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

Gianluca Bontempi,
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 "House Prices : Advanced Regression Techniques" Kaggle competition [1] by implementing and assessing different supervised learning algorithms and different methods of feature selection in the related regression 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 objective is to build a predictive model which is able to predict the final price of a home. The model has to be designed using the train.csv file which can be downloaded from the kaggle platform, it includes roughly 1500 labeled samples and 79 features. The student should register with the login "INFOF422Lastname" (e.g. INFOF422Bontempi) and accept the rules of the competition (notably no hidden additional accounts).

1. `https://www.kaggle.com/c/house-prices-advanced-regression-techniques`

3 Specifications

The student has to choose a learning method and a feature selection method among at least three alternatives.

For the learning method, the only packages that may be used are those seen during the exercise classes : `stats`, `nnet`, `tree`, `lazy`, and `e1071`, for linear models, neural networks, decision trees, nearest neighbours and SVM, respectively.

The accuracy of the regression models during the selection process should be assessed by using the root mean squared error between the logarithm of the predicted value and the logarithm of the observed sale price [2].

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 a feature selection procedure. This procedure must be detailed in the notebook. The text must contain the list of selected variables and the motivation of 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 feature 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 feature 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://www.kaggle.com/c/house-prices-advanced-regression-techniques#evaluation`

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 21st of May (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`).