Classification and Regression Trees

Stephen Jun Villejo

School of Statistics

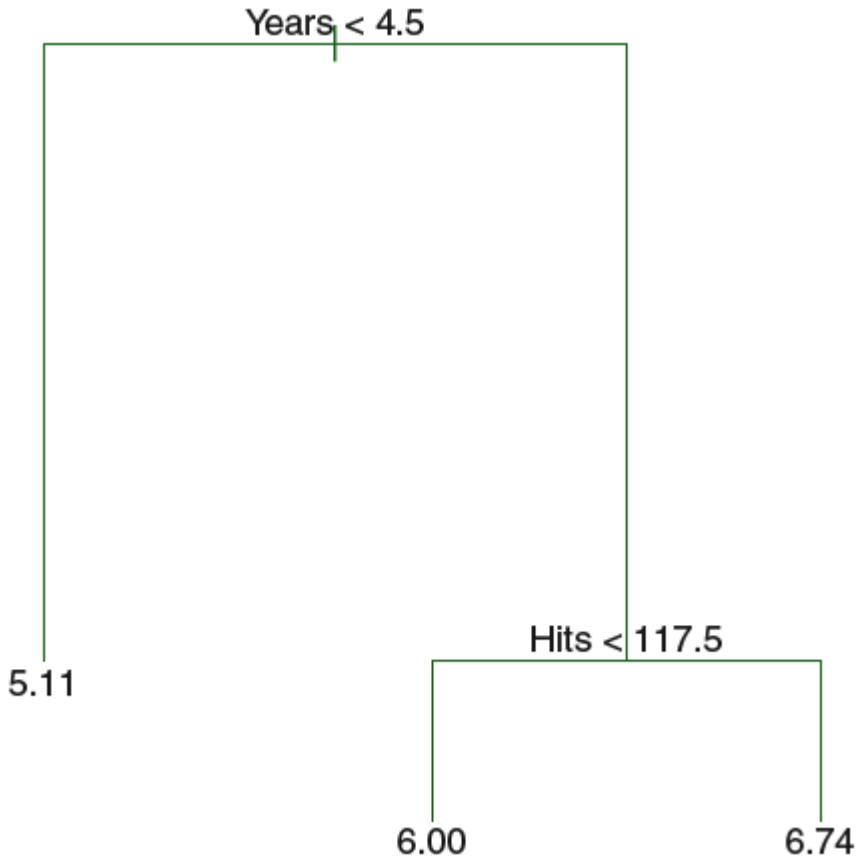
Tree-based Methods

- Tree-based methods may be used for regression and classification.
- If the response variable is continuous, the tree is called a **regression tree.** If the response is categorical, the tree is called a **classification tree**.
- These involve segmenting the possible values of the predictors (**predictor space**) into a number of simple regions.
- The mean or mode of the observations in a region is used as the predicted value of any observation falling in that region.

Regression Trees

Predicting Baseball Players' Salaries

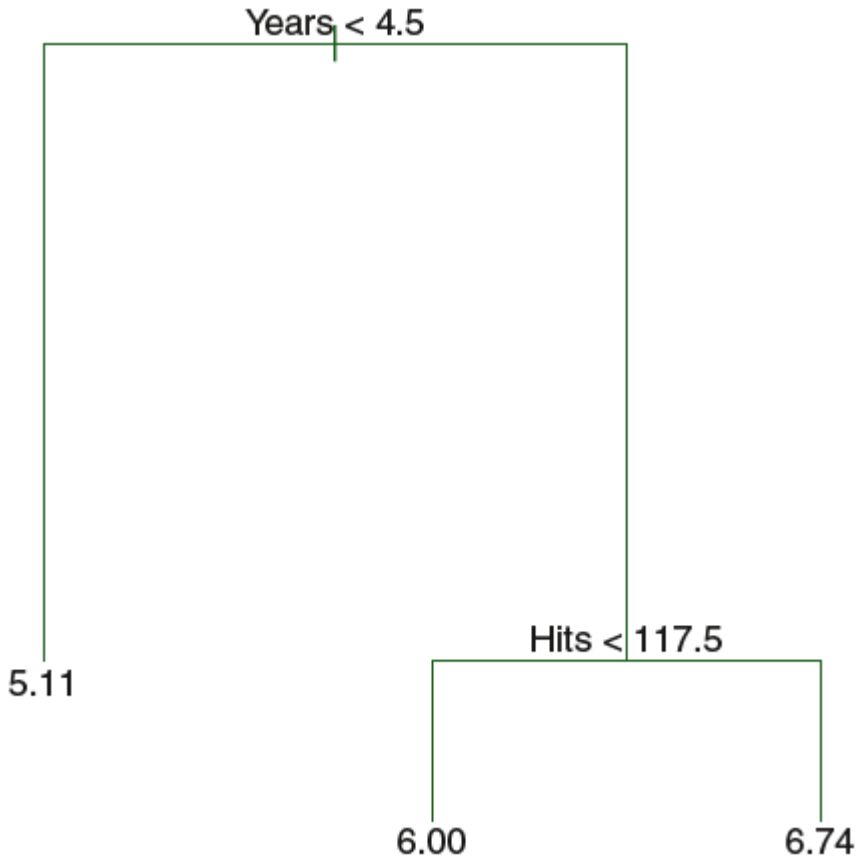
- Suppose we want to predict a baseball's `Salary` based on `Years` (the number of years that he has played in the major leagues) and `Hits` (the number of hits that he made in the previous year).
- Figure on the right shows a regression tree fit to the data.



Regression Trees

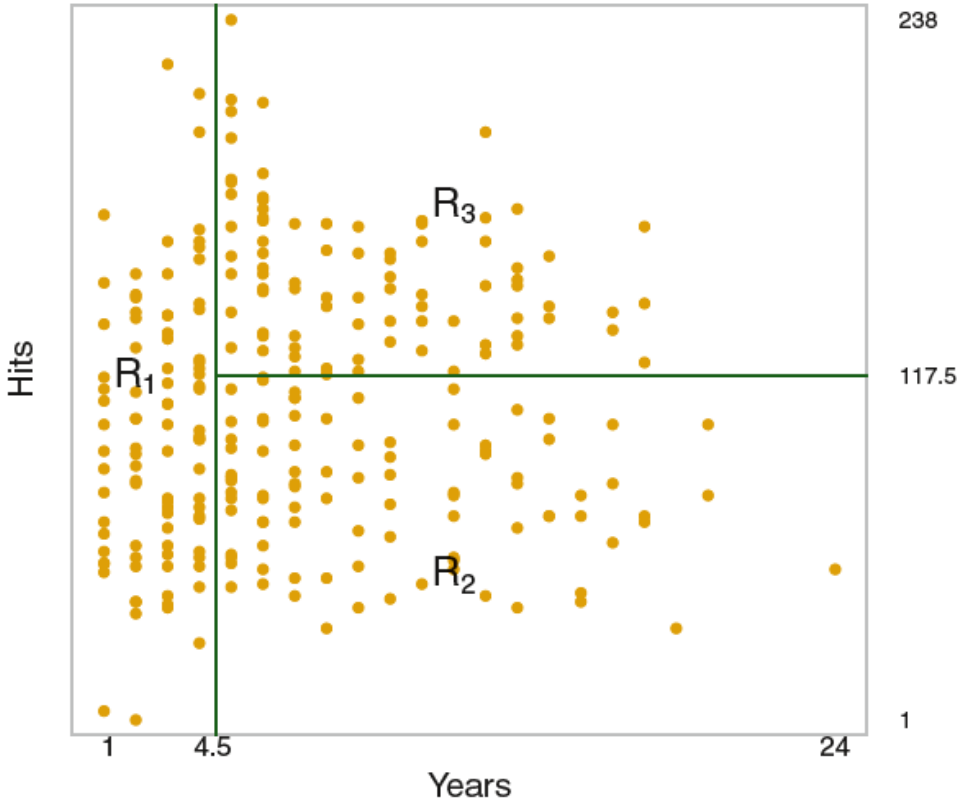
Predicting Baseball Players' Salaries

- The predicted salary for players with `Years` < 4.5 is $e^{5.11} = \$165{,}174$.
- Players with `Years` $\geq$ 4.5 are subdivided by the `Hits`.
- The tree partitions the predictor space into three regions.
- The three regions are called the **terminal nodes** or **leaves**.



Regression Trees

- Figure on the right shows how the partitioning of the predictor space is done.
- Based on the regression tree, `Years` is considered the most important factor.
- Also, among players who have been in the league for five or more years, the number of hits made in the previous year does affect salary.



Regression Trees

Steps in constructing a tree:

- Step 1: Divide the predictor space into distinct and non-overlapping regions.
- Step 2: For every observation that falls into a region, we make the same prediction, which is simply the mean of the response values for all observations in that region.

How do we construct the regions?

We use a *top-down, greedy* approach known as ***recursive binary splitting***.

Regression Trees

Recursive Binary Splitting

- It 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 the tree.
- It is *greedy* because at each step of the tree-building process, the best split is made.

What do we mean by "best"?

We consider all predictors and then choose the predictor and cutpoint for each of the predictors such that the resulting split will yield to a tree with small error – i.e., units in a region have response value close to each other.

Regression Trees

Tree Pruning

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

Why prune?

- The process described will produce good predictions on the *training set*, but may lead to poor performance for new set of observations, which we call a *test set*.
- Overfitting leads to small bias but high variance.
- Pruning will give smaller variance and better interpretation at the cost of a little bias.

Regression Trees

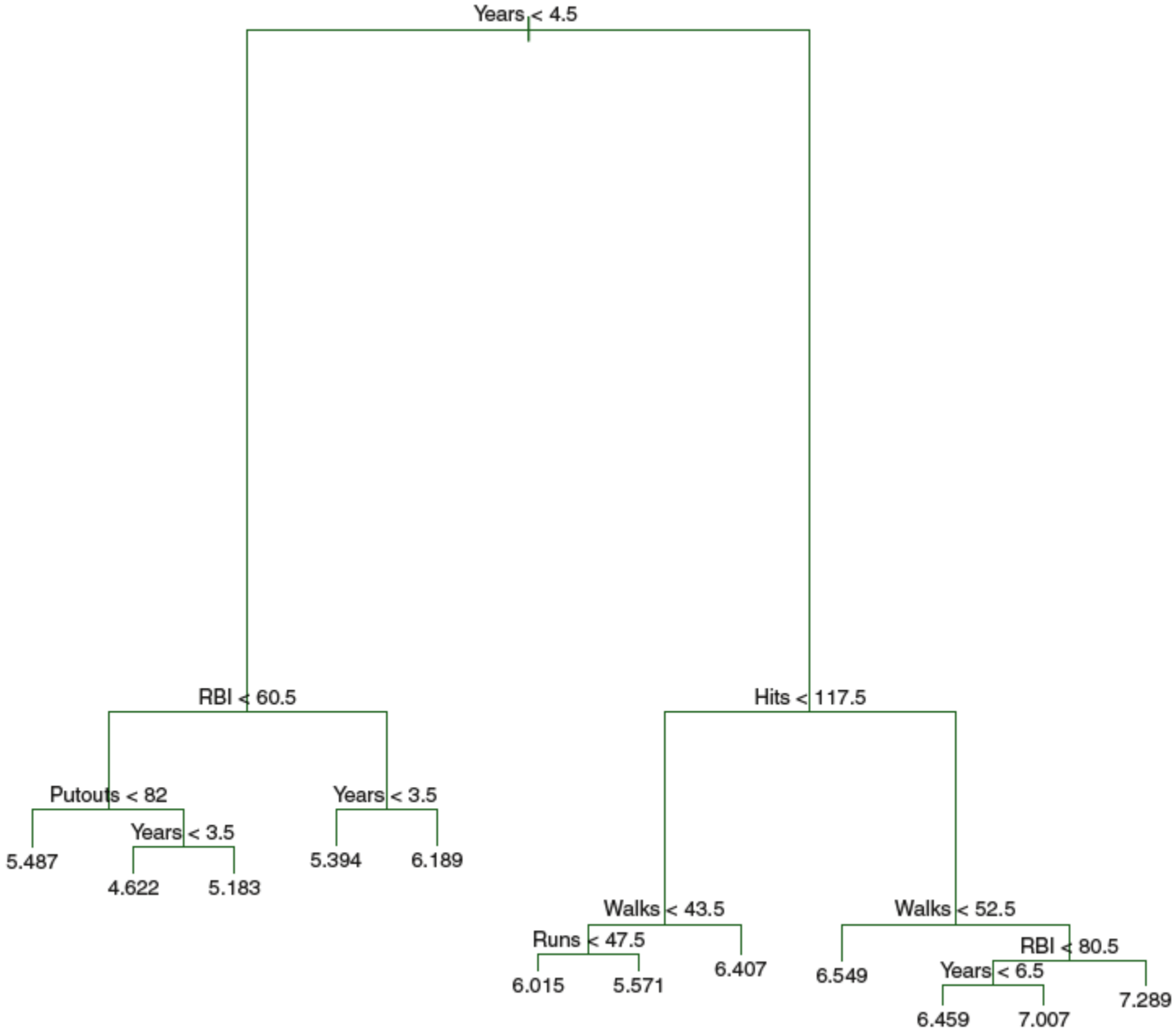
Tree Pruning

How do we choose the best pruned tree?

- Ideally, we obtain every possible sub-tree and choose the sub-tree with the smallest test error rate. This may be computationally cumbersome or infeasible.
- Instead, we give a penalty on trees with too many terminal nodes while still minimizing residual sum of squares (RSS).
- A larger tree will always give a smaller RSS, but this decline in RSS might be offset by the penalty, a price to pay for having a tree with too many terminal nodes.

Tree-based Methods

- Figure on the right shows the unpruned tree, also called the **maximal tree.**



R Exercise

Classification Trees

- A *classification tree* is very similar to a regression tree, except that it is used to predict a qualitative response rather than a quantitative one.
- Here, we predict that each observation belongs to the modal class of the observations 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.
- In regression trees, we minimize the RSS, i.e., the units in a region should be close each other. In classification trees, we have two preferable metrics that we minimize: **classification error rate**, **Gini index**, and **cross-entropy**.

Classification Trees

Classification error rate

- The classification error rate is the fraction of the training observations in that region that do not belong to the most common class.
- The classification error rate is given by:

$$E = 1 - \max_k(\hat{p}_{mk}).$$

Classification Trees

Gini index

- The Gini index is referred to as a measure of purity - a small value indicates that a node contains predominantly observations from a single class.

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

- $\hat{p}_{mk}$ represents the proportion of training observations in the mth region that are from the kth class.

Classification Trees

Cross-entropy index

- Like the Gini index, the cross-entropy will take on a small value if the mth node is pure.
- The Gini index and the cross-entropy are quite similar numerically.
- The Cross-entropy index is given by:

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

Tree-based methods

Advantages and Disadvantages of Trees

- Trees are very easy to explain to people. In fact, they are even easier to explain than linear regression.
- Some people believe that decision trees closely mirror human decision-making.
- Trees can be displayed graphically, and are easily interpreted even by a non-expert.
- Trees can easily handle qualitative predictors without the need to create dummy variables.
- Unfortunately, trees generally do not have the same level of predictive accuracy as some of the other regression and classification approaches.

R Exercise

Tree-based methods

- Trees generally do not have the same level of predictive accuracy as some of the other regression and classification approaches.
- There are many methods to substantially improve the predictive performance of trees. Some of these are : **bagging**, **random forests**, and **boosting**.

- (transfer this) Random Forests is an improvement of the bagging technique.

Bagging

Steps for Bagging:

Step 1: Take repeated random samples of observations from the *training set*.

Step 2: Construct a maximal tree on each resampled training set.

Step 3: Average all the predictions. The final prediction is the average of all the predictions.

For the classification problem, the overall prediction is the most commonly occurring class among all predictions. In other words, all the tress will take a **majority vote**.

Bagging

Note:

Bagging has been demonstrated to give impressive improvements in accuracy by combining together hundreds or thousands of trees into a single procedure.

Bagging

Out-of-Bag Error Estimation

- For each resampled (bootstrap) training set, there will be observations that will not be part of the sample. These observations are called ***out-of-bag* (OOB)** observations.
- Prediction of the ith response can be made using each of the trees which that observation was (OOB).
- The average of the predicted responses (or the majority vote if classification is the goal) leads to a single OOB prediction.
- The overall OOB error can then be computed for all observations.

Bagging

Variable Importance

- Bagging improves prediction at the expense of interpretability.
- One can obtain an overall summary of the importance of each predictor using the RSS (regression tree) or the Gini index (classification tree).
- We can get the total decrease of RSS or the Gini index for each predictor.

Random Forests

- Provides an improvement over bagged trees.
- As in bagging, we build a number of decision trees on bootstrapped training samples. But, each time a split in a tree is considered, a random sample of m predictors is choses as split candidates from the p full set of predictors.
- Typically, we choose the $\sqrt{p}$ as the value of m.

Random Forests

Why only a subset of predictors?

- Suppose there is one very strong predictor in the data set. Then in the collection of bagged trees, most or all of the trees will use this strong predictor in the top split. Hence, the trees will be similar, and the predictions will be highly correlated. In this setting, bagging will not lead to a substantial reduction in variance.
- Random forests *decorrelates* the trees, making the average of the trees less variable.

R Exercise