Labwork on "Ensemble learning"

Advanced machine learning
IMT Mines Alès - 2IA Department
Nicolas Sutton-Charani

Exercice 1: feature importance (result interpretation)

(a) Load the 'Boston' dataset (from sklearn.datasets).

(b) Validate the predictive power of a random forest and a xgb regressor both with 500 trees on the Boston dataset.

(c) Train the previous models on the whole dataset and evaluate them on the same data.

(d) Compute 2 types of feature importance according to the 2 regressors, plot them and conclude.

Exercice 2: models comparison

(a) Import the wine dataset from this link [1] and plot the corresponding correlation matrix with a clear and readable heatmap that respect the following constraints:

- the legend should include extreme correlations values (i.e. -1 and +1)
- negative/positive correlations should be represented respectively by cold/warm colours
- x-labels have to be rotated of 45°
- all variable names have to be clearly readable

(b) With the *sklearn* package set up a model comparison pipeline between a *dummy* classifier (predicting systematicaly the most frequent label of the training data), a decision tree, a kNN and a neural network models with a random sample containing 3/4 of the data for training and 1/4 for testing. Repeat this operation 100 times, draw the results in terms of means and boxplots and conclude.

N.B.: Unless specified in the question, the solver to be used for neural network is 'adam', the trees depths are fixed to 1, kNN models are tuned to 2 neighbors and neural network are limited to MLP of 1 hidden layer of 100 neurons computed with a maximum of 1000 iterations.

(c) Add hard and a soft votes approaches to the pipeline of question (b) and comment the results.

(d) Try a stacking approach (with the same 3 single classifiers as before), add it to the comparison of the previous question and comment the results.

[1] https://archive.ics.uci.edu/ml/datasets/Wine
https://drive.google.com/open?id=1GAw5nF0q0ibfhWKwcND9y4Y0kWQdnhCd

(e) Add a random forest, an AdaBoost, a gradient bossting and a XGBoost to the pipeline and comment the results.

(f) Tune the forest size (i.e. number of trees) with a grid-search approach and with a Bayes-search approach on the whole dataset.

(g) Realise a comparison between random forest and XGB with a 10-fold cross validation procedure after having tuned their sizes (number of weak classifiers) and their depths on the whole dataset. What is wrong with our tuning strategy?

(h) Realise the complete models comparison of previous question with a proper tuning step for all models' main hyperparameters.

***NB**:* Due to the long computation times involved in that question, no results are exepected, only the codes have to be written.

Exercice 3: order of ensemble learning

Should I train 1 forest of 10 trees, 2 forests of 5 trees or 5 forest of 2 trees?

(a) Load the 'BreastCancer' dataset (from sklearn.datasets on python or from the mlbench package on R).

(b) Answer the main question with a model comparison (involving simple vote aggregations) and plot the results.

(c) Do the same thing replacing the forest approach with stacking and boosting approaches.