Assignment 3: Random Forests

June 2, 2022

1 Trees, Random Forests and Correlation - 5 Points

1. Using the Bias-Variance Trade off, derive the variance of a random forest:

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2, \tag{1}$$

where B is the number of trees and σ^2 is the variance. Evaluate and diagnose the negative correlation case.

2. Show that as the tree number B gets larger, the OOB error approaches the $N-$fold Cross-Validation error.

3. Show that the sampling correlation between a pair of random forest trees at a point x is given by:

$$\rho(x) = \frac{Var_Z[E_{\Theta|Z}T(x;\Theta(Z))]}{Var_Z[E_{\Theta|Z}T(x;\Theta(Z)) + E_{\Theta|Z}Var_Z[T(x;\Theta(Z))}\tag{2}$$

4. Evaluate a random forest model to the credit risk dataset to explore the sensitivity to the parameter m. Plot both the OOB error and the test error against the parameter m.

2 General Aspects - 5 Points

1. Why bagging is based on random sampling with replacement? Would bagging still reduce a forecast's variance if sampling were without replacement?

2. Suppose that your training set is based on highly overlap labels.
   (a) Does this make bagging prone to overfitting, or just ineffective? Explain.
   (b) Is out-of-bag accuracy generally reliable in sequential datasets, like a time series? Why?

3. Build an ensemble of estimator, where the base estimator is a decision tree.
   (a) How is this ensemble different from a RF
   (b) How could you use sklearn no build a RF. Which parameters you need to tune?

4. Consider the relation between a RF, the number of trees it is composed of, and the number of features utilized.
   (a) Could you envision a relation between the minimum number of trees needed in a RF and the number of features utilized?
   (b) Could the number of trees be too small for the number of features used?
   (c) Could the number of trees be too high for the number of observations avaliable?