procedure is encouraged. (**3 points**)
2. implement in the R language a model selection procedure. This procedure must be detailed in the notebook. The text must mention the different (and at least three) models which have been taken into consideration and the procedure used for model assessment and selection. The use of formulas, tables and pseudo-code to describe the model selection procedure is encouraged. (**3 points**)
3. implement in the R language a procedure implementing a combination of models strategy. This procedure must be detailed in the notebook. The text should mention the different models taken into consideration as well as the techniques used for the combination. The use of formulas, tables and pseudo-code to describe the ensemble selection procedure is encouraged. (**3 points**)
4. On the basis of the procedure described in the previous steps the student must compute the predictions for the competition and submit them via the Kaggle website. The name of the student should appear in the official leader board of the competition. (**1 point**)

2. `https://github.com/Yannael/info-f-422/tree/master/Project`

5 Report and deadline

— The report must be a notebook named 'report.ipynb', containing a detailed description of the procedures which led to the to choice of the learning algorithm.
— The notebook and results (as a CSV file) must be sent before the 7th of August (23 :59) by an email with the Subject : `INFOF422 project` to Yann-Aël Le Borgne (email : `yleborgn@ulb.ac.be`) and cc to Fabrizio Carcillo (email : `fcarcill@ulb.ac.be`).

5.1 Note on the report structure

The report should be a summary of the goal, methodology and main results of your work. The notebook format is chosen to make the report results fully reproducible, but it should be structured and written as if it was a Latex or Word report.

While we do not enforce a specific structure, we recommend the following :
— Abstract - a short summary of your report
— Introduction - with dataset description, goals, and an overview of the report structure
— Feature selection : Methodology and main results
— Model selection : Methodology and main results
— Ensemble techniques : Methodology and main results
— Discussion and conclusion : Summary of your work, and discussion of what worked well, not well, why, what insights you got from the analyses you made.

We emphasize again that the report should not be a long list of experiments and results, but a structured summary (report) of your analyses. For the sake of clarity, you can put some functions in additional R scripts, and source them in the notebook.

5.2 Note on allowed package

We intentionally limit the packages that you can use, as it is more important in the report that you show your understanding of three out of the five classes of basic learning algorithms (linear regression, decision trees, KNN, neural nets, SVM) and how tuning their parameters can improve the accuracy, rather that testing many different methods as black boxes.

Other packages can however be used, in particular those related to plotting and data frame management (ggplot2, dplyr, data.tables, ...), or those closely related to the classes of models seen during the hands-on classes (MASS, ridge, ...). However packages like caret (which automates feature selection and ensembles), randomforest or xgboost (that implement ensembles) or others of the kind are not allowed. One of the goal of the project is to make you implement these methods from scratch, and discuss what are the underlying technical challenges.

The list of accepted/not accepted packages is maintained at https ://github.com/Yannael/info-f-422/tree/master/Project. Feel free to ask us if you want to use another packages, and we will let you know, and update the list accordingly.