Tree-Based Methods

HI 743

Ryan Gallagher

Department of Health Informatics and Administration
Zilber College of Public Health
University of Wisconsin - Milwaukee

March 13th, 2025

Overview

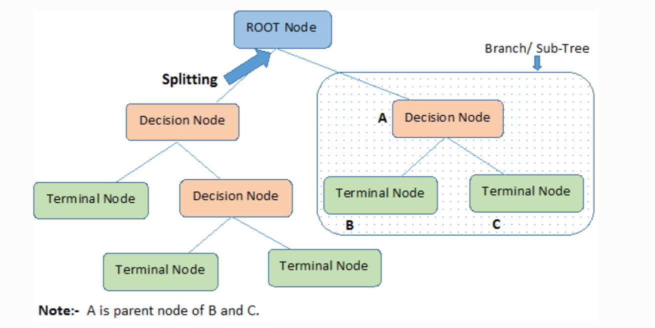
1. Regression Trees
2. Tree Pruning
3. Cross-Validation
4. Classification Trees
5. Trees vs. Linear Models

Introduction to Decision Trees

- Tree-based methods are non-parametric approaches used for both classification and regression tasks.
- They partition the feature space into distinct regions and make predictions based on the majority class (classification) or average response (regression).
- Decision trees provide an intuitive and interpretable way to model relationships between variables.

Tree Structure and Terminology

- Each split in the tree creates two branches, dividing the predictor space into regions.
- The final partitions are known as **terminal nodes** or **leaves**.
- Internal nodes define the decision rules based on feature values.
- The process of growing a tree continues until stopping criteria (such as minimum node size) are met.



Regression Trees

- Regression trees are used when the response variable is continuous.
- The predictor space is recursively split into distinct and non-overlapping regions.
- Each split is chosen to minimize the residual sum of squares (RSS) within each region.
- The predicted value for each region is the mean response of the training observations in that region.

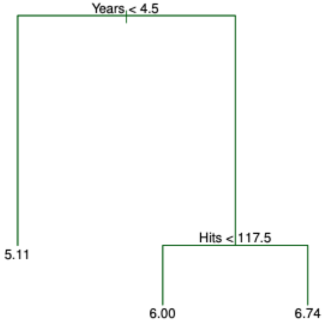
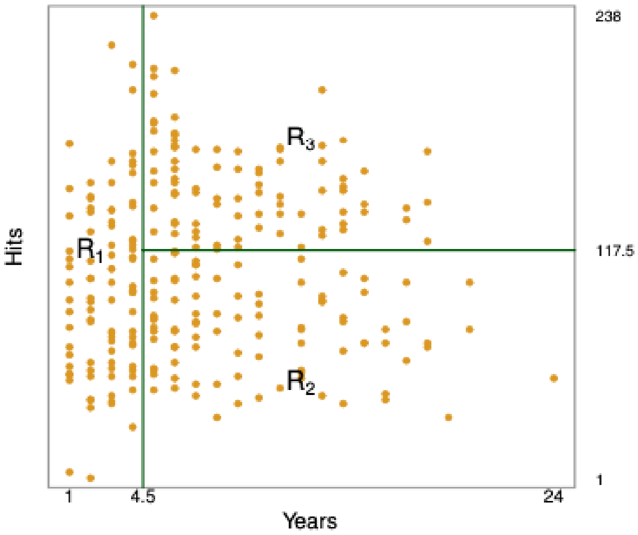


FIGURE 8.1. *For the* `Hitters` *data, a regression tree for predicting the log salary of a baseball player, based on the number of years that he has played in the major leagues and the number of hits that he made in the previous year. At a given internal node, the label (of the form $X_j < t_k$) indicates the left-hand branch emanating from that split, and the right-hand branch corresponds to $X_j \geq t_k$. For instance, the split at the top of the tree results in two large branches. The left-hand branch corresponds to* `Years<4.5`*, and the right-hand branch corresponds to* `Years>=4.5`*. The tree has two internal nodes and three terminal nodes, or leaves. The number in each leaf is the mean of the response for the observations that fall there.*



238
117.5
1
R1
R2
R3
Hits
1
4.5
24
Years

Tree Pruning

- **Decision trees tend to overfit the training data**, resulting in high variance and poor generalization to unseen data.
- **Pruning** is a technique used to simplify trees by removing branches that do not improve predictive performance.
- **Cost complexity pruning** (also known as *weakest link pruning*) selects a subtree that minimizes a balance between the RSS and tree complexity:

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

- The parameter α controls the trade-off between model complexity and fit (tuning parameter).

Types of Pruning

- **Pre-pruning** (early stopping): Stops tree growth when a criterion is met (e.g., minimum number of observations in a node).
  - Prevents overly complex trees but risks missing meaningful structure.
- **Post-pruning** (cost complexity pruning): Grows a large tree and prunes back using cross-validation to select the best subtree.
  - More computationally intensive but generally leads to better models.

Cross-Validation

- **Cross-validation** is a technique used to estimate model performance and avoid overfitting.
- The data is **split into multiple subsets** (folds), and the model is trained and tested across these folds.
- In pruning, cross-validation helps determine the optimal complexity parameter α by selecting the subtree that minimizes prediction error.
- Common choices include **k-fold cross-validation** (e.g., 10-fold CV) to balance bias and variance.

Cross-Validation for Pruning & Advantages

Cross Validation:

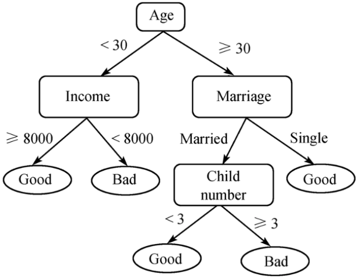
- The optimal subtree is selected using cross-validation.
- A sequence of pruned trees is generated using different values of α.
- The tree minimizing the cross-validation error is selected.

Advantages of Pruning:

- Reduces overfitting, leading to better generalization to unseen data.
- Produces simpler and more interpretable models.
- Reduces variance, improving stability of the predictions.

Classification Trees

- Used for predicting categorical responses rather than continuous values.
- Each observation is assigned to the most common class in the corresponding terminal node.
- Provides both class predictions and class probabilities for interpretability.
- Recursive binary splitting is used to partition the feature space.



Evaluation Measures for Classification Trees

- **Classification Error Rate:** Measures misclassification frequency but is not sensitive enough for tree growth.
- **Gini Index:** Measures total variance across classes:

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{2}$$

- **Entropy:** Measures uncertainty:

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk} \tag{3}$$

- Both Gini Index and Entropy provide better sensitivity to node purity compared to classification error.

Comparing Classification Trees with Other Methods

- Classification trees provide an intuitive and interpretable model.
- However, they tend to have higher variance and lower accuracy compared to ensemble methods.
- Alternative classification methods include:
  - Logistic Regression (for linear decision boundaries)
  - K-Nearest Neighbors (for non-linear decision boundaries)
  - Support Vector Machines (for high-dimensional spaces)

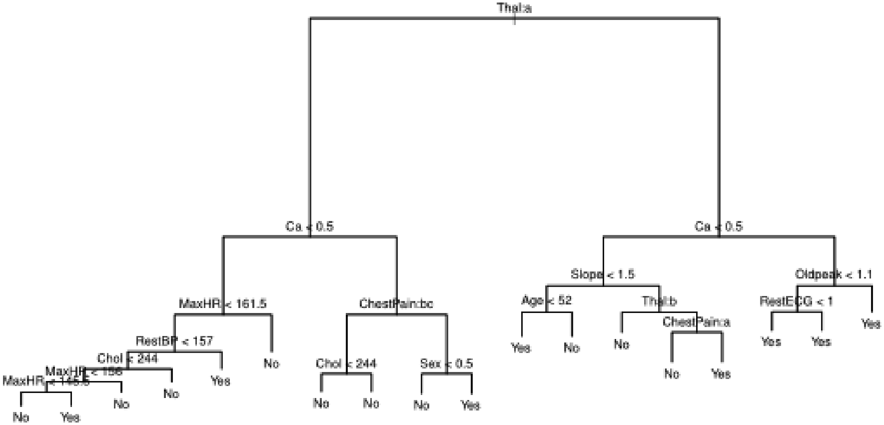


Figure: Hitters Classification Tree (unpruned)

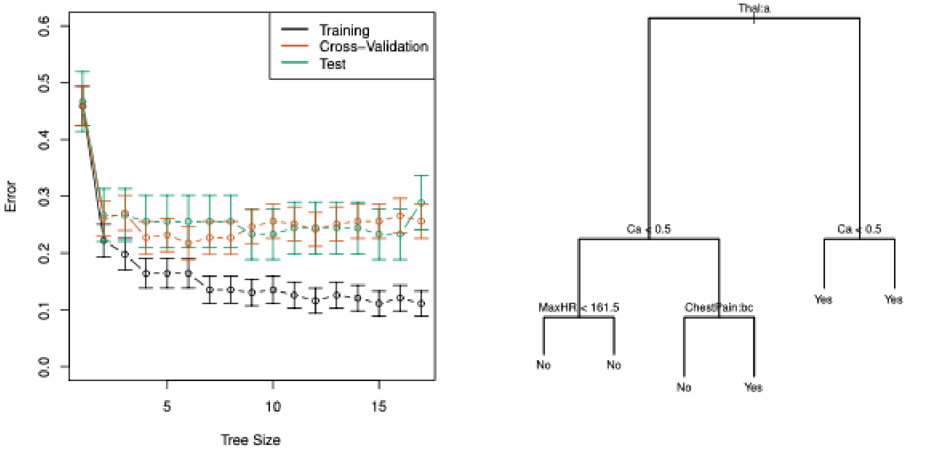


FIGURE 8.6. Heart *data.* Top: *The unpruned tree.* Bottom Left: *Cross-validation error, training, and test error, for different sizes of the pruned tree.* Bottom Right: *The pruned tree corresponding to the minimal cross-validation error.*

Trees vs. Linear Models

- Linear models assume a linear relationship:

$$f(X) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j \tag{4}$$

- Trees partition the predictor space into regions and fit a constant in each region:

$$f(X) = \sum_{m=1}^{M} c_m \cdot 1(X \in R_m) \tag{5}$$

- Trees work well for capturing complex, nonlinear relationships, while linear models excel when a linear structure is appropriate.

When to Use Trees vs. Linear Models

- Use **linear models** when:
  - The relationship between predictors and response is approximately linear.
  - Interpretability and inferential understanding are important.
  - The number of predictors is small and well-structured.

- Use **decision trees** when:
  - The relationship between predictors and response is highly nonlinear.
  - There are complex interactions between features.
  - Handling missing data and categorical variables directly is beneficial.