Classification & Regression Trees

Prof. Sharon McNicholas

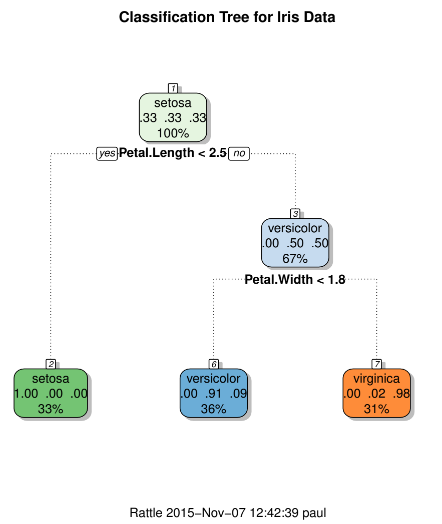
STATS 780/CSE 780

Introduction

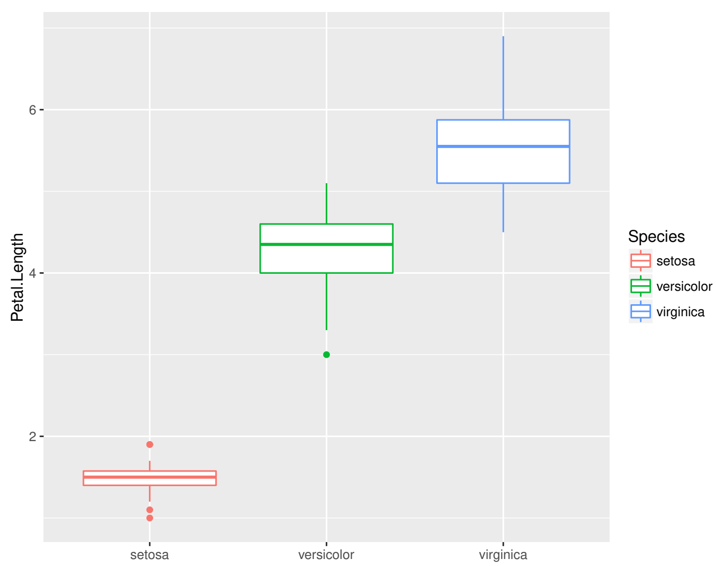
- Today, we will take a first look at classification and regression trees.
- The definitive source for material pertaining to classification trees is probably (still) the famous text by Breiman et al. (1984).[a]
- Conceptually, classification trees are quite simple.
- However, there are some subtleties.

[a] Breiman L., Friedman J.H., Olshen R.A. and Stone, C.J. (1984). *Classification and Regression Trees*. Chapman & Hall/CRC.

Iris Data: Classification Tree



Iris Data: Box Plot



A Classification Tree

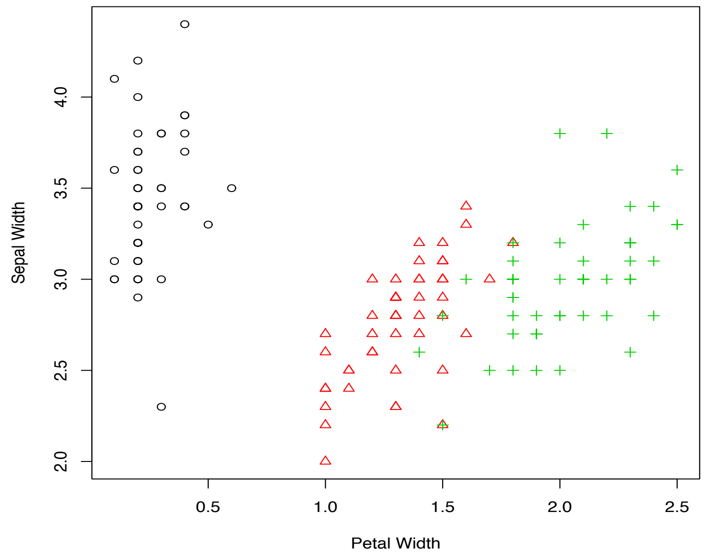
- Starts at the top (root) and recursively partitions data based on the best splitter.
- As we saw with the iris tree, a splitter is a variable (together with a rule).
- What it means to be the "best" splitter will be discussed, *inter alia*, shortly.
- I think that going in-depth on pruning is not as helpful but you can consult Hastie et al. (2009).[a]
- A classification tree is also called a decision tree.

[a]Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*. Second Edition. Springer: New York.

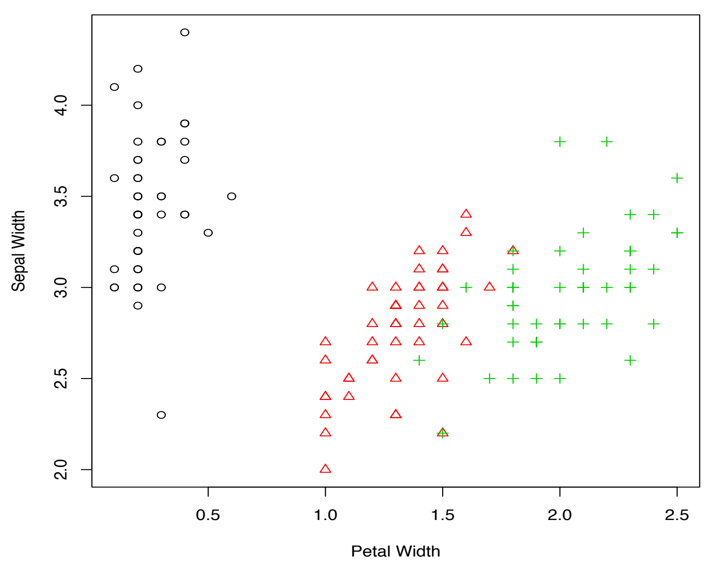
Splitting: Notation

- Consider the notation of Hastie et al. (2009) so that a node m represents a region R_m with N_m observations.
- When one considers the relationship between a tree and the space of observations, this is natural.
- Before proceeding, consider a tree built using two variables from the iris data — two variables so that we can visualize the partitioning of the space of observations.

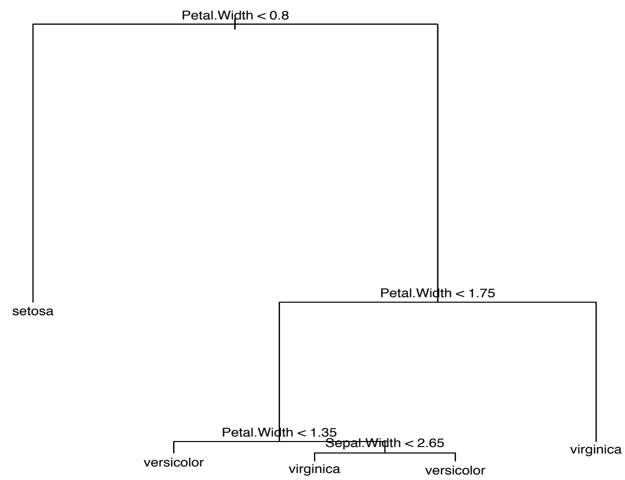
Iris: Sepal Width and Petal Width



Iris: Sepal Width and Petal Width



Iris: Sepal Width and Petal Width



Returning to Notation

- Recall that a node m represents a region R_m with N_m observations.
- Again, remaining with the notation of Hastie et al. (2009), the proportion of observations from class g in node m is

$$\hat{p}_{mg} = \frac{1}{N_m} \sum_{\mathbf{x}_i \in R_m} \mathbb{I}(y_i = g).$$

- All observations in node m are classified into the class with the largest proportion of observations; in other words, the majority class; or, in mathematical language, class $g^* = \arg\max_k \hat{p}_{mg}$.
- This is all just a formalization of what we discussed when we looked at the tree.

Growing the Tree

- We need to decide how to come up with splits (or splitters).
- We could think about the misclassification error, i.e.,

$$1 - \hat{p}_{mg^*}.$$

- Or the Gini index, i.e.,

$$\sum_{g=1}^{G} \hat{p}_{mg}(1 - \hat{p}_{mg}).$$

- Or the information (also called deviance or cross-entropy), i.e.,

$$-\sum_{g=1}^{G} \hat{p}_{mg} \log \hat{p}_{mg}.$$

Growing the Tree contd.

- Note that when used in this way, misclassification error, Gini index, and information are referred to as impurity measures.
- In this vernacular, we want nodes that are pure (or as pure as possible).
- The classification rate is generally not the impurity measure of choice.
- We can think a little about why.
- The Gini index has appealing interpretations.

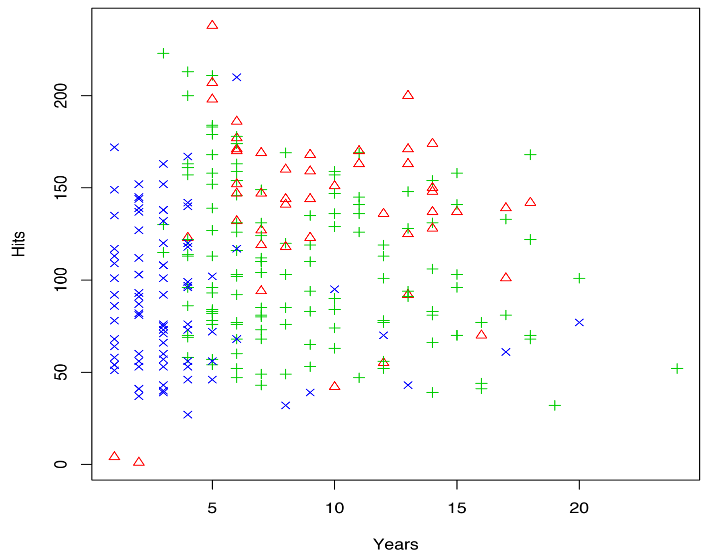
Regression Trees

- Now, we will look at regression trees.
- They proceed in an analogous fashion to classification trees.
- However, we now have a regression problem as opposed to a classification problem, i.e., we want to predict Y based on $X_1, \ldots, X_p$.
- Following James et al. (2013), let's start with an example on Major League Baseball (MLB) data.

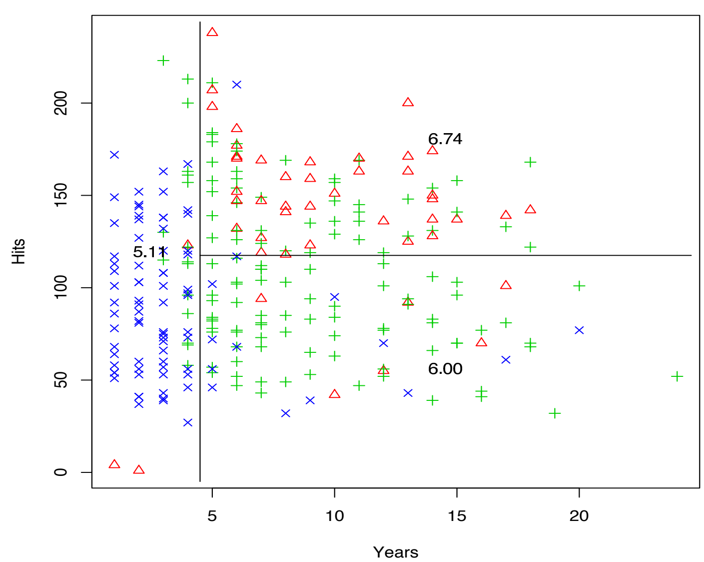
The `Hitters` Data

- MLB data from the 1986 and 1987 seasons.
- Suppose we are interested in predicting `Salary` based on `Hits` and `Years`.
- After looking at the data, it seems clear that `logSalary` is a better option.
- Here is a scatter plot, coloured by `logSalary`, of `Years` versus `Hits`.

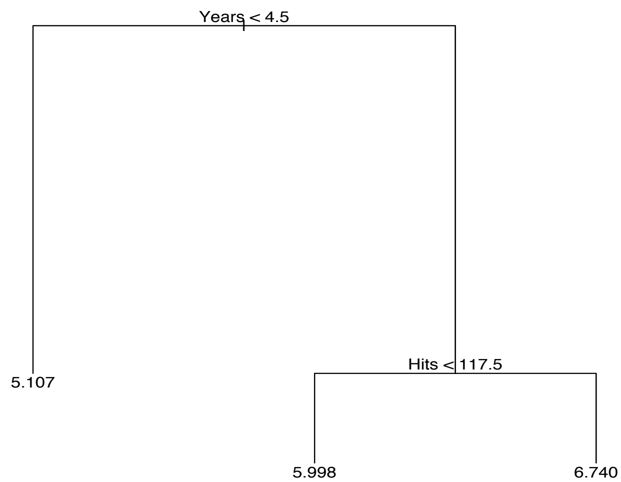
Hitters: Hits and Years



Hitters: Hits and Years



Hitters: Hits and Years



Interpretation

- Like before, we have a partitioning of the data into regions.
- But how do we interpret the number associated with each region?
- Having seen this example — and we will return to it later — we will now see how the tree is grown.
- As with classification trees, we need a splitter at each step.

Growing the Tree

- Borrowing the notation of James et al. (2013), we need to choose a value s and a variable X_j to split data into the regions

$$R_1(j,s) = \{\mathbf{X} \mid X_j < s\} \quad \text{and} \quad R_2(j,s) = \{\mathbf{X} \mid X_j \geq s\}.$$

- This choice (i.e., the choice of j and s) is made to minimize

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

where $\hat{y}_{R_k}$ is the mean of (training) observations in R_k, for $k = 1, 2$.

- Further splits continue in this fashion.

Comments

- Trees have lots of advantages.

- For one, they are very easy to understand.

- They mimic decision processes, e.g., of a physician perhaps.

- They are very easy (natural, even) to visualize.

- But they are not as good as some other methods.

Comments cont.

- Ensembles of trees, however, can lead to markedly improved performance.
- I want to do boosting, bagging, and random forests next.
- But I think some of you probably do not know what the bootstrap is.
- So we are going to spend some time learning about the bootstrap.
- Then, in the next class, we will move on to boosting, bagging, and random forests.
- Now, some more examples in R.