Week 7: Decision Trees for Regression

Kevin Lee

Department of Statistics
Western Michigan University

Decision Trees

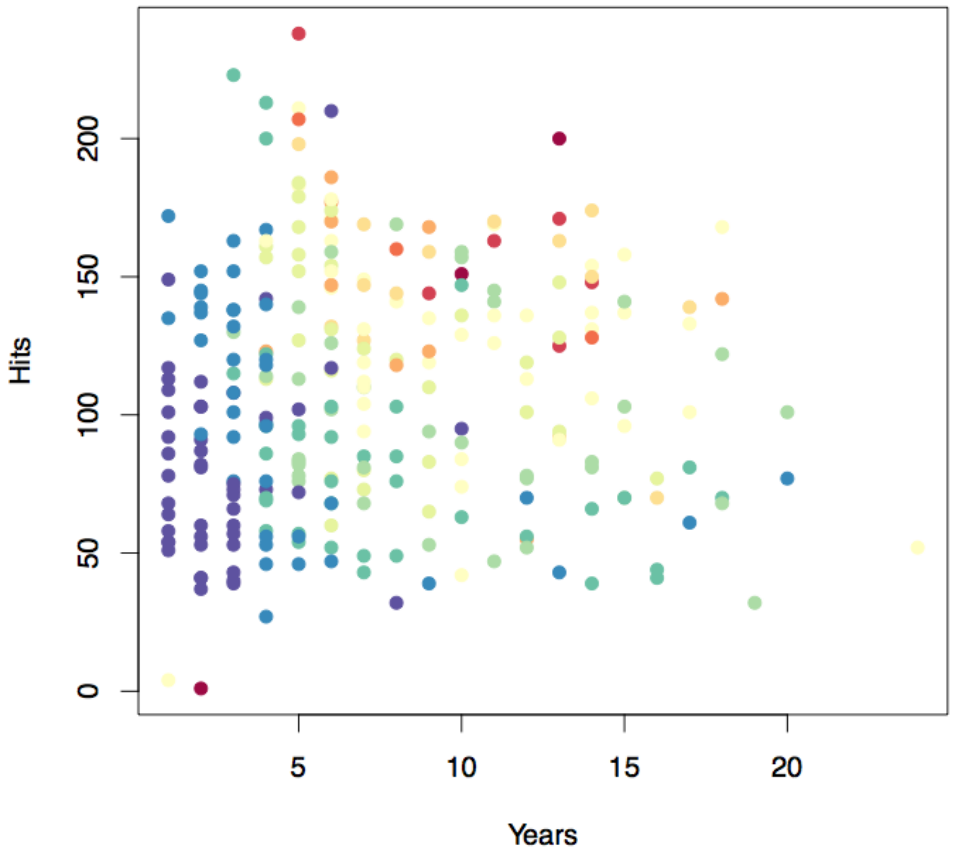
- Decision Trees are versatile Machine Learning algorithms that can perform both regression and classification tasks.
- Decision Trees methods are very powerful algorithms, capable of fitting complex data sets.

Partitioning the Predictor Space

- Divide the predictor space (i.e. all the possible values for X_1, X_2, ..., X_p) into distinct regions, R_1, R_2, ..., R_J).
- Then for every $\boldsymbol{X}$ that falls in a particular region R_j, we make the same prediction.
- Since the set of splitting rules used to segment the predictor space can be summarized in a tree, these type of approaches are known as Decision Trees.

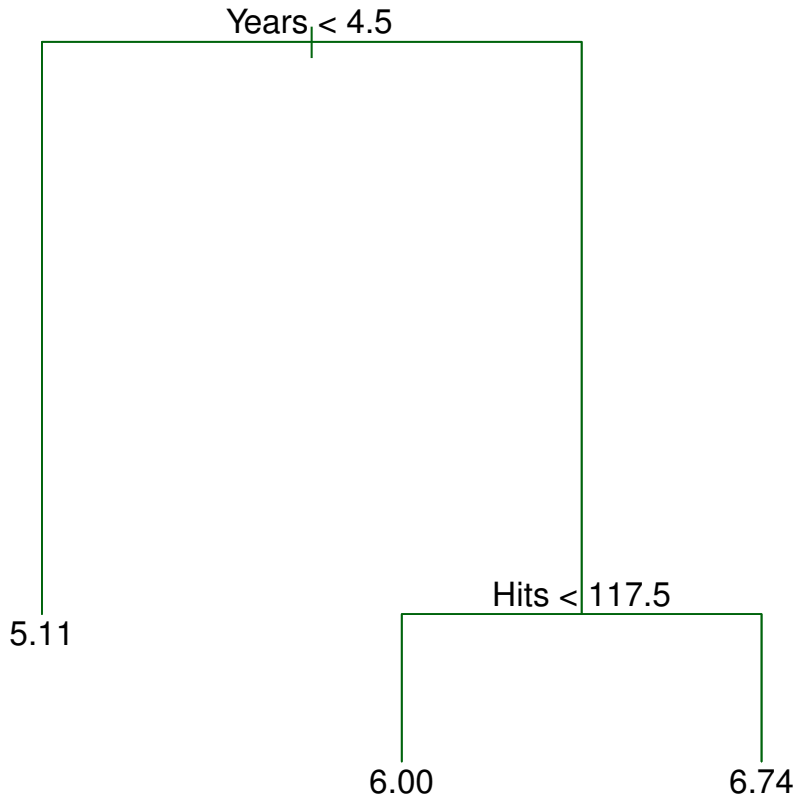
Baseball Salary Data

Player's salary is color-coded from low (blue, green) to high (yellow, red).



Decision Tree for Baseball Salary Data

- The decision tree has two internal nodes and three terminal nodes, or leaves.
- The predicted salary is the number in each terminal nodes. It is the mean of the response for the observations that fall in that region.



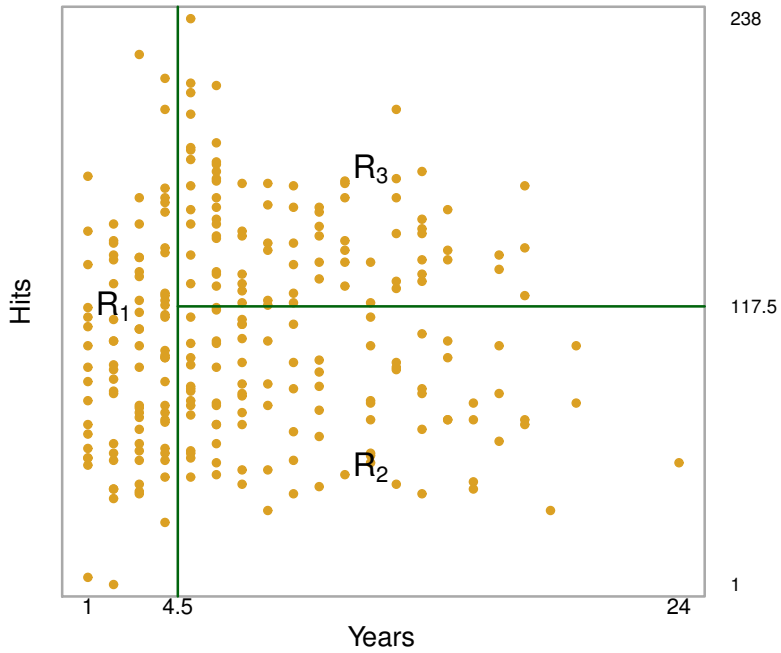
Another Way of Visualizing the Decision Tree

- The decision tree stratifies or segments the players into three regions of predictor space:

  $R_1 = \{X \mid Years < 4.5\}$

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

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



Interpretation of Results

- `Years` is the most important factor in determining Salary, and players with more experience earn higher salaries than less experienced players.

- 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` last year tend to have higher salaries.

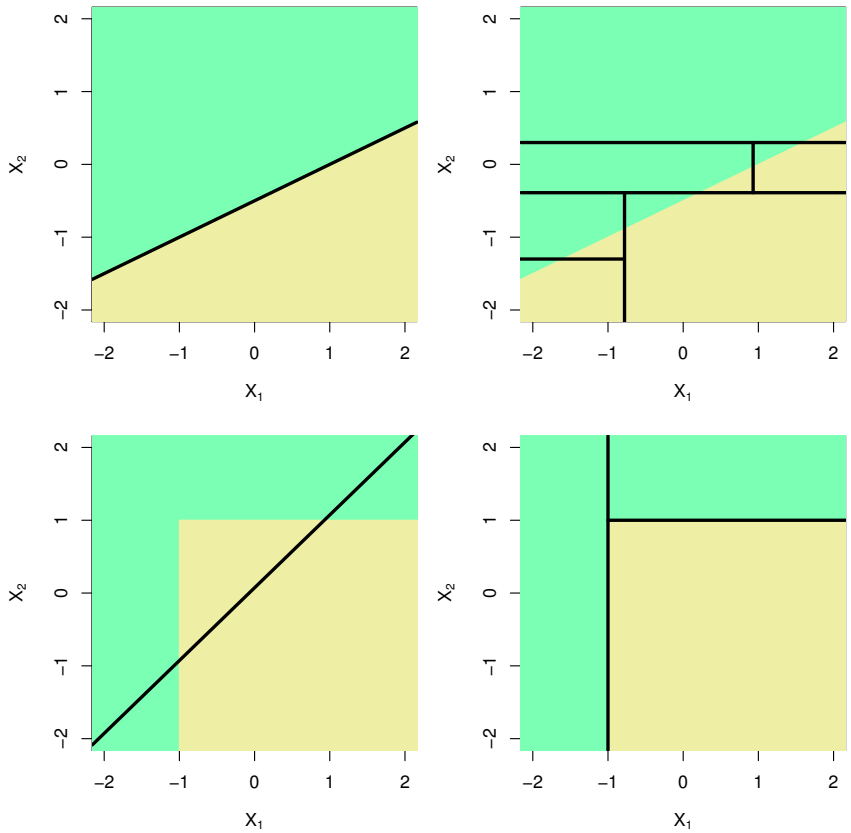
Decision Trees for Classification

- Very similar to a regression tree, except that it is used to predict a qualitative response rather than a quantitative one.

- For a classification tree, we predict that each observation belongs to the most commonly occurring class of training observations in the region to which it belongs.

Trees vs. Linear Models

- If the relationship between the predictors and response is linear, then linear models would outperform regression trees.
- On the other hand, if the relationship between the predictors is non-linear, then decision trees would outperform linear models.



Pros and Cons of Decision Trees

- Pros:
  - Trees are very easy to explain to people (probably even easier than linear regression).
  - Trees can be plotted graphically, and are easily interpreted even by non-expert.
  - They work fine on both classification and regression problems.
- Cons:
  - Trees don't have the same prediction accuracy as some of the more complicated approaches.
  - Trees can be very non-robust. In other words, a small change in the data cause a large change in the final estimated tree.