Decision Trees Part 2

Previously on CSCI 4622

Dpar = {0, 1, ..., 7}   3

p_c = FRACTION IN CLASS C

Last time we saw how to construct Decision Tree classifiers with binary categorical data

- Partition training data according to binary tree
- For each node, split on a single feature so as to minimize impurity in the result child nodes

One impurity measure is **Entropy**: $\sum_c -p_c \log_2 p_c$

CROSS-ENTROPY

- One way to quantify reduction in impurity is through Information Gain

$$IG(D_{par}) = I(D_{par}) - \frac{|D_{left}|}{|D_{par}|} I(D_{left}) - \frac{|D_{right}|}{|D_{par}|} I(D_{right})$$

- Evaluate IG for each possible split on a feature, choose the one with largest IG

Some Questions

- We've looked at binary categorical features. What happens if the features are continuous?
- Are there other measures of impurity besides entropy?
- How well do these things perform, anyway?

Splitting With Continuous Features

- How could we do this in a naïve way, and at what cost?

n TRAINING EXAMPLES

CHOOSE EACH ONE & COMPUTE IG

$\Rightarrow O(n^2)$

- What could we do that would be smarter?

SORT DATA BY WIND FEATURE

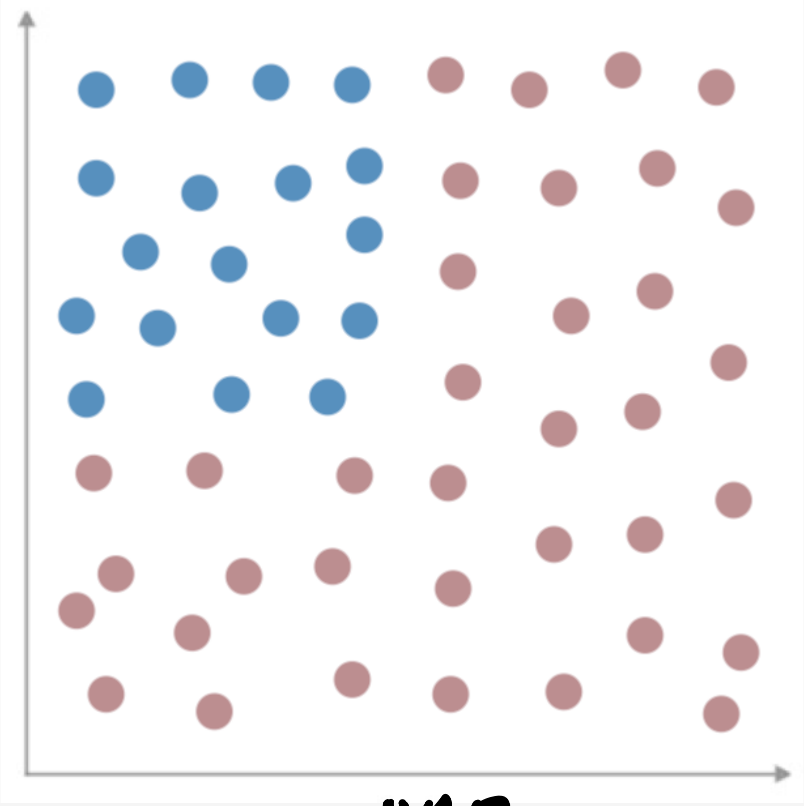
WIND: 5 × 10 × 15 × 20 × 25 × 31 ×

$P_0 = 0$, $P_1 = 1$ $\quad$ $P_0 = .5$, $P_1 = .5$ $\quad$ $P_0 = 2/3$, $P_1 = 1/3$

$O(n) + \boxed{O(n \log n)}$

+VE - BLUE

SUN



WIND

What Other Impurity Measures Exist?

- The obvious, **misclassification error**

+ + −
+ − ⇒ predict all +ve

$MCE = \frac{2}{5}$

let p = fraction of points that are positive

$MCE = \min(p, 1-p)$

p small ⇒ pred −ve

p large ⇒ pred +ve

What Other Impurity Measures Exist?

- The less obvious, the **Gini Index**

$$1 - \sum_c p_c^2$$

$$p$$

$$(1-p)$$

$$= 1 - p^2 - (1-p)^2$$

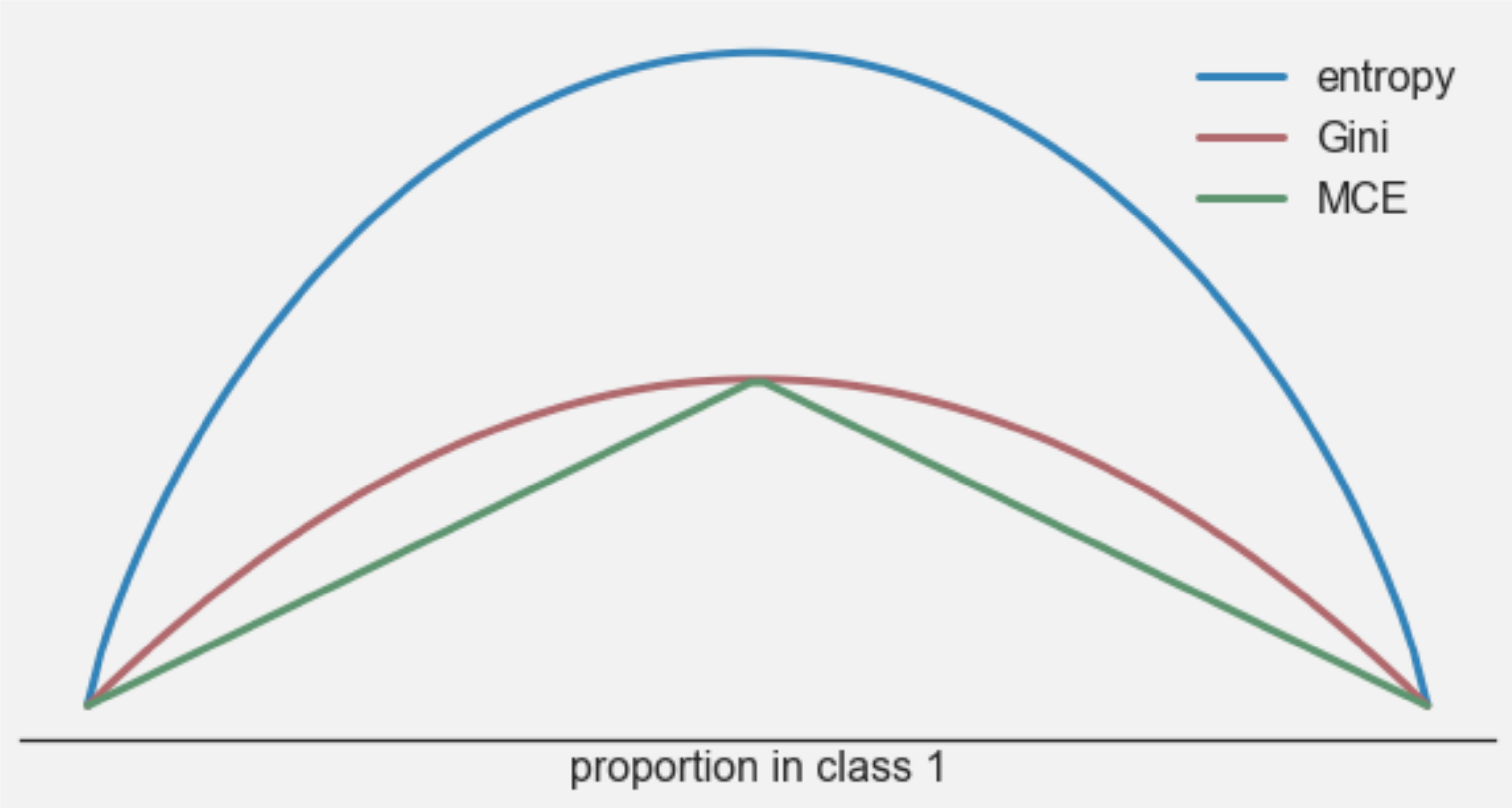
$$= 2p(1-p)$$

PURE $p = 0$ or 1 $\quad G \to 0$

IMPURE $p = 1/2$ $\quad G \to \frac{1}{2}$

What Other Impurity Measures Exist?

- So we have Entropy, MCE, and Gini Index. Which one is better?



What Other Impurity Measures Exist?

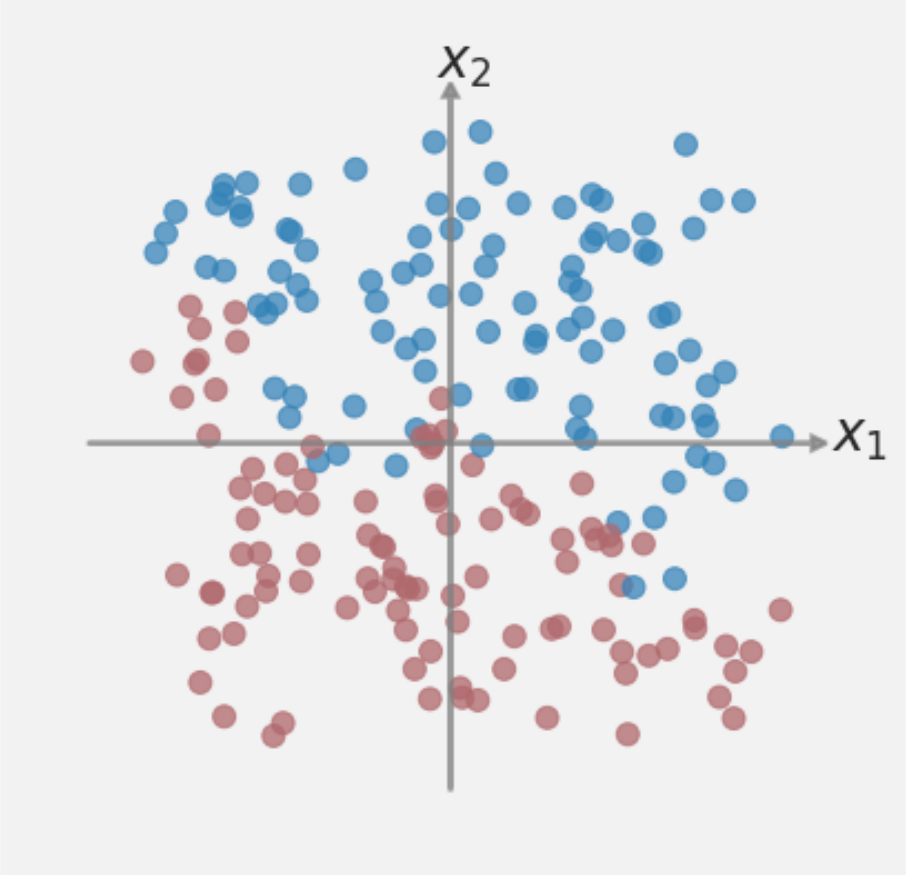
- So we have Entropy, MCE, and Gini Index. Which one is better?

Answer: Ehh, unclear

- Gini and Entropy are differentiable
- Gini is slightly cheaper because Entropy uses logs
- In the end, it doesn't really matter
- Gini used by CART
- Entropy used by C4.5
- MCE sometimes used in post-pruning strategies

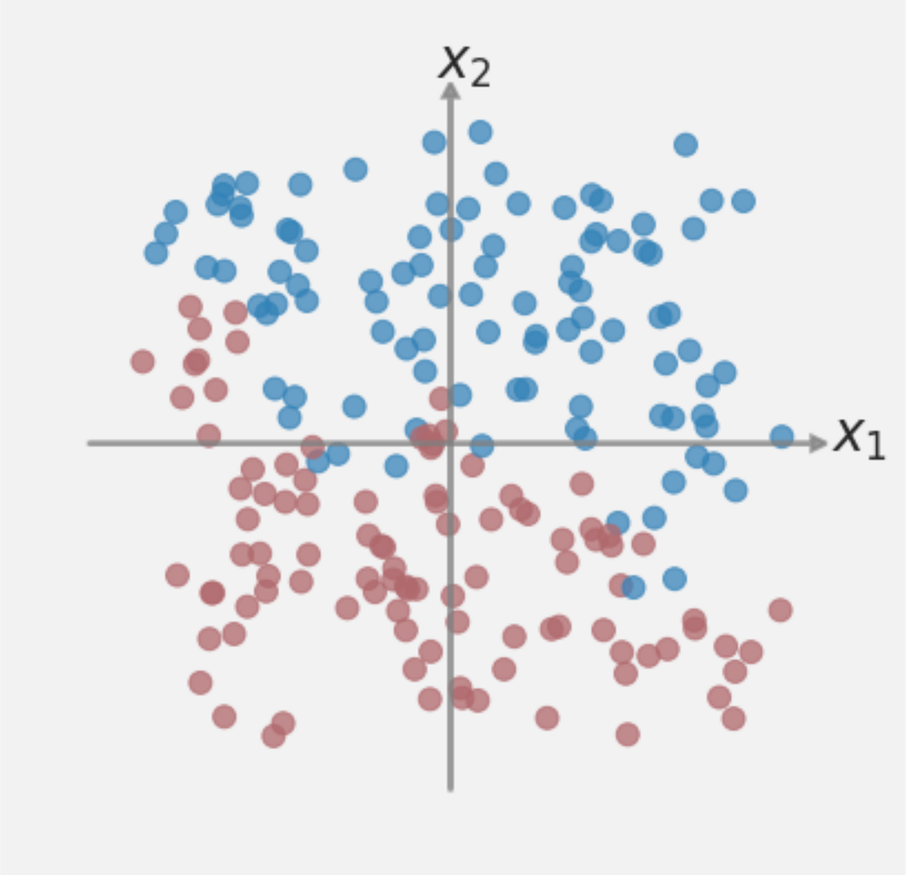
So How Well Do These Things Work?

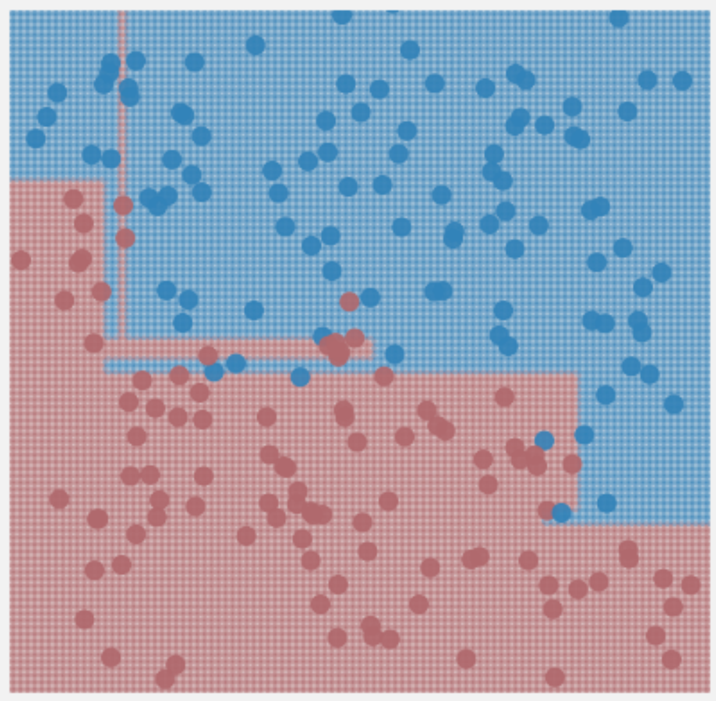
- In general, Decision Tree Classifiers are very high-variance methods



So How Well Do These Things Work?

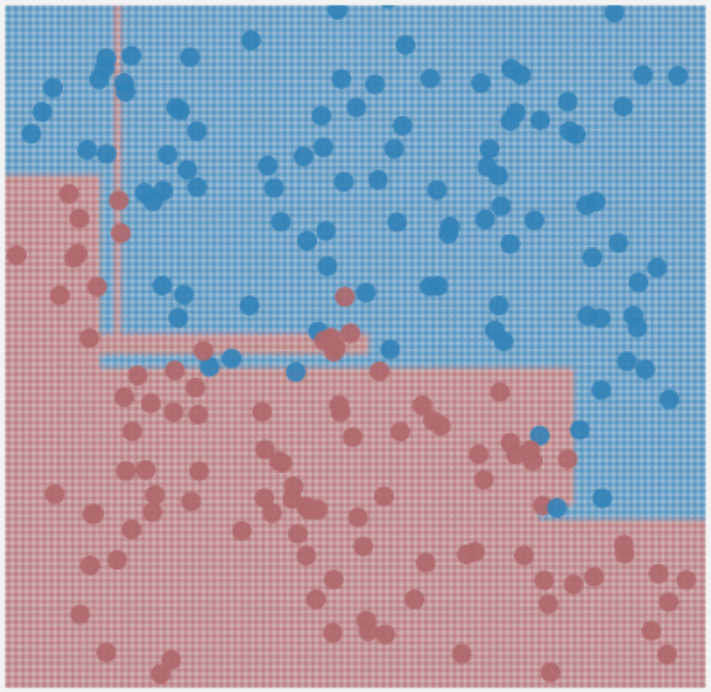
- In general, Decision Tree Classifiers are very high-variance methods and tend to overfit





Combating Overfitting

- How could we combat this?



Combating Overfitting

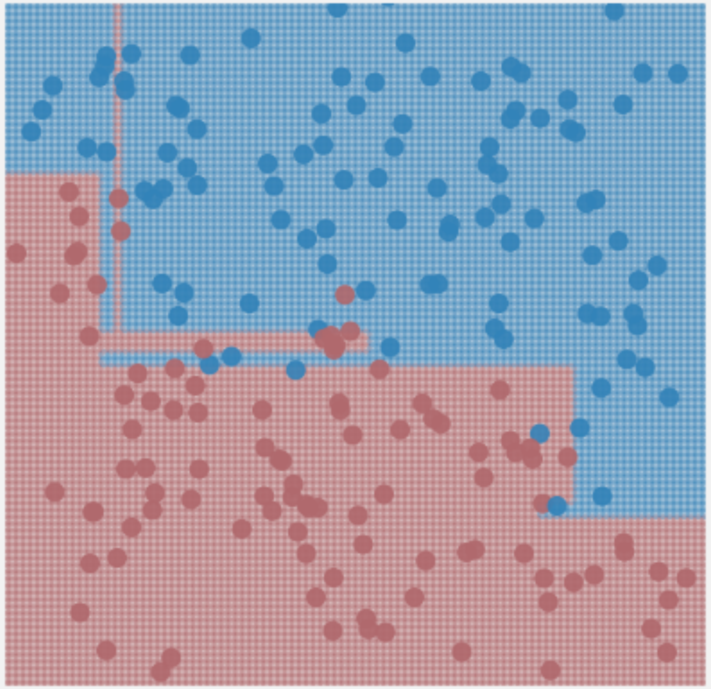
- How could we combat this?

Prepruning:

- Don't let the tree have too many levels
- Stopping conditions

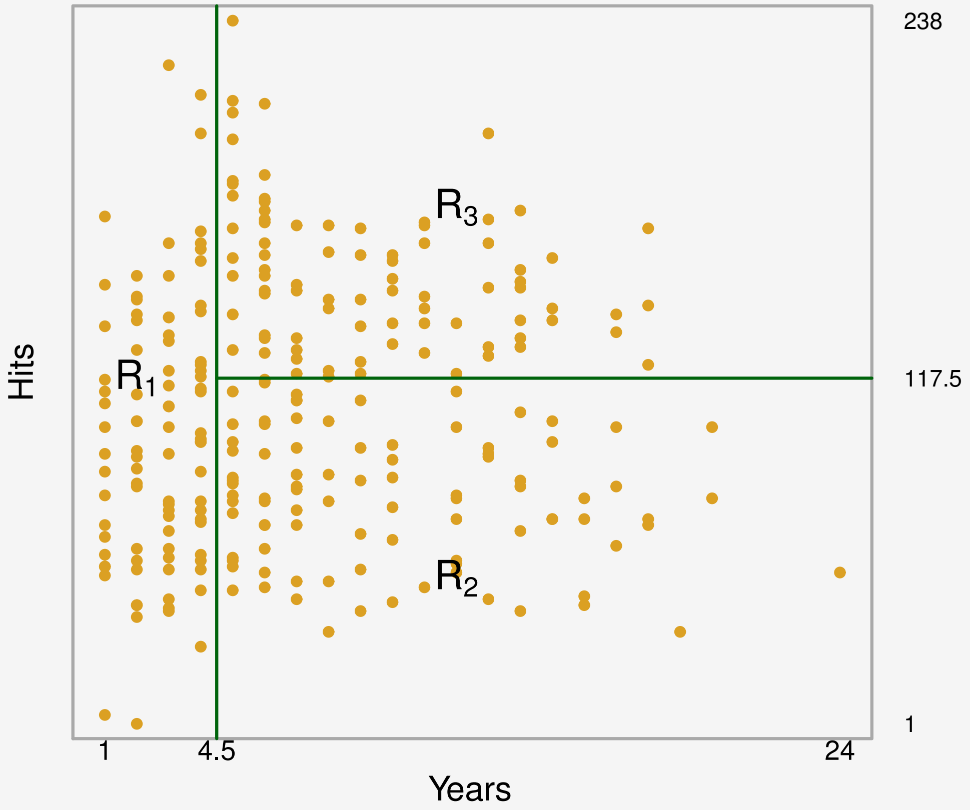
Postpruning:

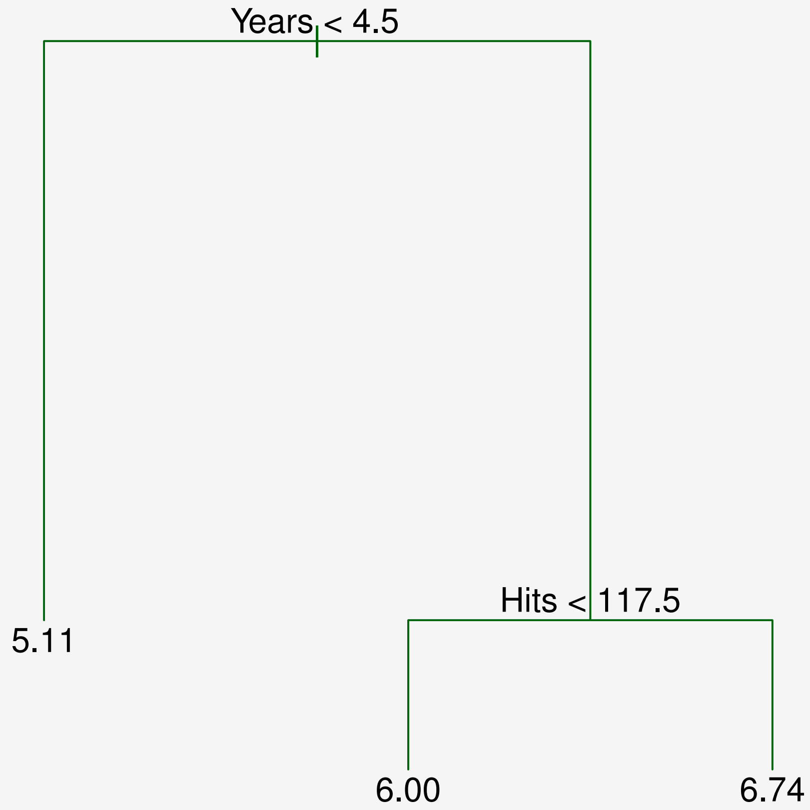
- Build the complete tree
- Go back and remove vertices that don't affect performance too much



Bonus: Regression Trees

- Suppose you want to PREDICT the salary of a MLB player based on two features: how many hits they average a year and how many years they've been in the league



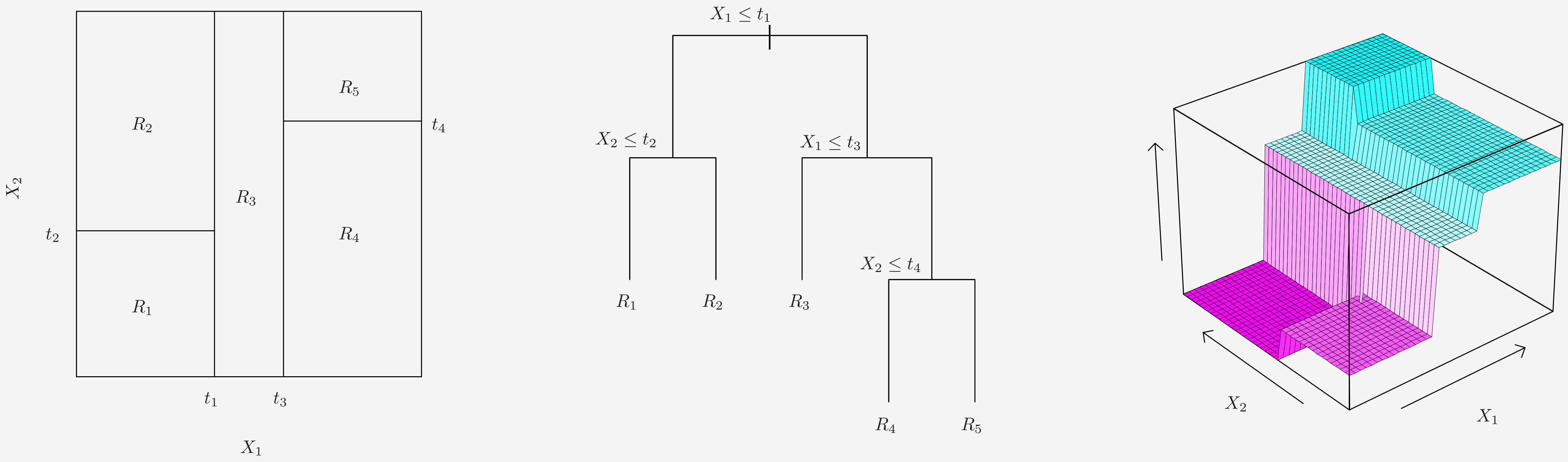


Bonus: Regression Trees

- Suppose you want to PREDICT the salary of a MLB player based on two features: how many hits they average a year and how many years they've been in the league
- Perform the usual binary splitting by feature
- Instead of classification, predict the response based on average response in leaf node

Bonus: Regression Trees

- Instead of classification, predict the response based on average response in leaf node



Bonus: Regression Trees

- What measure do we use to decide which feature to split on?

- The goal is to find boxes that minimize the $RSS = \sum_{j=1}^{J} \sum_{i \in R_j} \left(y_i - \hat{y}_{R_j}\right)^2$

- Consider splitting the data on a node into two boxes. Choose feature and split to minimize

$$\sum_{i:\; x_i \in R_1(j,s)} \left(y_i - \hat{y}_{R_1}\right)^2 + \sum_{i:\; x_i \in R_2(j,s)} \left(y_i - \hat{y}_{R_2}\right)^2$$

where $R_1(j,s) = \{X \mid X_j < s\}$ and $R_2(j,s) = \{X \mid X_j \geq s\}$

Where Do We Go From Here?

- Unfortunately, Decision Tree Classifiers are prone to overfitting
- On their own, tend to do worse than other classifiers we've seen

But **WITH THEIR POWERS COMBINED**!

Next up, Ensemble Methods.

- Take various weaker classifiers, and combine them to get strong classifiers
- Bagging and Random Forests
- Boosted Decision Trees and the AdaBoost Algorithm