Random Forest for Classification Problems

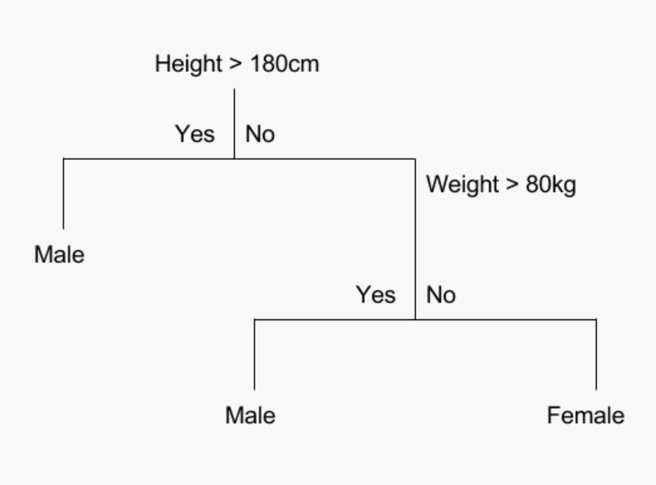
Raphael, Arkadiusz and Burak

Uni Bonn

January 17, 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 splitof 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

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

$$D = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$$

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.

$$y = f(x) + \epsilon \text{ and } \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$$
$$D = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$$

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.

$$y = f(x) + \epsilon \text{ and } \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$$
$$D = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$$

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

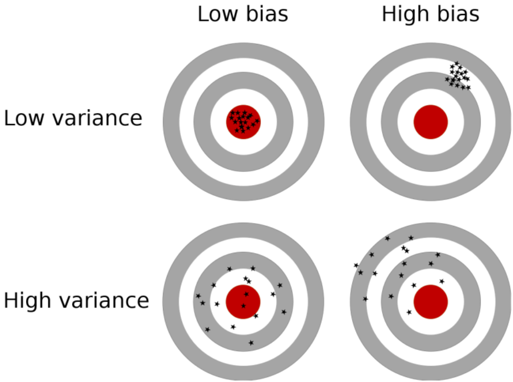


Figure: Illustration of bias-variance trade-off [3]

Decision trees generally have low bias and high variance [1].

Bias-Variance Trade-off
Illustration

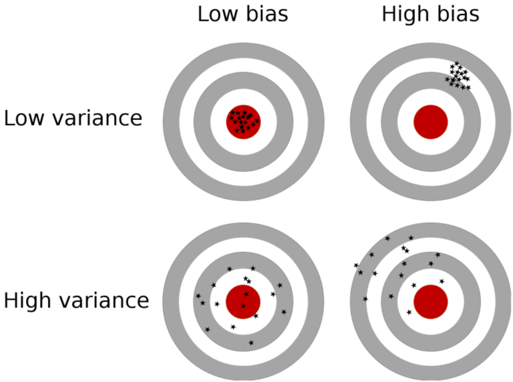
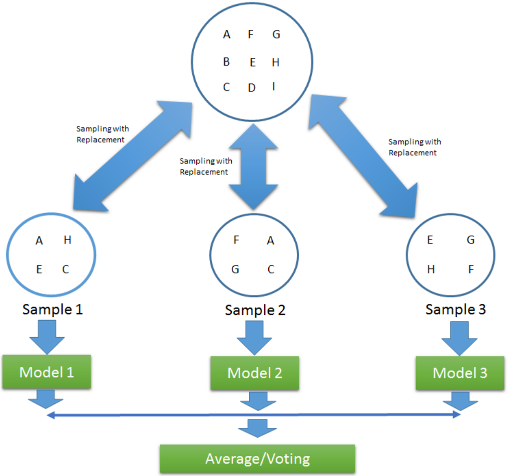


Figure: Illustration of bias-variance trade-off [3]

Decision trees generally have low bias and high variance [1].

Bagging

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



Random Forest

An ensemble of randomly trained decision trees, so in other words random forest was defined by L. Breiman:

Theorem

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:

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: $\lfloor \frac{n}{3} \rfloor$

Random Forest: algorithm

Algorithm 1: Random Forest for Regression or Classification

1. For $b = 1$ to B:
   1. Draw a bootstrap sample θ_b of size N from the training data.
   2. 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:
      1. Select m variables at random from the n variables
      2. Pick the best variable/split-point among the m
      3. Split the node into two daughter nodes
2. Output the ensemble of trees $\{T_{\theta_b}\}_1^B$

Section 2

Mathematical Concept

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

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

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

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

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

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

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

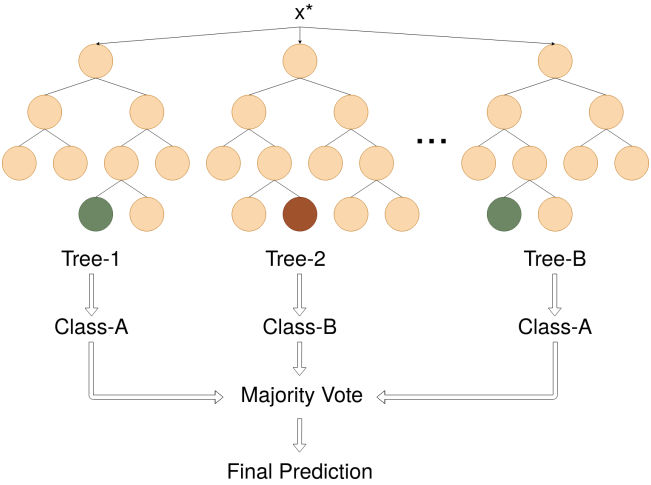
$$\boldsymbol{RF}_{D,\theta_1,\theta_2,...,\theta_B}(x^*) = \underset{c \in Y}{argmax} \sum_{b=1}^{B} 1(\hat{T}_b(x^*) = c)$$

Soft Voting

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

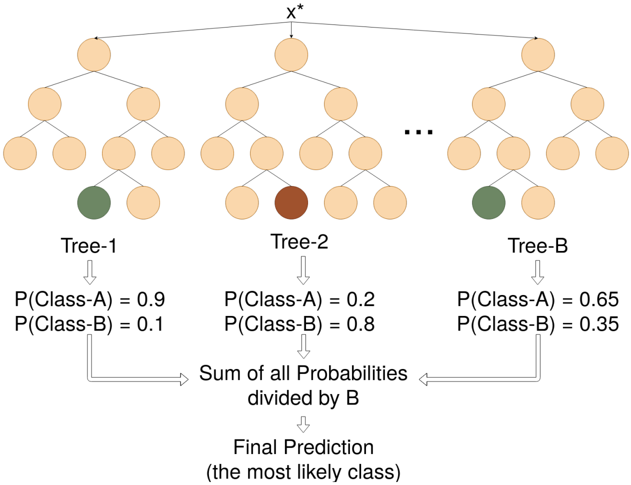
where $\hat{p}_{D,\theta_b}(Y = c|X = x^*)$ is the probability estimates of a tree.

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

Mathematical Concept

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

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

Mathematical Concept
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, there exists a model that minimizes the generalization error and can be derived analitically independent of the model [2].
Conditioning the expected generalization error on X gives:

$$\mathbb{E}_{X,Y}\{L(Y,\phi_\beta(X))\} = \mathbb{E}_X\{\mathbb{E}_{Y|X}\{L(Y,\phi_\beta(X))\}\}$$

Point-wise minimization of inner term yields:

$$\phi_\beta = \underset{c \in Y}{argmin}\, \mathbb{E}_{Y|X=x}\{L(Y,c)\}$$

Bayes Model ϕ_β is best attainable model.

$\boldsymbol{Err}(\phi_\beta(X))$ is the irreducible error.

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, there exists a model that minimizes the generalization error and can be derived analitically independent of the model [2].
Conditioning the expected generalization error on X gives:

$$\mathbb{E}_{X,Y}\{L(Y, \phi_\beta(X))\} = \mathbb{E}_X\{\mathbb{E}_{Y|X}\{L(Y, \phi_\beta(X))\}\}$$

Point-wise minimization of inner term yields:

$$\phi_\beta = \underset{c \in Y}{argmin}\, \mathbb{E}_{Y|X=x}\{L(Y, c)\}$$

- Bayes Model ϕ_β is best attainable model.
- $\boldsymbol{Err}(\phi_\beta(X))$ is the irreducible error.
- **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, there exists a model that minimizes the generalization error and can be derived analitically independent of the model [2].
Conditioning the expected generalization error on X gives:

$$\mathbb{E}_{X,Y}\{L(Y,\phi_\beta(X))\} = \mathbb{E}_X\{\mathbb{E}_{Y|X}\{L(Y,\phi_\beta(X))\}\}$$

Point-wise minimization of inner term yields:

$$\phi_\beta = \underset{c \in Y}{argmin}\, \mathbb{E}_{Y|X=x}\{L(Y,c)\}$$

Bayes Model ϕ_β is best attainable model.

$\boldsymbol{Err}(\phi_\beta(X))$ is the irreducible error.

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, there exists a model that minimizes the generalization error and can be derived analitically independent of the model [2].
Conditioning the expected generalization error on X gives:

$$\mathbb{E}_{X,Y}\{L(Y, \phi_\beta(X))\} = \mathbb{E}_X\{\mathbb{E}_{Y|X}\{L(Y, \phi_\beta(X))\}\}$$

Point-wise minimization of inner term yields:

$$\phi_\beta = \underset{c \in Y}{argmin}\, \mathbb{E}_{Y|X=x}\{L(Y, c)\}$$

Bayes Model ϕ_β is best attainable model.

$\boldsymbol{Err}(\phi_\beta(X))$ is the irreducible error.

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, there exists a model that minimizes the generalization error and can be derived analitically independent of the model [2].
Conditioning the expected generalization error on X gives:

$$\mathbb{E}_{X,Y}\{L(Y, \phi_\beta(X))\} = \mathbb{E}_X\{\mathbb{E}_{Y|X}\{L(Y, \phi_\beta(X))\}\}$$

Point-wise minimization of inner term yields:

$$\phi_\beta = \underset{c \in Y}{argmin}\, \mathbb{E}_{Y|X=x}\{L(Y, c)\}$$

Bayes Model ϕ_β is best attainable model.

$\boldsymbol{Err}(\phi_\beta(X))$ is the irreducible error.

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, there exists a model that minimizes the generalization error and can be derived analitically independent of the model [2].
Conditioning the expected generalization error on X gives:

$$\mathbb{E}_{X,Y}\{L(Y,\phi_\beta(X))\} = \mathbb{E}_X\{\mathbb{E}_{Y|X}\{L(Y,\phi_\beta(X))\}\}$$

Point-wise minimization of inner term yields:

$$\phi_\beta = \underset{c \in Y}{argmin}\, \mathbb{E}_{Y|X=x}\{L(Y,c)\}$$

Bayes Model ϕ_β is best attainable model.

$\boldsymbol{Err}(\phi_\beta(X))$ is the irreducible error.

Result: Ensembling has no effect on Bayes Error.

Mathematical Concept
Bias

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

$Bias^2$ 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 $\mathbb{E}_{D,\Theta}\{\boldsymbol{RF}_{D,\Theta}(X))\}$.

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

$Bias^2$ 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 $\mathbb{E}_{D,\Theta}\{\boldsymbol{RF}_{D,\Theta}(X))\}$.

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

$Bias^2$ 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 $\mathbb{E}_{D,\Theta}\{\boldsymbol{RF}_{D,\Theta}(X))\}$.

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

$Bias^2$ 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 $\mathbb{E}_{D,\Theta}\{\boldsymbol{RF}_{D,\Theta}(X))\}$.

Mathematical Concept

Bias: The Expected Value of Random Forest

Random Forest Estimator for regression shares the same idea with soft-voting classification;

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

Taking expectation gives;

$$\begin{aligned}\mathbb{E}_{D,\Theta}\{\boldsymbol{RF}_{D,\Theta}(X)\} &= \mathbb{E}_{D,\Theta}\{\frac{1}{B}\sum_{b=1}^{B} T_{D,\theta_b}(X)\} \\ &= \frac{1}{B}\sum_{b=1}^{B} \mathbb{E}_{D,\theta_b}\{T_{D,\theta_b}(X)\} \\ &= \mu_{D,\hat{\theta}}(X)\end{aligned}$$

where $\mu_{D,\hat{\theta}}(X)$ is the average prediction of all ensembled trees.

Mathematical Concept

Bias: The Expected Value of Random Forest

Random Forest Estimator for regression shares the same idea with soft-voting classification;

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

Taking expectation gives;

$$\begin{aligned}
\mathbb{E}_{D,\Theta}\{\boldsymbol{RF}_{D,\Theta}(X)\} &= \mathbb{E}_{D,\Theta}\{\frac{1}{B}\sum_{b=1}^{B} T_{D,\theta_b}(X)\} \\
&= \frac{1}{B}\sum_{b=1}^{B} \mathbb{E}_{D,\theta_b}\{T_{D,\theta_b}(X)\} \\
&= \mu_{D,\hat{\theta}}(X)
\end{aligned}$$

where $\mu_{D,\hat{\theta}}(X)$ is the average prediction of all ensembled trees.

Random Forest Estimator for regression shares the same idea with soft-voting classification;

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

Taking expectation gives;

$$\begin{aligned}
\mathbb{E}_{D,\Theta}\{\boldsymbol{RF}_{D,\Theta}(X)\} &= \mathbb{E}_{D,\Theta}\{\frac{1}{B}\sum_{b=1}^{B} T_{D,\theta_b}(X)\} \\
&= \frac{1}{B}\sum_{b=1}^{B} \mathbb{E}_{D,\theta_b}\{T_{D,\theta_b}(X)\} \\
&= \mu_{D,\hat{\theta}}(X)
\end{aligned}$$

where $\mu_{D,\hat{\theta}}(X)$ is the average prediction of all ensembled trees.

Random Forest Estimator for regression shares the same idea with soft-voting classification;

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

Taking expectation gives;

$$\begin{aligned}
\mathbb{E}_{D,\Theta}\{\boldsymbol{RF}_{D,\Theta}(X)\} &= \mathbb{E}_{D,\Theta}\{\frac{1}{B}\sum_{b=1}^{B} T_{D,\theta_b}(X)\} \\
&= \frac{1}{B}\sum_{b=1}^{B} \mathbb{E}_{D,\theta_b}\{T_{D,\theta_b}(X)\} \\
&= \mu_{D,\hat{\theta}}(X)
\end{aligned}$$

where $\mu_{D,\hat{\theta}}(X)$ is the average prediction of all ensembled trees.

Mathematical Concept

Bias Comparison

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

$Bias^2$ of Tree

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

$Bias^2$ of a Random Forest

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

Result: Ensembling trees does not necessarily decrease $Bias^2$.

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

$Bias^2$ of Tree

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

$Bias^2$ of a Random Forest

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

Result: Ensembling trees does not necessarily decrease $Bias^2$.

Mathematical Concept

Variance: Correlation Coefficient

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

For any two trees $T_{D,\theta'}$ and $T_{D,\theta''}$ trained with the same data D and different growing parameters θ' and θ'', we can define the correlation coefficient as follows

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

$\rho(X)$ is close to 1

- Highly correlated trees.
- Randomization has no significant effect.

$\rho(X)$ is close to 0

- Non-correlated trees
- Trees are perfectly random.

Mathematical Concept

Variance: Correlation Coefficient

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

For any two trees $T_{D,\theta'}$ and $T_{D,\theta''}$ trained with the same data D and different growing parameters θ' and θ'', we can define the correlation coefficient as follows

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

$\rho(X)$ is close to 1

- Highly correlated trees.
- Randomization has no significant effect.

$\rho(X)$ is close to 0

- Non-correlated trees
- Trees are perfectly random.

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

For any two trees $T_{D,\theta'}$ and $T_{D,\theta''}$ trained with the same data D and different growing parameters θ' and θ'', we can define the correlation coefficient as follows

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

$\rho(X)$ is close to 1

- Highly correlated trees.
- Randomization has no significant effect.

$\rho(X)$ is close to 0

- Non-correlated trees
- Trees are perfectly random.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Result: Ensembling trees decreases the variance.

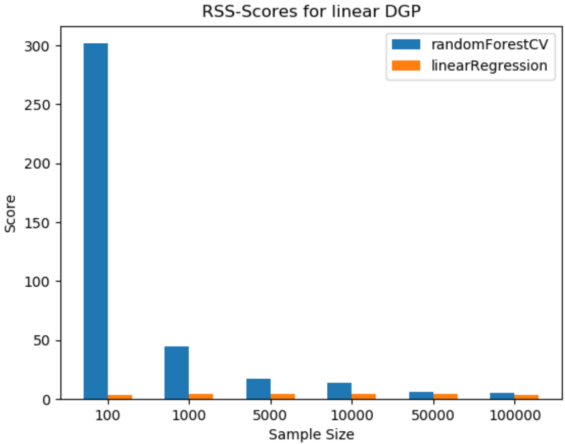
Simulation Study

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

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon, \tag{1}$$

whereas

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


RSS-Scores for linear DGP
randomForestCV
linearRegression
Score
300
250
200
150
100
50
0
100
1000
5000
10000
50000
100000
Sample Size

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

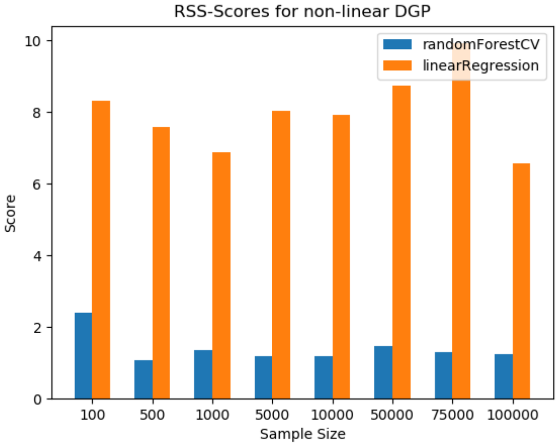
$$y = \beta_0 + \beta_1 \mathbb{1}(x_1 \geq 0, x_2 \geq 0) + \beta_2 \mathbb{1}(x_1 \geq 0, x_2 < 0) + \beta_3 \mathbb{1}(x_1 < 0) + \epsilon, \quad (2)$$

whereas

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

are the same as in the previous DGP.

Simulation Study
Non-Linear DGP Results



Real Data

Data: Titanic data

Method used:

1. Random Forest
2. AdaBoost
3. Gradient Boosting Classifier

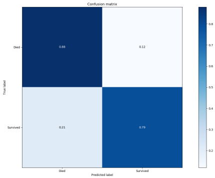
Goal: Given features of passengers predict which passengers survived the Titanic shipwreck

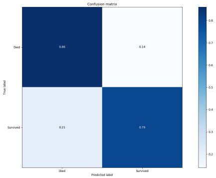
Real Data: results

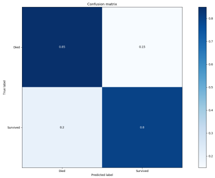
Random Forest
Accuracy: 84,32%

AdaBoost
Accuracy: 82.8%

Gradient Boosting
Accuracy: 82,8%







Contact Data

- Burak Balaban:
- Raphael Redmer: ra.redmer@outlook.com
- Arkadiusz Modzelewski: