Bagging

The Bootstrap aggregation or **bagging** its a general procedure for reducing the **variance** of statistical learning methods, which means bagging can be used with any model we discuss before but given that linear models tend to be low variance **bagging** is frequently used in the context of Decision Trees.

The idea of averaging a set observations results in a reduced **variance** resembling the mean of the data, so then to reduce the variance for a given model is :

- Building sets of training data
- Fitting a model for each set
- Averaging out the prediction results

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

- B separate training sets
- $\hat{f}^{1}, \dots \hat{f}^{b}$ are the fitted model for each training set b
- For classification it's simply **majority vote**, the overall prediction is the most commonly occurring class

Out-of-Bag Error Estimation

This is a alternative to estimating the **test error** without using Cross-Validation, it's slightly a more optimistic than **LOOCV**. For intuition when performing The Bootstrap with replacement an observation x_1 the probability of it being selected is $\frac{1}{n}$ out of n number of observations, the probability of it not being selected in the bootstrapped sub-data set is simply $1 - \frac{1}{n}$, given that we do the process n times will result :

$$\left(1 - \frac{1}{n}\right)^{n}$$

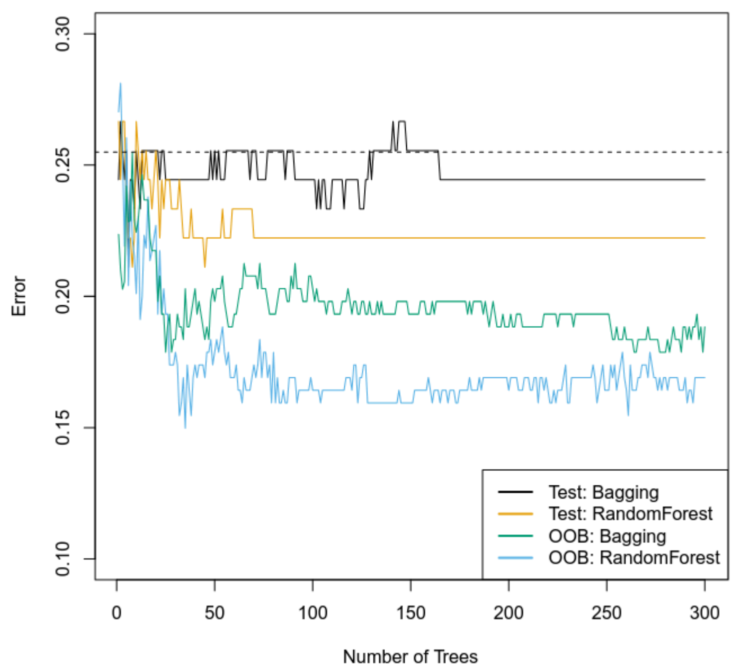
Calculating the limits of this expression results in :

$$lim_{n\to\infty}\left(1 - \frac{1}{n}\right)^{n} = \frac{1}{\mathrm{e}} \approx \frac{1}{3}$$

Which means that $\frac{1}{3}$ one third of the observations are not used to fit a given tree in the bagging process, building the logic these are new unseen data for that fitted tree B_1.

We can use those called **Out-of-Bag** observations as a test error estimate for each fitted tree B and average out the results(in regression) or voting majority(in classification)

Note : we detect the **OOB** observations for each tree individually



- As noted before the **OOB** test error estimation is more optimistic than the cross-validation one but still a valuable metric when it's computationally expansive to perform the cross validation

Variable Importance Measure

A clear disadvantage to **Bagging Trees** is the loss of interpretibility which is a strong point in Decision Trees which also allow us to perform feature selection and graphical representation, so the bagging process gives us accuracy at the expense of interpretability.

The importance of a feature is calculated with **RSS** or **Gini-index** so for reach tree B we calculate the decrease on **RSS** or **Gini-index** and average out the results for all the trees

Variance Reduction

The **bagging** procedure mostly effect the **Variance** that's why it's used on high variance learners(Decision Trees, ANN), with a slight increase in the **bias** since the bootstrap with replacement train only on $\frac{2}{3}$ of the data(covered in **OOB** above).

Bias Unchanged mostly :

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

- Base Learner $\hat{f}_b$ with high variance

The expected bias is :

$$\text{Bias}^2 = [f(x) - \mathbb{E}[\hat{f}_b(x)]]^2$$

Since the base learners are identically distributed :

$$\mathbb{E}[f_{bag}(x)] = \mathbb{E}[\hat{f}_b(x)]$$

- The bagged results bias is equal to a single model bias

Variance Reduced :

$$\text{Var}[f_{bag}(x)] = \text{Var}\left[\frac{1}{B}\sum_{b=1}^{B}\hat{f}_b(x)\right] = \frac{1}{B^2}\sum_{b=1}^{B}\text{Var}[\hat{f}_b(x)] = \frac{1}{B}\text{Var}[\hat{f}_b(x)]$$

- Averaging independent models results reduce the variance

Effect on Correlation

When talking about variance reduction the limiting factor is the correlation between the models, the variance decomposition can be expressed as :

$$\text{Var}\left(\frac{1}{B}\sum_{b=1}^{B}\hat{f}_b(x)\right) = \frac{1}{B}(1-\rho)\sigma^2 + \rho\sigma^2$$

- This is a average pairwise correlation between base learners, where ρ the correlation factor
- If $\rho = 0$ represent a perfect uncorrelated models $\text{Var(ensemble)} = \frac{1}{B}\sigma^2$

The correlation between models can be from:

- The Bootstrap sampling overlap the $\frac{2}{3}$ of the original data
- Base learners learning form the same underlying distributed of the original data set
- Decision Trees selecting the same splits features + Greedy approach of trees