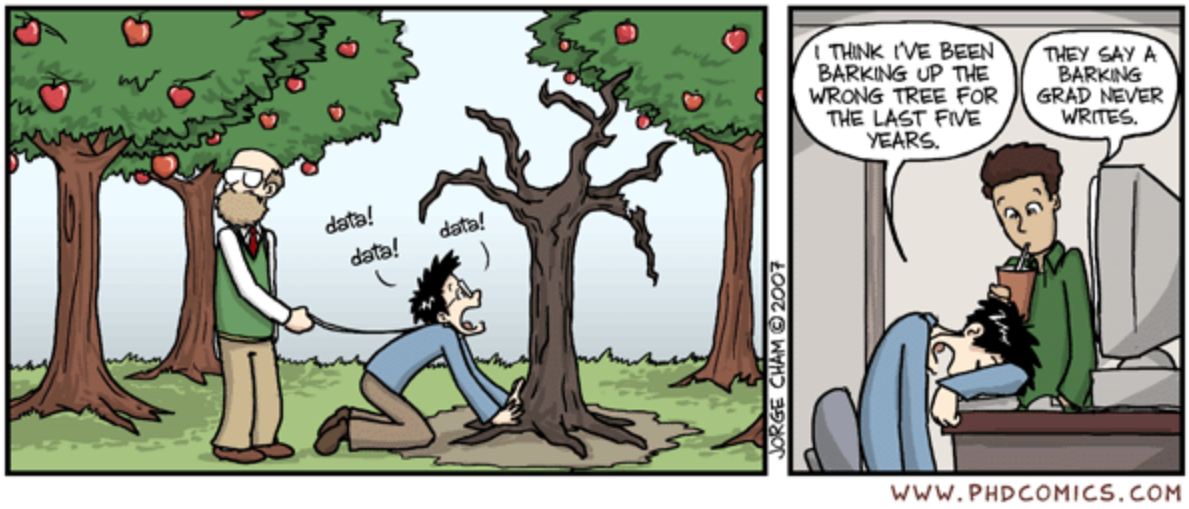
Chapter 8: Tree-Based Methods

We will introduce *tree-based* methods for regression and classification.

The set of splitting rules can be summarized in a tree ⇒ “decision trees”.

Combining a large number of trees can often result in dramatic improvements in prediction accuracy at the expense of interpretation.

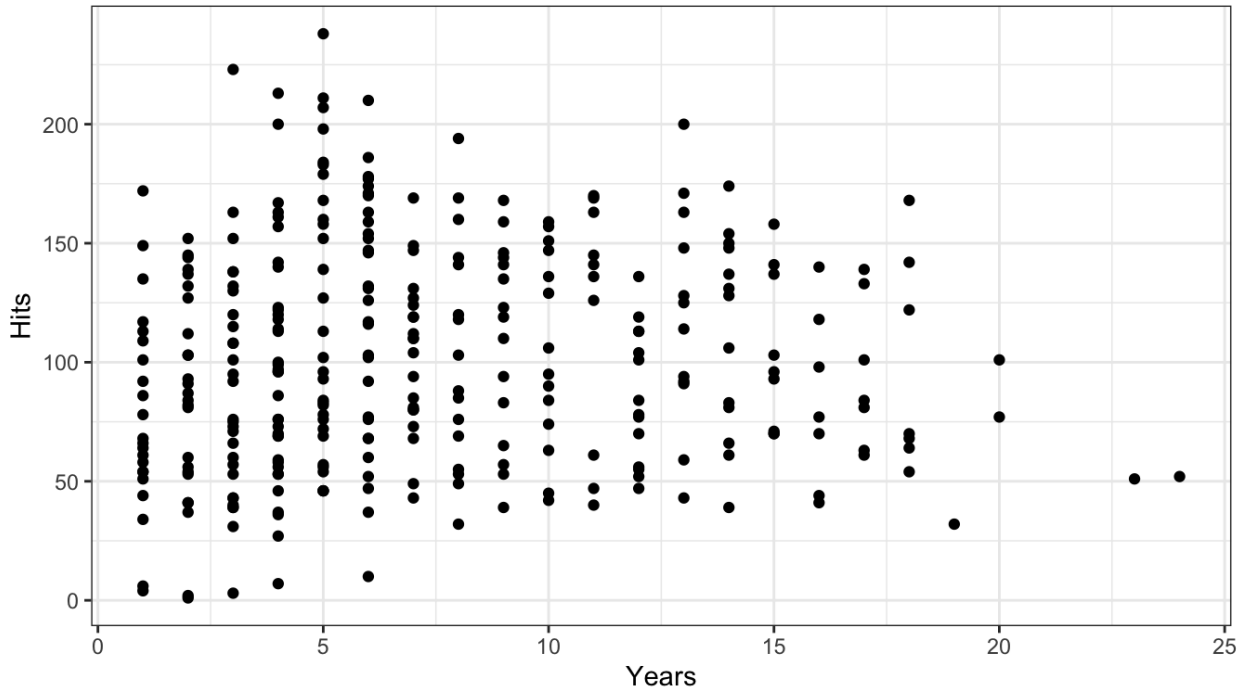


Credit: http://phdcomics.com/comics.php?f=852

Decision trees can be applied to both regression and classification problems. We will start with regression.

1 Regression Trees

Example: We want to predict baseball salaries using the `Hittters` data set based on `Years` (the number of years that a player has been in the major leagues) and `Hits` (the number of hits he made the previous year).



The predicted salary for players is given by the mean response value for the players in that box. Overall, the tree segments the players into 3 regions of predictor space.

We now discuss the process of building a regression tree. There are to steps:

1.

2.

How do we construct the regions $R_1, \ldots, R_J$?

The goal is to find boxes $R_1, \ldots, R_J$ that minimize the RSS.

The approach is *top-down* because

The approach is *greedy* because

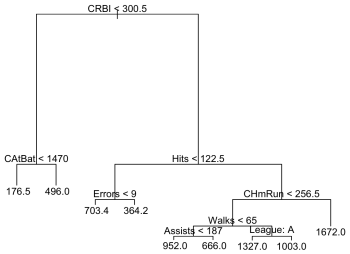
In order to perform recursive binary splitting,

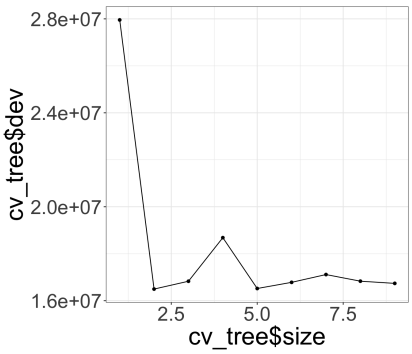
The process described above may produce good predictions on the training set, but is likely to overfit the data.

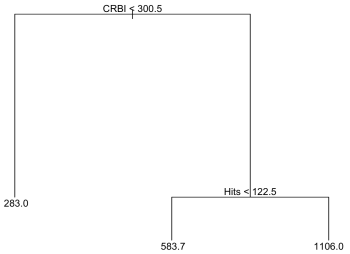
A smaller tree, with less splits might lead to lower variance and better interpretation at the cost of a little bias.

A strategy is to grow a very large tree T_0 and then *prune* it back to obtain a *subtree*.

Algorithm for building a regression tree:







2 Classification Trees

A *classification tree* is very similar to a regression tree, except that it is used to predict a categorical response.

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

The task of growing a classification tree is quite similar to the task of growing a regression tree.

It turns out that classification error is not sensitive enough.

When building a classification tree, either the Gini index or the entropy are typically used to evaluate the quality of a particular split.

3 Trees vs. Linear Models

Regression and classification trees have a very different feel from the more classical approaches for regression and classification.

Which method is better?

3.1 Advantages and Disadvantages of Trees