DATASCI 347 Machine Learning

Lecture 12: Classification tree and bagging

Ruoxuan Xiong

Suggested reading: ISL Chapter 8

Regression tree

- **Two main steps** in constructing regression trees
  1. Partition the feature space into J **distinct and non-overlapping** regions, $R_1, R_2, \cdots, R_J$
  2. Make the **same** prediction for every observation in region R_j: Mean of the training observations in R_j

- Tree pruning to avoid overfitting

Classification tree

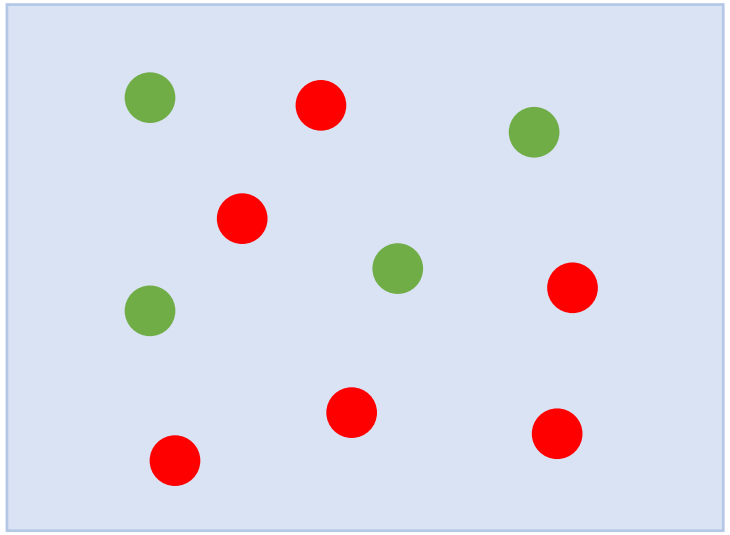
- Classification trees work much like regression trees. Instead,
- In step 1, minimize the classification error rate (rather than RSS)
- In step 2, predict the response by **majority vote**, i.e. pick *the most common class* in every region

Classification losses: The 0–1 loss

- The 0–1 loss or misclassification rate in region m:

$$\sum_{i \in R_m} 1(Y_i \neq \hat{Y}_{R_m})$$

- Example:
  - $\hat{Y}_{R_m} = \text{red}$
  - $\sum_{i \in R_m} 1(Y_i \neq \hat{Y}_{R_m}) = 4$



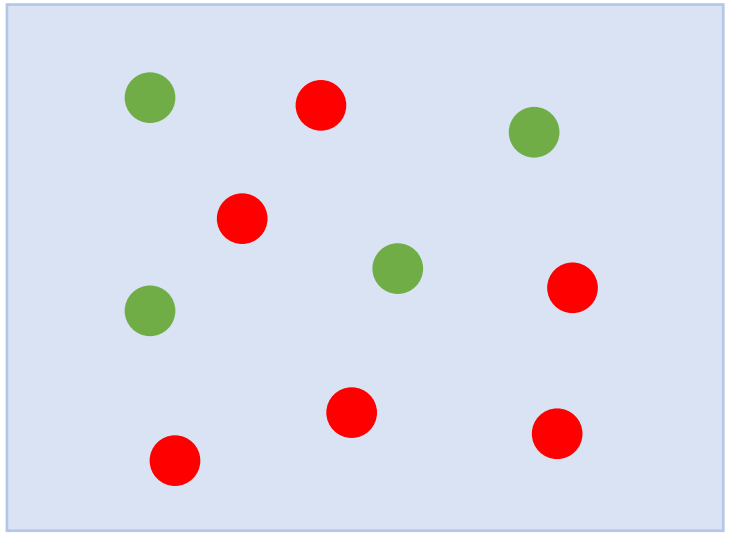
Region m

Classification losses: Gini index

- The Gini index in region m

$$G_m = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

- p_{mk}: proportion of training observations in the mth region that are from kth class
- Example:
  - $\hat{p}_{m,\text{red}} = \frac{6}{10} = 0.6$
  - $\hat{p}_{m,\text{green}} = \frac{4}{10} = 0.4$
  - $G_m = 0.6(1 - 0.6) + 0.4(1 - 0.4) = 0.48$



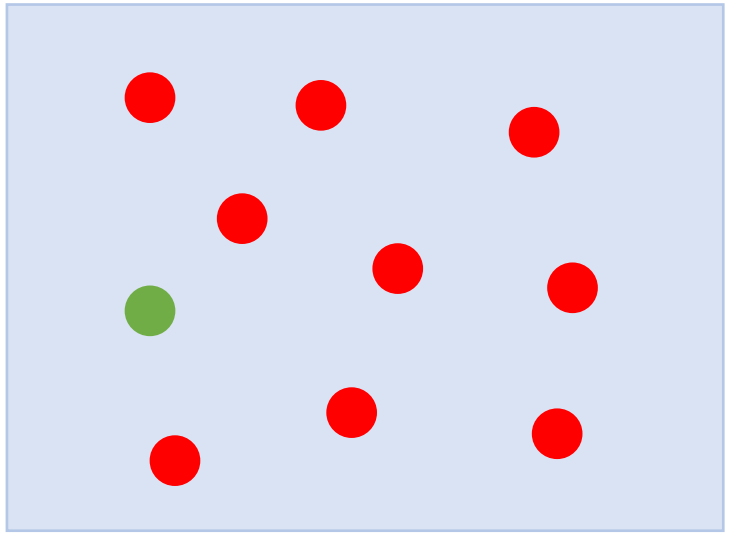
Region m

Classification losses: Gini index

- The Gini index in region m

$$G_m = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

- p_{mk}: proportion of training observations in the mth region that are from kth class
- Example:
  - $\hat{p}_{m,\text{red}} = \frac{9}{10} = 0.9$
  - $\hat{p}_{m,\text{green}} = \frac{1}{10} = 0.1$
  - $G_m = 0.9(1 - 0.9) + 0.1(1 - 0.1) = 0.18$
- G_m is a measure of node purity
  - G_m is small if all $\hat{p}_{mk}$'s are close to zero or one



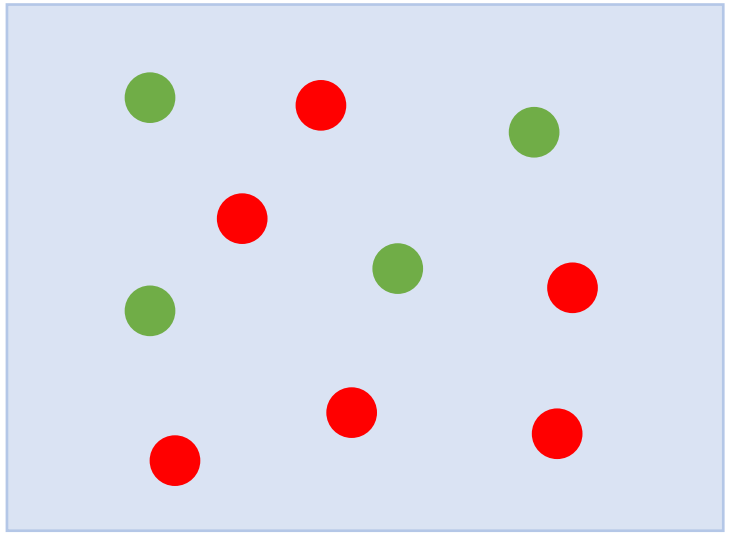
Region m

Classification losses: Entropy

- The entropy in region m

$$D_m = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$$

- p_{mk}: proportion of training observations in the mth region that are from kth class
- Example:
  - $\hat{p}_{m,\text{red}} = \frac{6}{10} = 0.6$
  - $\hat{p}_{m,\text{green}} = \frac{4}{10} = 0.4$
  - $D_m = -0.6 \log 0.6 - 0.4 \log 0.4 = 0.673$



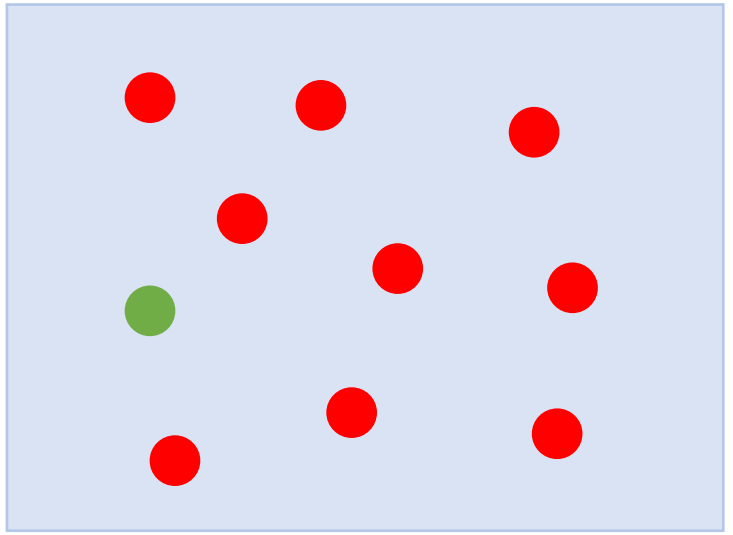
Region m

Classification losses: Entropy

- The entropy in region m

$$D_m = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$$

- p_{mk}: proportion of training observations in the mth region that are from kth class
- Example:
  - $\hat{p}_{m,\text{red}} = \frac{9}{10} = 0.9$
  - $\hat{p}_{m,\text{green}} = \frac{1}{10} = 0.1$
  - $D_m = -0.9 \log 0.9 - 0.1 \log 0.1 = 0.461$
- D_m is another measure of purity
  - D_m is small if all $\hat{p}_{mk}$'s are close to zero or one



Region m

Classification losses

- The 0–1 loss or misclassification rate in region m (prune tree)

$$\sum_{i \in R_m} 1(Y_i \neq \hat{Y}_{R_m})$$

- The Gini index in region m (evaluate the quality of a split)

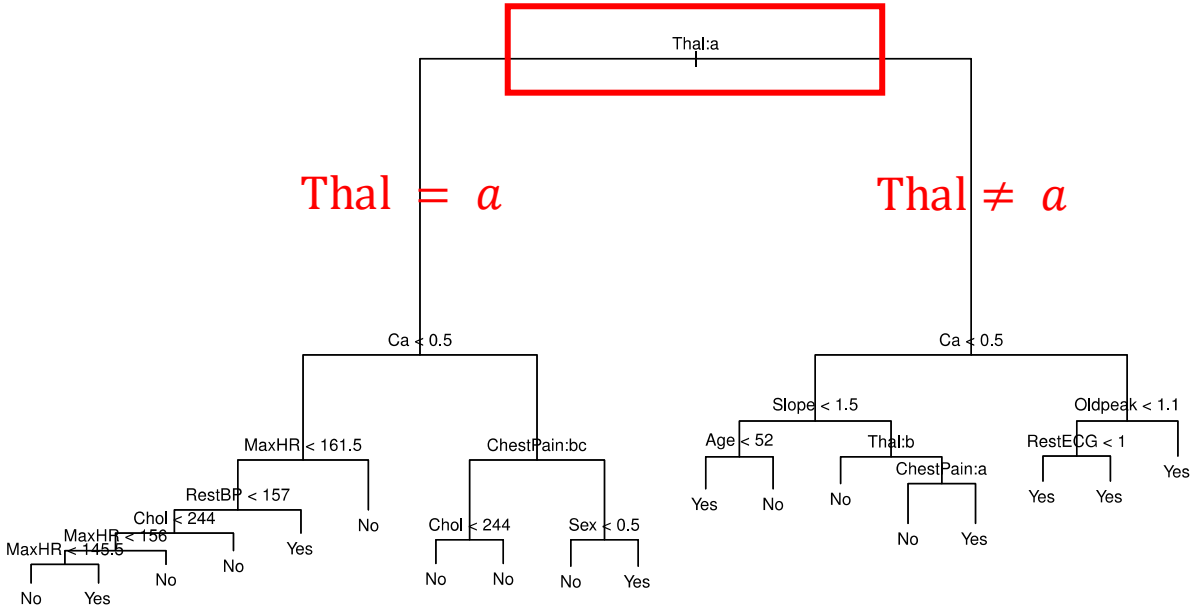
$$G_m = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

- The entropy in region m (evaluate the quality of a split)

$$D_m = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$$

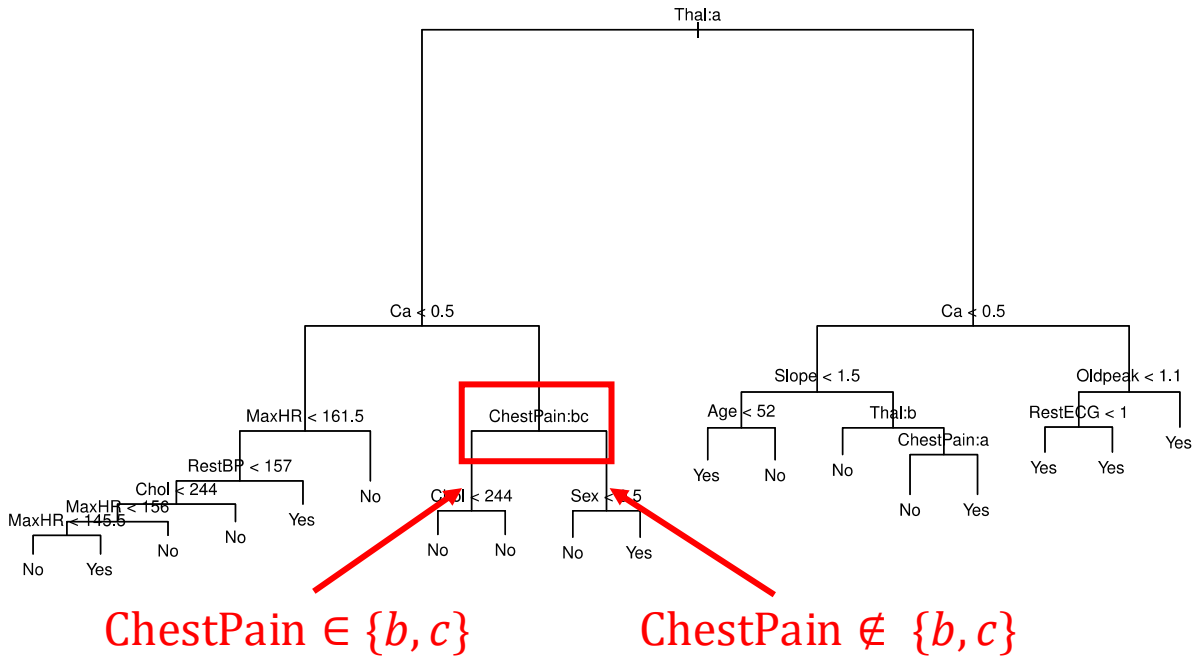
Example: Predicting heart disease

- 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
- Some predictors are qualitative
  - Thal (Thallium stress test)
  - ChestPain
  - Sex



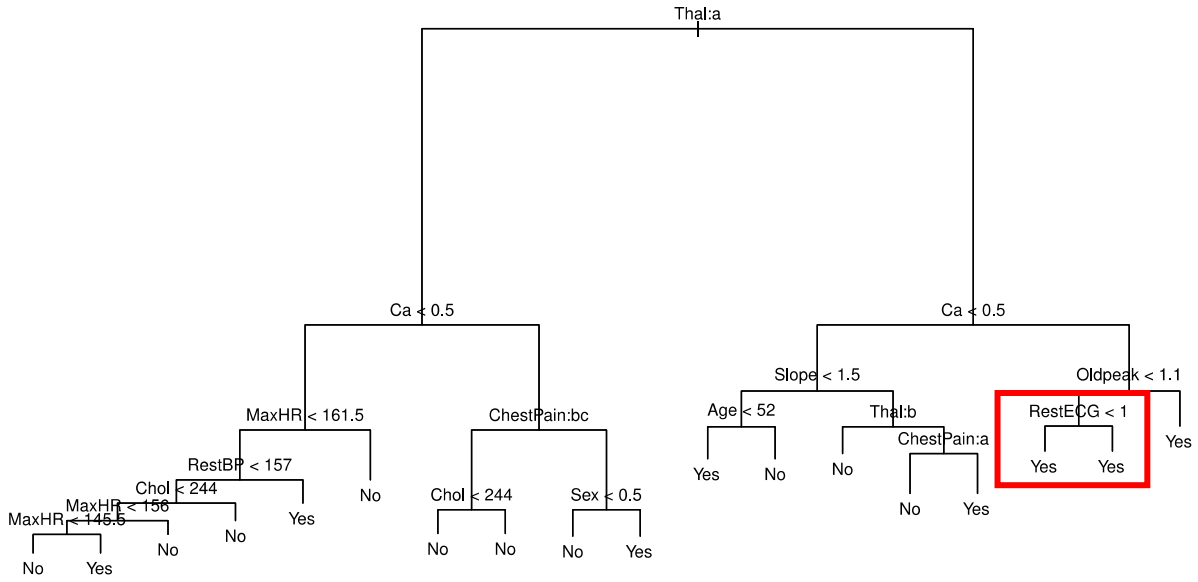
Example: Predicting heart disease

- 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
- Some predictors are qualitative
  - Thal (Thallium stress test)
  - ChestPain
  - Sex



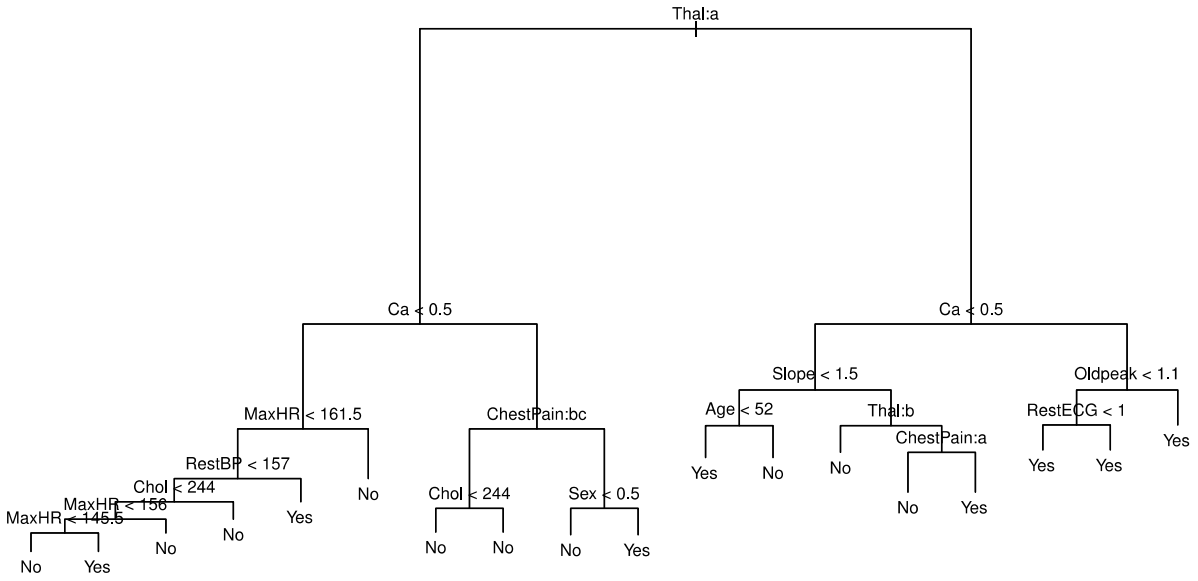
Example: Predicting heart disease

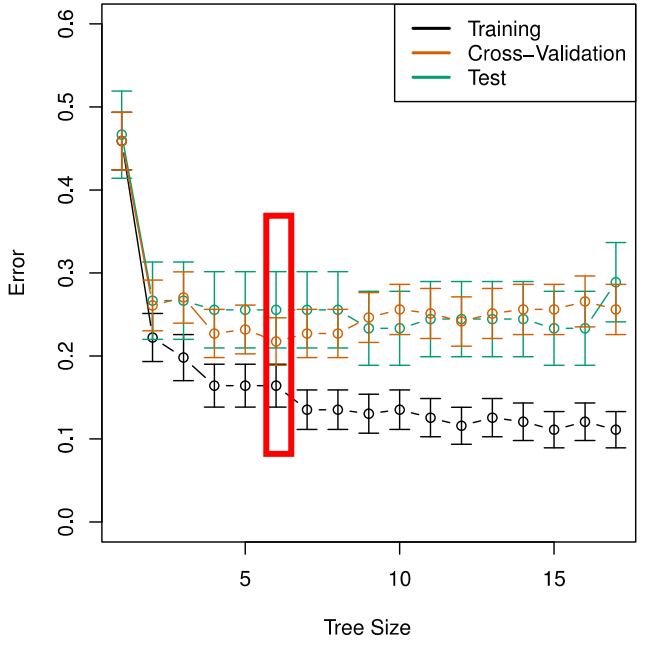
- Predict whether a patient with chest pain has heart disease based on Age, Sex, Chol (a cholesterol measure), and other heart and lung function measures
- Some terminal nodes have the same predicted value
  - Reason: Increased node purity
- Example
  - RestECG ≥ 1: $\frac{9}{9}$ with Yes
  - RestECG < 1: $\frac{7}{11}$ with Yes



Example: Predicting heart disease

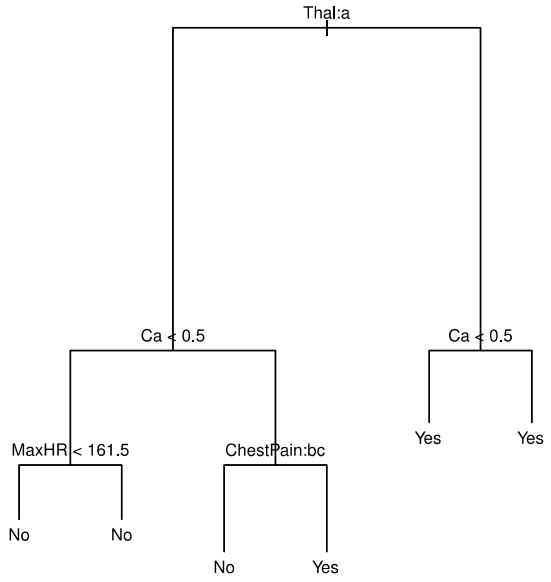
- Predict whether a patient with chest pain has heart disease based on Age, Sex, Chol (a cholesterol measure), and other heart and lung function measures
- Cross validation to prune tree

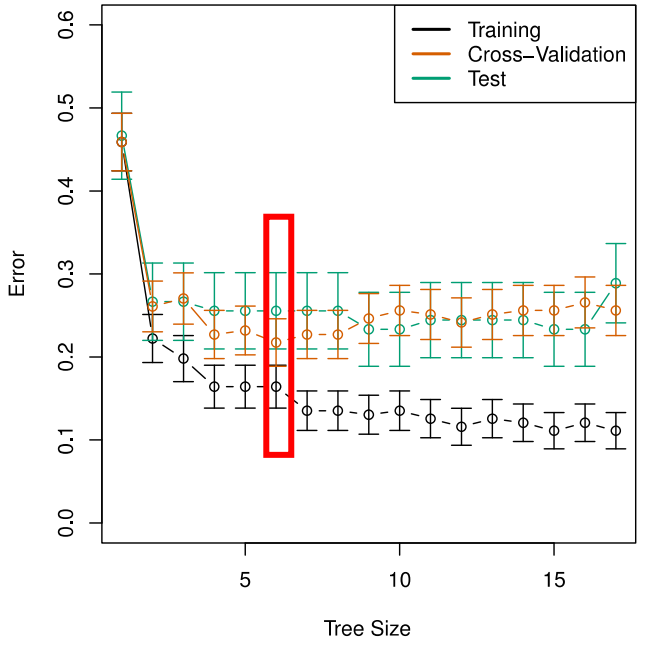




Example: Predicting heart disease

- 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
- Pruned tree after cross-validation:





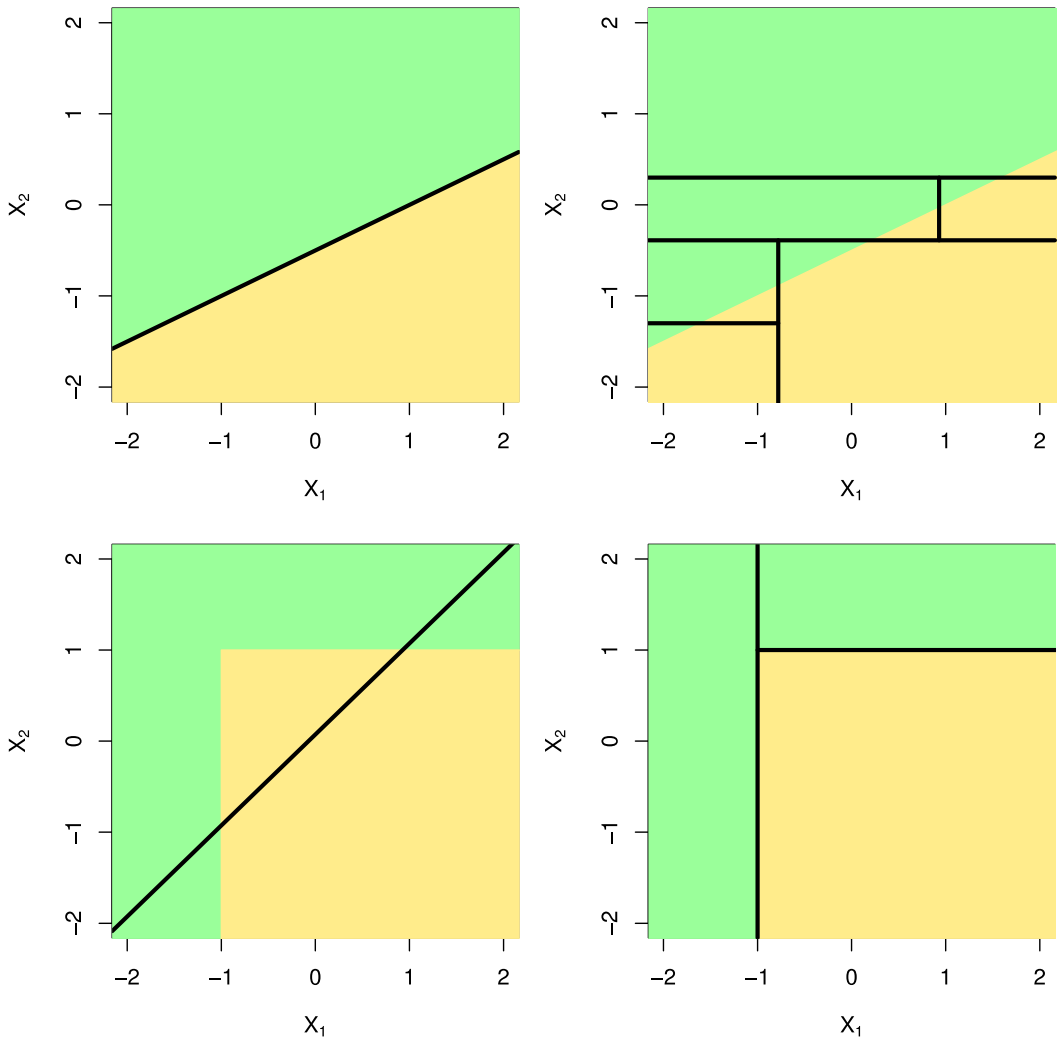
Tree vs. linear models

- Linear model

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$$

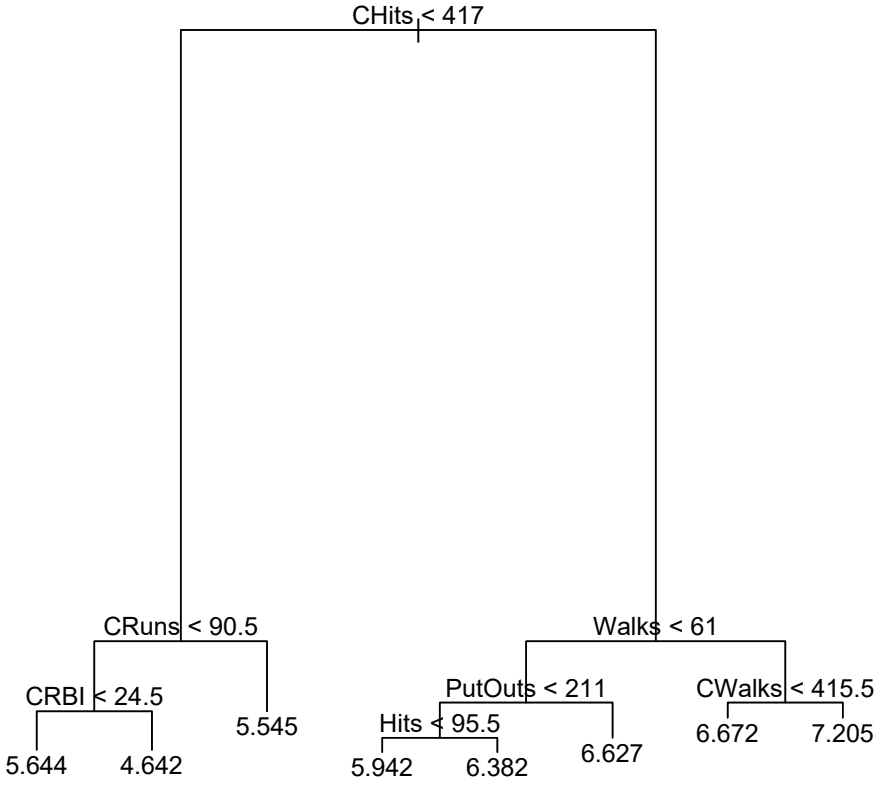
- Regression/Classification tree model

$$f(X) = \sum_{m=1}^{M} c_m 1(X \in R_m)$$

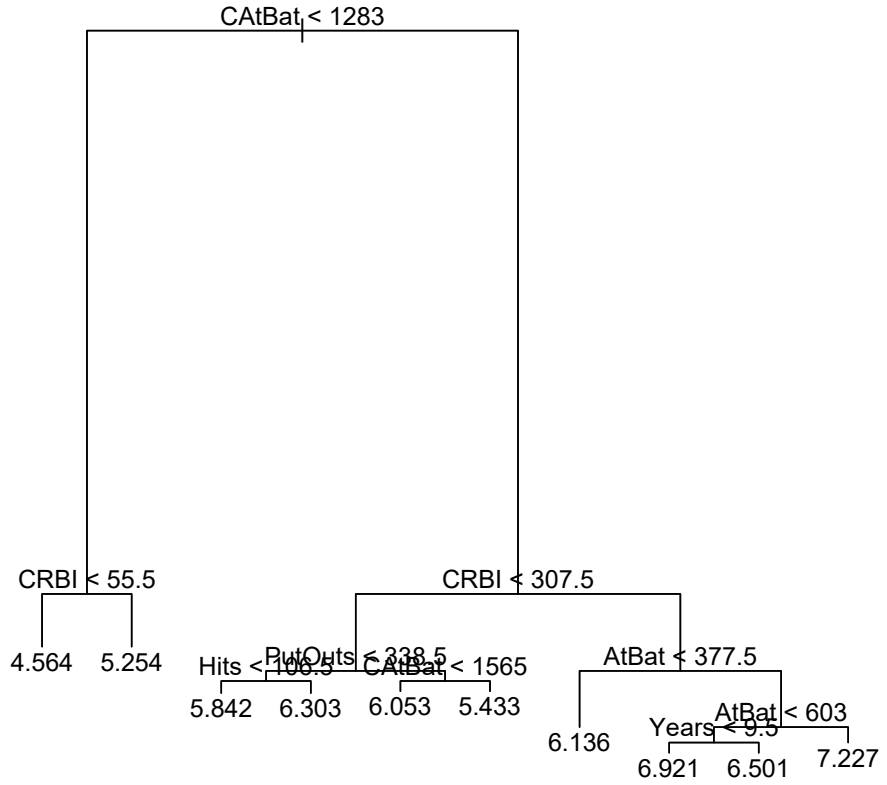


Decision tree has a high variance

- **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