Modern High Dimensional Nonlinear Regression

Alexander Quispe

October 1, 2021

Tree-based Methods

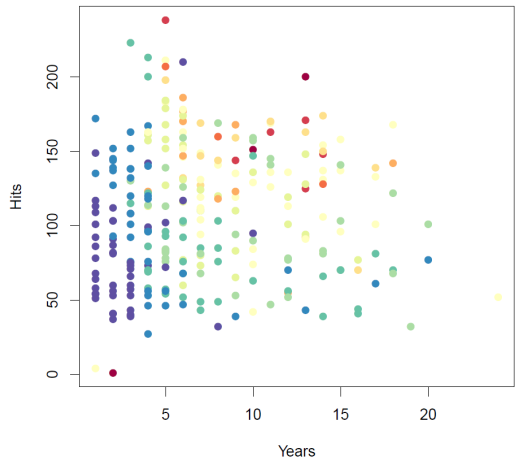
These notes are based on the *An Introduction to Statistical Learning* book and Chernozhukov's lecture notes for 14.38 course at MIT.

- We will work on tree-based methods for regression but not for classification.
- These involve **stratifying** or **segmenting** the predictor space into a number of simple regions.
- The set of splitting rules used to segment the predictor space can be summarized in a tree. This is called **decision-tree methods**.

Baseball salary data: how would you stratify it?

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

Figure: Salary-Hits-Years in Baseball



Decision tree for these data

Figure: Decision tree for baseball data

Details of previous figure

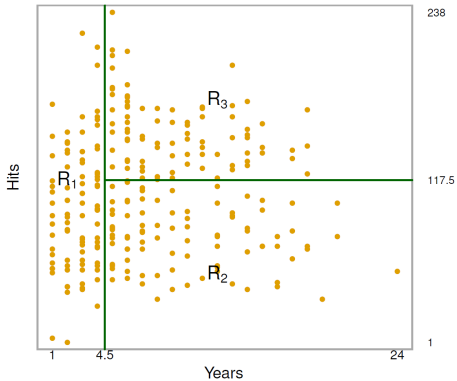
At a given node:

- $X_j < t_k$ indicates the left-hand branch emanating from that split
- $X_j >= t_k$ is the right-hand branch
- The left-hand branch corresponds to $Years < 4.5$
- 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.

Results

- $R_1 = \{X | Years < 4.5\}$
- $R_2 = \{X | Years >= 4.5, Hits < 117.5\}$
- $R_3 = \{X | Years >= 4.5, Hits > 117.5\}$

Figure: Segments



Terminology for Trees

- In keeping with the **tree** analogy, the regions R_1, R_2, and R_3 are known as **terminal nodes**
- 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**.
- In the hitters tree, the two internal nodes are indicated by the text $\mathit{Years} < 4.5$ and $\mathit{Hits} < 117.5$.

Interpretation of Results

- Years is the most important factor in determining **Salary**, and players with less experience earn lower salaries than more experienced players.
- Given that a player is less experienced, the number of Hits that he made in the previous year seems to play little role in his **Hits Salary**.
- But among players who have been in the major leagues for five or more years, the number of Hits made in the previous year does affect Salary, and players who made more **Hits Salary Hits** last year tend to have higher salaries.

Tree-building process

- We divide the predictor space — that is, the set of possible values for $X_1, X_2, \ldots, X_p$ into J distinct and non-overlapping regions $R_1, R_2, \ldots R_J$
- 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.
- The goal is to find boxes $R1, \ldots, RJ$ that minimize the RSS, given by

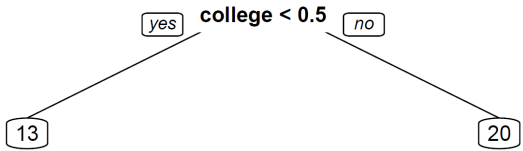
$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2 \tag{1}$$

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

Top-down, greedy approach

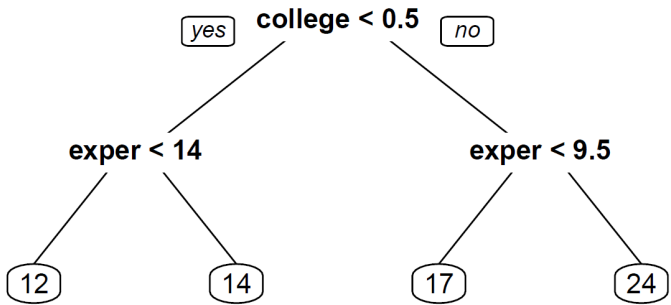
- **Top-down** because it begins at the top of the tree and then successively splits the predictor space
- **Greedy** because at each step of the tree-building process, the best split is made at that particular step, rather than looking ahead and picking a split that will lead to a better tree in some future step.
- Example using wage data:

Figure: Wage example - Tree 1.



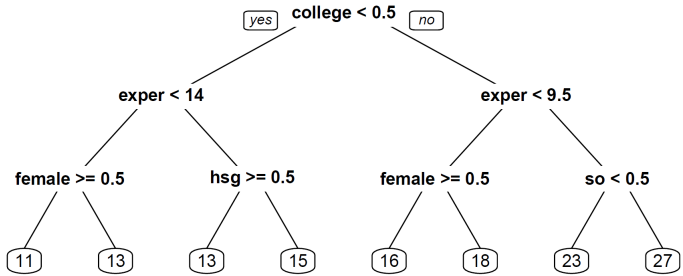
Top-down, greedy approach

Figure: Wage example - Tree 2



Top-down, greedy approach

Figure: Wage example - Tree 3



Pruning Regression Trees

- First, the deeper we grow the tree, the better is our approximation to the regression function $g(Z)$
- The deeper the tree, the noisies our estimate $\hat{g}(Z)$ becomes, since there are fewer observations per terminal node to estimate the predicted value for this node.
- From a prediction point of view, we can try to find the right depth or the structure of the tree by cross-validation. For example, in the wage example the tree of depth 2 performs better in terms of cross-validated *MSE* than the trees of depth 3 or 1. The process of cutting down the branches of the tree to improve predictive performance is called **"Pruning the Tree"**.

Cost complexity pruning

we consider a sequence of trees indexed by a nonnegative tuning parameter α. For each value of α there corresponds a subtree $T \subset T0$ such that

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

1. T indicates the number of terminal nodes of the tree T
2. R_m is the rectangle, corresponding to the mth terminal node
3. $\hat{y}_{R_m}$ is the mean of the training observations in R_m

Cost complexity pruning

Figure: Tree algorithm

1. Use recursive binary splitting to grow a large tree on the training data, stopping only when each terminal node has fewer than some minimum number of observations.
2. Apply cost complexity pruning to the large tree in order to obtain a sequence of best subtrees, as a function of α.
3. Use K-fold cross-validation to choose α. For each $k = 1, \ldots, K$:
   - 3.1 Repeat Steps 1 and 2 on the $\frac{K-1}{K}$th fraction of the training data, excluding the kth fold.
   - 3.2 Evaluate the mean squared prediction error on the data in the left-out kth fold, as a function of α.

   Average the results, and pick α to minimize the average error.
4. Return the subtree from Step 2 that corresponds to the chosen value of α.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013)