DATASCI 347 Machine Learning

Lecture 13: Bagging

Ruoxuan Xiong

Suggested reading: ISL Chapter 8

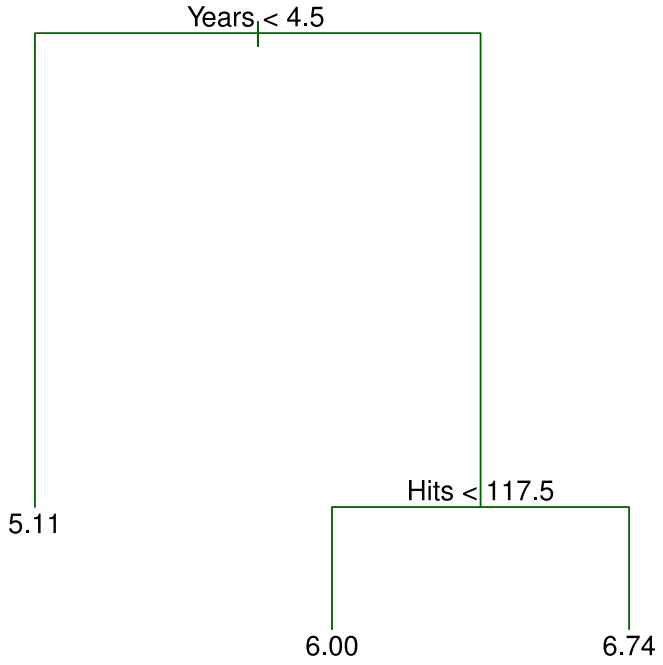
Decision tree

- **Tree construction**
  - Partition the feature space into J **distinct and non-overlapping** regions, $R_1, R_2, \cdots, R_J$
  - Regression tree: Mean of the training observations in R_j as the predicted value for every point in region R_j
  - Classification tree: Pick *the most common class* of the training observations in R_j as the predicted value for every point in region R_j

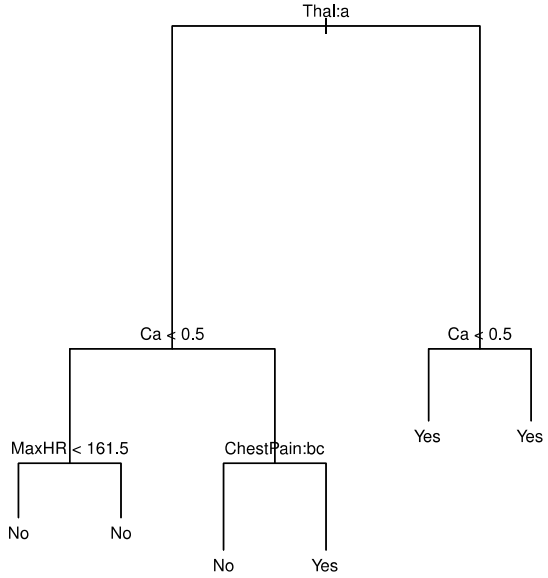
- **Tree pruning** to avoid overfitting, e.g., use cost complexity pruning

Example of regression and classification trees

- Predict a baseball player's salary

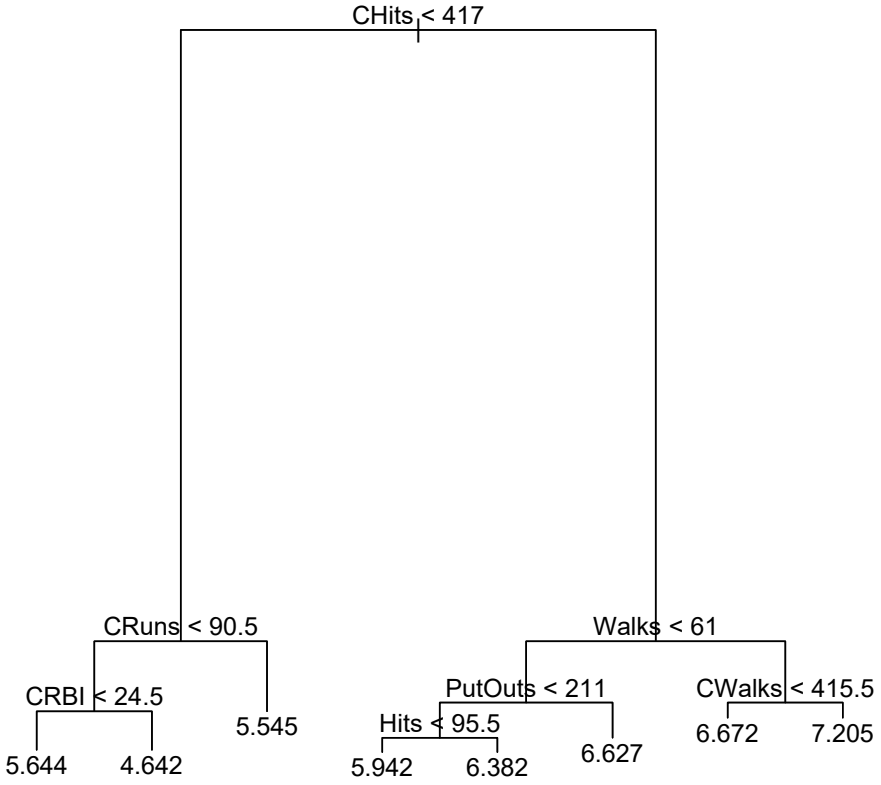


- Predict heart disease (yes or no)

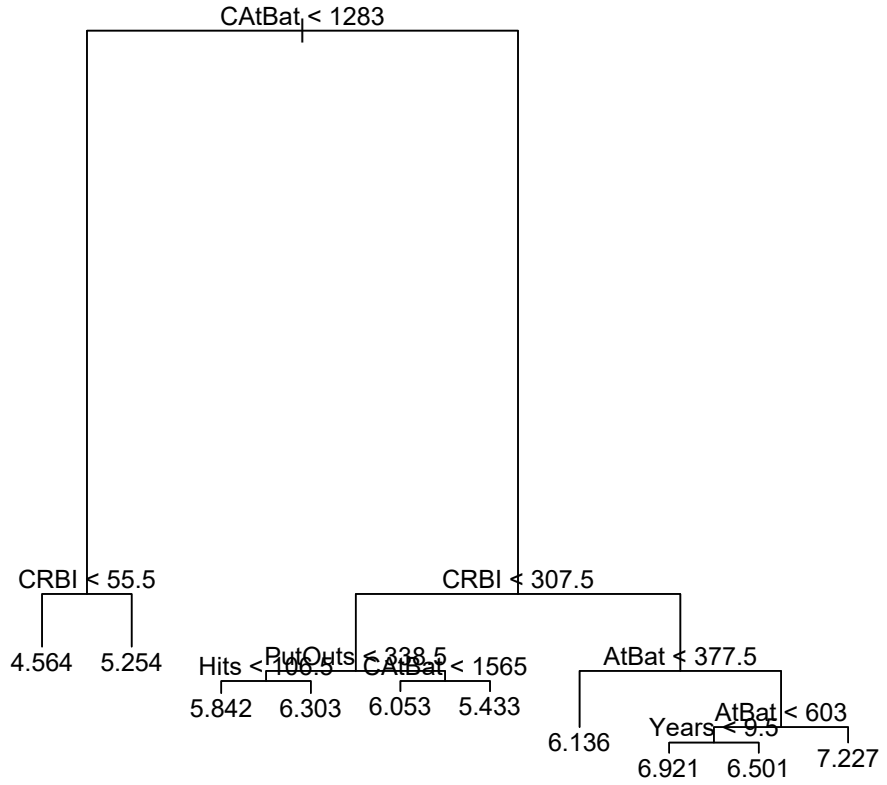


Decision tree has a high variance

- **Example:** Predicting a baseball player's salary
  - Split the training data into two equal-sized parts at random creates disparity



Subsample 1



Subsample 2

Bagging

- Bagging is a way to reduce such variance
- **Idea: Bootstrap aggregation**
- **Example**: Estimate the mean of Z

Z_1	1.03
Z_2	1.56
Z_3	2.37
Z_4	2.13
Z_5	2.47

$$\bar{Z} = 1.91$$

$$\text{Var}(\bar{Z}) = \frac{\sigma^2}{n} = \frac{1}{5} = 0.2$$

Data generating process: $Z \sim N(2,1)$

Toy example

- Suppose we have many independent sampling of data sets

Data set 1

$Z_1^{(1)}$	1.03
$Z_2^{(1)}$	1.56
$Z_3^{(1)}$	2.37
$Z_4^{(1)}$	2.13
$Z_5^{(1)}$	2.47

$\bar{Z}^{(1)} = 1.91$
$\mathrm{Var}(\bar{Z}^{(1)}) = 0.2$

Data set 2

$Z_1^{(2)}$	3.44
$Z_2^{(2)}$	3.06
$Z_3^{(2)}$	2.42
$Z_4^{(2)}$	2.40
$Z_5^{(2)}$	-0.78

$\bar{Z}^{(2)} = 2.11$
$\mathrm{Var}(\bar{Z}^{(2)}) = 0.2$

Data set 3

$Z_1^{(3)}$	-0.13
$Z_2^{(3)}$	2.28
$Z_3^{(3)}$	2.09
$Z_4^{(3)}$	2.72
$Z_5^{(3)}$	1.40

$\bar{Z}^{(3)} = 1.67$
$\mathrm{Var}(\bar{Z}^{(3)}) = 0.2$

Data set 4

$Z_1^{(4)}$	0.94
$Z_2^{(4)}$	1.84
$Z_3^{(4)}$	1.92
$Z_4^{(4)}$	2.49
$Z_5^{(4)}$	2.37

$\bar{Z}^{(4)} = 1.91$
$\mathrm{Var}(\bar{Z}^{(4)}) = 0.2$

$$\bar{Z}_{agg} = (\bar{Z}^{(1)} + \bar{Z}^{(2)} + \bar{Z}^{(3)} + \bar{Z}^{(4)})/4 = 1.90$$

$$\mathrm{Var}(\bar{Z}_{agg}) = \frac{0.2}{4} = 0.05$$

Toy example

- In practice, we only have one training data set
- How can we create many data sets? **Idea: Bootstrap**

Z_1	1.03
Z_2	1.56
Z_3	2.37
Z_4	2.13
Z_5	2.47

Sampling with replacement

Sample #1

Z_1	1.03
Z_2	1.56
Z_1	1.03
Z_5	2.47
Z_4	2.13

Sample #2

Z_4	2.13
Z_1	1.03
Z_3	2.37
Z_2	1.56
Z_3	2.37

Sample #3

Z_5	2.47
Z_2	1.56
Z_3	2.37
Z_2	1.56
Z_1	1.03

Sample #4

Z_5	2.47
Z_3	2.37
Z_3	2.37
Z_1	1.03
Z_2	1.56

Bagging to reduce variance

- Estimate the mean on each bootstrap sampling set

Sample #1

Z_1	1.03
Z_2	1.56
Z_5	2.47
Z_5	2.47
Z_4	2.13

$\bar{Z}^{(1)} = 1.93$

Sample #2

Z_4	2.13
Z_1	1.03
Z_3	2.37
Z_2	1.56
Z_3	2.37

$\bar{Z}^{(2)} = 1.89$

Sample #3

Z_5	2.47
Z_2	1.56
Z_3	2.37
Z_2	1.56
Z_1	1.03

$\bar{Z}^{(3)} = 1.80$

Sample #4

Z_5	2.47
Z_3	2.37
Z_3	2.37
Z_1	1.03
Z_2	1.56

$\bar{Z}^{(4)} = 1.96$

Toy example

- Average all estimates

$$\bar{Z}^{(1)} = 1.93 \qquad \bar{Z}^{(2)} = 1.89 \qquad \bar{Z}^{(3)} = 1.80 \qquad \bar{Z}^{(4)} = 1.96$$

$$\bar{Z}_{bag} = (\bar{Z}^{(1)} + \bar{Z}^{(2)} + \bar{Z}^{(3)} + \bar{Z}^{(4)})/4 = 1.90$$

- This is called **bagging** (**B**ootstrap **agg**regat**ing**)
  - Bagging amounts to averaging the fits from B "weakly" dependent data sets, which would reduce the variance by approximately a factor $\frac{1}{B}$

Bagging for decision trees

- Estimate a decision tree model $f(x)$ using bootstrap

X_1	Y_1
X_2	Y_2
X_3	Y_3
X_4	Y_4
X_5	Y_5

Sampling with replacement

Sample #1

X_1	Y_1
X_2	Y_2
X_1	Y_1
X_5	Y_5
X_4	Y_4

Sample #2

X_4	Y_4
X_1	Y_1
X_3	Y_3
X_2	Y_2
X_3	Y_3

Sample #3

X_5	Y_5
X_2	Y_2
X_3	Y_3
X_2	Y_2
X_1	Y_1

Sample #4

X_5	Y_5
X_3	Y_3
X_3	Y_3
X_1	Y_1
X_2	Y_2

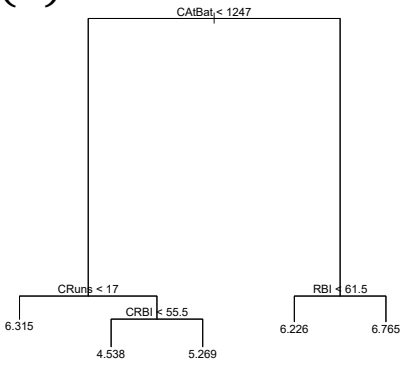
Bagging for decision trees

- Estimate a decision tree model $f(x)$ using bootstrap

Sample #1

X_1	Y_1
X_2	Y_2
X_1	Y_1
X_5	Y_5
X_4	Y_4

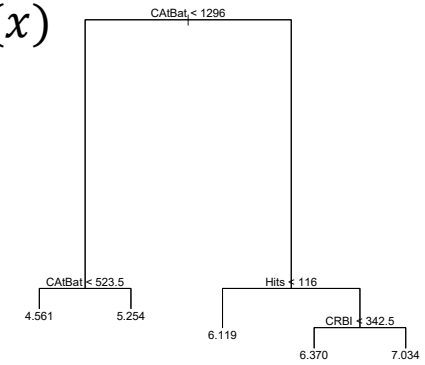
$\hat{f}^1(x)$



Sample #2

X_4	Y_4
X_1	Y_1
X_3	Y_3
X_2	Y_2
X_3	Y_3

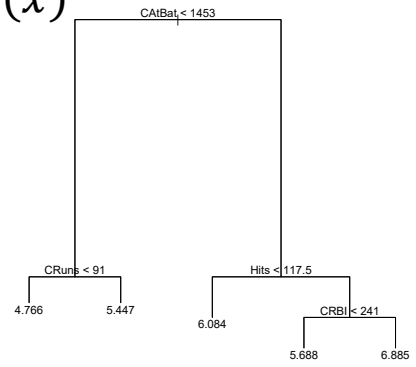
$\hat{f}^2(x)$



Sample #3

X_5	Y_5
X_2	Y_2
X_3	Y_3
X_2	Y_2
X_1	Y_1

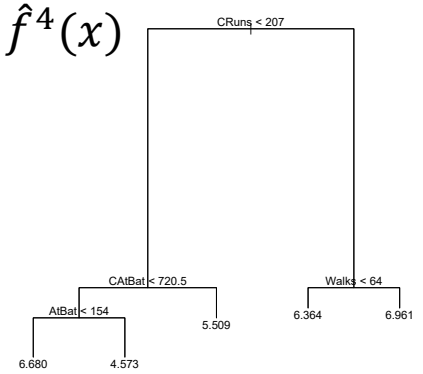
$\hat{f}^3(x)$



Sample #4

X_5	Y_5
X_3	Y_3
X_3	Y_3
X_1	Y_1
X_2	Y_2

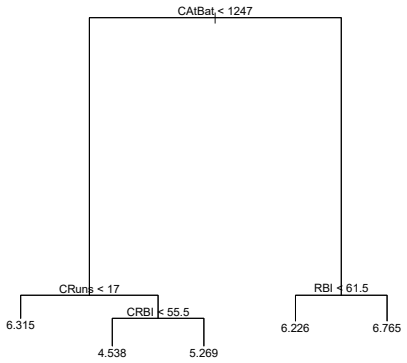
$\hat{f}^4(x)$



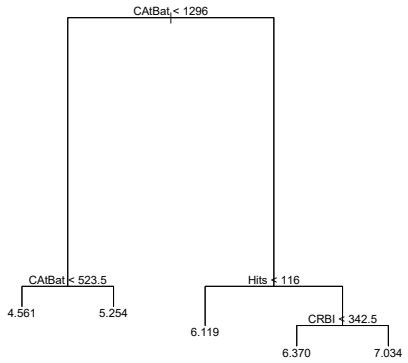
Bagging to reduce variance

- Average all the predictions

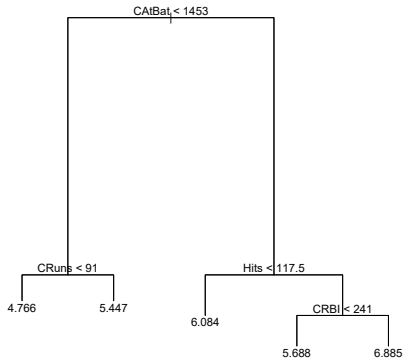
$\hat{f}^1(x)$



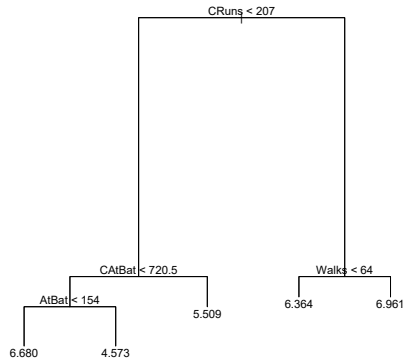
$\hat{f}^2(x)$



$\hat{f}^3(x)$



$\hat{f}^4(x)$



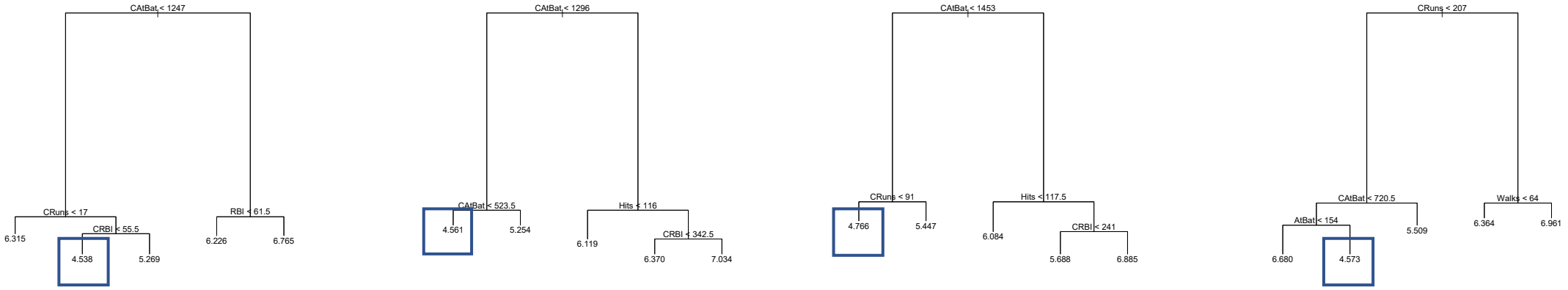
$$\hat{f}_{bag}(x) = \frac{1}{4}\{\hat{f}^1(x) + \hat{f}^2(x) + \hat{f}^3(x) + \hat{f}^4(x)\}$$

- If we have B bootstrapped samples, $\hat{f}_{bag}(x) = \frac{1}{B}\{\hat{f}^1(x) + \hat{f}^2(x) + \cdots + \hat{f}^B(x)\}$

Example

	AtBat	Hits	HmRun	Runs	RBI	Walks	Years	CAtBat	CHits	CHmRun	CRuns	CRBI	CWalks	League	Division	PutOuts	Assists	Errors	Salary	NewLeague
-Andy Allanson	293	66	1	30	29	14	1	293	66	1	30	29	14	A	E	446	33	20	NA	A

$\hat{f}^1(x)$ $\hat{f}^2(x)$ $\hat{f}^3(x)$ $\hat{f}^4(x)$



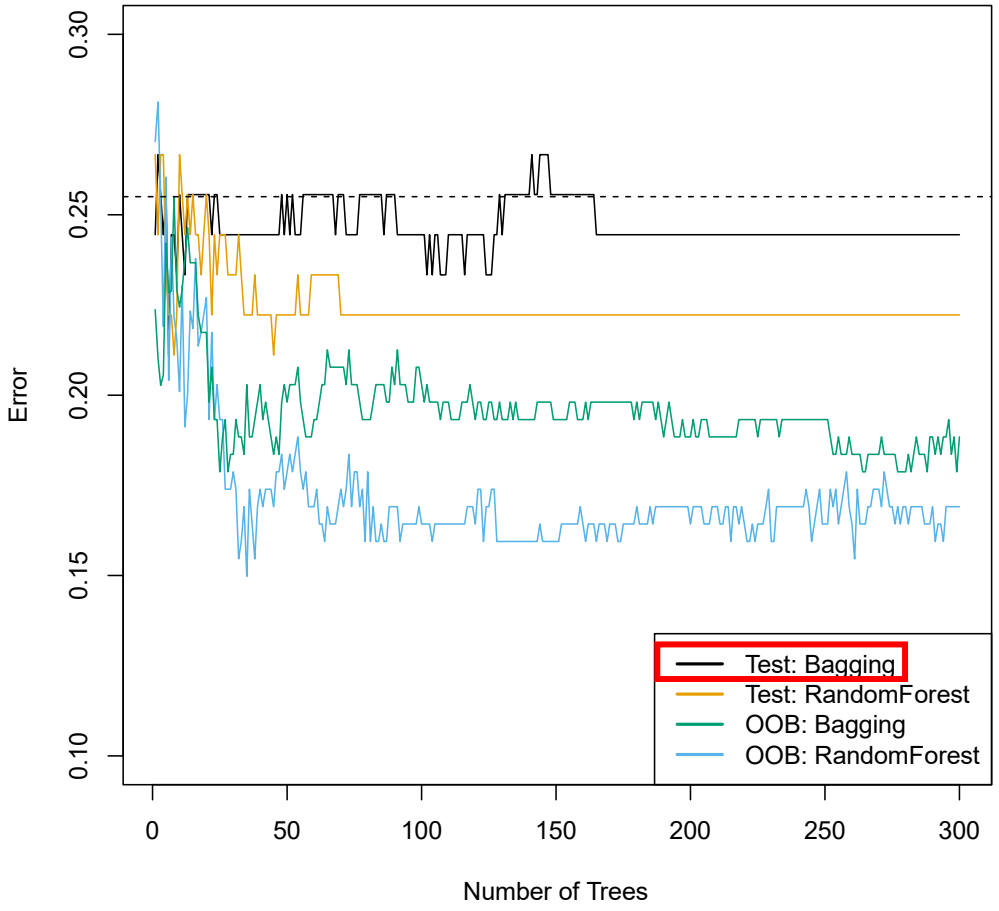
$\hat{f}^1(x) = 4.538$ $\hat{f}^2(x) = 4.561$ $\hat{f}^3(x) = 4.766$ $\hat{f}^4(x) = 4.573$

$$\hat{f}_{bag}(x) = \frac{1}{4}\{\hat{f}^1(x) + \hat{f}^2(x) + \hat{f}^3(x) + \hat{f}^4(x)\} = (4.538 + 4.561 + 4.766 + 4.573)/4 = 4.6095$$

- If the problem is classification, how should we aggregate the predictions?

Example: Predicting heart disease

- **Example:** Predict whether a patient with chest pain has heart disease based on Age, Sex, Chol (a cholesterol measure), and other heart and lung function measures
- Dash line: Single tree
- Bagging outperforms a single decision tree
- The number of trees B does not matter after some threshold
- In practice, $B = 100$ is sufficient
  - When error has settled down



Out-of-bag (OOB) error estimation

- **Cross-validation:** To estimate the test error of a bagging estimate, we could use cross-validation
- How should we perform cross-validation with Bootstrap?
- Each time we draw a bootstrap sample, we only use 63% of the observations
  - Related to Problem 1 in Homework 2
  - We can show that an observation is not in the bootstrap sample is $\left(1-\frac{1}{n}\right)^n \approx \frac{1}{e} = 0.37$
- **Idea**: Use the rest of the observations as a **hold out set**

Out-of-bag (OOB) error estimation

- **Idea**: Use the rest of the observations as a **hold out set**
- **Out-of-bag (OOB) error:**
  - For each sample X_i, find the prediction $\hat{Y}_i^b$ for all bootstrap samples b which do not contain X_i
  - Around $0.37B$ of them. Average these predictions to obtain $\hat{Y}_i^{oob}$
- **Example**: For the observation X_4, predict $\hat{Y}_4^b$

$$\hat{Y}_4^{oob} = \frac{1}{2}(\hat{Y}_4^3 + \hat{Y}_4^4)$$

$\hat{Y}_4^1$
Sample #1

X_1	Y_1
X_2	Y_2
X_1	Y_1
X_5	Y_5
X_4	Y_4

$\hat{Y}_4^2$
Sample #2

X_4	Y_4
X_1	Y_1
X_3	Y_3
X_2	Y_2
X_3	Y_3

$\hat{Y}_4^3$
Sample #3

X_5	Y_5
X_2	Y_2
X_3	Y_3
X_2	Y_2
X_1	Y_1

$\hat{Y}_4^4$
Sample #4

X_5	Y_5
X_3	Y_3
X_3	Y_3
X_1	Y_1
X_2	Y_2

Out-of-bag (OOB) error estimation

- **Out-of-bag (OOB) error:**
  - **Step 1:** For each sample X_i, find the prediction $\hat{Y}_i^b$ for all bootstrap samples b which do not contain X_i. These should be around $0.37B$ of them. Average these predictions to obtain $\hat{Y}_i^{oob}$
  - **Step 2:** Compute the error $(Y_i - \hat{Y}_i^{oob})^2$
  - **Step 3:** Average the errors over all observations $i = 1, \cdots, n$
- **Example:**

$$\frac{1}{5}\{(Y_1 - \hat{Y}_1^{oob})^2 + (Y_2 - \hat{Y}_2^{oob})^2 + \cdots + (Y_5 - \hat{Y}_5^{oob})^2\}$$

Sample #1

X_1	Y_1
X_2	Y_2
X_1	Y_1
X_5	Y_5
X_4	Y_4

Sample #2

X_4	Y_4
X_1	Y_1
X_3	Y_3
X_2	Y_2
X_3	Y_3

Sample #3

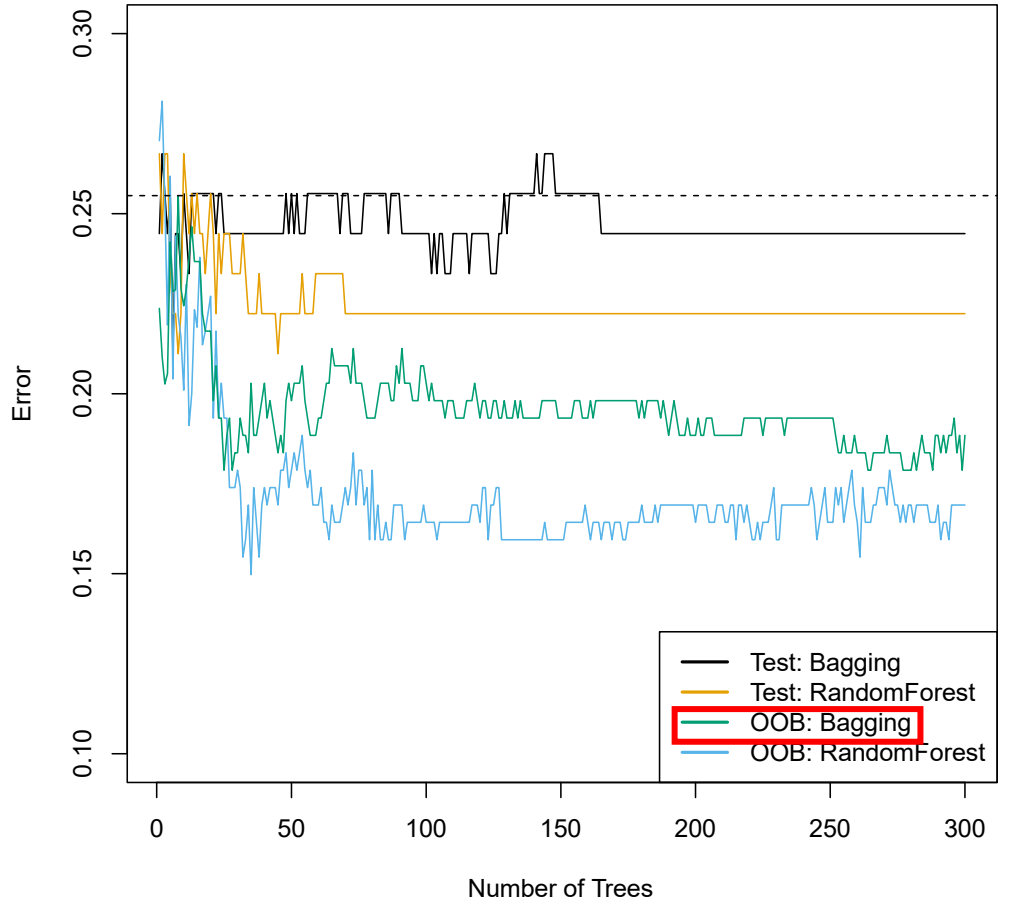
X_5	Y_5
X_2	Y_2
X_3	Y_3
X_2	Y_2
X_1	Y_1

Sample #4

X_5	Y_5
X_3	Y_3
X_3	Y_3
X_1	Y_1
X_2	Y_2

Out-of-bag (OOB) error

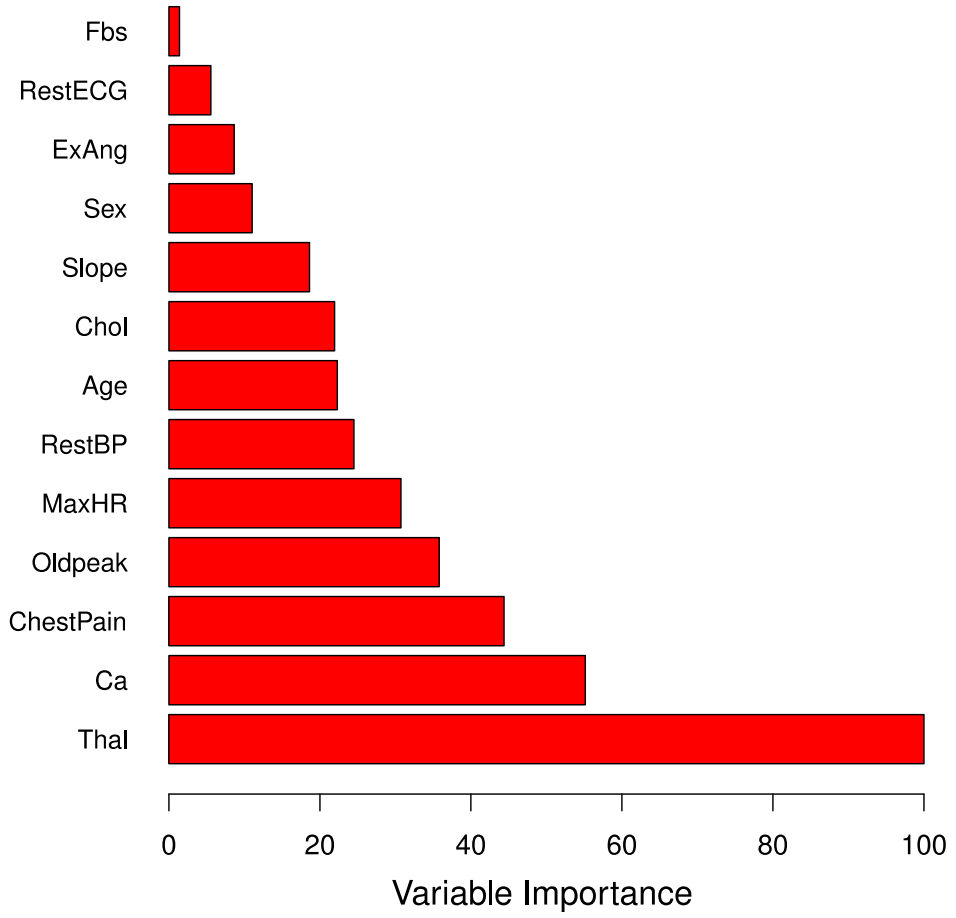
- **Example:** Predict whether a patient with chest pain has heart disease
  - OOB error follows a similar trend to test error



Bagging decision trees

- **Disadvantage**: Loss of interpretability
  - Every time we fit a decision tree to a Bootstrap sample, we get a different tree
- **Solution: Variable importance**
  - For each predictor, add up the total amount by which the RSS (or Gini index) decreases every time we use the predictor in
  - Average the total over each Bootstrap estimate $T^1, \cdots, T^B$

- **Example:** Predicting heart disease



Recall the classification tree to predict heart disease

- **Example**: Predict whether a patient with chest pain has heart disease based on Age, Sex, Chol (a cholesterol measure), and other heart and lung function measures