Random Forest for Classification Problems

Raphael Redmer, Arkadiusz Modzelewski,
Burak Balaban

University of Bonn
Research Module in Econometrics and Statistics

January 20, 2020

Overview

A tree is grown starting from the root node by repeatedly using the following steps on each node (also called binary splitting)[1]:

(i) Find best split s for each feature X_m

(ii) Find the best split of the node

(iii) Repeat until stopping criterion got satisfied

A tree is grown starting from the root node by repeatedly using the following steps on each node (also called binary splitting)[1]:

(i) Find best split s for each feature X_m

(ii) Find the best split of the node

(iii) Repeat until stopping criterion got satisfied
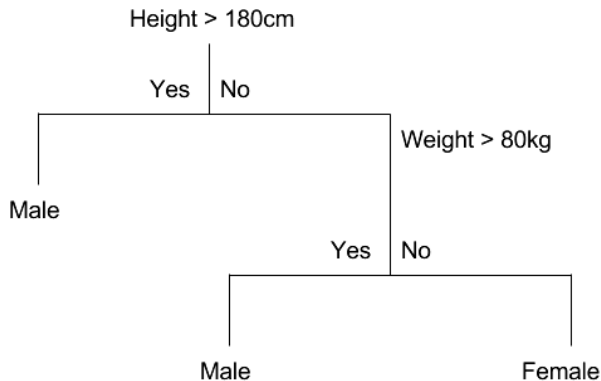
Decision Tree
Tree Building Process

A tree is grown starting from the root node by repeatedly using the following steps on each node (also called binary splitting)[1]:

(i) Find best split s for each feature X_m

(ii) Find the best split of the node

(iii) Repeat until stopping criterion got satisfied

Decision Tree
Purity Measures

Gini Measure

$$i(t) = \sum_{c \in C} p(c|t)(1 - p(c|t))$$

Information Entropy

$$i(t) = \sum_{c \in C} p(c|t)log(p(c|t))$$

where C is the set of classes c and t a node of the tree.

Bias Variance Trade-off
The Expected Generalization Error

$y = f(x) + \epsilon$ and $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$

Estimate of $f(x)$: $\hat{f}(x)$

The expected generalization error: **$Err(\hat{f}(x))$**

The decomposition of a model's expected generalization error is

$$Err(\hat{f}(x)) = \sigma_\epsilon^2 + [Bias(\hat{f}(x))]^2 + Var(\hat{f}(x))$$

σ_ϵ^2 is irreducible and independent of the model.

Trade-off between bias and variance.

Aim: Decrease variance while keeping bias unincreased.

Bias Variance Trade-off
The Expected Generalization Error

$y = f(x) + \epsilon$ and $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$

Estimate of $f(x)$: $\hat{f}(x)$

The expected generalization error: $\textbf{\textit{Err}}(\hat{f}(x))$

The decomposition of a model's expected generalization error is

$$\textbf{\textit{Err}}(\hat{f}(x)) = \sigma_\epsilon^2 + [\textit{Bias}(\hat{f}(x))]^2 + \textit{Var}(\hat{f}(x))$$

σ_ϵ^2 is irreducible and independent of the model.

Trade-off between bias and variance.

Aim: Decrease variance while keeping bias unincreased.

Bias Variance Trade-off
The Expected Generalization Error

$y = f(x) + \epsilon$ and $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$

Estimate of $f(x)$: $\hat{f}(x)$

The expected generalization error: **$Err(\hat{f}(x))$**

The decomposition of a model's expected generalization error is

$$Err(\hat{f}(x)) = \sigma_\epsilon^2 + [Bias(\hat{f}(x))]^2 + Var(\hat{f}(x))$$
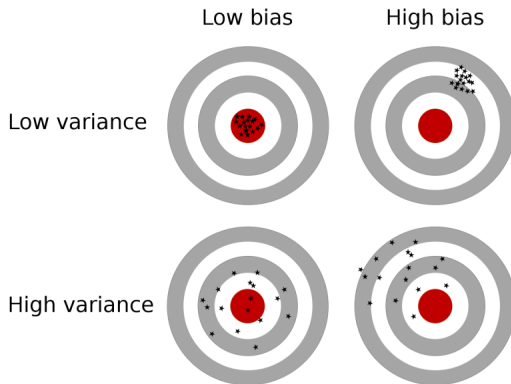
σ_ϵ^2 is irreducible and independent of the model.

Trade-off between bias and variance.

Aim: Decrease variance while keeping bias unincreased.
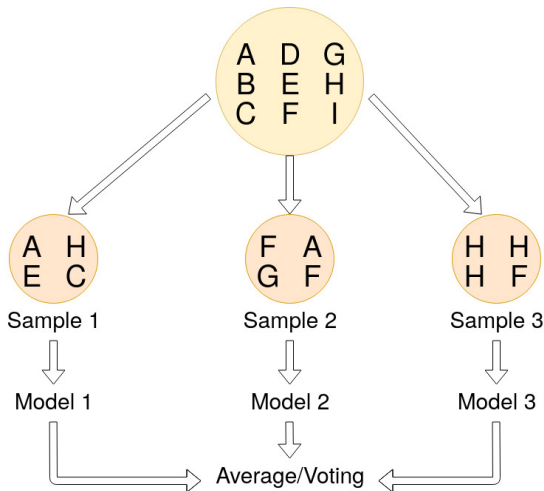
Bias-Variance Trade-off
Illustration

Decision trees generally have low bias and high variance.

Bagging is:

1. created for methods with high variance
2. reduces variance and gives better predictions
3. improvement of bagging: Random Forest

Bagging
Ilustration

Random Forest
Definition

Definition (by L.Breiman [2])

A random forest is a classifier consisting of a collection of tree-structured classifiers $\hat{T}_{\theta_b}(\mathbf{x})$, $b = 1, ..., B$ where the θ_b are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input $\mathbf{x}$.

Random Forest is an extension and improvement over bagging:

1. Like in bagging, multiple decision trees are built

2. Improvement: an injection of randomness is made

Random Forest
Randomness in the model

Two key concepts that makes decision forest "random" are [3]:

1. Random sampling of training data points when building trees

2. Random subsets of variables considered when splitting nodes.
   Recommended number of variables:

   a. For classification: $\lfloor \sqrt{n} \rfloor$

   b. For regression: $\lfloor \frac{n}{3} \rfloor$

Random Forest
Algorithm

Algorithm 1: Random Forest for Regression or Classification [3]

1. For $b = 1$ to B:

   a. Draw a bootstrap sample D_b of size N from the training data.

   b. Grow the Random Forest tree T_{D_b, θ_b} to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached:

      i. Select m variables denoted by θ_b at random from the n variables

      ii. Pick the best variable/split-point among the m

      iii. Split the node into two daughter nodes

2. Output the ensemble of trees $\{T_{D_b, \theta_b}\}_{b=1}^{B}$

Section 2

Mathematical Concept

Mathematical Concept
Prediction Mechanism

Let $D = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$ and
$T_{D,\theta}$ is a fully grown tree trained on set D with using parameters θ.
Random Forest estimate of an observation x^* is

Majority Voting

$$RF_{D,\theta_1,\theta_2,...,\theta_B}(x^*) = \underset{c \in Y}{argmax} \sum_{b=1}^{B} \mathbb{1}(\hat{T}_b(x^*) = c)$$

Soft Voting

$$RF_{D,\theta_1,\theta_2,...,\theta_B}(x^*) = \underset{c \in Y}{argmax} \frac{1}{B} \sum_{b=1}^{B} \hat{p}_{D,\theta_b}(Y = c | X = x^*)$$

where $\hat{p}_{D,\theta_b}(Y = c | X = x^*)$ is the probability estimates of a tree.

Mathematical Concept
Prediction Mechanism

Let $D = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$ and
$T_{D,\theta}$ is a fully grown tree trained on set D with using parameters θ.
Random Forest estimate of an observation x^* is

Majority Voting

$$RF_{D,\theta_1,\theta_2,...,\theta_B}(x^*) = \underset{c \in Y}{argmax} \sum_{b=1}^{B} \mathbb{1}(\hat{T}_b(x^*) = c)$$

Soft Voting

$$RF_{D,\theta_1,\theta_2,...,\theta_B}(x^*) = \underset{c \in Y}{argmax} \frac{1}{B} \sum_{b=1}^{B} \hat{p}_{D,\theta_b}(Y = c | X = x^*)$$

where $\hat{p}_{D,\theta_b}(Y = c | X = x^*)$ is the probability estimates of a tree.

Mathematical Concept
Prediction Mechanism

Let $D = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$ and
$T_{D,\theta}$ is a fully grown tree trained on set D with using parameters θ.
Random Forest estimate of an observation x^* is

Majority Voting

$$RF_{D,\theta_1,\theta_2,...,\theta_B}(x^*) = \underset{c \in Y}{argmax} \sum_{b=1}^{B} \mathbb{1}(\hat{T}_b(x^*) = c)$$

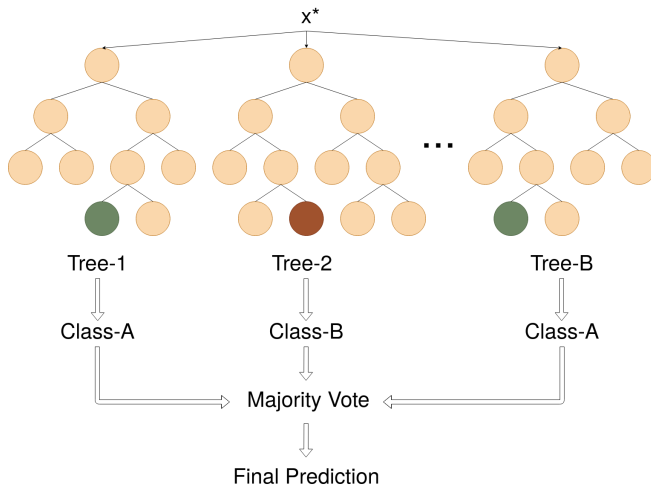### Soft Voting

$$RF_{D,\theta_1,\theta_2,...,\theta_B}(x^*) = \underset{c \in Y}{argmax} \frac{1}{B} \sum_{b=1}^{B} \hat{p}_{D,\theta_b}(Y = c | X = x^*)$$

where $\hat{p}_{D,\theta_b}(Y = c | X = x^*)$ is the probability estimates of a tree.
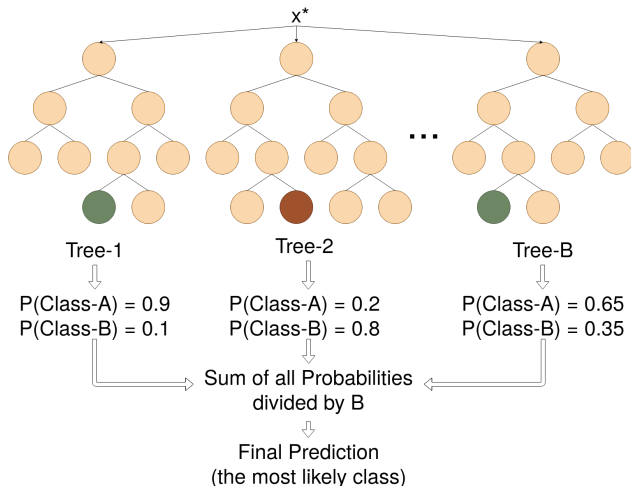
Given $D = (X, Y)$,
the expected generalization error of $T_{D,\theta}$ is

$$\textbf{\textit{Err}}(T_{D,\theta}(X)) = \mathbb{E}_{X,Y}\{L(Y, T_{D,\theta}(X))\}$$

where $L(Y, T_{D,\theta}(X))$ is the loss function.

The decomposition of $\textbf{\textit{Err}}(T_{D,\theta})$ is

$$\textbf{\textit{Err}}(T_{D,\theta}(X)) = \textbf{\textit{Err}}(\phi_\beta(X)) + [Bias(T_{D,\theta}(X))]^2 + Var(T_{D,\theta}(X))$$

similarly

$$\textbf{\textit{Err}}(\textbf{\textit{RF}}_{D,\Theta}(X)) = \textbf{\textit{Err}}(\phi_\beta(X)) + [Bias(\textbf{\textit{RF}}_{D,\Theta}(X))]^2 + Var(\textbf{\textit{RF}}_{D,\Theta}(X))$$

where $\Theta = \{\theta_1, \theta_2, ..., \theta_B\}$.

Mathematical Concept
The Expected Generalization Error of $T_{D,\theta}$

Given $D = (X, Y)$,
the expected generalization error of $T_{D,\theta}$ is

$$\textbf{\textit{Err}}(T_{D,\theta}(X)) = \mathbb{E}_{X,Y}\{L(Y, T_{D,\theta}(X))\}$$

where $L(Y, T_{D,\theta}(X))$ is the loss function.

The decomposition of $\textbf{\textit{Err}}(T_{D,\theta})$ is

$$\textbf{\textit{Err}}(T_{D,\theta}(X)) = \textbf{\textit{Err}}(\phi_\beta(X)) + [Bias(T_{D,\theta}(X))]^2 + Var(T_{D,\theta}(X))$$

similarly

$$\textbf{\textit{Err}}(\textbf{\textit{RF}}_{D,\Theta}(X)) = \textbf{\textit{Err}}(\phi_\beta(X)) + [Bias(\textbf{\textit{RF}}_{D,\Theta}(X))]^2 + Var(\textbf{\textit{RF}}_{D,\Theta}(X))$$

where $\Theta = \{\theta_1, \theta_2, ..., \theta_B\}$.

Mathematical Concept
The Expected Generalization Error of $T_{D,\theta}$

Given $D = (X, Y)$,
the expected generalization error of $T_{D,\theta}$ is

$$\textbf{\textit{Err}}(T_{D,\theta}(X)) = \mathbb{E}_{X,Y}\{L(Y, T_{D,\theta}(X))\}$$

where $L(Y, T_{D,\theta}(X))$ is the loss function.

The decomposition of $\textbf{\textit{Err}}(T_{D,\theta})$ is

$$\textbf{\textit{Err}}(T_{D,\theta}(X)) = \textbf{\textit{Err}}(\phi_\beta(X)) + [Bias(T_{D,\theta}(X))]^2 + Var(T_{D,\theta}(X))$$

similarly

$$\textbf{\textit{Err}}(\textbf{\textit{RF}}_{D,\Theta}(X)) = \textbf{\textit{Err}}(\phi_\beta(X)) + [Bias(\textbf{\textit{RF}}_{D,\Theta}(X))]^2 + Var(\textbf{\textit{RF}}_{D,\Theta}(X))$$

where $\Theta = \{\theta_1, \theta_2, ..., \theta_B\}$.

$$Err(RF_{D,\Theta}(X)) = \textbf{\textit{Err}}(\phi_\beta(X)) + [Bias(RF_{D,\Theta}(X))]^2 + Var(RF_{D,\Theta}(X))$$

Theoretically, given the probability distribution of P(X,Y)
Bayes Model ϕ_β i.e the best possible model can be derived
and $\textbf{\textit{Err}}(\phi_\beta)$ can be calculated [4].

For comparison of $\textbf{\textit{Err}}(T_{D,\theta}(X))$ and $\textbf{\textit{Err}}(RF_{D,\Theta}(X))$
the residual error is the same.

Result: Ensembling has no effect on Bayes Error.

Mathematical Concept
Residual Error

$$Err(RF_{D,\Theta}(X)) = \boldsymbol{Err}(\phi_\beta(X)) + [Bias(RF_{D,\Theta}(X))]^2 + Var(RF_{D,\Theta}(X))$$

Theoretically, given the probability distribution of P(X,Y)
Bayes Model ϕ_β i.e the best possible model can be derived
and $\boldsymbol{Err}(\phi_\beta)$ can be calculated [4].

For comparison of $\boldsymbol{Err}(T_{D,\theta}(X))$ and $\boldsymbol{Err}(RF_{D,\Theta}(X))$
the residual error is the same.

Result: Ensembling has no effect on Bayes Error.

Mathematical Concept
Residual Error

$$Err(RF_{D,\Theta}(X)) = \textbf{\textit{Err}}(\phi_\beta(X)) + [Bias(RF_{D,\Theta}(X))]^2 + Var(RF_{D,\Theta}(X))$$

Theoretically, given the probability distribution of P(X,Y)
Bayes Model ϕ_β i.e the best possible model can be derived
and $\textbf{\textit{Err}}(\phi_\beta)$ can be calculated [4].

For comparison of $\textbf{\textit{Err}}(T_{D,\theta}(X))$ and $\textbf{\textit{Err}}(RF_{D,\Theta}(X))$
the residual error is the same.

Result: Ensembling has no effect on Bayes Error.

Mathematical Concept
Residual Error

$$Err(RF_{D,\Theta}(X)) = \mathbf{Err}(\phi_\beta(X)) + [Bias(RF_{D,\Theta}(X))]^2 + Var(RF_{D,\Theta}(X))$$

Theoretically, given the probability distribution of P(X,Y)
Bayes Model ϕ_β i.e the best possible model can be derived
and $\mathbf{Err}(\phi_\beta)$ can be calculated [4].

For comparison of $\mathbf{Err}(T_{D,\theta}(X))$ and $\mathbf{Err}(RF_{D,\Theta}(X))$
the residual error is the same.

Result: Ensembling has no effect on Bayes Error.

Mathematical Concept
Bias

$$Err(\textbf{RF}_{D,\Theta}(X)) = Err(\phi_\beta(X)) + [Bias(\textbf{RF}_{D,\Theta}(X))]^2 + Var(\textbf{RF}_{D,\Theta}(X))$$

With Squared Loss Function: $L(Y, T_{D,\theta}(X)) = \mathbb{E}_D\{(Y - T_{D,\theta}(X))^2\}$

$Bias^2$ of Tree

$$[Bias(T_{D,\theta}(X))]^2 = (\phi_\beta(X) - \mathbb{E}_D\{T_{D,\theta}(X)\})^2$$

$Bias^2$ of Random Forest

$$[Bias(\textbf{RF}_{D,\Theta}(X))]^2 = (\phi_\beta(X) - \mathbb{E}_{D,\Theta}\{\textbf{RF}_{D,\Theta}(X)\})^2$$

We need $\mathbb{E}_{D,\theta}\{T_{D,\theta}(X)\}$ and $\mathbb{E}_{D,\Theta}\{\textbf{RF}_{D,\Theta}(X)\}$ for comparison.

$$Err(RF_{D,\Theta}(X)) = Err(\phi_\beta(X)) + [Bias(\boldsymbol{RF}_{D,\Theta}(X))]^2 + Var(RF_{D,\Theta}(X))$$

With Squared Loss Function: $L(Y, T_{D,\theta}(X)) = \mathbb{E}_D\{(Y - T_{D,\theta}(X))^2\}$

$Bias^2$ of Tree

$$[Bias(T_{D,\theta}(X))]^2 = (\phi_\beta(X) - \mathbb{E}_D\{T_{D,\theta}(X)\})^2$$

$Bias^2$ of Random Forest

$$[Bias(\boldsymbol{RF}_{D,\Theta}(X))]^2 = (\phi_\beta(X) - \mathbb{E}_{D,\Theta}\{\boldsymbol{RF}_{D,\Theta}(X)\})^2$$

We need $\mathbb{E}_{D,\theta}\{T_{D,\theta}(X)\}$ and $\mathbb{E}_{D,\Theta}\{\boldsymbol{RF}_{D,\Theta}(X)\}$ for comparison.

Mathematical Concept
Bias

$$Err(RF_{D,\Theta}(X)) = Err(\phi_\beta(X)) + [Bias(RF_{D,\Theta}(X))]^2 + Var(RF_{D,\Theta}(X))$$

With Squared Loss Function: $L(Y, T_{D,\theta}(X)) = \mathbb{E}_D\{(Y - T_{D,\theta}(X))^2\}$

$Bias^2$ of Tree

$$[Bias(T_{D,\theta}(X))]^2 = (\phi_\beta(X) - \mathbb{E}_D\{T_{D,\theta}(X)\})^2$$

$Bias^2$ of Random Forest

$$[Bias(RF_{D,\Theta}(X))]^2 = (\phi_\beta(X) - \mathbb{E}_{D,\Theta}\{RF_{D,\Theta}(X)\})^2$$

We need $\mathbb{E}_{D,\theta}\{T_{D,\theta}(X)\}$ and $\mathbb{E}_{D,\Theta}\{RF_{D,\Theta}(X)\}$ for comparison.

Mathematical Concept
Bias

$$Err(\textbf{RF}_{D,\Theta}(X)) = Err(\phi_\beta(X)) + [Bias(\textbf{RF}_{D,\Theta}(X))]^2 + Var(\textbf{RF}_{D,\Theta}(X))$$

With Squared Loss Function: $L(Y, T_{D,\theta}(X)) = \mathbb{E}_D\{(Y - T_{D,\theta}(X))^2\}$

$Bias^2$ of Tree
$$[Bias(T_{D,\theta}(X))]^2 = (\phi_\beta(X) - \mathbb{E}_D\{T_{D,\theta}(X)\})^2$$

$Bias^2$ of Random Forest
$$[Bias(\textbf{RF}_{D,\Theta}(X))]^2 = (\phi_\beta(X) - \mathbb{E}_{D,\Theta}\{\textbf{RF}_{D,\Theta}(X)\})^2$$

We need $\mathbb{E}_{D,\theta}\{T_{D,\theta}(X)\}$ and $\mathbb{E}_{D,\Theta}\{\textbf{RF}_{D,\Theta}(X)\}$ for comparison.

Mathematical Concept
Bias: The Expected Value

We can define $\mathbb{E}_D\{T_{D,\theta}(X)\} = \mu_{D,\theta}(X)$

Random Forest Estimator for regression is;

$$\boldsymbol{RF}_{D,\Theta}(X) = \frac{1}{B}\sum_{b=1}^{B} T_{D,\theta_b}(X)$$

Taking expectation gives;

$$\mathbb{E}_{D,\Theta}\{\boldsymbol{RF}_{D,\Theta}(X)\} = \mathbb{E}_{D,\Theta}\{\frac{1}{B}\sum_{b=1}^{B} T_{D,\theta_b}(X)\}$$

$$= \frac{1}{B}\sum_{b=1}^{B}\mathbb{E}_{D,\theta_b}\{T_{D,\theta_b}(X)\} \qquad (\theta\text{'s are i.i.d.})$$

$$= \mu_{D,\theta}(X)$$

Mathematical Concept
Bias: The Expected Value

We can define $\mathbb{E}_D\{T_{D,\theta}(X)\} = \mu_{D,\theta}(X)$

Random Forest Estimator for regression is;

$$\boldsymbol{RF}_{D,\Theta}(X) = \frac{1}{B}\sum_{b=1}^{B} T_{D,\theta_b}(X)$$

Taking expectation gives;

$$\mathbb{E}_{D,\Theta}\{\boldsymbol{RF}_{D,\Theta}(X)\} = \mathbb{E}_{D,\Theta}\{\frac{1}{B}\sum_{b=1}^{B} T_{D,\theta_b}(X)\}$$

$$= \frac{1}{B}\sum_{b=1}^{B} \mathbb{E}_{D,\theta_b}\{T_{D,\theta_b}(X)\} \qquad (\theta\text{'s are i.i.d.})$$

$$= \mu_{D,\theta}(X)$$

Mathematical Concept
Bias: The Expected Value

We can define $\mathbb{E}_D\{T_{D,\theta}(X)\} = \mu_{D,\theta}(X)$

Random Forest Estimator for regression is;

$$\boldsymbol{RF}_{D,\Theta}(X) = \frac{1}{B} \sum_{b=1}^{B} T_{D,\theta_b}(X)$$

Taking expectation gives;

$$\mathbb{E}_{D,\Theta}\{\boldsymbol{RF}_{D,\Theta}(X)\} = \mathbb{E}_{D,\Theta}\{\frac{1}{B} \sum_{b=1}^{B} T_{D,\theta_b}(X)\}$$

$$= \frac{1}{B} \sum_{b=1}^{B} \mathbb{E}_{D,\theta_b}\{T_{D,\theta_b}(X)\} \qquad (\theta\text{'s are i.i.d.})$$

$$= \mu_{D,\theta}(X)$$

Mathematical Concept
Bias: The Expected Value

We can define $\mathbb{E}_D\{T_{D,\theta}(X)\} = \mu_{D,\theta}(X)$

Random Forest Estimator for regression is;

$$\boldsymbol{RF}_{D,\Theta}(X) = \frac{1}{B}\sum_{b=1}^{B} T_{D,\theta_b}(X)$$

Taking expectation gives;

$$\mathbb{E}_{D,\Theta}\{\boldsymbol{RF}_{D,\Theta}(X)\} = \mathbb{E}_{D,\Theta}\{\frac{1}{B}\sum_{b=1}^{B} T_{D,\theta_b}(X)\}$$

$$= \frac{1}{B}\sum_{b=1}^{B}\mathbb{E}_{D,\theta_b}\{T_{D,\theta_b}(X)\} \qquad (\theta\text{'s are i.i.d.})$$

$$= \mu_{D,\theta}(X)$$

Mathematical Concept
Bias Comparison

$$Err(RF_{D,\Theta}(X)) = Err(\phi_\beta(X)) + [Bias(RF_{D,\Theta}(X))]^2 + Var(RF_{D,\Theta}(X))$$

$Bias^2$ of Tree

$$[Bias(T_{D,\theta}(X))]^2 = (\phi_\beta(X) - \mu_{D,\theta}(X))^2$$

$Bias^2$ of a Random Forest

$$[Bias(RF_{D,\Theta}(X))]^2 = (\phi_\beta(X) - \mu_{D,\theta}(X))^2$$

Result: Ensembling trees has no effect on bias.

Mathematical Concept
Bias Comparison

$$Err(RF_{D,\Theta}(X)) = Err(\phi_\beta(X)) + [Bias(\textbf{RF}_{D,\Theta}(X))]^2 + Var(RF_{D,\Theta}(X))$$

$Bias^2$ of Tree

$$[Bias(T_{D,\theta}(X))]^2 = (\phi_\beta(X) - \mu_{D,\theta}(X))^2$$

$Bias^2$ of a Random Forest

$$[Bias(\textbf{RF}_{D,\Theta}(X))]^2 = (\phi_\beta(X) - \mu_{D,\theta}(X))^2$$

Result: Ensembling trees has no effect on bias.

$$Err(RF_{D,\Theta}(X)) = Err(\phi_\beta(X)) + [Bias(RF_{D,\Theta}(X))]^2 + Var(RF_{D,\Theta}(X))$$

$\forall T_{D,\theta'}, T_{D,\theta''}$ such that $\theta' \neq \theta''$, the correlation coefficient can be written as follows

$$\rho(X) = \frac{\mathbb{E}_{D,\theta',\theta''}\{T_{D,\theta'}(X)T_{D,\theta''}(X)\} - \mu_{D,\theta}^2(X)}{\sigma_{D,\theta}^2(X)}$$

where $\sigma_{D,\theta}^2(X) = \mathbb{V}_{D,\theta}\{T_{D,\theta}(X)\}$.

- Highly correlated trees $\implies \rho(X)$ is close to 1

- Non-correlated trees $\implies \rho(X)$ is close to 0

Mathematical Concept
Variance: Correlation Coefficient

$$Err(RF_{D,\Theta}(X)) = Err(\phi_\beta(X)) + [Bias(RF_{D,\Theta}(X))]^2 + Var(RF_{D,\Theta}(X))$$

$\forall T_{D,\theta'}, T_{D,\theta''}$ such that $\theta' \neq \theta''$, the correlation coefficient can be written as follows

$$\rho(X) = \frac{\mathbb{E}_{D,\theta',\theta''}\{T_{D,\theta'}(X)T_{D,\theta''}(X)\} - \mu_{D,\theta}^2(X)}{\sigma_{D,\theta}^2(X)}$$

where $\sigma_{D,\theta}^2(X) = \mathbb{V}_{D,\theta}\{T_{D,\theta}(X)\}$.

- Highly correlated trees $\implies \rho(X)$ is close to 1

- Non-correlated trees $\implies \rho(X)$ is close to 0

Mathematical Concept
Variance: Correlation Coefficient

$$Err(\textbf{RF}_{D,\Theta}(X)) = \textbf{Err}(\phi_\beta(X)) + [Bias(\textbf{RF}_{D,\Theta}(X))]^2 + Var(\textbf{RF}_{D,\Theta}(X))$$

$\forall T_{D,\theta'}, T_{D,\theta''}$ such that $\theta' \neq \theta''$, the correlation coefficient can be written as follows

$$\rho(X) = \frac{\mathbb{E}_{D,\theta',\theta''}\{T_{D,\theta'}(X)T_{D,\theta''}(X)\} - \mu^2_{D,\theta}(X)}{\sigma^2_{D,\theta}(X)}$$

where $\sigma^2_{D,\theta}(X) = \mathbb{V}_{D,\theta}\{T_{D,\theta}(X)\}$.

- Highly correlated trees $\implies \rho(X)$ is close to 1

- Non-correlated trees $\implies \rho(X)$ is close to 0

Mathematical Concept
Variance Comparison

$$Err(\pmb{RF}_{D,\Theta}(X)) = Err(\phi_\beta(X)) + [Bias(\pmb{RF}_{D,\Theta}(X))]^2 + Var(\pmb{RF}_{D,\Theta}(X))$$

Variance of Random Forest

$$\mathbb{V}_{D,\Theta}\{\pmb{RF}_{D,\Theta}(X)\} = \rho(X)\sigma_{D,\theta}^2(X) + \frac{1-\rho(X)}{B}\sigma_{D,\theta}^2(X)$$

As $B \to \infty$, $\mathbb{V}_{D,\Theta}\{\pmb{RF}_{D,\Theta}(X)\}$ converges to $\rho(X)\sigma_{D,\theta}^2(X)$.

Due to randomization $\rho(X) < 1$

$$\implies \mathbb{V}_{D,\Theta}\{\pmb{RF}_{D,\Theta}(X)\} = \rho(x)\sigma_{D,\theta}^2(X) < \sigma_{D,\theta}^2(X) = \mathbb{V}_{D,\theta}\{T_{D,\theta}(X)\}.$$

Result: Ensembling trees decreases the variance.

Mathematical Concept
Variance Comparison

$$Err(\textbf{RF}_{D,\Theta}(X)) = Err(\phi_\beta(X)) + [Bias(\textbf{RF}_{D,\Theta}(X))]^2 + Var(\textbf{RF}_{D,\Theta}(X))$$

Variance of Random Forest

$$\mathbb{V}_{D,\Theta}\{\textbf{RF}_{D,\Theta}(X)\} = \rho(X)\sigma^2_{D,\theta}(X) + \frac{1-\rho(X)}{B}\sigma^2_{D,\theta}(X)$$

As $B \to \infty$, $\mathbb{V}_{D,\Theta}\{\textbf{RF}_{D,\Theta}(X)\}$ converges to $\rho(X)\sigma^2_{D,\theta}(X)$.

Due to randomization $\rho(X) < 1$

$$\implies \mathbb{V}_{D,\Theta}\{\textbf{RF}_{D,\Theta}(X)\} = \rho(x)\sigma^2_{D,\theta}(X) < \sigma^2_{D,\theta}(X) = \mathbb{V}_{D,\theta}\{T_{D,\theta}(X)\}.$$

Result: Ensembling trees decreases the variance.

Mathematical Concept
Variance Comparison

$$Err(RF_{D,\Theta}(X)) = Err(\phi_\beta(X)) + [Bias(RF_{D,\Theta}(X))]^2 + Var(\mathbf{RF}_{D,\Theta}(X))$$

Variance of Random Forest

$$\mathbb{V}_{D,\Theta}\{\mathbf{RF}_{D,\Theta}(X)\} = \rho(X)\sigma_{D,\theta}^2(X) + \frac{1-\rho(X)}{B}\sigma_{D,\theta}^2(X)$$

As $B \to \infty$, $\mathbb{V}_{D,\Theta}\{\mathbf{RF}_{D,\Theta}(X)\}$ converges to $\rho(X)\sigma_{D,\theta}^2(X)$.

Due to randomization $\rho(X) < 1$

$$\implies \mathbb{V}_{D,\Theta}\{\mathbf{RF}_{D,\Theta}(X)\} = \rho(x)\sigma_{D,\theta}^2(X) < \sigma_{D,\theta}^2(X) = \mathbb{V}_{D,\theta}\{T_{D,\theta}(X)\}.$$

Result: Ensembling trees decreases the variance.

Mathematical Concept
Variance Comparison

$$Err(\boldsymbol{RF}_{D,\Theta}(X)) = Err(\phi_\beta(X)) + [Bias(\boldsymbol{RF}_{D,\Theta}(X))]^2 + Var(\boldsymbol{RF}_{D,\Theta}(X))$$

Variance of Random Forest

$$\mathbb{V}_{D,\Theta}\{\boldsymbol{RF}_{D,\Theta}(X)\} = \rho(X)\sigma^2_{D,\theta}(X) + \frac{1-\rho(X)}{B}\sigma^2_{D,\theta}(X)$$

As $B \to \infty$, $\mathbb{V}_{D,\Theta}\{\boldsymbol{RF}_{D,\Theta}(X)\}$ converges to $\rho(X)\sigma^2_{D,\theta}(X)$.

Due to randomization $\rho(X) < 1$

$$\implies \mathbb{V}_{D,\Theta}\{\boldsymbol{RF}_{D,\Theta}(X)\} = \rho(x)\sigma^2_{D,\theta}(X) < \sigma^2_{D,\theta}(X) = \mathbb{V}_{D,\theta}\{T_{D,\theta}(X)\}.$$

Result: Ensembling trees decreases the variance.

Section 3

Simulation Study

Simulation Study
Linear DGP

The linear DGP generates the data tuples (y, x_1, x_2, x_3) as follows:
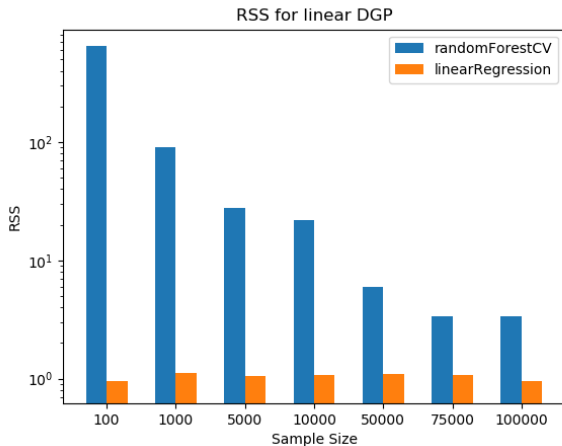
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon,$$

whereas

- $(\beta_0, \beta_1, \beta_2, \beta_3) = (0.3, 5, 10, 15)$
- $x_1, x_2, x_3 \sim \mathcal{N}(0, 3)$
- $\epsilon \sim \mathcal{N}(0, 1)$

Simulation Study
Linear DGP Results

RSS for linear DGP

Simulation Study
Non-Linear DGP

The non-linear DGP generates the data tuples (y, x_1, x_2) as follows:

$$y = \beta_0 + \beta_1 \mathbb{1}(x_1 \geq 0, x_2 \geq 0) + \beta_2 \mathbb{1}(x_1 \geq 0, x_2 < 0) + \beta_3 \mathbb{1}(x_1 < 0) + \epsilon,$$
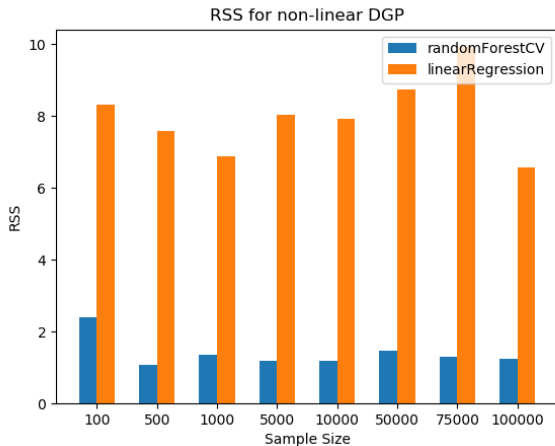
whereas

- $(\beta_0, \beta_1, \beta_2, \beta_3) = (0.3, 5, 10, 15)$
- $x_1, x_2 \sim \mathcal{N}(0, 3)$
- $\epsilon \sim \mathcal{N}(0, 1)$

are the same as in the previous DGP.

Simulation Study
Non-Linear DGP Results

RSS for non-linear DGP

Section 4

Real Data

Real Data
Outline

Data: Titanic data

Method used:

1. Random Forest
2. Adaptive Boosting
3. Gradient Boosting

Goal: Predict which passengers survived the Titanic shipwreck given characteristics of passengers

Real Data
Results

Random Forest

Accuracy: 84.32%

Adaptive Boosting

Accuracy: 82.8%

Gradient Boosting

Accuracy: 82.8%

	D_{pred}	S_{pred}
D_{real}	88%	12%
S_{real}	21%	79%

	D_{pred}	S_{pred}
D_{real}	86%	14%
S_{real}	21%	79%

	D_{pred}	S_{pred}
D_{real}	85%	15%
S_{real}	20%	80%

where D stands for died and S is for survived.

Contact Data

- Burak Balaban
  /burakbalaban/
  burak.balaban@uni-bonn.de

- Raphael Redmer
  /RaRedmer/
  ra.redmer@outlook.com

- Arkadiusz Modzelewski
  /ArcadiusM/
  arcadius.modzelewski@gmail.com

The link to presentation's repository:

References

📄 L. Breiman et al. *Classification and Regression Trees*. The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis, 1984. ISBN: 9780412048418. URL: https://books.google.de/books?id=JwQx-WOmSyQC.

📄 Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

📄 Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York, 2001.

📄 Gilles Louppe. "Understanding random forests". In: *University of Liège* (2014).