

20) Decision Trees

Vitor Kamada

January 2020

Tables, Graphics, and Figures from:

James et al. (2017): Ch 8.1

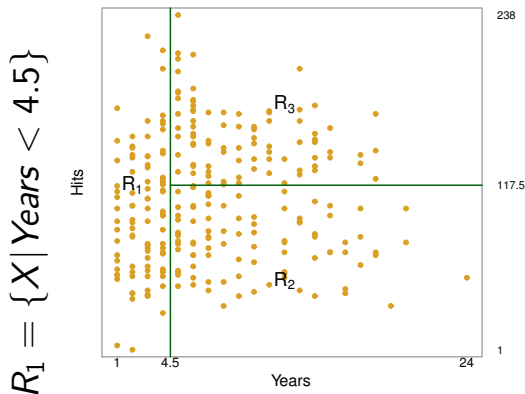
Hitters Data



$$e^{5.11} \cong \$165K, e^6 \cong \$402K, e^{6.74} \cong \$845K$$

Three-Region Partition

$$R_3 = \{X | \text{Years} \geq 4.5, \text{Hits} \geq 117.5\}$$



$$R_2 = \{X | \text{Years} \geq 4.5, \text{Hits} < 117.5\}$$

OLS vs Trees (Impurity Measure)

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

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

$$\hat{c}_m = \frac{1}{N_m} \sum_{x_i \in R_m} y_i$$

$$Q_m(T) = \frac{1}{N_m} \sum_{x_i \in R_m} (y_i - \hat{c}_m)^2$$

Top-down Greedy (Recursive Binary Splitting)

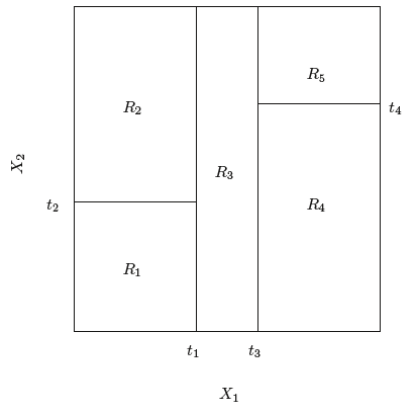
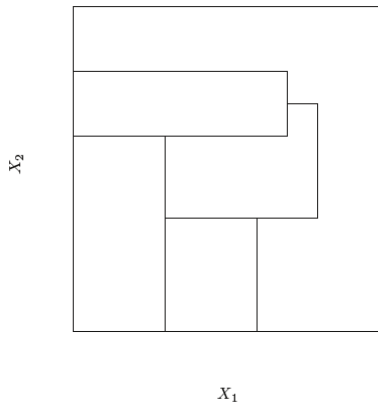
$$\sum_{j=1}^J \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

$$R_1(j, s) = \{X | X_j < s\}$$

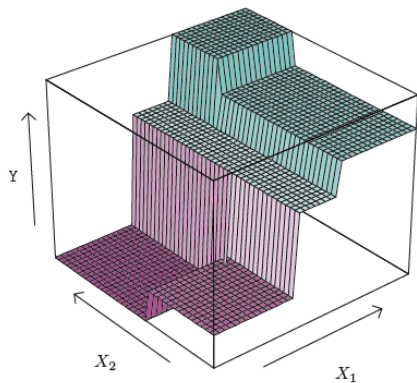
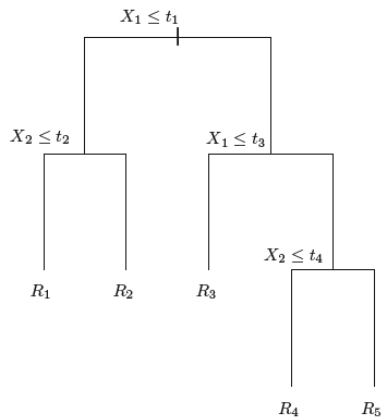
$$R_2(j, s) = \{X | X_j \geq s\}$$

$$\sum_{i: X_i \in R_1(j, s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i: X_i \in R_2(j, s)} (y_i - \hat{y}_{R_2})^2$$

No Recursive Binary Splitting vs Recursive Binary Splitting



Tree and Perspective Plot

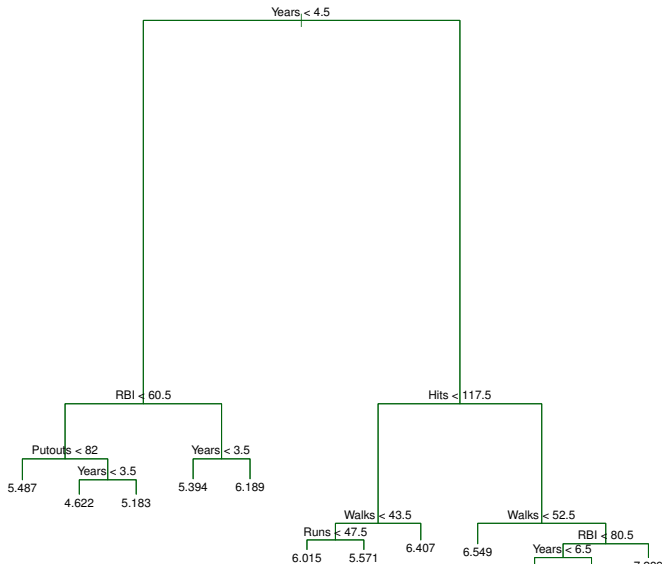


Cost Complexity Pruning (Weakest Link Pruning)

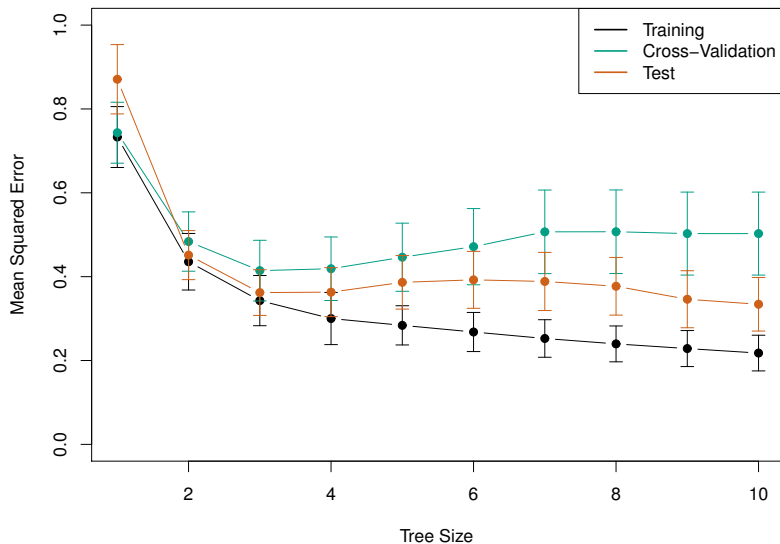
$$\sum_{m=1}^{|T|} \sum_{x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T|$$

$|T| = \#$ of terminal nodes of the tree

Unpruned Tree (Top-down Greedy Splitting)



Six-Fold Cross-Validation for Pruning Tree



Training Error Rate, Gini Index, and Entropy

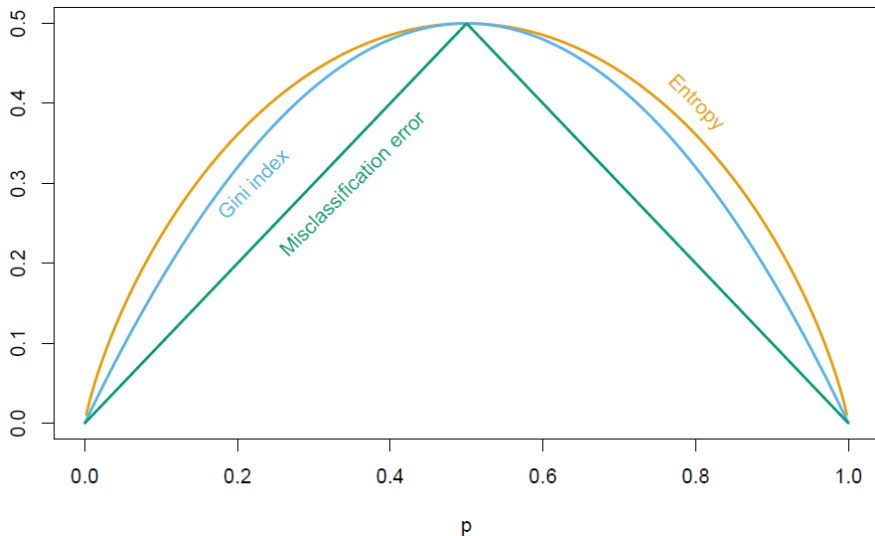
$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k)$$

$$E = 1 - \max_k (\hat{p}_{mk})$$

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

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

Gini Index and Entropy are more sensitive to changes in the node probabilities



Heart Data Set

303 patients

AHD: Yes for heart disease based on an angiographic test

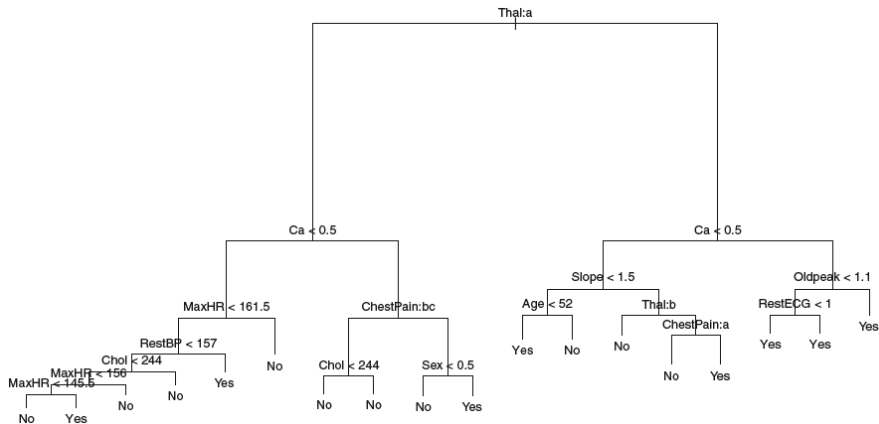
Thal: Thallium stress test, nuclear imaging shows how blood flows into heart

ChestPain: angina, atypical angina, non-anginal pain, and asymptomatic

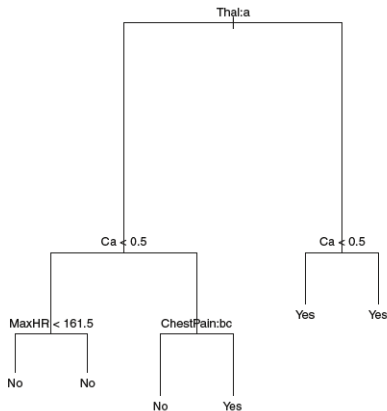
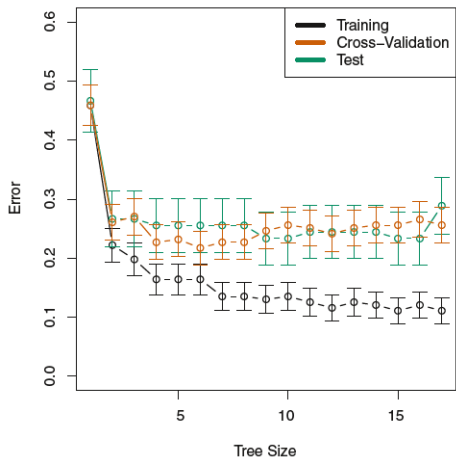
RestECG: Electrocardiograms

Heart Data: Unpruned Tree

Normal < |Thal:a| < Fixed or Reversible Defects



Pruned Tree (Minimal Cross-Validation Error)



Linear vs Non-linear True Decision Boundary

