19) Bagging, Random Forests, and Boosting

Vitor Kamada

January 2020

Reference

Tables, Graphics, and Figures from:

1) Hastie et al. (2017): Ch 10 and 15

2) James et al. (2017): Ch 8.2

Bagging

Set of n independent observations $Z_1, \ldots, Z_n$, each with variance σ^2

$$Var(\bar{Z}) = \frac{\sigma^2}{n}$$

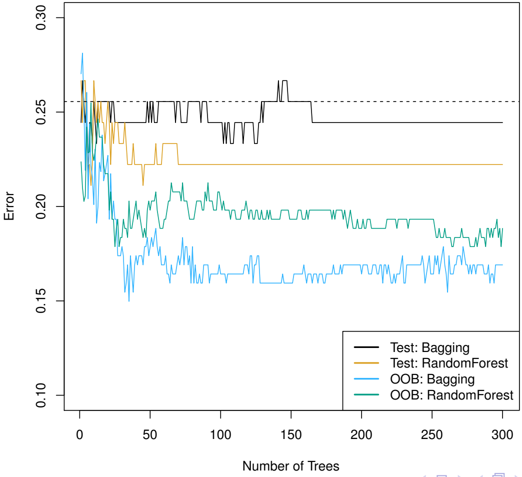
$$\hat{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}^{*b}(x)$$

Out-of-Bag Error Estimation (OOB)

Each bagged tree makes use $\frac{2}{3}$ of the observations

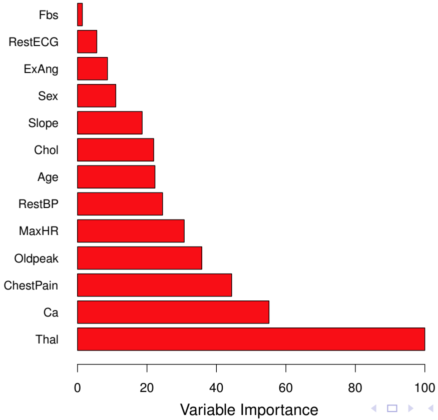
$\frac{1}{3}$ of the observations not used to fit a bagged tree are the out-of-bag (OOB) observations

Heart Data: Single Classification Tree vs Bagging, Random Forest, and Out-of-Bag



The Mean Decrease in Gini Index for each Variable, relative to the Largest

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$



The split only uses m predictors ($m \approx \sqrt{p}$)

Bagging: $m = p$

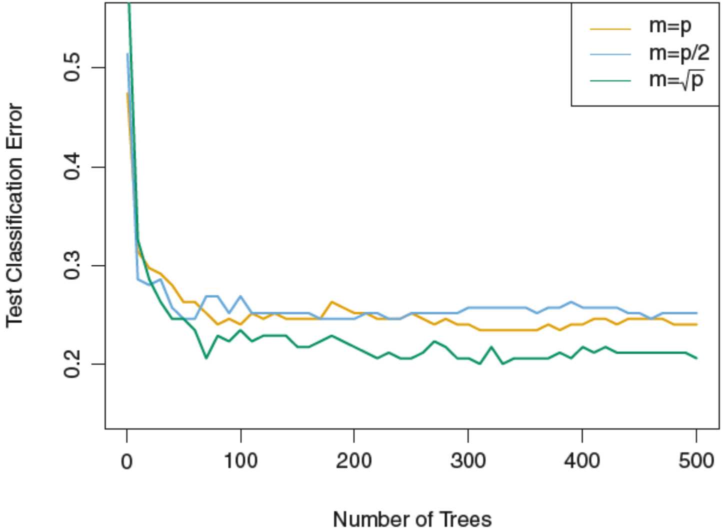
$$Var(X_1 + X_2)$$

$$Var(X_1) + Var(X_2) + 2Cov(X_1, X_2)$$

$$4Var(X) \text{ or } 2Var(X)$$

Gene Expression Data Set with 500 Predictors

A single tree has an error rate of 45.7%



Boosting

For $b = 1, 2, ..., B$, repeat:

a) Fit a tree $\hat{f}^b$ with d splits ($d + 1$ terminal nodes) to the training data (X, r)

b) Update $\hat{f}$ by adding in a shrunken version of the new tree:

$$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x)$$

c) Update the residuals:

$$r_i \leftarrow r_i - \lambda \hat{f}^b(x_i)$$

Boosting ($\lambda = 0.01$) vs Random Forests

Test error rate for a single tree is 24%