Lecture 18: Tree-Based Methods

Nayel Bettache

Department of Statistics and Data Science, Cornell University

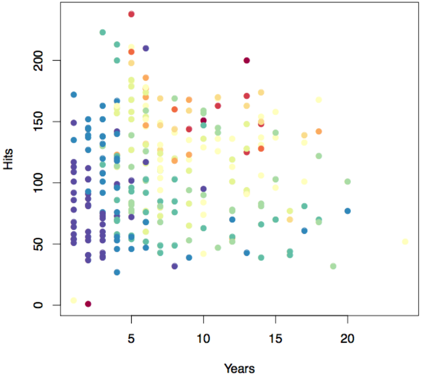
Tree-Based Methods

- Tree-based method can be applied for both regression and classification.
- Since the method can be summarized in a tree, these types of approaches are known as decision tree methods.
- Tree-based methods are simple and useful for interpretation.
- However, they typically are not competitive with the best supervised learning approaches.
- We will introduce bagging, random forests, and boosting to combine multiple trees to improve the performance.

The Basics of Decision Trees

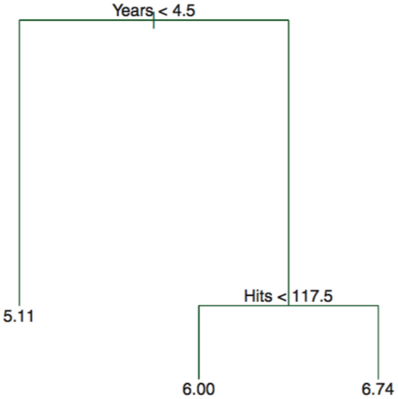
We first consider regression problems, and then move on to classification.

Consider the baseball salary data (Hitters data): We want to predict the salary of a baseball player, based on the number of years that he has played in the leagues and the number of hits that he made in the previous year.



Salary is color-coded from low (blue, green) to high (yellow,red)

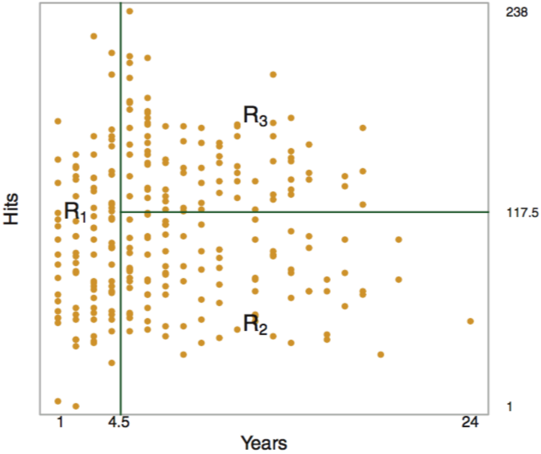
What does Decision Tree Look Like?



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 $X_j \geq t_k$. The number in each leaf (external node) is the mean of the response for the observations that fall there.

The three-region partition for the Hitters data set

Overall, the tree stratifies or segments the players into three regions of predictor space: $R_1 = \{X | Years < 4.5\}$, $R_2 = \{X | Years >= 4.5, Hits < 117.5\}$, and $R3 = \{X | Years >= 4.5, Hits >= 117.5\}$.



Terminology for Trees

- In keeping with the tree analogy, the regions R_1, R_2, and R_3 are known as **terminal nodes** or **leaves**.
- Decision trees are typically drawn upside down, in the sense that the leaves are at the bottom of the tree.
- The points along the tree where the predictor space is split are referred to as **internal nodes**.
- We refer to the segments of the trees that connect the nodes as **branches**.

How to Build a Regression Tree?

- Step 1: We divide the predictor space (the set of possible values for $X_1, X_2, ..., X_p$) into J distinct and non-overlapping regions, $R_1, R_2, ..., R_J$.
- Step 2: For every observation that falls into the region R_j, we make the same prediction, which is simply the mean of the response values for the training observations in R_j.

How to Construct Regions $R_1, R_2, ..., R_J$?

- In theory, the regions could have any shape. However, we choose to divide the predictor space into high-dimensional rectangles, or boxes, for simplicity and for ease of interpretation of the resulting predictive model.
- The goal is to find boxes $R_1, R_2, ..., R_J$ that minimize the RSS, given by

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

where $\hat{y}_{R_j}$ is the mean response for the training observations within the jth box.

How to Construct Regions $R_1, R_2, ..., R_J$?

- Unfortunately, it is computationally infeasible to consider every possible partition of the feature space into J boxes.
- For this reason, we take a **top-down, greedy** approach that is known as **recursive binary splitting**.
- The approach is top-down because it begins at the top of the tree and then successively splits the predictor space; each split is indicated via two new branches further down on the tree.
- It is greedy because at each step of the tree-building process, the best split is made at that particular step, rather than looking ahead and picking a split that will lead to a better tree in some future step.

How to Construct Regions $R_1, R_2, ..., R_J$?

- We first select the predictor X_j and the cutpoint s such that splitting the predictor space into the regions $\{X|X_j < s\}$ and $\{X|X_j >= s\}$ leads to the greatest possible reduction in RSS.
- That is, we consider all predictors $X_1, ..., X_p$, and all possible values of the cutpoint s for each of the predictors, and then choose the predictor and cutpoint such that the resulting tree has the lowest RSS.
- Mathematically, we seek the value of j and s that minimize the equation

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

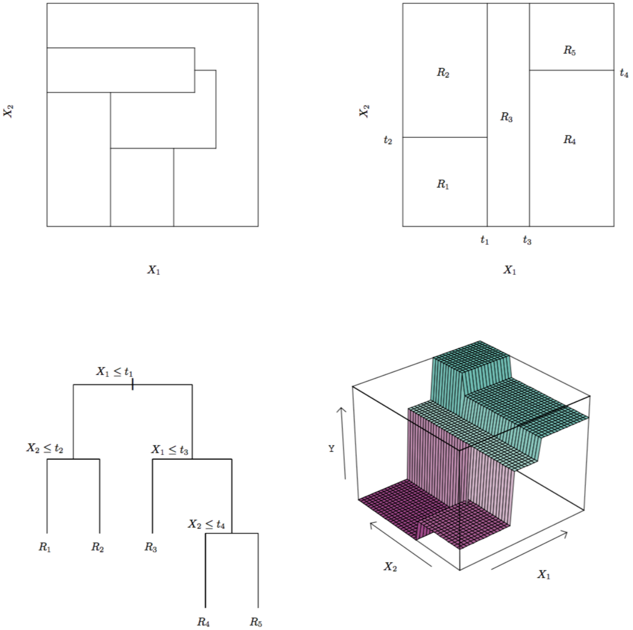
where $R_1(j, s) = \{X|X_j < s\}$ and $R_2(j, s) = \{X|X_j \geq s\}$, $\hat{y}_{R_1}$ and $\hat{y}_{R_2}$ are the mean response for the training data in $R_1(j, s)$ and $R_2(j, s)$.

How to Construct Regions $R_1, R_2, ..., R_J$?

- Next, we repeat the process, looking for the best predictor and best cutpoint in order to split the data further so as to minimize the RSS within each of the resulting regions.
- However, this time, instead of splitting the entire predictor space, we split one of the two previously identified regions. We now have three regions.
- Again, we look to split one of these three regions further, so as to minimize the RSS. The process continues until a stopping criterion is reached; for instance, we may continue until no region contains more than five observations.

Once $R_1, R_2, ..., R_J$ is given, we just predict the response for a given test observation using the mean of the training observations in the region.

An Example with 5 Regions



An Example with 5 Regions

- Top Left: A partition of two-dimensional feature space that could not result from recursive binary splitting.
- Top Right: The output of recursive binary splitting on a two-dimensional example.
- Bottom Left: A tree corresponding to the partition in the top right panel.
- Bottom Right: A perspective plot of the prediction surface corresponding to that tree.