DECISION TREES

Chapter 08 (part 01)

Outline

- The Basics of Decision Trees
  - Regression Trees
  - Classification Trees
  - Pruning Trees
  - Trees vs. Linear Models
  - Advantages and Disadvantages of Trees

Partitioning Up the Predictor Space

- One way to make predictions in a regression problem is to divide the predictor space (i.e. all the possible values for for $X_1,X_2,\ldots,X_p$) into distinct regions, say R_1, $R_2,\ldots,R_k$

- Then for every X that falls in a particular region (say R_j) we make the same prediction,

REGRESSION TREES

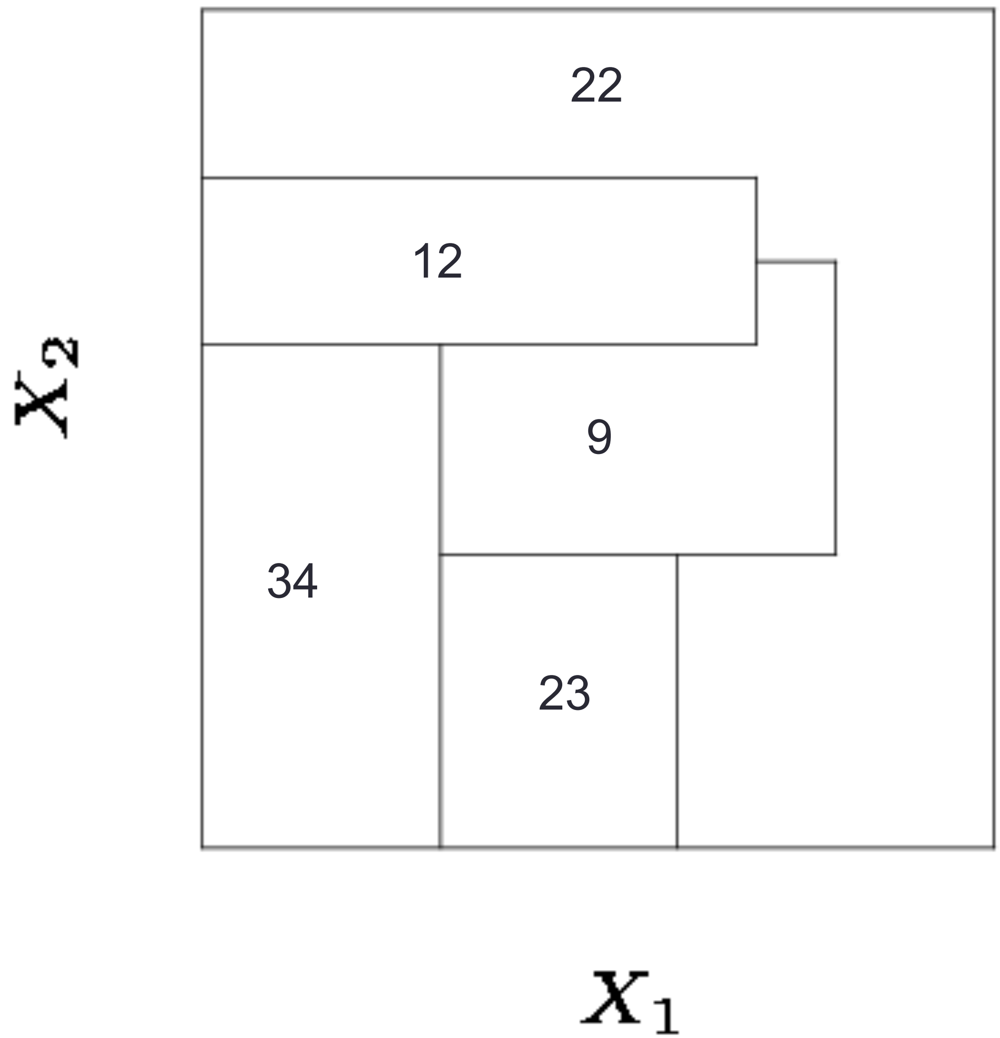
Regression Trees

- Suppose for example we have two regions R_1 and R_2 with $\hat{Y}_1 = 10, \hat{Y}_2 = 20$

- Then for any value of X such that $X \in R_1$ we would predict 10, otherwise if $X \in R_2$ we would predict 20.

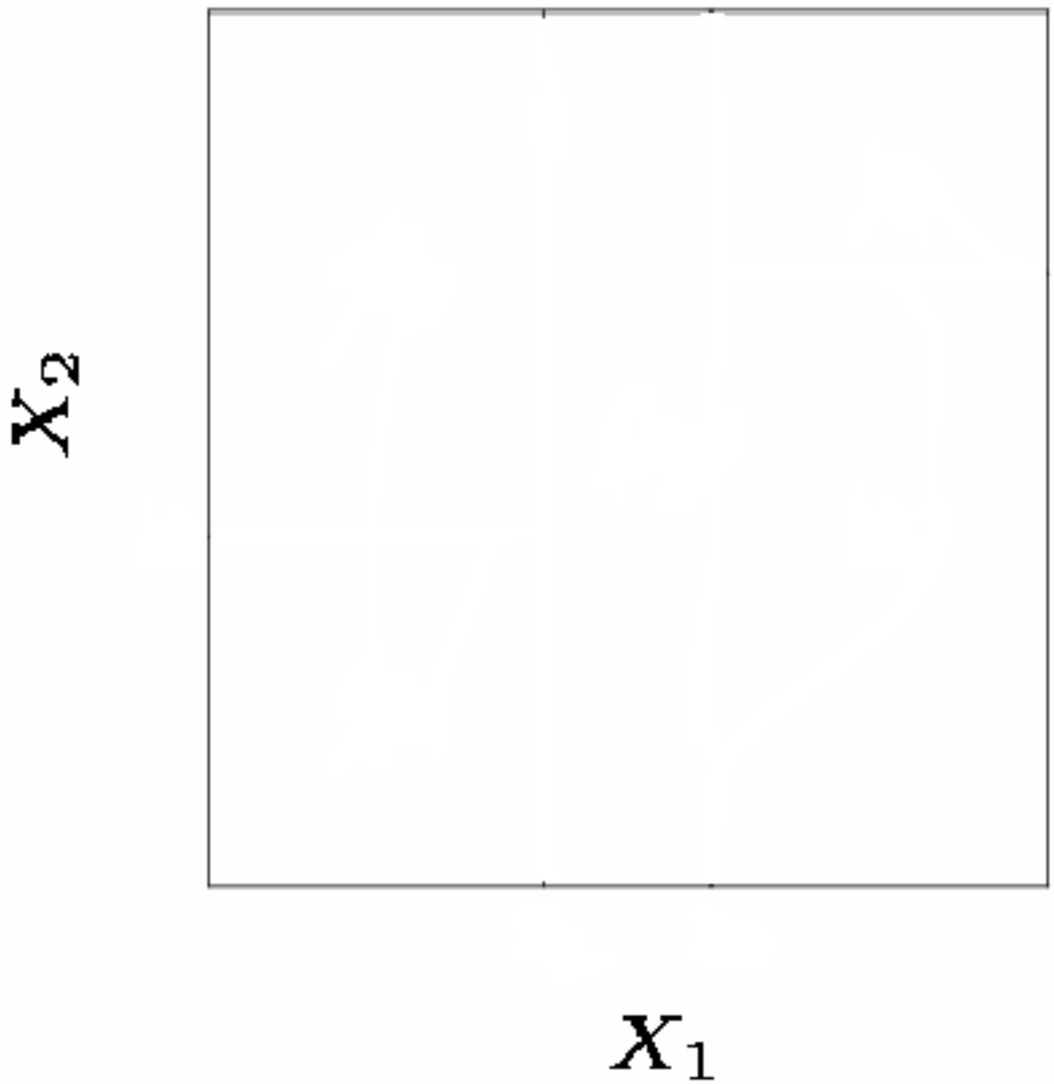
The General View

- Here we have two predictors and five distinct regions
- Depending on which region our new X comes from we would make one of five possible predictions for Y.



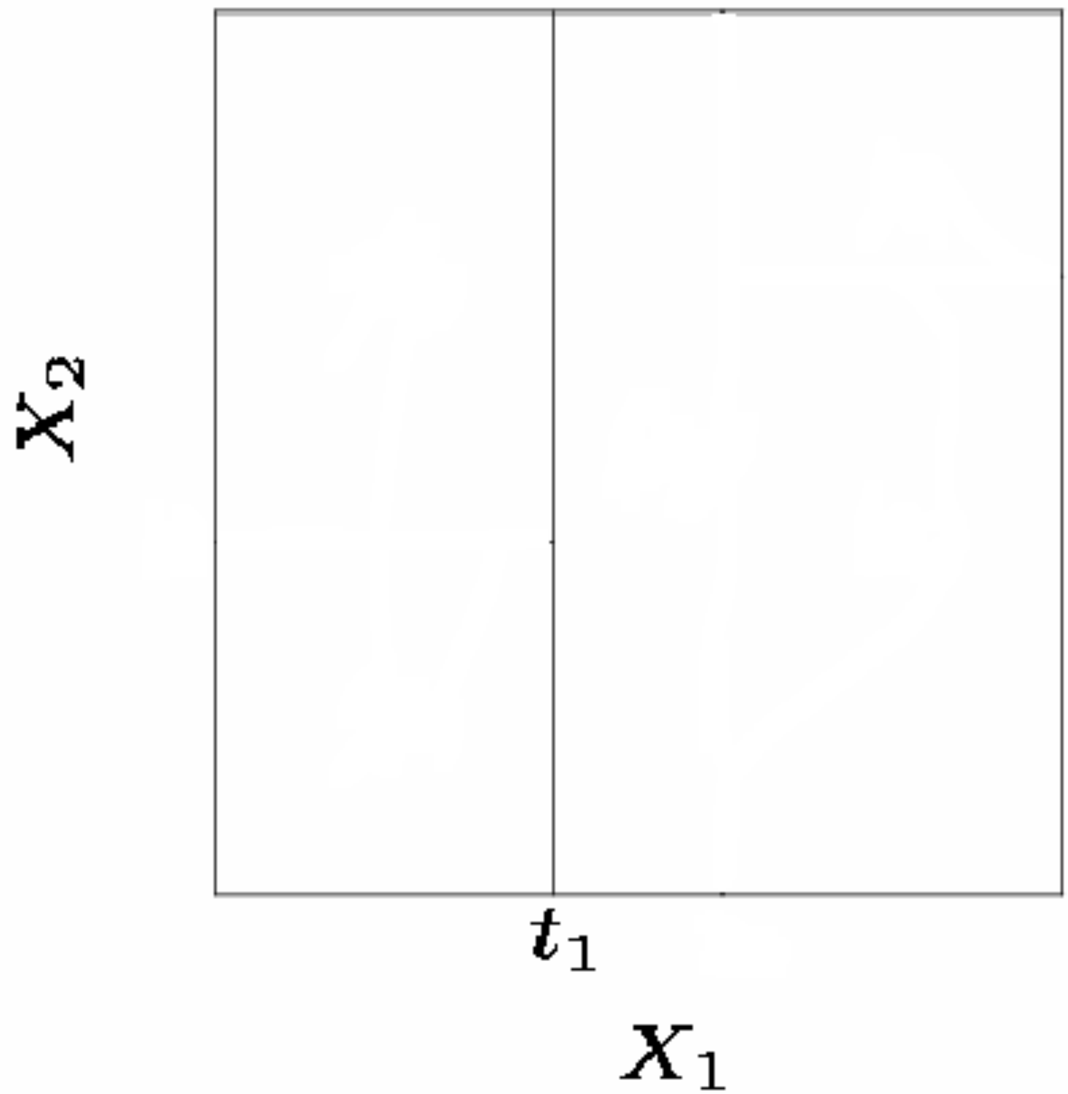
Splitting the X Variables

- Generally we create the partitions by iteratively splitting one of the X variables into two regions



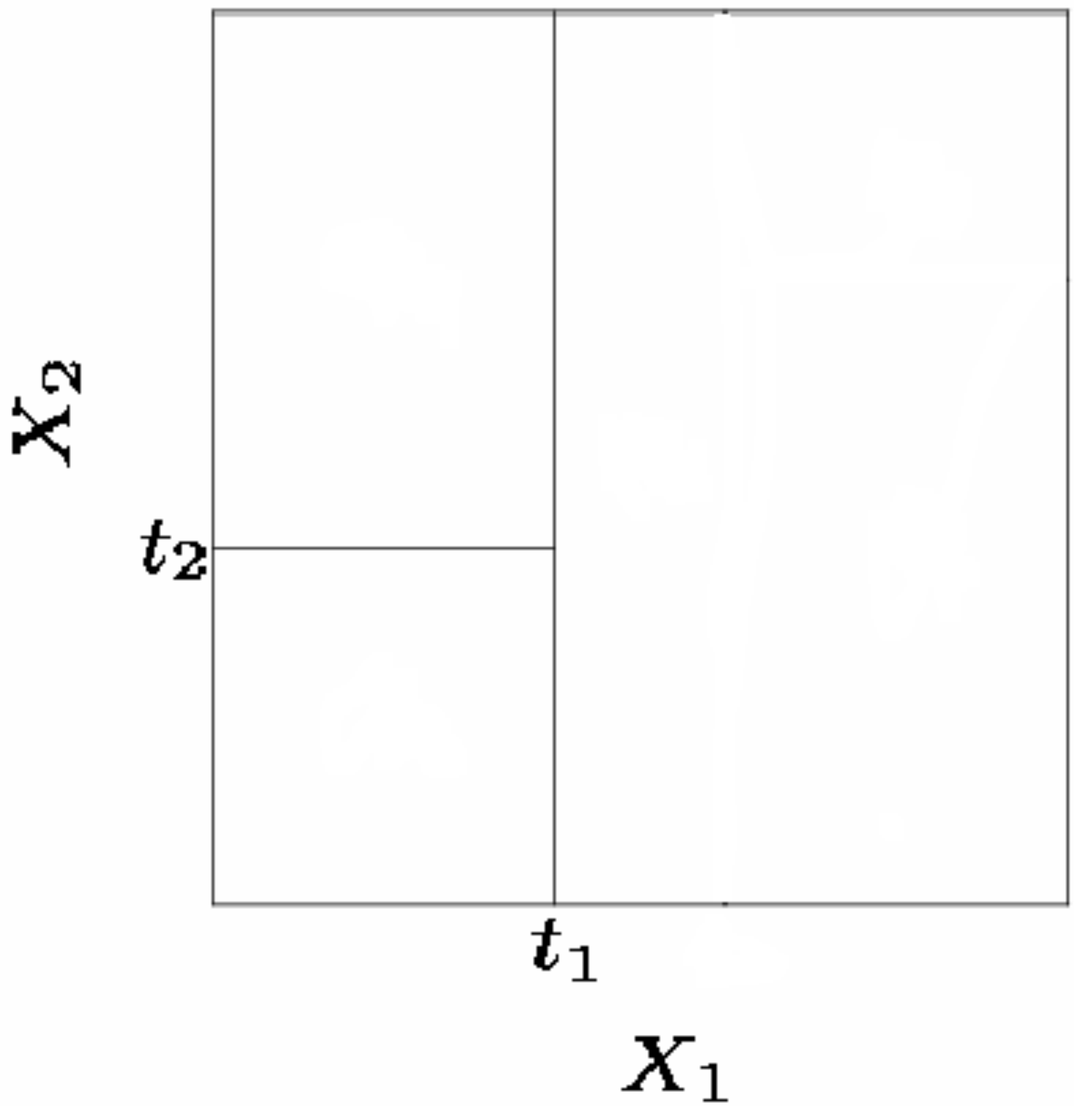
Splitting the X Variable

1. First split on $X_1=t_1$



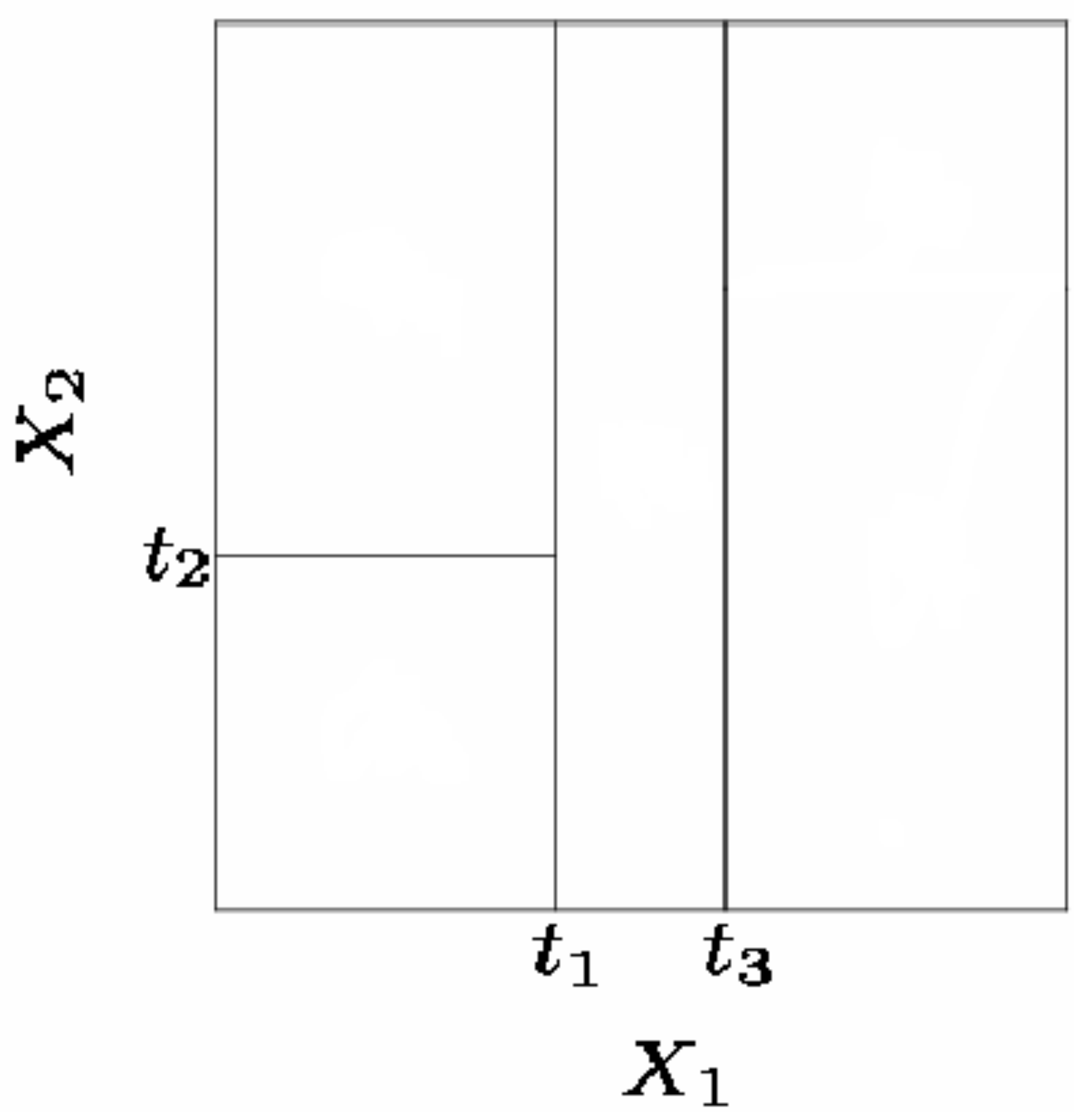
Splitting the X Variable

1. First split on $X_1=t_1$
2. If $X_1<t_1$, split on $X_2=t_2$



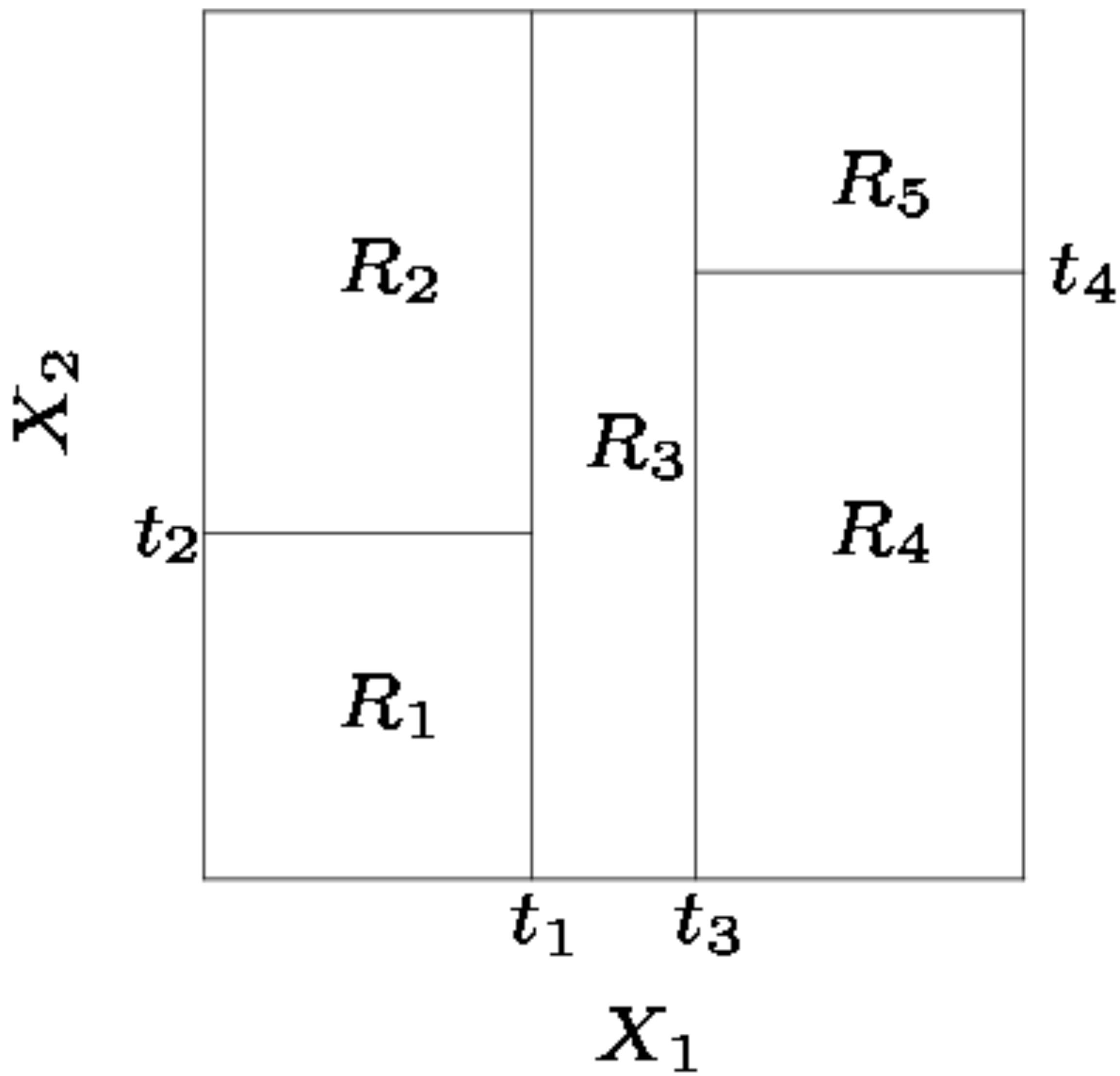
Splitting the X Variable

1. First split on $X_1=t_1$
2. If $X_1<t_1$, split on $X_2=t_2$
3. If $X_1>t_1$, split on $X_1=t_3$

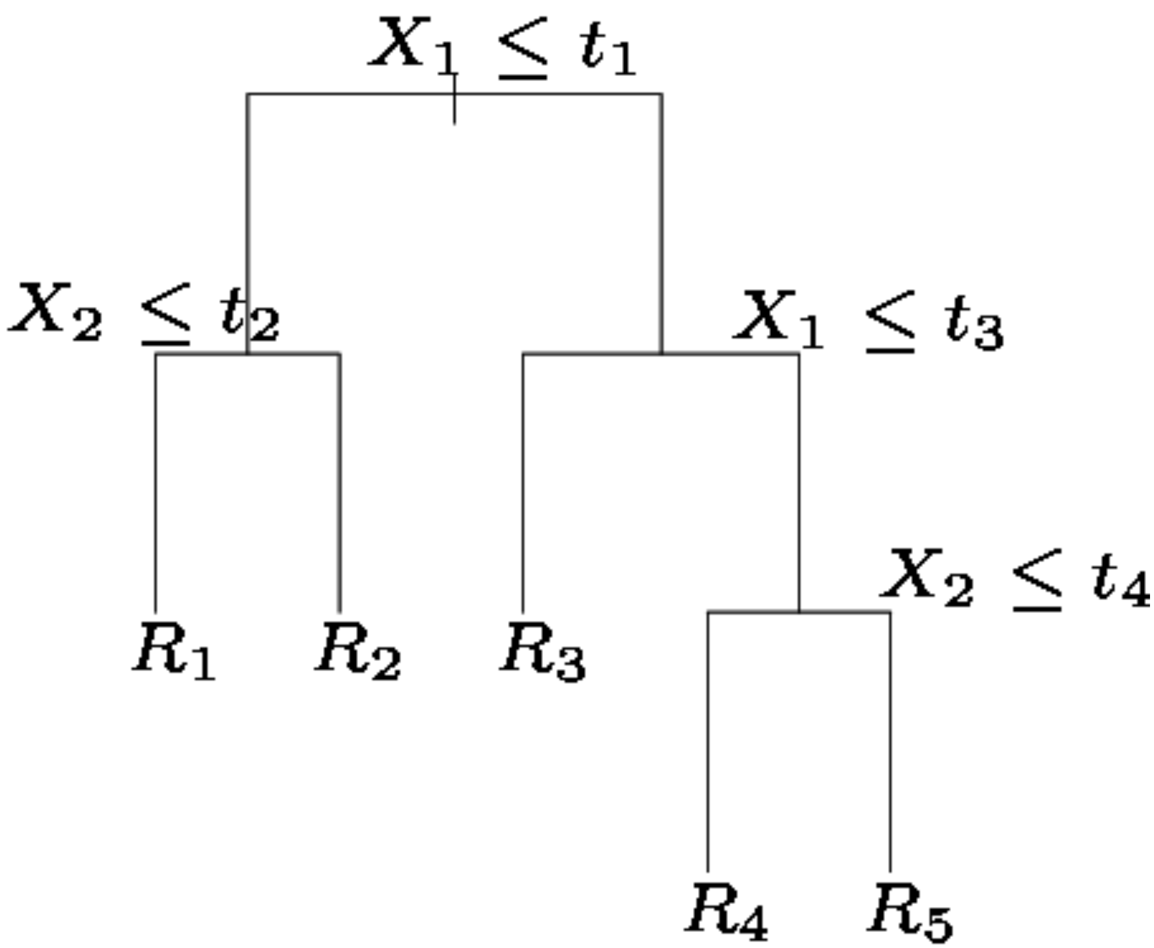


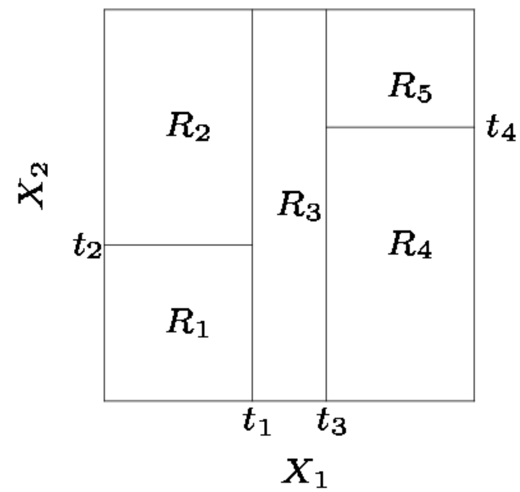
Splitting the X Variable

1. First split on $X_1=t_1$
2. If $X_1<t_1$, split on $X_2=t_2$
3. If $X_1>t_1$, split on $X_1=t_3$
4. If $X_1>t_3$, split on $X_2=t_4$



Splitting the X Variable



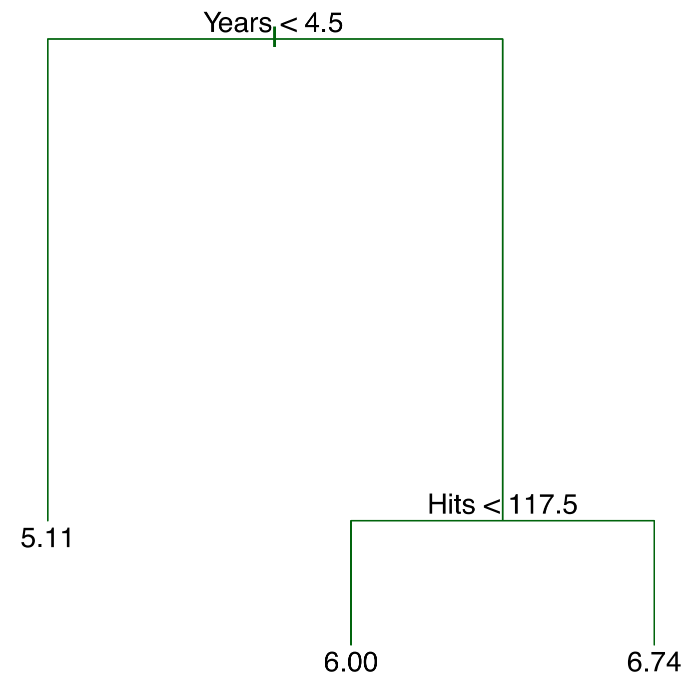


- When we create partitions this way we can always represent them using a tree structure.
- This provides a very simple way to explain the model to a non-expert i.e. your boss!

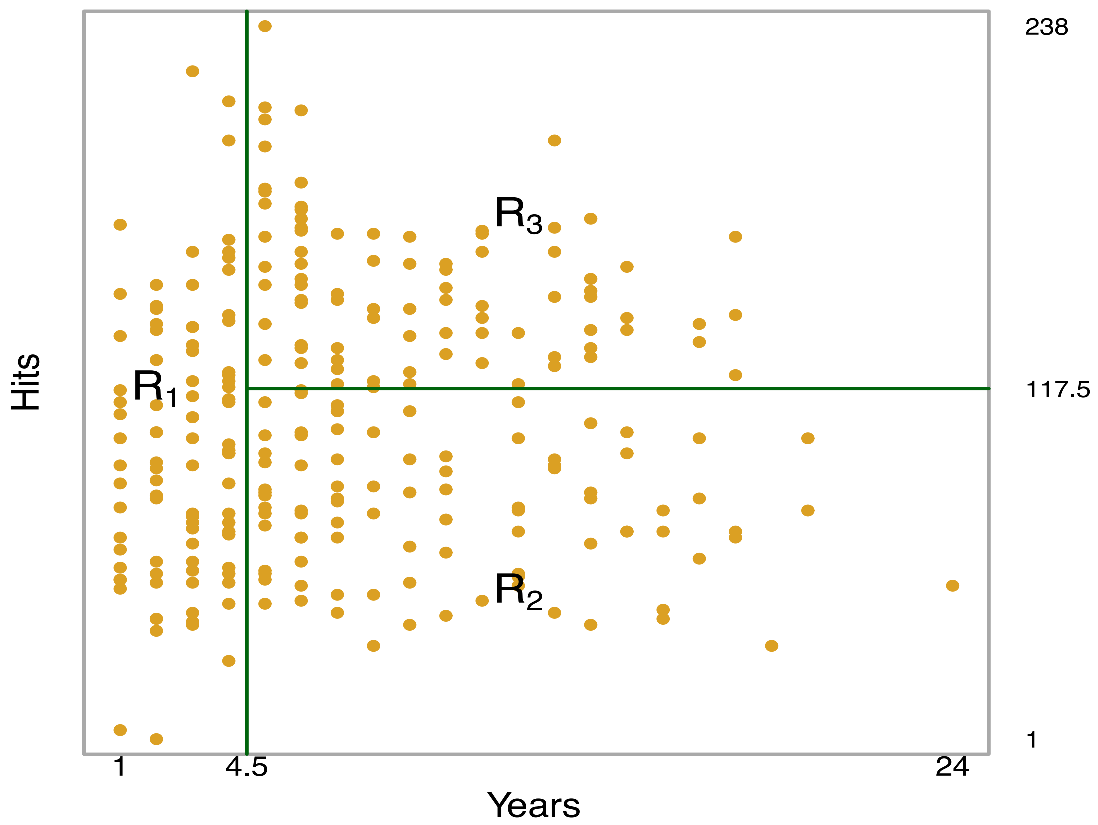
Example: Baseball Players' Salaries

- The predicted Salary is the number in each leaf node. It is the mean of the response for the observations that fall there
- Note that Salary is measured in 1000s, and log-transformed
- The predicted salary for a player who played in the league for more than 4.5 years and had less than 117.5 hits last year is

$$\$1000 \times e^{6.00} = \$402,834$$



Another way of visualizing the decision tree…



Some Natural Questions

1. Where to split? i.e. how do we decide on what regions to use i.e. $R_1, R_2, \ldots, R_k$ or equivalently what tree structure should we use?

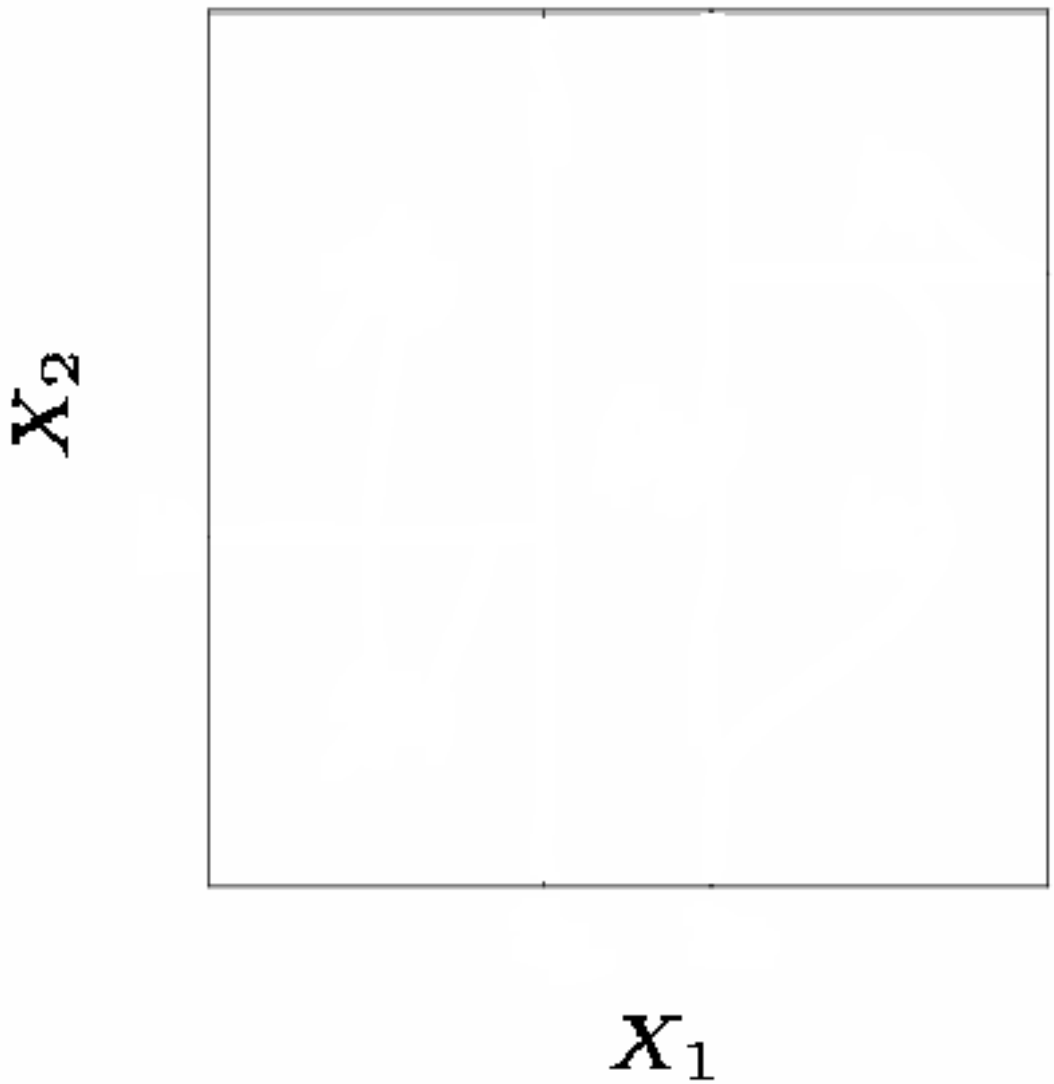
2. What values should we use for $\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_k$?

1. What values should we use for $\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_k$?

- Simple!
- For region R_j, the best prediction is simply the average of all the responses from our training data that fell in region R_j.

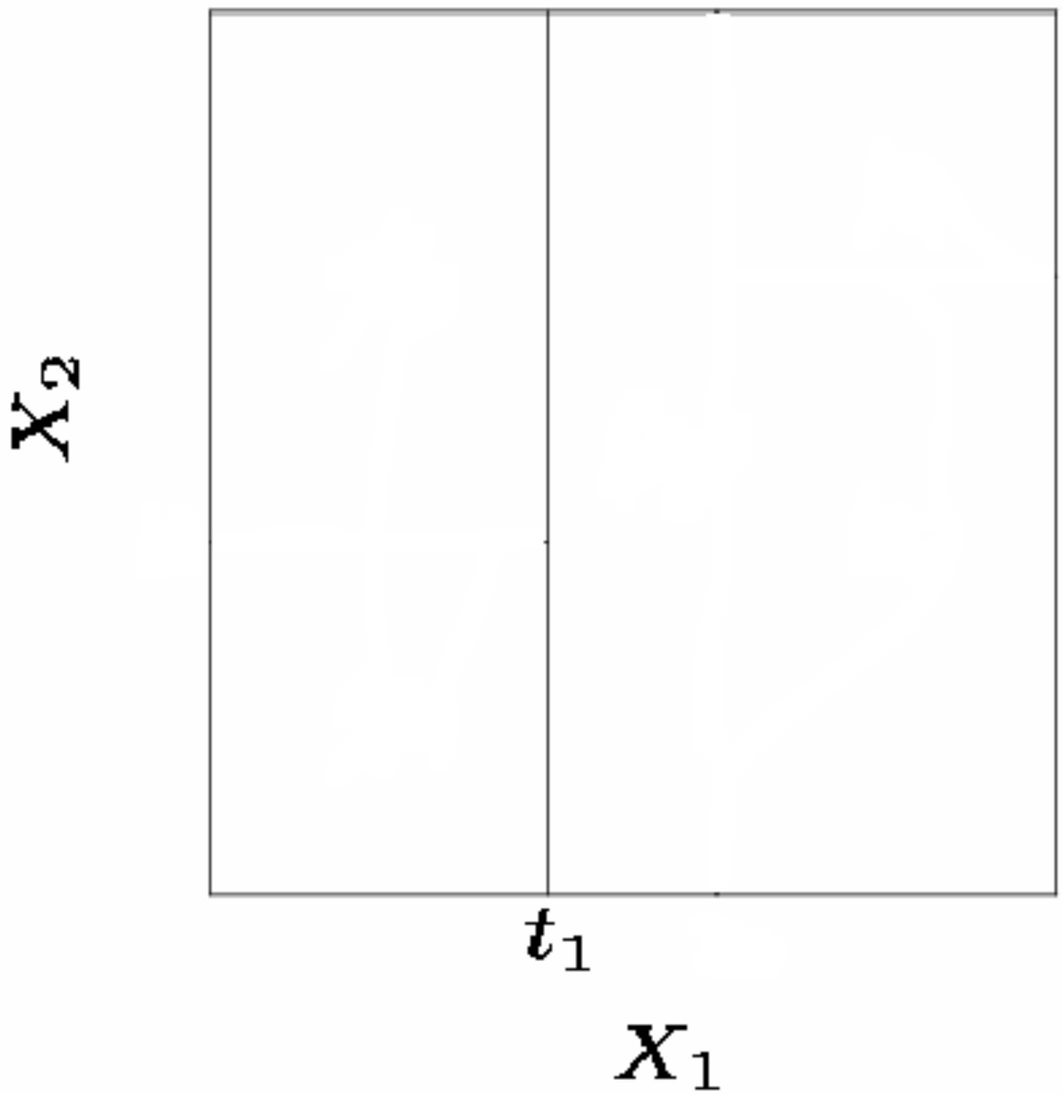
2. Where to Split?

- We consider splitting into two regions, X_j>s and X_j<s for all possible values of s and j.
- We then choose the s and j that results in the lowest MSE on the training data.



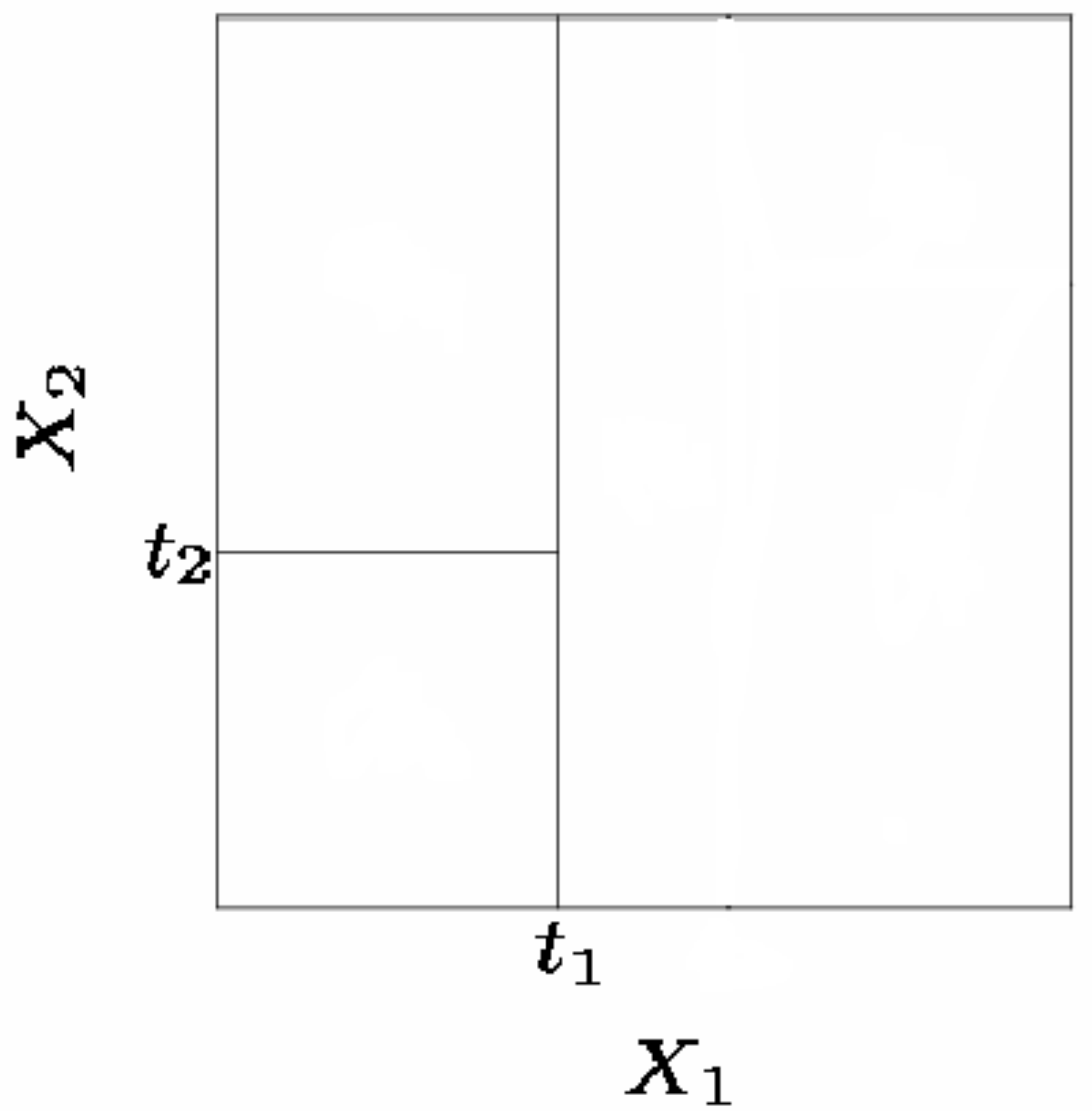
Where to Split?

- Here the optimal split was on X_1 at point t_1.
- Now we repeat the process looking for the next best split except that we must also consider whether to split the first region or the second region up.
- Again the criteria is smallest MSE.



Where to Split?

- Here the optimal split was the left region on X_2 at point t_2.
- This process continues until our regions have too few observations to continue e.g. all regions have 5 or fewer points.

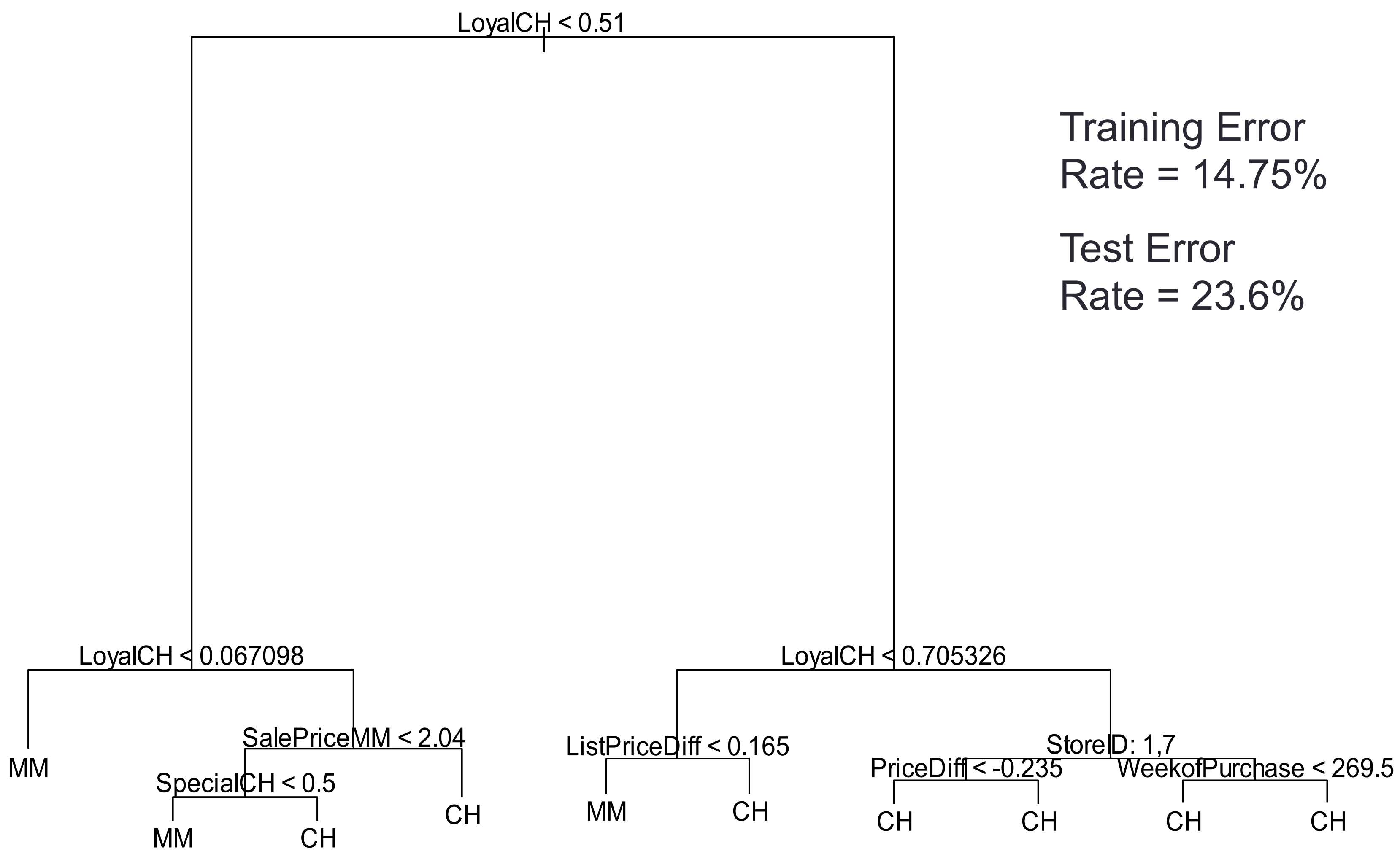


CLASSIFICATION TREES

Growing a Classification Tree

- A classification tree is very similar to a regression tree except that we try to make a prediction for a categorical rather than continuous Y.
- For each region (or node) we predict the most common category among the training data within that region.
- The tree is grown (i.e. the splits are chosen) in exactly the same way as with a regression tree except that minimizing MSE no longer makes sense.
- There are several possible different criteria to use such as the “gini index” and “cross-entropy” but the easiest one to think about is to minimize the error rate.

Example: Orange Juice Preference



Training Error Rate = 14.75%

Test Error Rate = 23.6%

TREE PRUNING

Improving Tree Accuracy

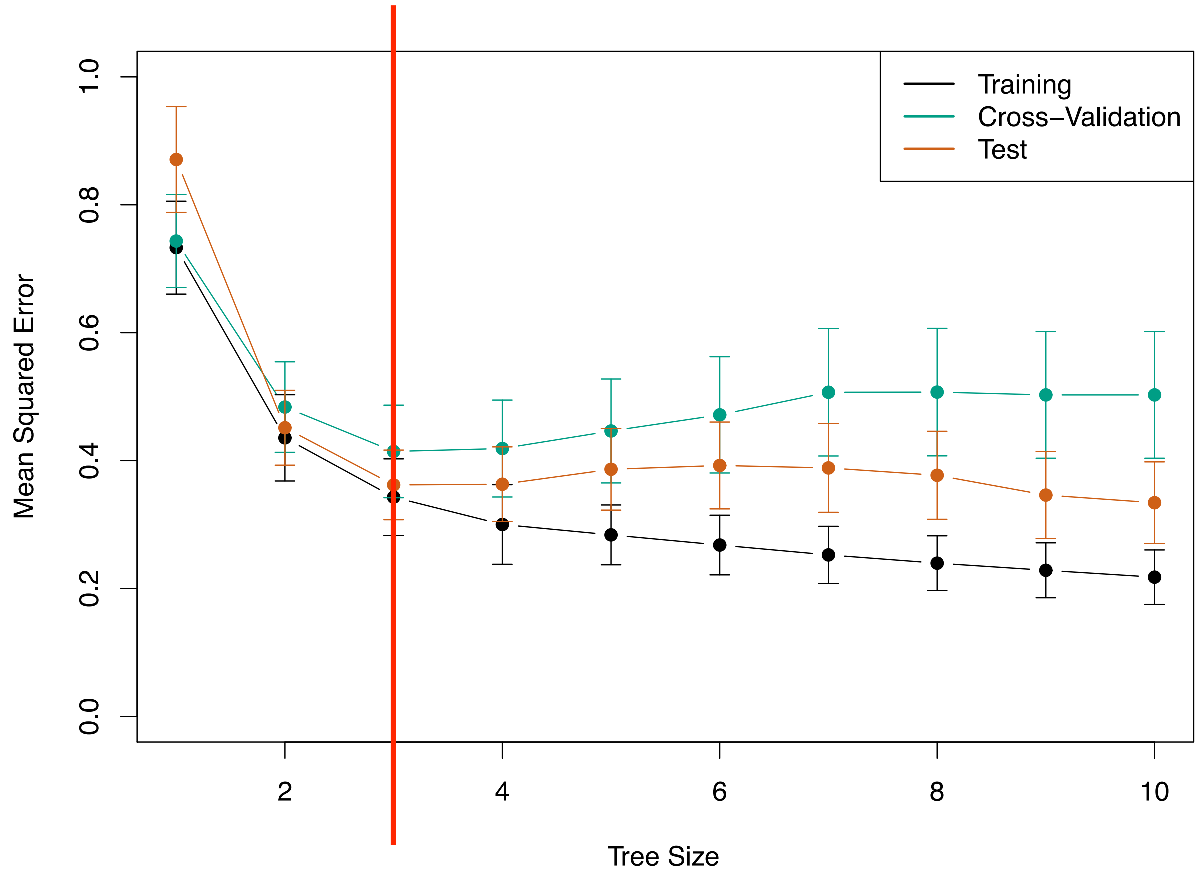
- A large tree (i.e. one with many terminal nodes) may tend to over fit the training data in a similar way to neural networks without a weight decay.

- Generally, we can improve accuracy by "pruning" the tree i.e. cutting off some of the terminal nodes.

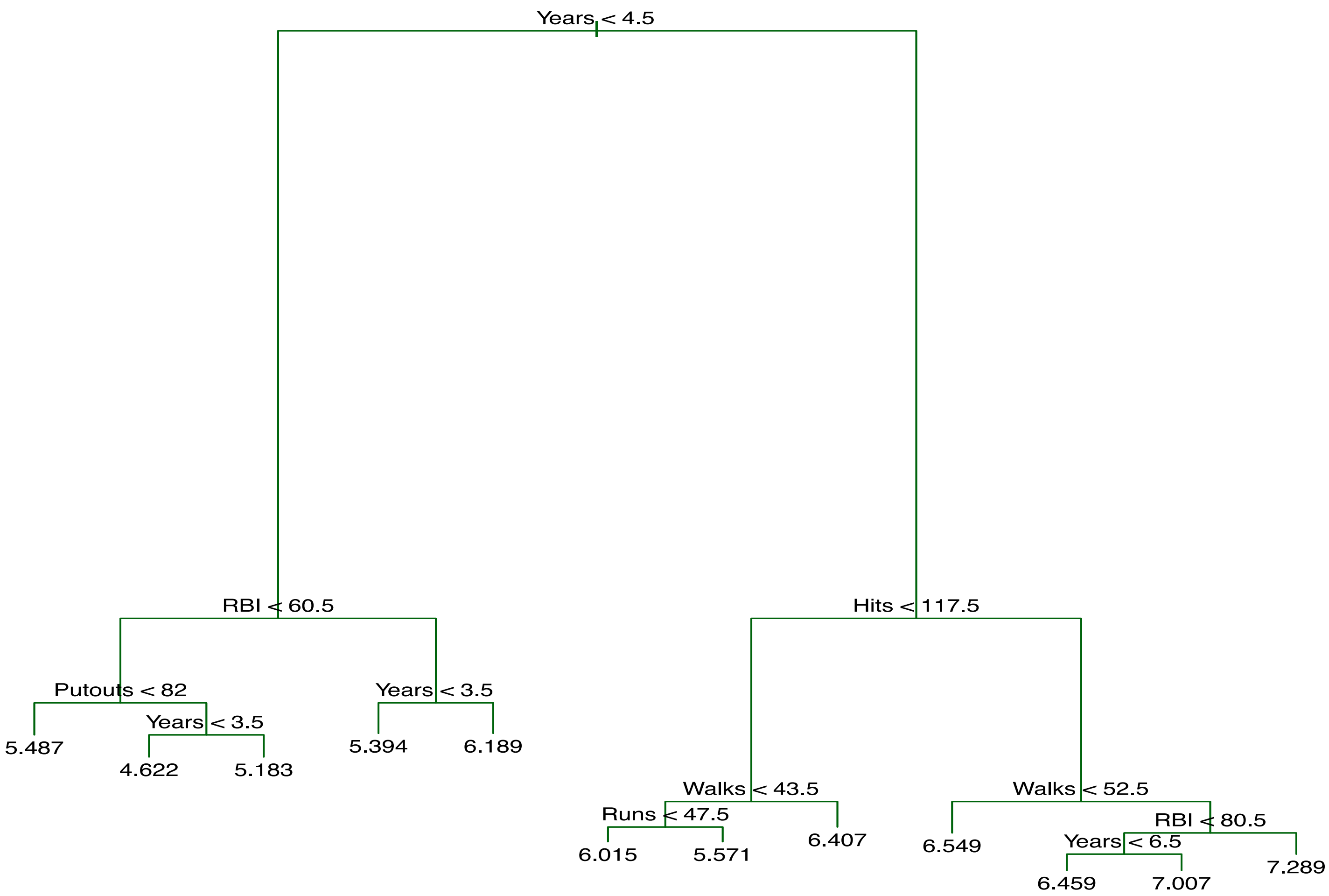
- How do we know how far back to prune the tree? We use **cross validation** to see which tree has the lowest error rate.

Example: Baseball Players' Salaries

- The minimum cross validation error occurs at a tree size of 3

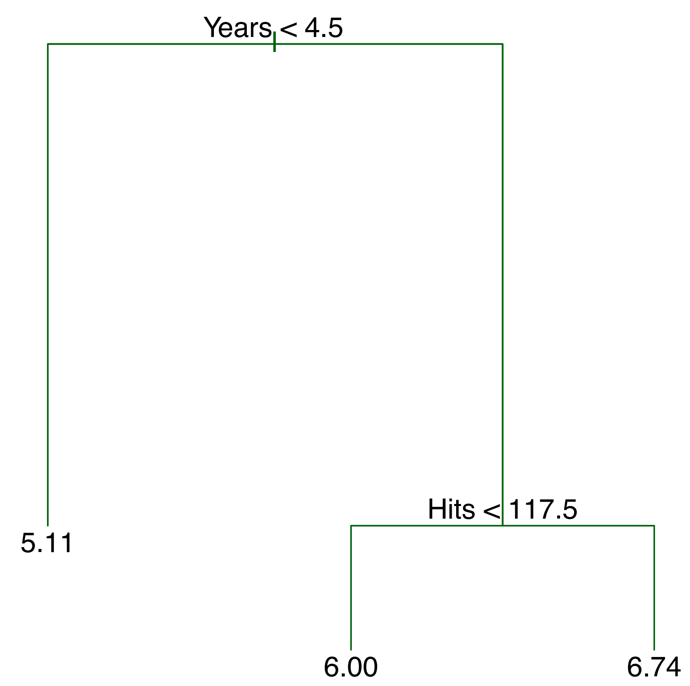


Example: Baseball Players' Salaries

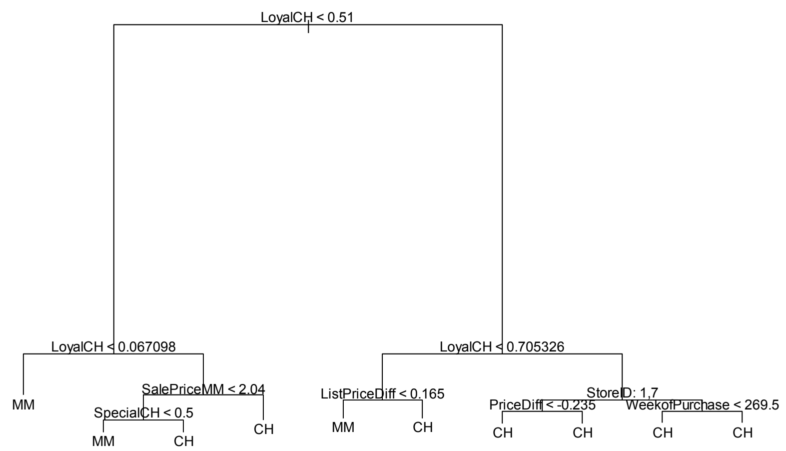


Example: Baseball Players' Salaries

- Cross Validation indicated that the minimum MSE is when the tree size is three (i.e. the number of leaf nodes is 3)



Example: Orange Juice Preference

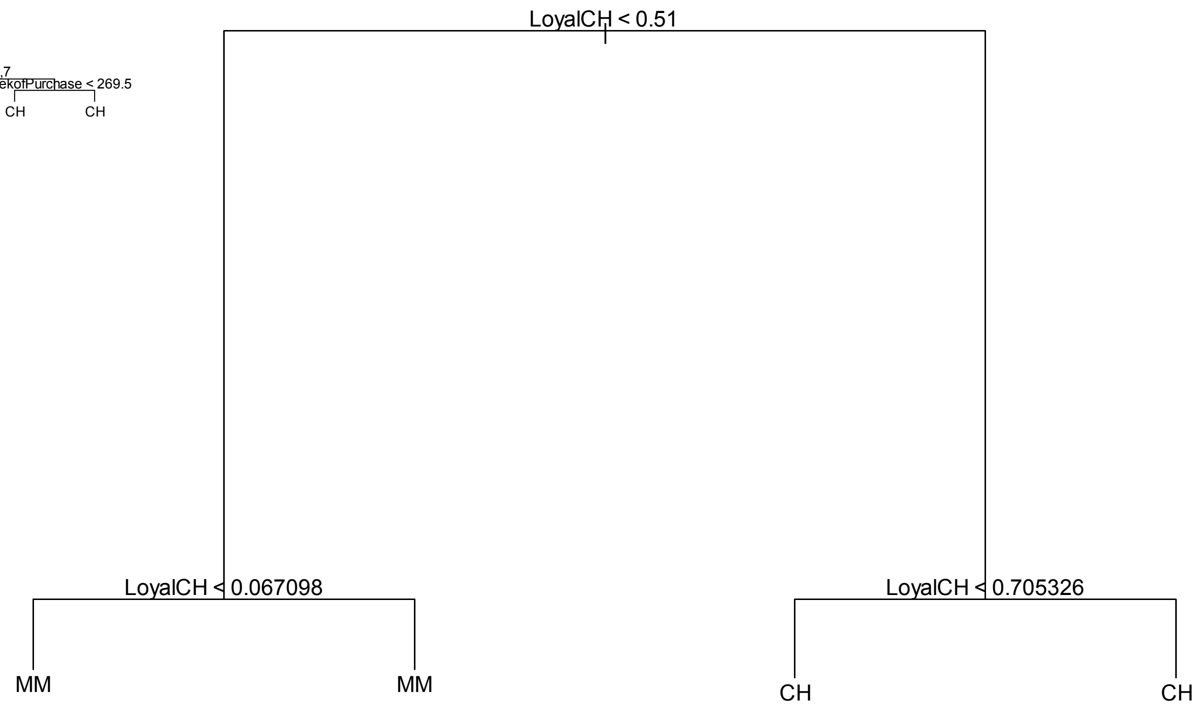


Full Tree Training Error Rate = 14.75%

Full Tree Test Error Rate = 23.6%

Pruned Tree

CV Tree Error Rate = 22.5%



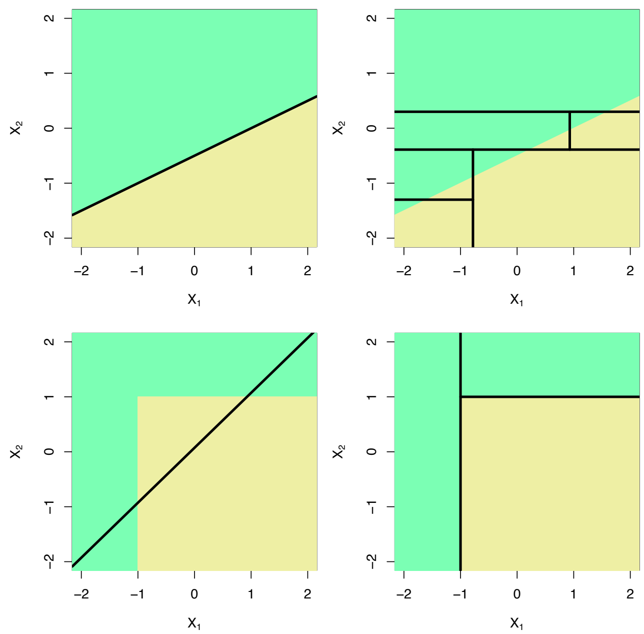
TREES VS. LINEAR MODELS

Trees vs. Linear Models

- Which model is better?
  - If the relationship between the predictors and response is linear, then classical linear models such as linear regression would outperform regression trees
  - On the other hand, if the relationship between the predictors is non-linear, then decision trees would outperform classical approaches

Trees vs. Linear Model: Classification Example

- Top row: the true decision boundary is linear
  - Left: linear model (good)
  - Right: decision tree
- Bottom row: the true decision boundary is non-linear
  - Left: linear model
  - Right: decision tree (good)



ADVANTAGES AND DISADVANTAGES OF TREES

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 that we examine in this course