Random Forest for Classification Problems

Raphael Redmer, Arkadiusz Modzelewski,
Burak Balaban

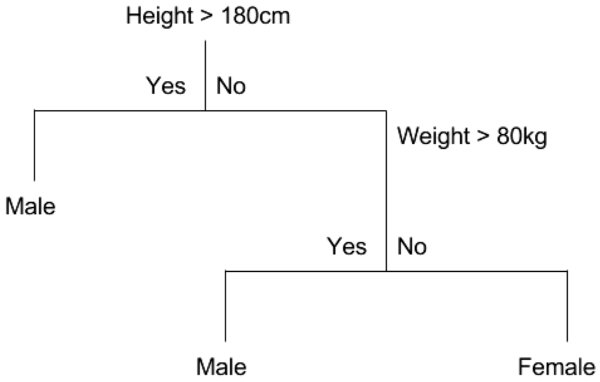
University of Bonn
Research Module in Econometrics and Statistics

January 20, 2020

Overview

Decision Tree

Example



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):

(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):

(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):

(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

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: $\boldsymbol{Err}(\hat{f}(x))$

The decomposition of a model's expected generalization error is

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

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: $\boldsymbol{Err}(\hat{f}(x))$

The decomposition of a model's expected generalization error is

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

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: $\boldsymbol{Err}(\hat{f}(x))$

The decomposition of a model's expected generalization error is

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

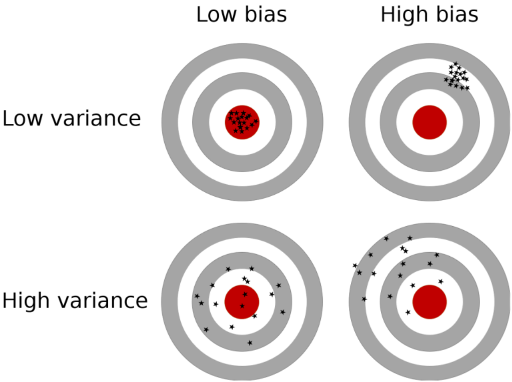
σ_ϵ^2 is irreducible and independent of the model.

Trade-off between bias and variance.

Aim: Decrease variance while keeping bias unincreased.

Bias-Variance Trade-off

Illustration



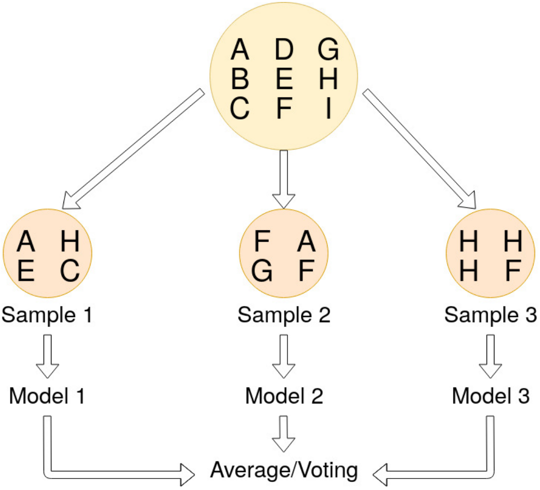
Decision trees generally have low bias and high variance.

Bagging

Bagging is:

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

Bagging
Illustration



Random Forest

Definition

Definition (by L.Breiman)

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

Two key concepts that makes decision forest "random" are:

1. Random sampling of training data points when building trees
2. Random subsets of features considered when splitting nodes. Recommended number of variables:
   a. For classification: $\lfloor \sqrt{n} \rfloor$
   b. For regression: $\lfloor \frac{n}{3} \rfloor$

Algorithm 1: Random Forest for Regression or Classification

1. For $b = 1$ to B:
   a. Draw a bootstrap sample θ_b of size N from the training data.
   b. Grow the Random Forest tree T_{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 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 $\{T_{\theta_b}\}_{b=1}^{B}$

Section 2

Mathematical Concept

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

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

Soft Voting

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

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

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

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

Soft Voting

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

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

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

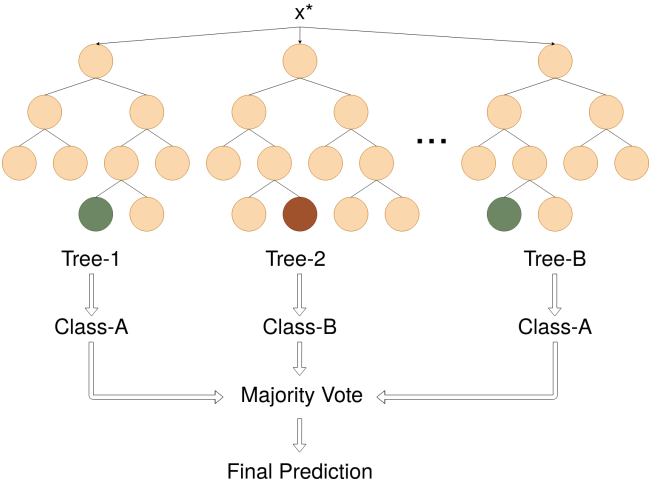
Soft Voting

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

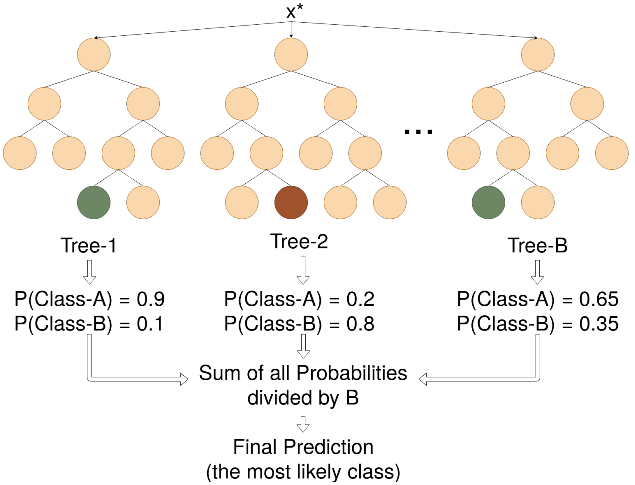
where $\hat{p}_{D,\theta_b}(Y = c | X = x^*)$ is the probability estimates of a tree.

Mathematical Concept

Majority Voting Illustration



Mathematical Concept
Soft Voting Illustration



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

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

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

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

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

similarly

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

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

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

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

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

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

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

similarly

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

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

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

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

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

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

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

similarly

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

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

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

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 [1].

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

Result: Ensembling has no effect on Bayes Error.

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

Theoretically, given the probability distribution of P(X,Y) Bayes Model ϕ_β i.e the best possible model can be derived and $Err(\phi_\beta)$ can be calculated [1].

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

Result: Ensembling has no effect on Bayes Error.

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

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 [1].

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

Result: Ensembling has no effect on Bayes Error.

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

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 [1].

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

Result: Ensembling has no effect on Bayes Error.

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

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.

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

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.

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

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

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

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.

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;

$$\begin{aligned}\mathbb{E}_{D,\Theta}\{\boldsymbol{RF}_{D,\Theta}(X)\} &= \mathbb{E}_{D,\Theta}\{\frac{1}{B}\sum_{b=1}^{B} T_{D,\theta_b}(X)\} \\ &= \frac{1}{B}\sum_{b=1}^{B} \mathbb{E}_{D,\theta_b}\{T_{D,\theta_b}(X)\} \qquad (\theta\text{'s are i.i.d.}) \\ &= \mu_{D,\theta}(X)\end{aligned}$$

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;

$$\begin{aligned}\mathbb{E}_{D,\Theta}\{\boldsymbol{RF}_{D,\Theta}(X)\} &= \mathbb{E}_{D,\Theta}\{\frac{1}{B}\sum_{b=1}^{B} T_{D,\theta_b}(X)\} \\ &= \frac{1}{B}\sum_{b=1}^{B} \mathbb{E}_{D,\theta_b}\{T_{D,\theta_b}(X)\} \qquad (\theta\text{'s are i.i.d.}) \\ &= \mu_{D,\theta}(X)\end{aligned}$$

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;

$$\begin{aligned}
\mathbb{E}_{D,\Theta}\{\boldsymbol{RF}_{D,\Theta}(X)\} &= \mathbb{E}_{D,\Theta}\{\frac{1}{B}\sum_{b=1}^{B} T_{D,\theta_b}(X)\} \\
&= \frac{1}{B}\sum_{b=1}^{B} \mathbb{E}_{D,\theta_b}\{T_{D,\theta_b}(X)\} \qquad (\theta\text{'s are i.i.d.}) \\
&= \mu_{D,\theta}(X)
\end{aligned}$$

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

$Bias^2$ of Tree

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

$Bias^2$ of a Random Forest

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

Result: Ensembling trees has no effect on bias.

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

$Bias^2$ of Tree

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

$Bias^2$ of a Random Forest

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

Result: Ensembling trees has no effect on bias.

Mathematical Concept

Variance: Correlation Coefficient

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

$\forall T_{D,\theta'}, T_{D,\theta''}$ 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: Correlation Coefficient

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

$\forall T_{D,\theta'}, T_{D,\theta''}$ 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

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

$\forall T_{D,\theta'}, T_{D,\theta''}$ 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

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

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.

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

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.

Mathematical Concept

Variance Comparison

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

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.

Mathematical Concept
Variance Comparison

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

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

The linear DGP generates the data tuples (y, x_1, x_2, x_3) as follows:

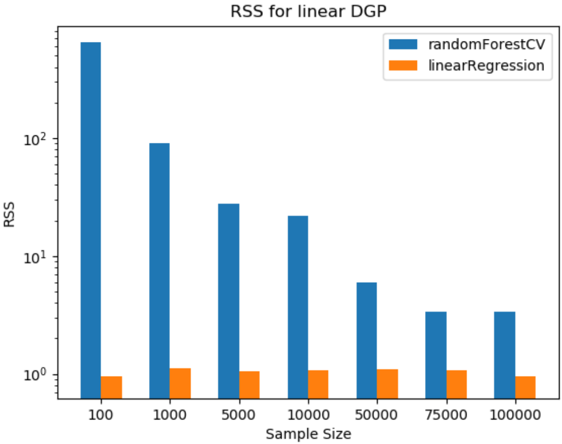
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon,$$

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



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

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

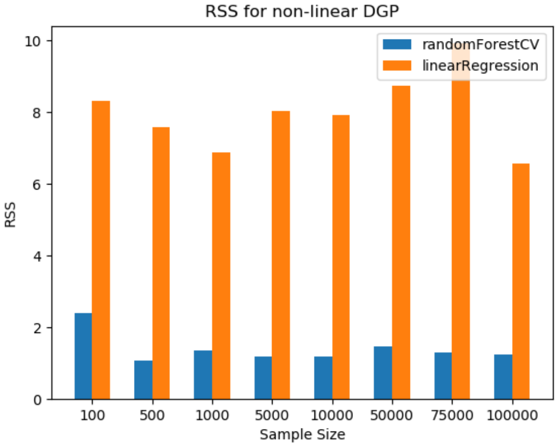
whereas

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

are the same as in the previous DGP.

Simulation Study

Non-Linear DGP Results



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%

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

Adaptive Boosting

Accuracy: 82.8%

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

Gradient Boosting

Accuracy: 82.8%

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

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: