INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE

IN4402: Aplicaciones de Probabilidades y Estadística

Classification and Regression Trees (CART)

ANDRÉS FERNÁNDEZ

CLASSIFICATION AND REGRESSION TREES

INTRODUCTION AND MAIN CONCEPTOS

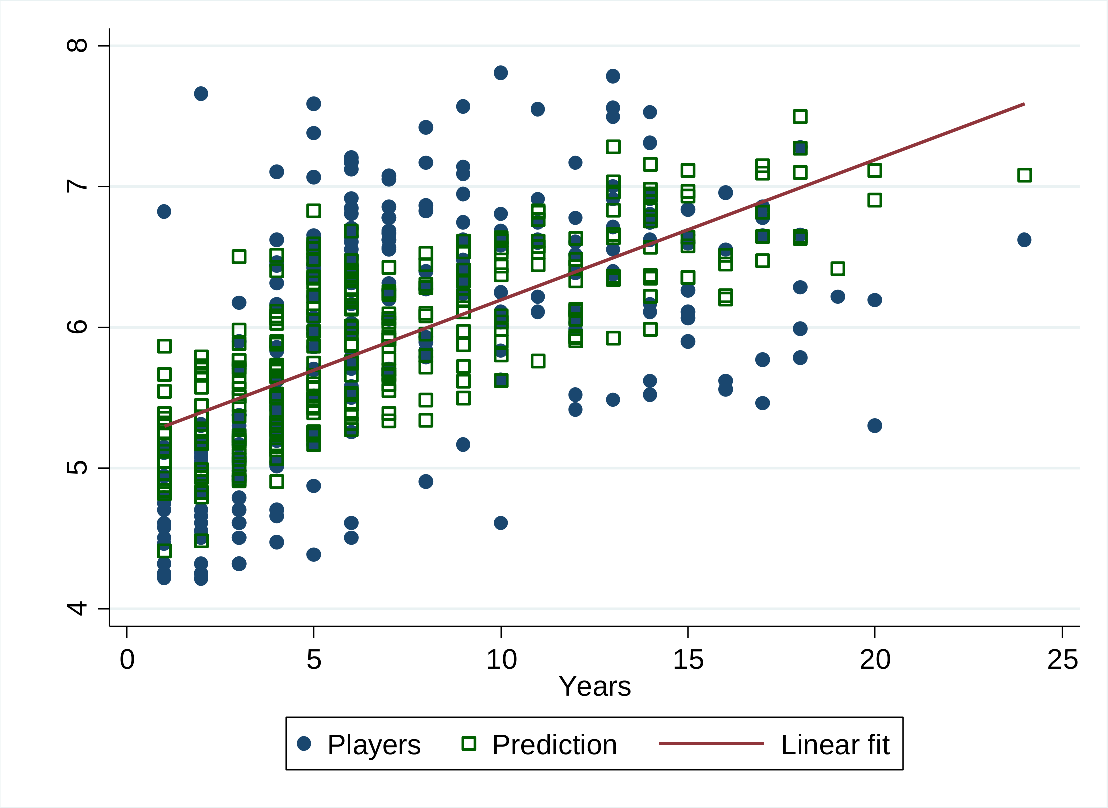
- ML algorithms *split data into subregions* in order to classify or predict

- **Trees** are the *graphical expression* of this process
  - They are built following a question-answer structure over a database
  - It is clear and easy to interpret

CLASSIFICATION AND REGRESSION TREES
INTRODUCTION AND MAIN CONCEPTOS

- Let's try to predict a baseball player salary from some characteristics
- In **Linear regression (OLS)** we estimate the parameters (β_0, β_k) that minimizes the residual sum of squares (RSS)

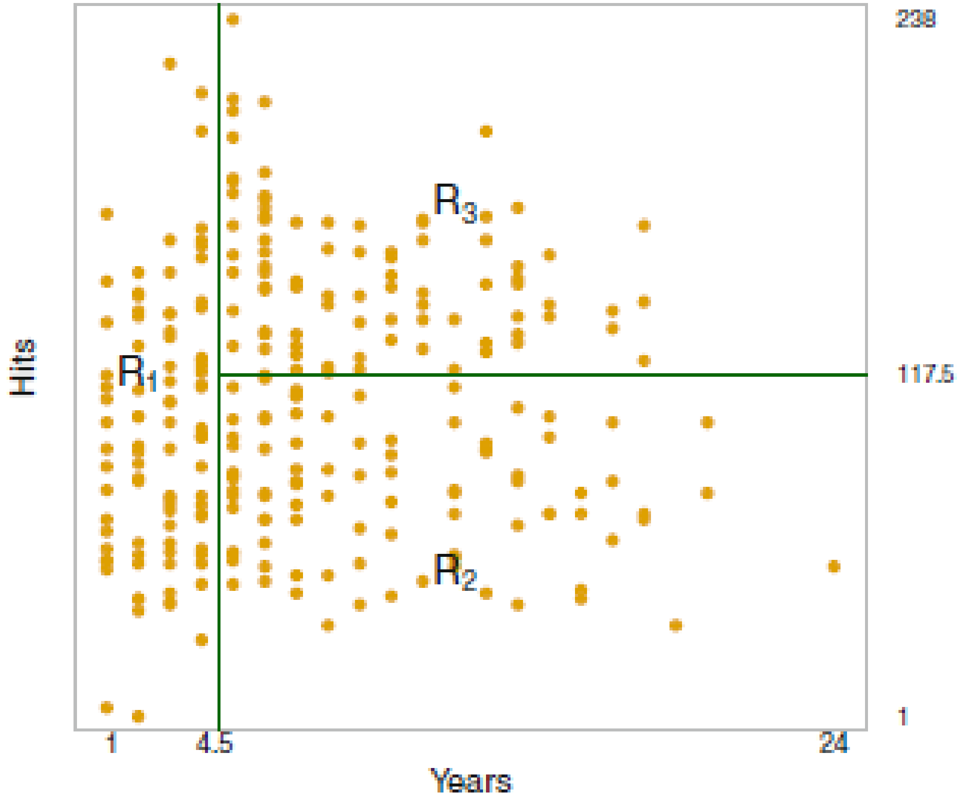
	lsalary
hits	0.009
	(0.001)**
years	0.098
	(0.008)**
_cons	4.275
	(0.118)**
R2	0.48
N	263



CLASSIFICATION AND REGRESSION TREES
INTRODUCTION AND MAIN CONCEPTOS

- Let's try to predict a baseball player salary from some characteristics
- In **tree-based algorithms** we split data into regions, and then every region is averaged to predict the outcome variable

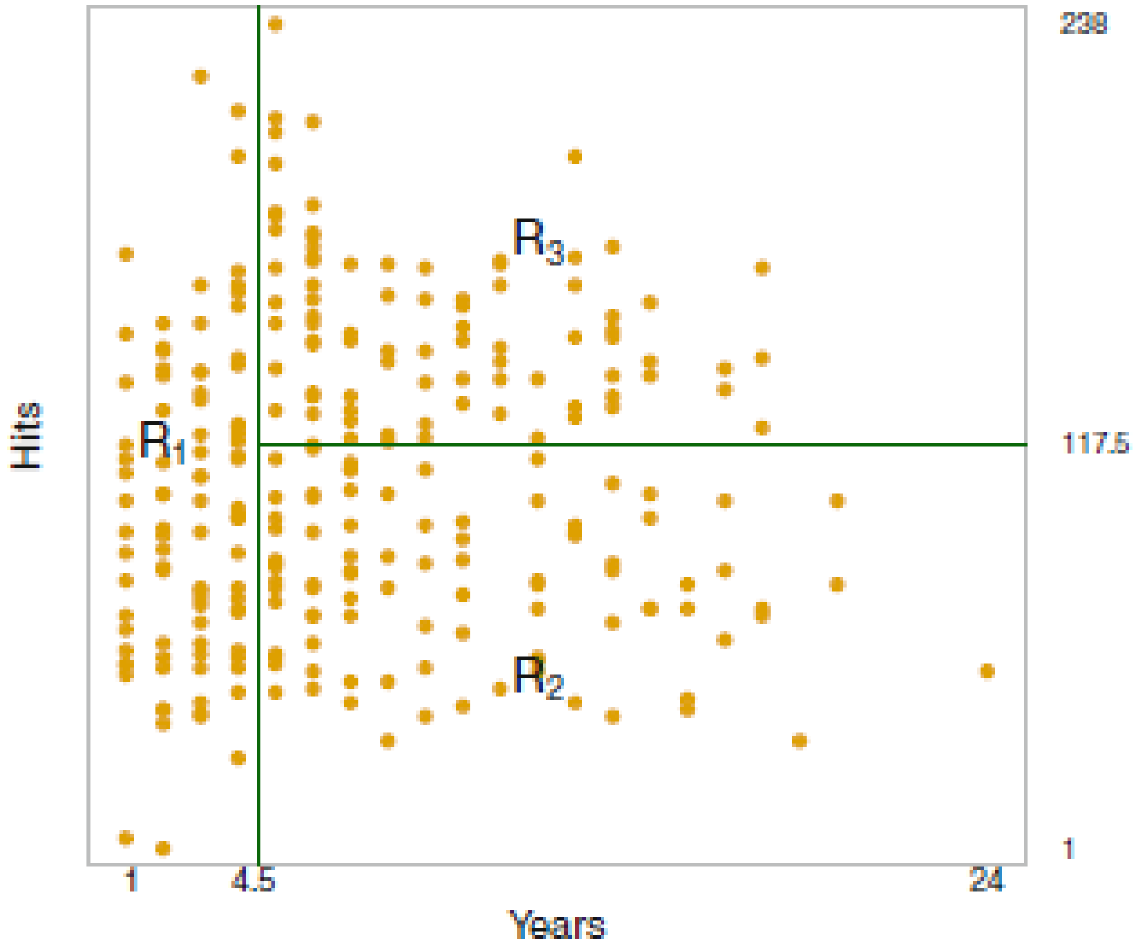
Region	Predicted LogSalary	Predicted Salary
R1	5.11	$165,174
R2	6.00	$402,834
R3	6.74	$845,346



Source: James, Witten, Hastie & Tibshirani (2013) An Introduction to Statistical Learning: with applications in R. New York: Springer

CLASSIFICATION AND REGRESSION TREES

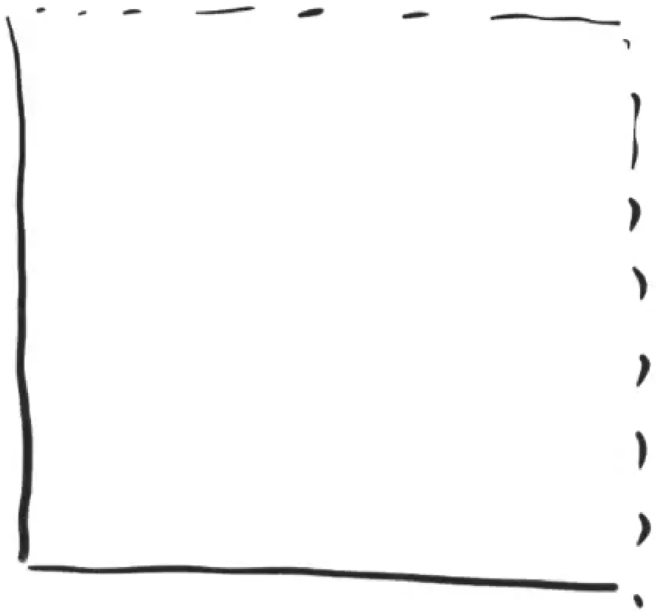
INTRODUCTION AND MAIN CONCEPTOS



Region	Prediction
R1	$165,174
R2	$402,834
R3	$845,346

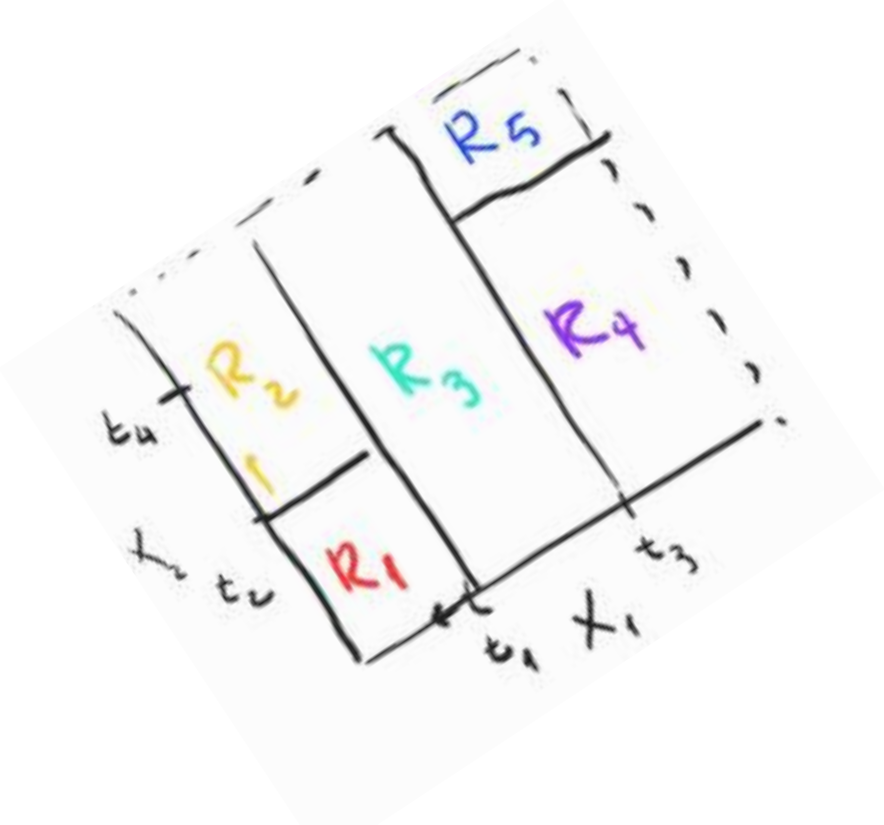
Source: James, Witten, Hastie & Tibshirani (2013) An Introduction to Statistical Learning: with applications in R. New York: Springer

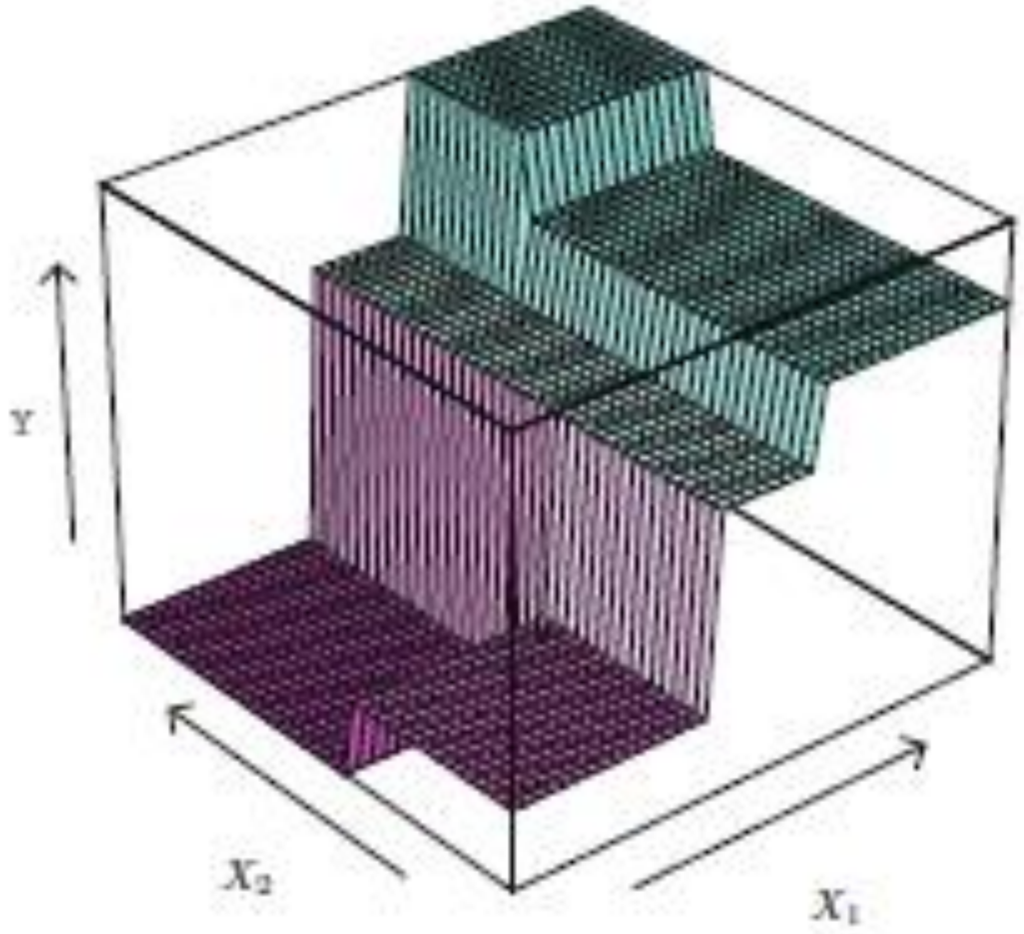
- Another example with five regions



CLASSIFICATION AND REGRESSION TREES
INTRODUCTION AND MAIN CONCEPTOS

- Another example with five regions





Source: James, Witten, Hastie & Tibshirani (2013) An Introduction to Statistical Learning: with applications in R. New York: Springer

INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE

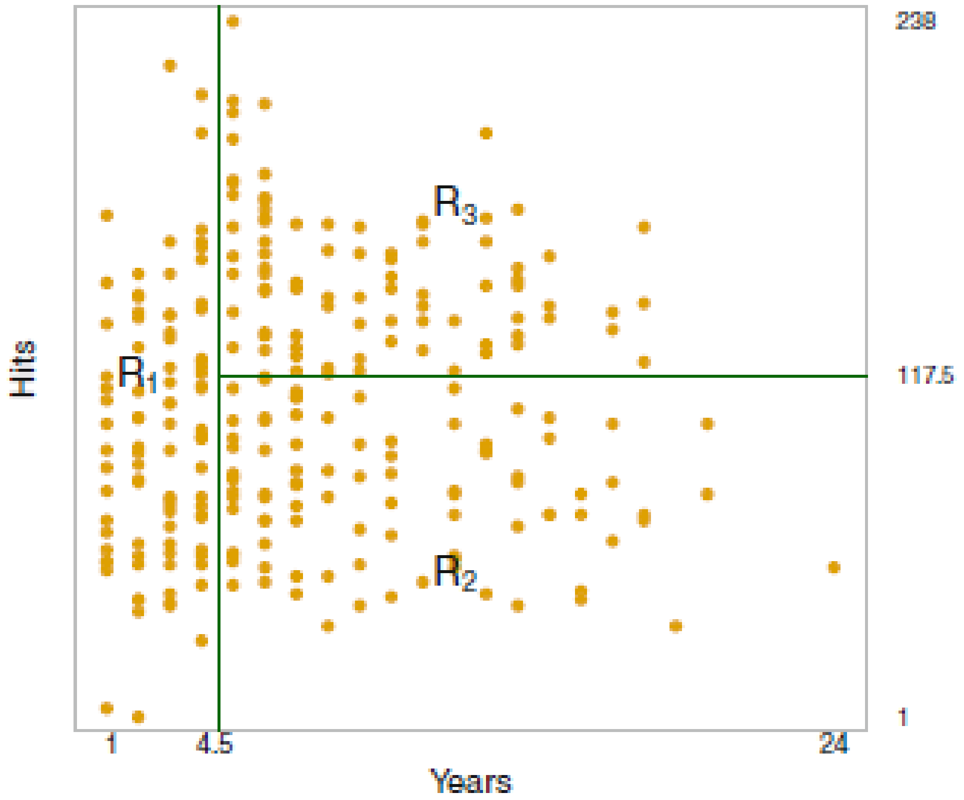
IN4402: Aplicaciones de Probabilidades y Estadística

SPLITTING AND PRUNING TREES

ANDRÉS FERNÁNDEZ

- Let's try to predict a baseball player salary from some characteristics
- In **tree-based algorithms** we split data into regions, and then every region is averaged to predict the outcome variable

Region	Predicted LogSalary	Predicted Salary
R1	5.11	$165,174
R2	6.00	$402,834
R3	6.74	$845,346



- How does the machine know…
  - That **years = 4.5** and **hits = 117.5** are the best splitting points?

  - Goal is to find regions $R_1, \ldots, R_J$, generated by cut-points $x_1, \ldots, x_J$ that minimize residual sum of squares (RSS)

$$\sum_{j=i}^{J} \sum_{i \in R_j} \left(y_i - \hat{y}_{R_j} \right)^2$$

  - We take a *top-down* ***greedy*** approach for *recursive binary splitting*

- How does the machine know… (CLASSIFICATION)
  - **Purity** of nodes (if in a certain group division all observations are yes or no)
  - **Gini** index of impurity of region m for variables k:

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

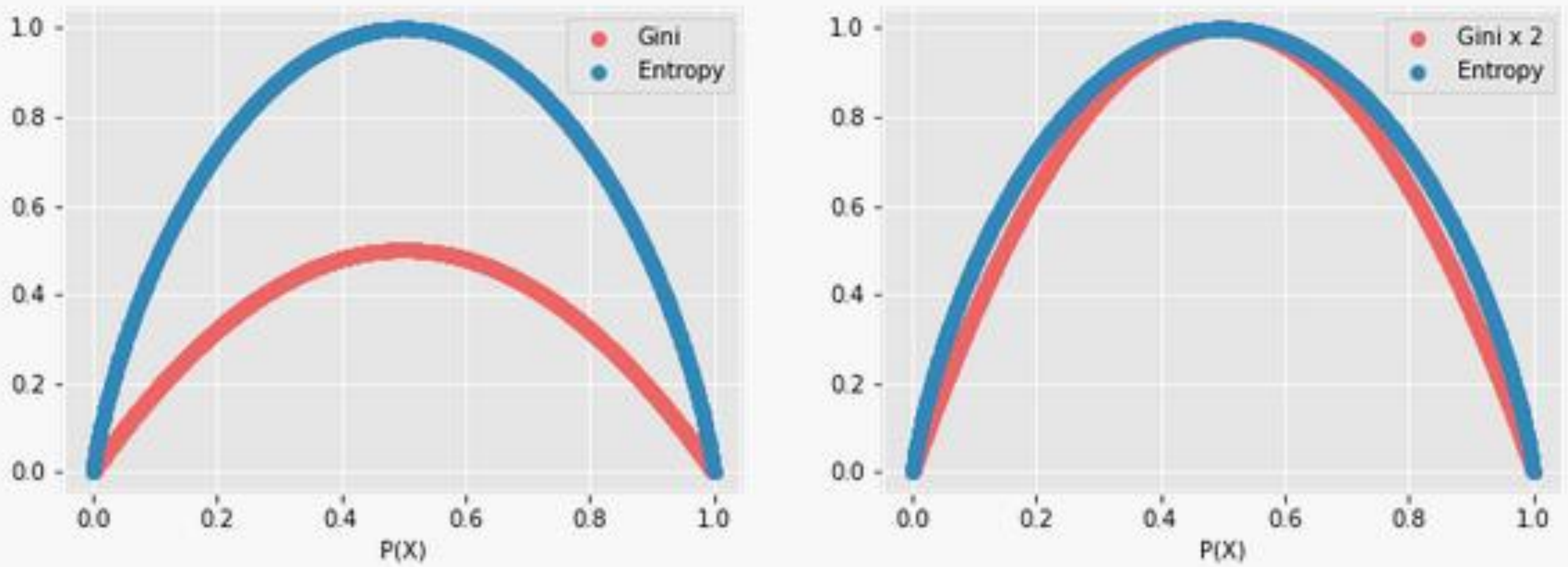
  - The algorithm chooses the variable that provides the lowest impurity
  - If a new node gives higher impurity, then the tree stops
  - "A variable giving 50/50 split in groups does not give information at all."

- How does the machine know… (CLASSIFICATION)
  - **Purity** of nodes (if in a certain group division all observations are yes or no)
  - **Entropy** measure of region m for variables k

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

  - The algorithm chooses the variable that provides the lowest impurity
  - If a new node gives higher impurity, then the tree stops
  - “A variable giving 50/50 split in groups does not give information at all.”

CLASSIFICATION



Gini Index and Entropy

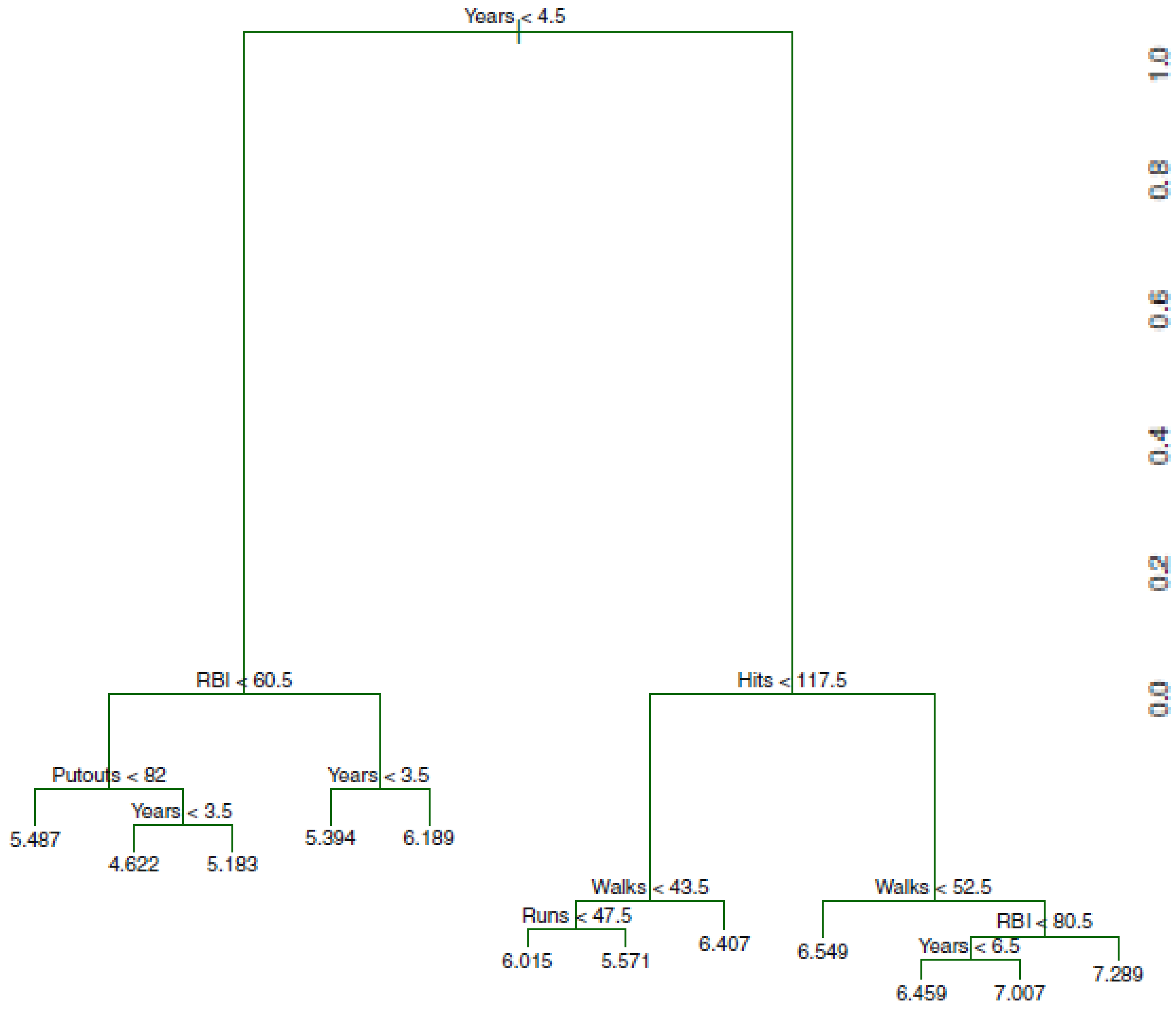
- How does the machine know…
  - When **to stop** splitting?
  - Stopping criterion are defined *a priori*, according to some dimension:
    - That every subregion contains more than five observations

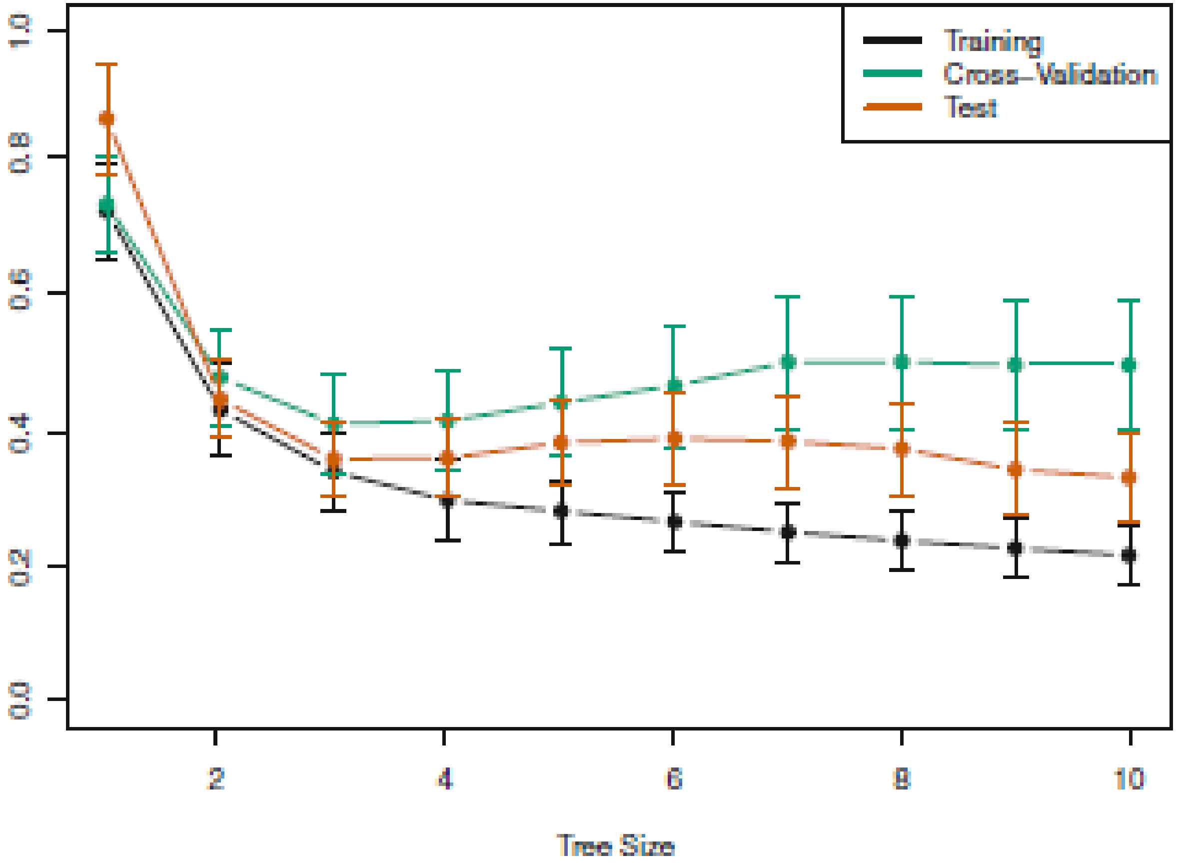
  - If we stop too "along the way" we might overfit the data
  - If we stop too "early" we might underfit the data
    - Stopping criterions or "tree pruning" – that each gain on RSS decrease is above some threshold (ie. The marginal gain for splitting overcomes the loose on overfitting risk)
  - What if we are loosing a "very good deal" after some splitting?

- Tree Pruning strategies
  - We could elaborate very large trees and **prune it back** to obtain adequate *subtree*
    - Evaluation of every subtree might be too much
  - **Cost complexity pruning (weakest link pruning)**
    - Let's use a *tuning* parameter. Subtree T has $|T|$ number of terminal nodes (leaves)

$$\sum_{m=1}^{|T|} \sum_{i:x_i \in R_m} \left(y_i - \hat{y}_{R_m}\right)^2 + \alpha |T| \quad \text{con } \alpha \geq 0$$

CLASSIFICATION AND REGRESSION TREES
INTRODUCTION AND MAIN CONCEPTOS





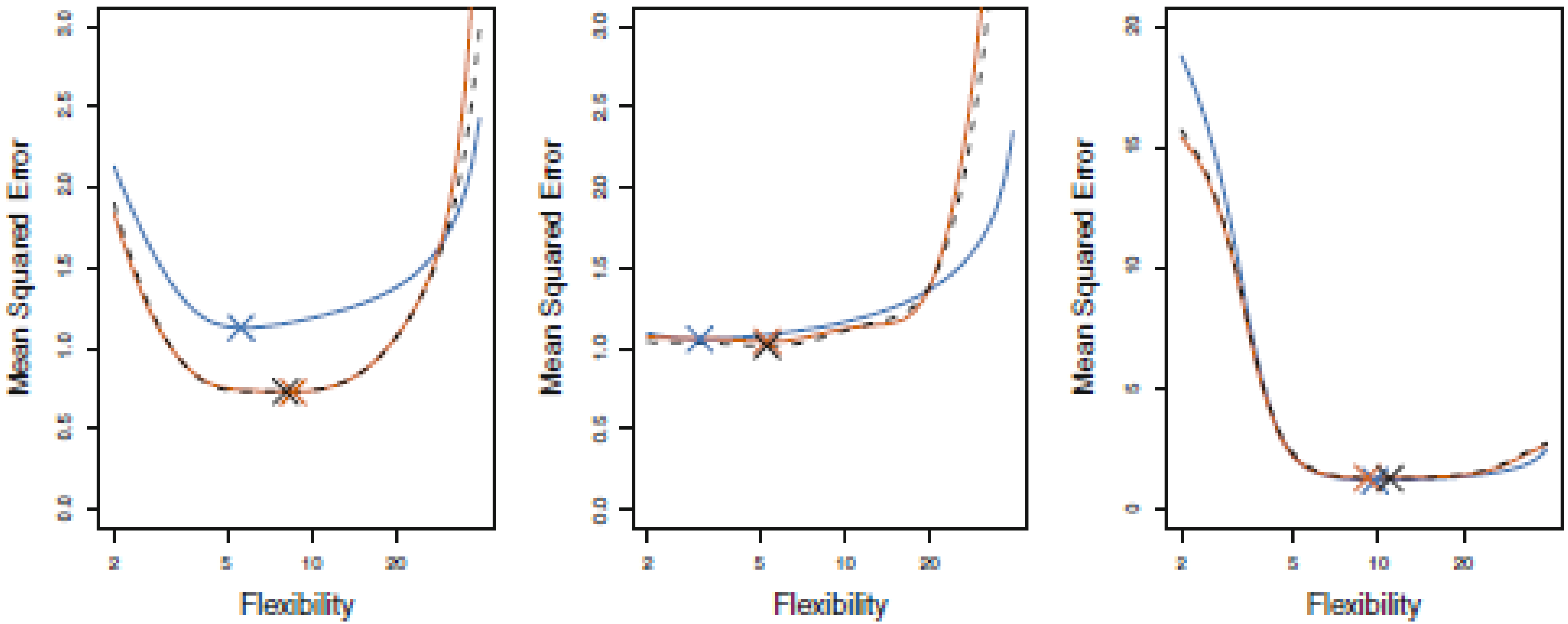
Source: James, Witten, Hastie & Tibshirani (2013) An Introduction to Statistical Learning: with applications in R. New York: Springer

- **K-fold cross-validation (or k-fold CV)**
  - It approximates the *test MSE*.
  - Divide the **training** sample in K groups of similar size
  - We use **all but Kth** sample to train and the K sample to evaluate
  - We average the results

Commonly used K= 5 or K= 10

CLASSIFICATION AND REGRESSION TREES
INTRODUCTION AND MAIN CONCEPTOS

K-fold CV is an approximation to test MSE. But it might be more useful for pointing out some parameters



K-fold is better tan leaving one point out of sample (overfitting)

- **Algorithm**
- 1. Use recursive binary splitting to grow a large tree on the training data, stopping only when each terminal node has fewer than some minimum number of observations.
- 2. Apply *cost complexity pruning* to the large tree in order to obtain a sequence of best subtrees, as a function of α.
- 3. Use K-fold cross-validation to choose α and number of leaves.
- 4. Return the subtree from Step 2 that corresponds to the chosen value of α.

INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE

IN4402: Aplicaciones de Probabilidades y Estadística

COMPARING TREES VS LINEAR MODELS

ANDRÉS FERNÁNDEZ

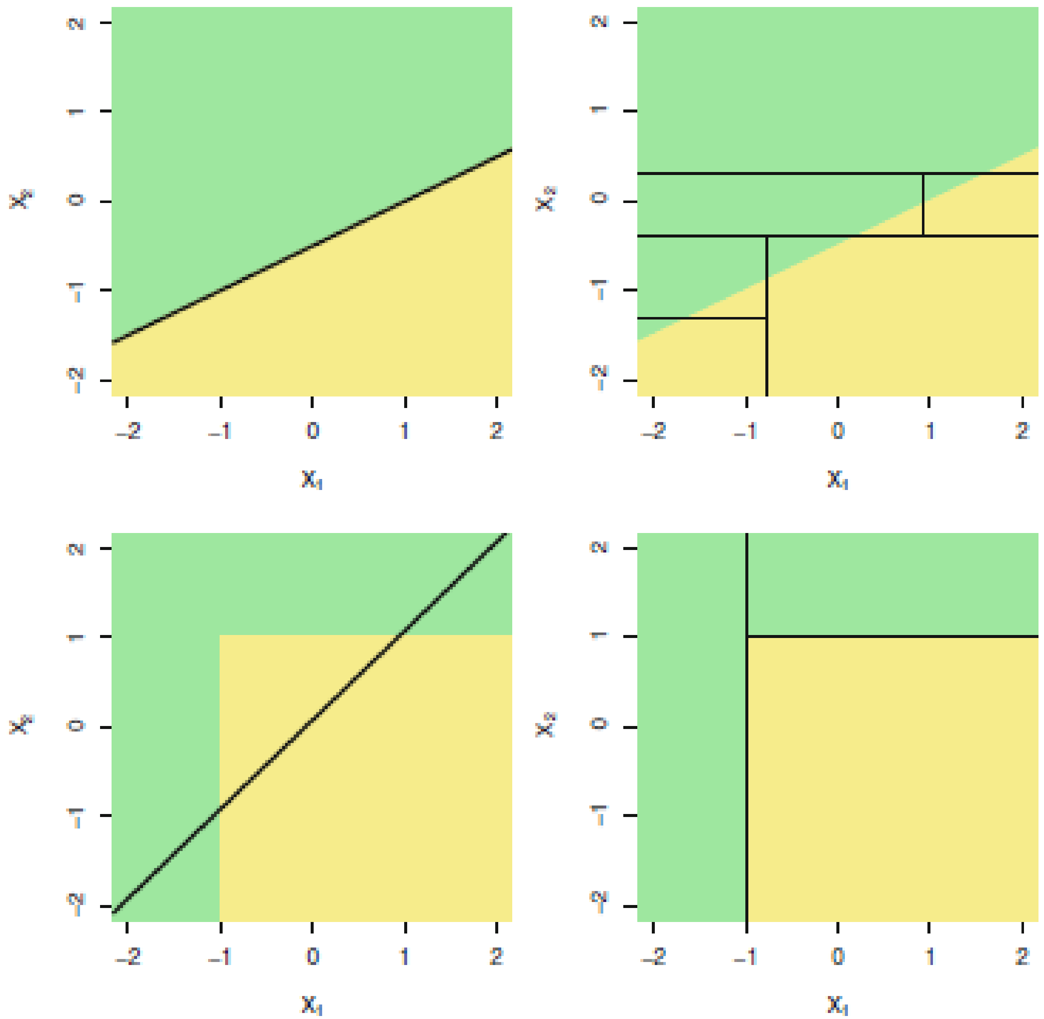
- **Comparing**
- **Linear**

$$f(X) = \beta_0 + \sum_{k=1}^{p} X_k \beta_k$$

- **Trees**

$$f(X) = \sum_{m=1}^{p} c_m \cdot 1_{(X \in R_m)}$$

- **Comparing**



- **Advantages of Trees**
  - Easy to interpret and explain
  - Might be closer to human decisión-making approach than other methods
  - Graphical explanation
  - Easy to use on qualitative predictions (binary outcomes)

- **Disadvantages**
  - Lower level of predictive accuracy than other ML methods
  - Non-robust: small changes in data can cause large change in final estimations

- Another advantage: they can be aggregated between them to improve performance: we say hello to **random forests**