Introduction to Tree-Based Methods

Predictive Modeling & Statistical Learning

Gaston Sanchez

Introduction

Tree-based methods

Aliases

- Decision Trees.
- Segmentation Trees.
- Partition Trees.

Tree-based methods

Aliases

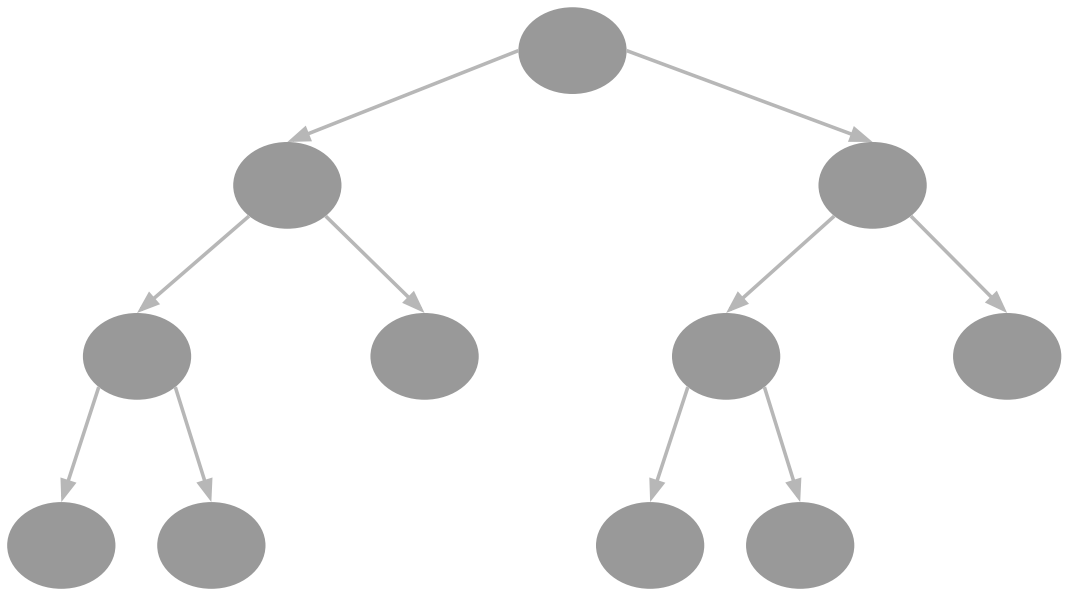
- Decision Trees.
- Segmentation Trees.
- Partition Trees.

Depending on the response variable Y

- Regression trees for quantitative Y
- Classification trees for categorical Y

Why trees?

Because their output can be visualized with a tree-like structure (inverted tree with root at the top, and descending branches)



Attractiveness

Tree-based Methods are probably the most popular supervised learning technique

Advantages

- Able to handle both types of responses
  - quantitative
  - categorical
- Able to handle any type of predictors:
  - continuous
  - ordinal
  - nominal
  - binary
- Even able to handle missing data

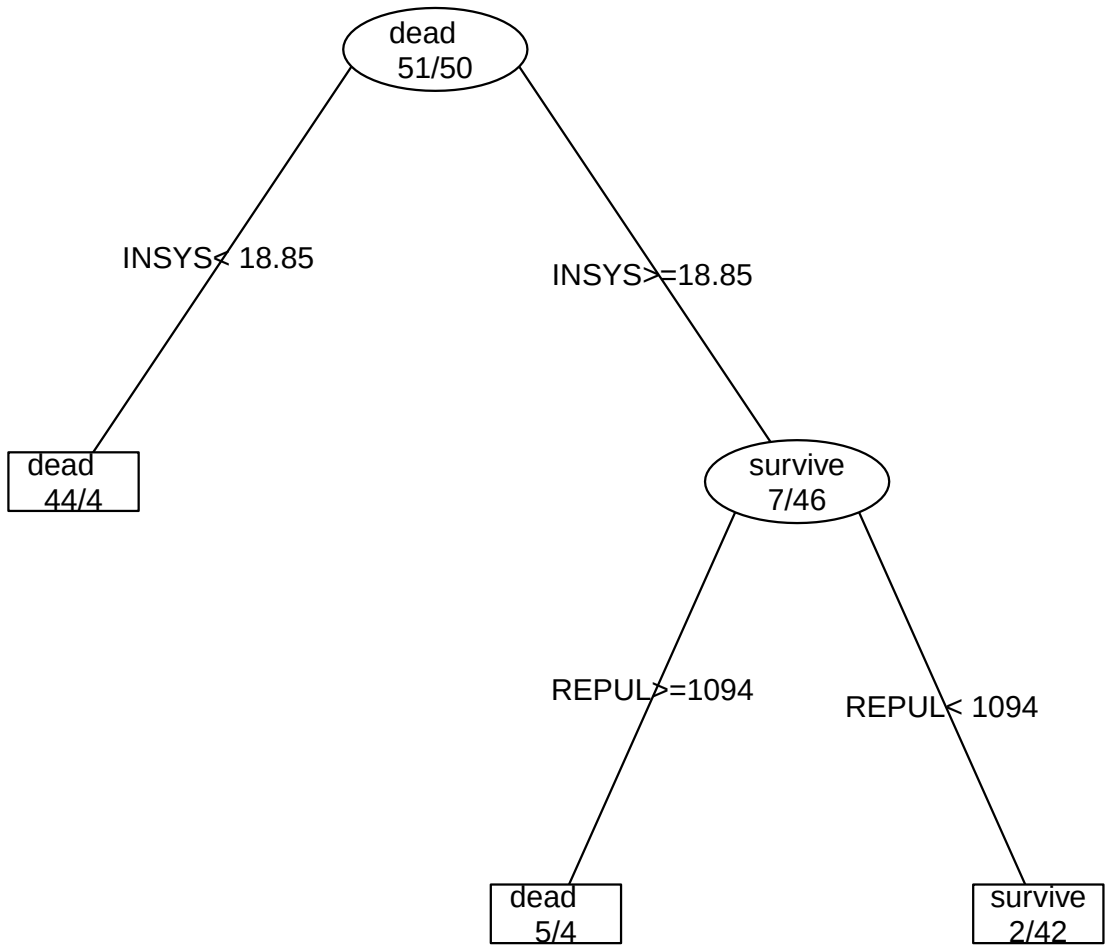
"Sexiness" of trees

User Experience (UX) attractiveness

- Eye-catching and impactful visual display
- The tree diagram can be read as a flowchart
- Easily interpretable
- Mimic the way (some) decisions are made
- Most users easily understand the tree and its decisions

You don't have to know much about statistics/machine-learning to read and understand the output.

Classification Tree for `infarctus` data



"Sexiness" of trees

Compare with logistic regression output:

```
Call:
glm(formula = PRONO ~ ., family = binomial, data = infarctus)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.55511  -0.40687  -0.01749   0.39964   2.69534

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.338e+00  9.549e+00   0.140    0.889
FRCAR       -4.745e-02  8.989e-02  -0.528    0.598
INCAR        5.782e+00  5.319e+00   1.087    0.277
INSYS       -1.102e-01  3.935e-01  -0.280    0.780
PRDIA       -3.904e-02  1.950e-01  -0.200    0.841
PAPUL       -1.511e-01  2.331e-01  -0.648    0.517
PVENT       -5.421e-02  7.891e-02  -0.687    0.492
REPUL        1.081e-05  3.951e-03   0.003    0.998

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 140.006  on 100  degrees of freedom
Residual deviance:  58.642  on  93  degrees of freedom
AIC: 74.642

Number of Fisher Scoring iterations: 6
```

About decision trees

Building a tree involves successively dividing the set of objects with the help of predictors $X_1, X_2, \ldots, X_p$, with the goal of obtaining final segments as much homogeneous as possible with respect to the response variable Y.

The successive divisions (or splits) of the set of objects is based on the principle of **recursive partitioning**.

About decision trees

"We start by choosing the variable which, by its categories, provides the best separation of the objects in each class, thus creating subsets, called nodes, each containing the largest possible proportion of objects in a single class.

The same operation is then repeated on each new node obtained, until no further separation of the objects is posible or desirable.

The construction is such that each of the terminal nodes (the leaves) mainly consists of the individuals of a single class. The set of rules for all the leaves forms the classification model." (Tuffery, 2011, p. 313)

About decision trees

- Non-parametric methods (no specification of model parameters)
- Almost no assumptions about data distributions
- Algorithmic principle: **recursive partitioning**
- Copes well with non-linear effects (e.g. interactions)

Overall Comparison

	Model assumptions	Estimated probabilities	Interpretable	Flexible
LDA	yes	yes	yes	no
QDA	yes	yes	yes	a bit
LogReg	yes	yes	yes	no
k-NN	no	no*	no	yes
Trees	no	yes	yes	somewhat

* can be derived from voting-frequency scheme

A bit of history

About decision trees

- Originally developed in the 1960s
- First neglected by statisticians
- Rethought independently by computer scientists and statisticians in the 1980s
- Regain interest with works led by Leo Breiman, as well as Ross Quinlan
- Quickly gained popularity in industry (e.g. marketing, health, finance)
- Further expanded during the 1990s and 2000s

Morgan and Sonquist's AID

- **AID** (Automatic Interaction Detection) by James Morgan and John Sonquist (1963 & 1964).
- Social scientists at the *Institute for Social Research* (University of Michigan).
- Paper published in the *Journal of the American Statistical Association* in 1963.
- Cited similar ideas of William Belson (1959) for a segmentation approach in Marketing.

Morgan and Sonquist's AID

- First regression tree (Y quantitative).
- Interest: looking for interactions through binary partitions.
- Suffered criticism from stats community.
- Improved over the 70s (renamed SEARCH).
- SEARCH was in service from about 1971 to 1996.

Morgan, Messenger and Mandell's THAID

- **THAID** (THeta Automatic Interaction Detection) by Morgan, Messenger and Mandell (1972).
- Extends AID ideas to classification (Y categorical).
- AID and THAID did not initially attract much interest in the stats community.
- Initial criticisms due to overfitting.
- Idea of *Recursive Partitioning* did gain attention in computer science.

Kass' CHAID

- **CHAID** (Chi-Squared Automatic Interaction Detection) by Gordon Kass (1980).
- Motivated by bad reputation of AID.
- Kass wanted to render AID safe.
- Classification tree for Y nominal.
- Splitting criterion based on chi-square test.

Breiman et al's CART

- **CART** (Classification And Regression Trees)
- Developed throughout 1974-1984 by statisticians
  - Leo Breiman (Berkeley)
  - Jerome Friedman (Stanford)
  - Charles Stone (Berkeley)
  - Richard Olshen (Stanford)
- CART book published in 1984.
- Commercialized software by Salford Systems.

Breiman et al's CART

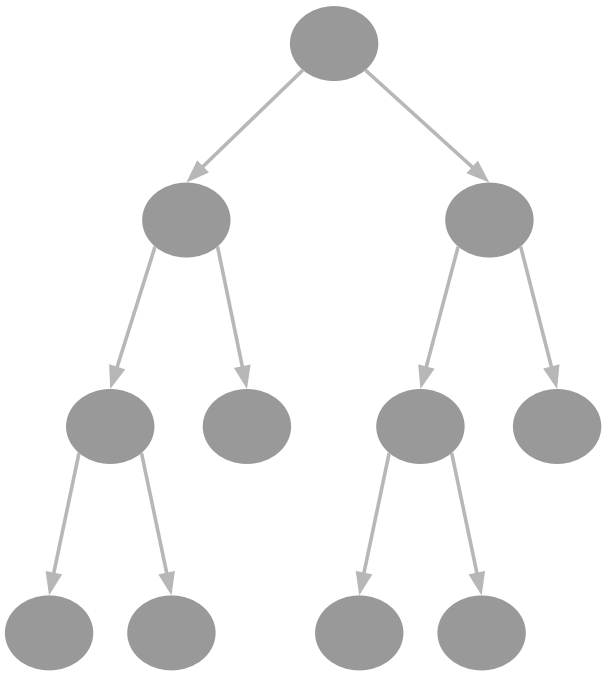
- Breiman and Friedman independently "reinvented the wheel" (1973)
- Began to use tree methods for classification.
- Stone and Olshen joined forces later.
- Olshen was an early user in medical applications.
- 1980: Idea about writing a book together.

Ross Quinlan's algorithms

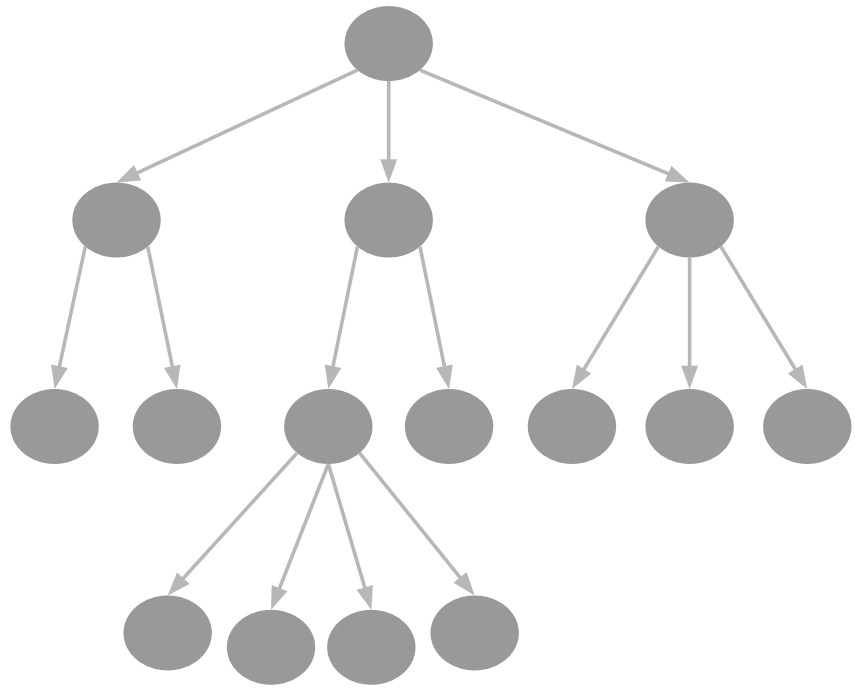
- ID3 (1986)
- M5
- C4.5 (1993) extends ID3, using entropy-based measure of node impurity.
- Solved problems of over and under fitting.
- Trees for both quantitative and categorical responses.
- Obtained conditions for all recursive partitioning techniques to be Bayes risk consistent.
- Commercialized software.

Vocabulary

Types of trees

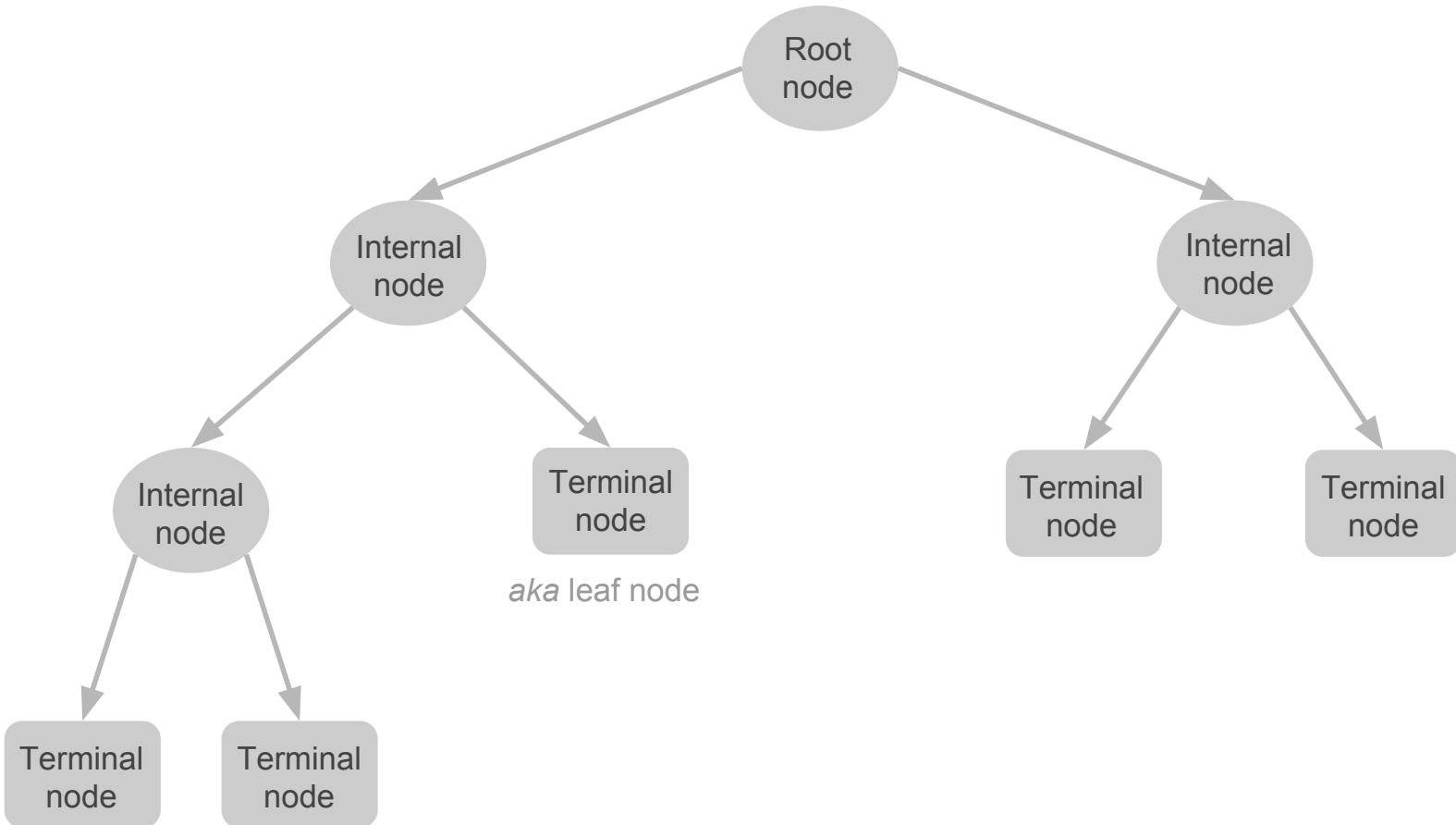


Binary tree
(2-way splits)

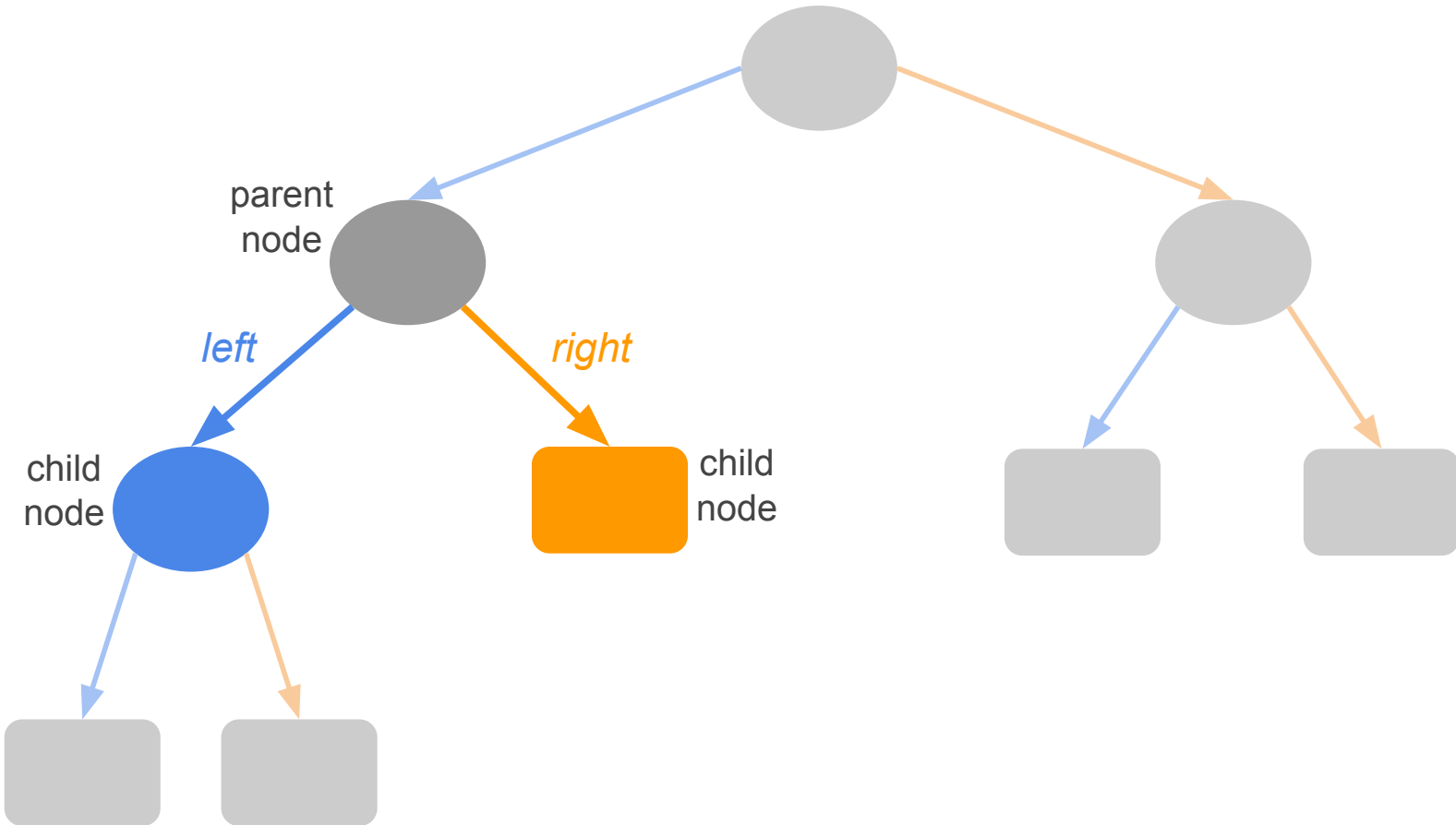


N-ary tree
(n-way splits)

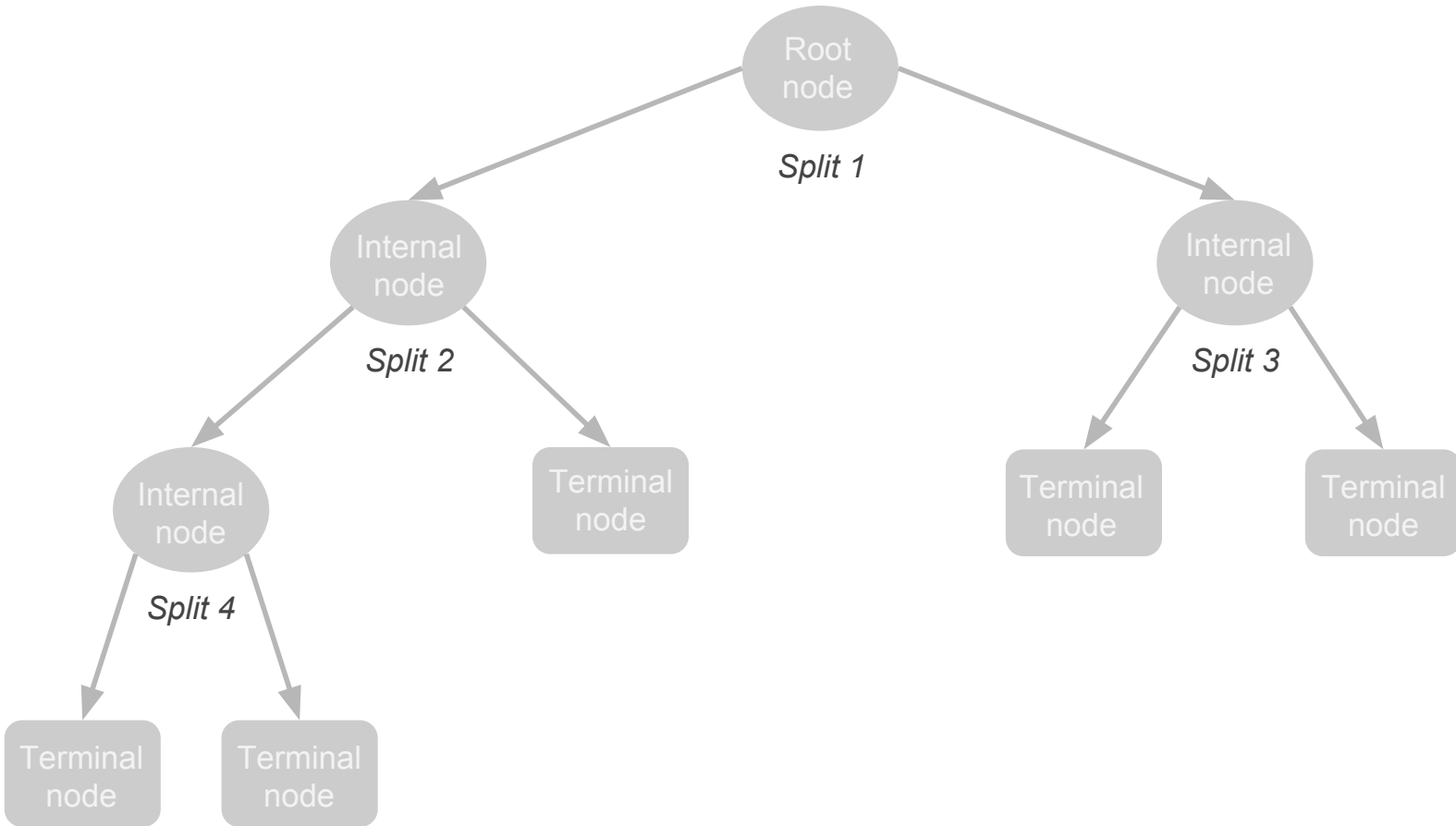
Some Vocabulary



Some Vocabulary



Some Vocabulary

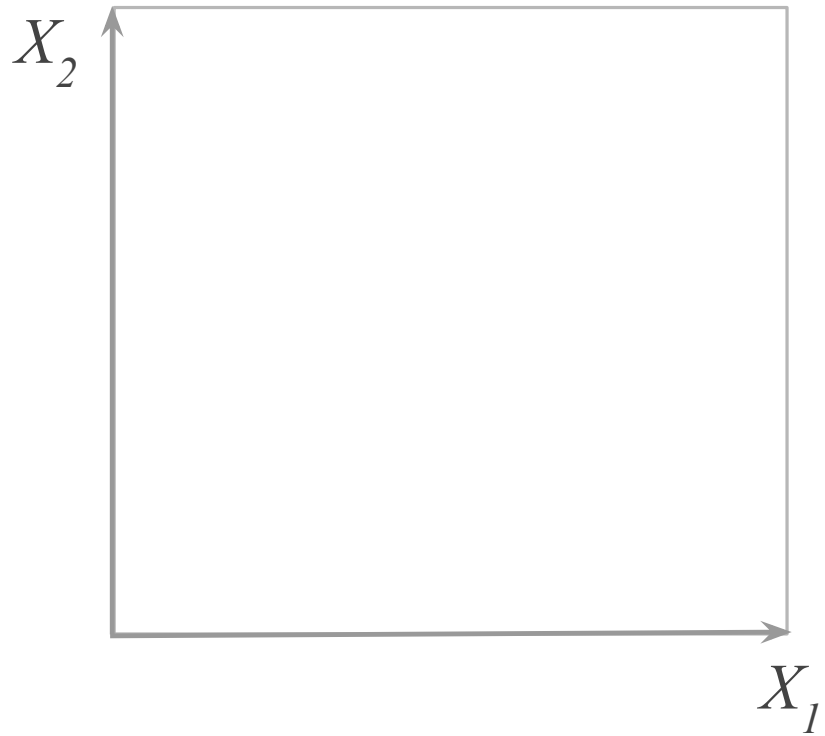


Regression Trees
Appetizer

Splitting the predictors

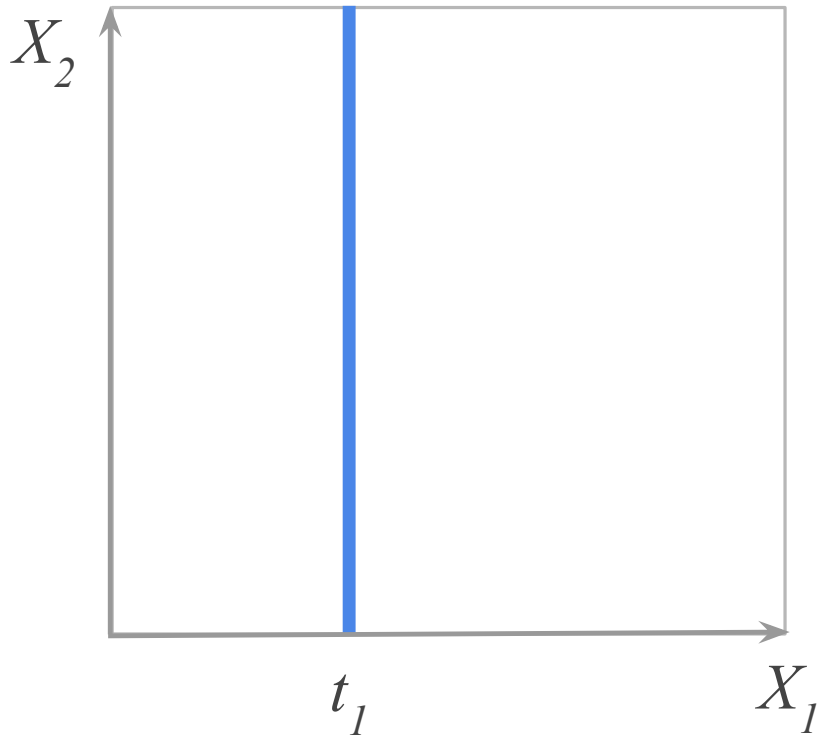
Let's assume we have two quantitative predictors.

Generally we create the partitions by iteratively splitting one of the predictors into two regions.



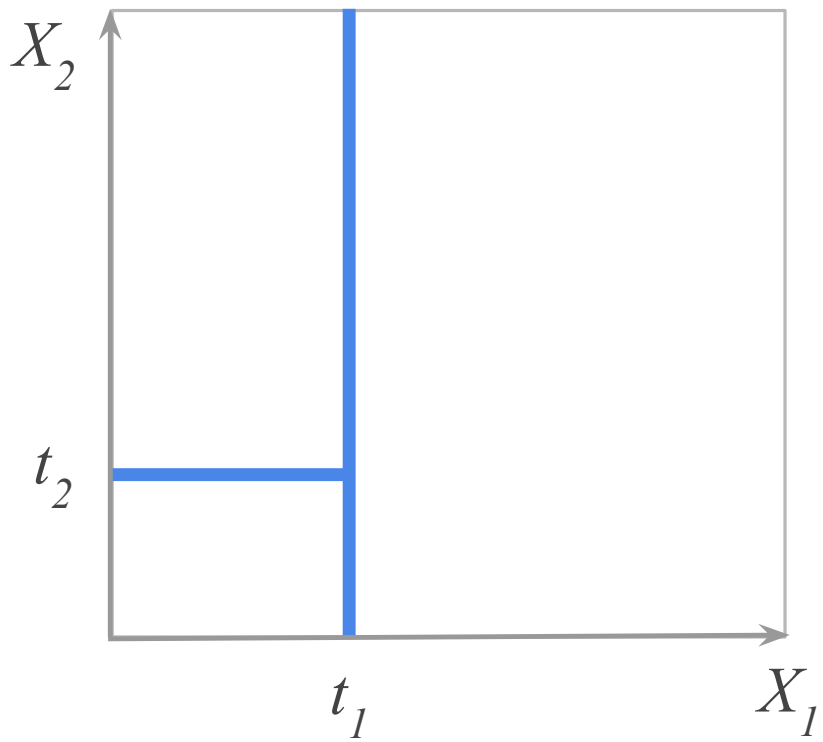
Splitting the predictors

- First split on $X_1 = t_1$



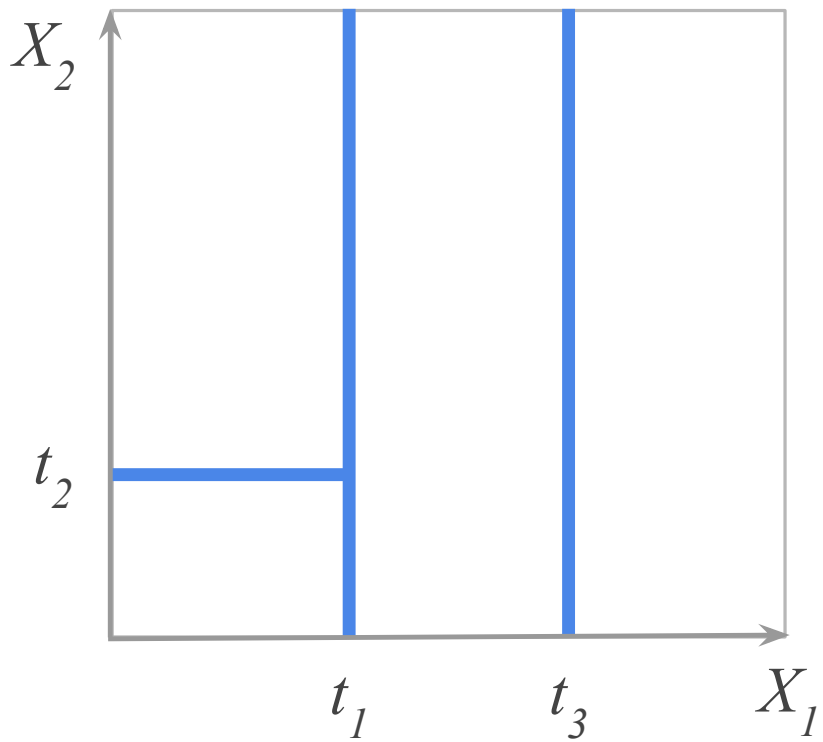
Splitting the predictors

- First split on $X_1 = t_1$
- If $X_1 \leq t_1$, split on $X_2 = t_2$



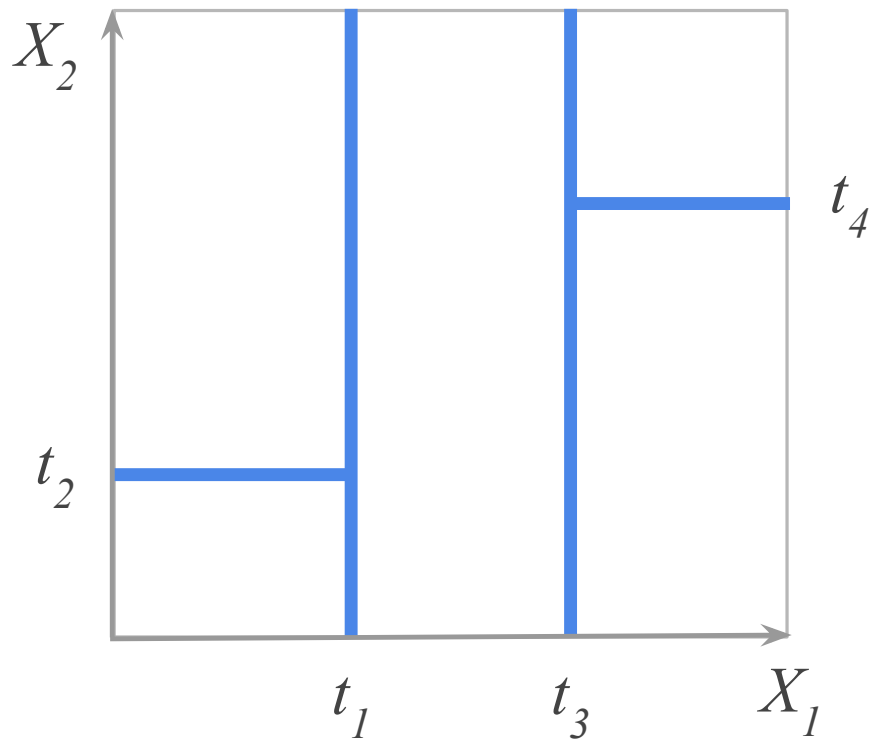
Splitting the predictors

- First split on $X_1 = t_1$
- If $X_1 \leq t_1$, split on $X_2 = t_2$
- If $X_1 > t_1$, split on $X_1 = t_3$



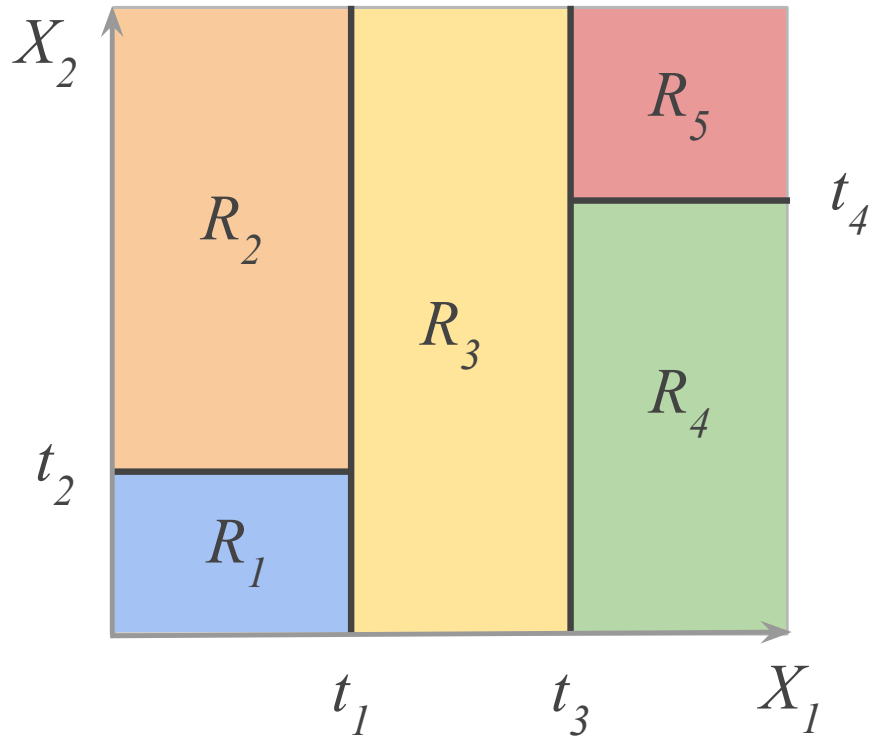
Splitting the predictors

- First split on $X_1 = t_1$
- If $X_1 \leq t_1$, split on $X_2 = t_2$
- If $X_1 > t_1$, split on $X_1 = t_3$
- If $X_1 > t_3$, split on $X_2 = t_4$

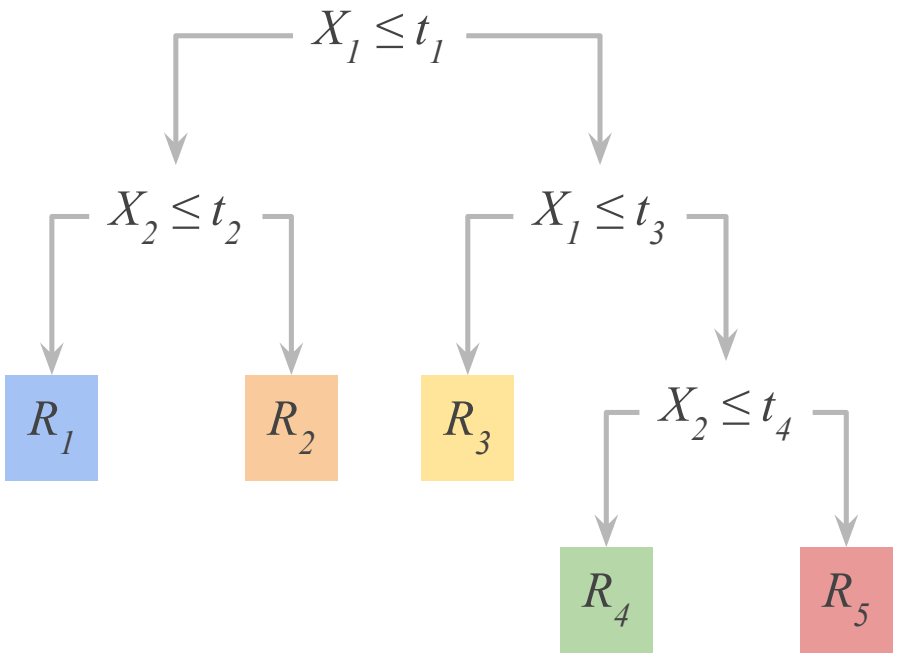


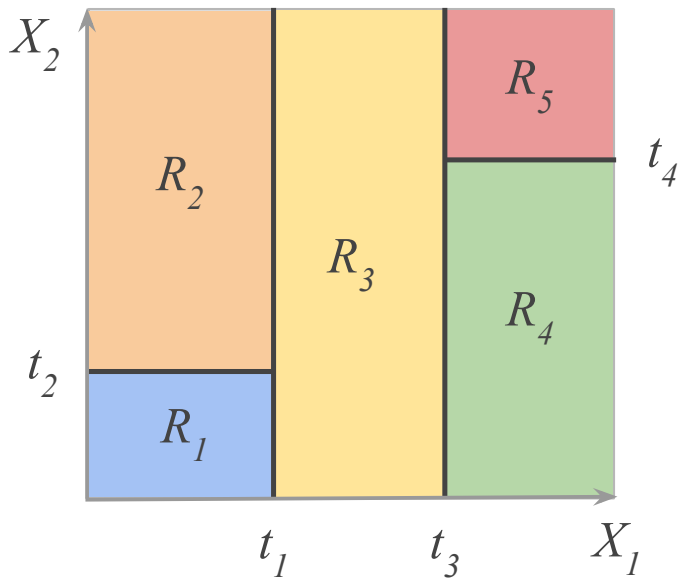
Splitting the predictors

- First split on $X_1 = t_1$
- If $X_1 \leq t_1$, split on $X_2 = t_2$
- If $X_1 > t_1$, split on $X_1 = t_3$
- If $X_1 > t_3$, split on $X_2 = t_4$



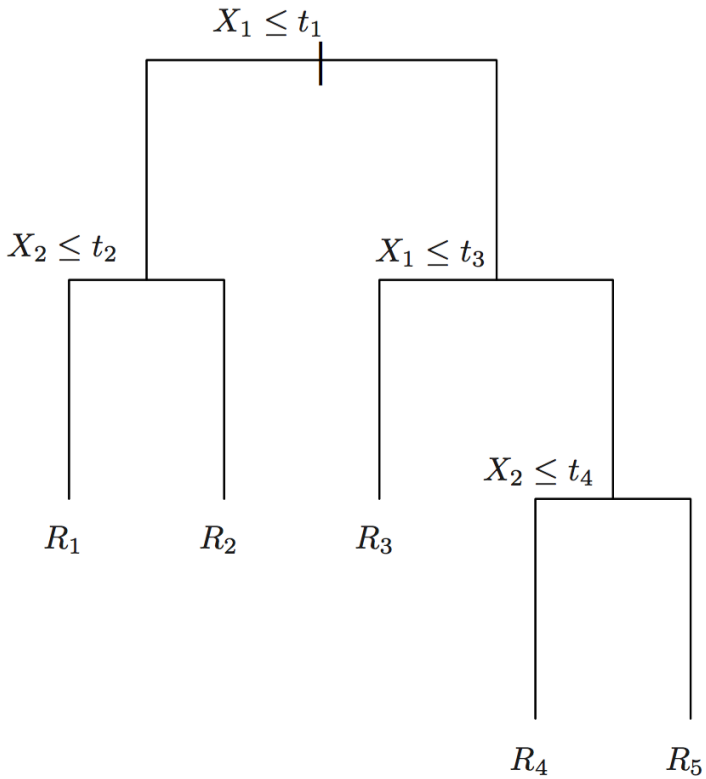
Splitting the predictors and tree display

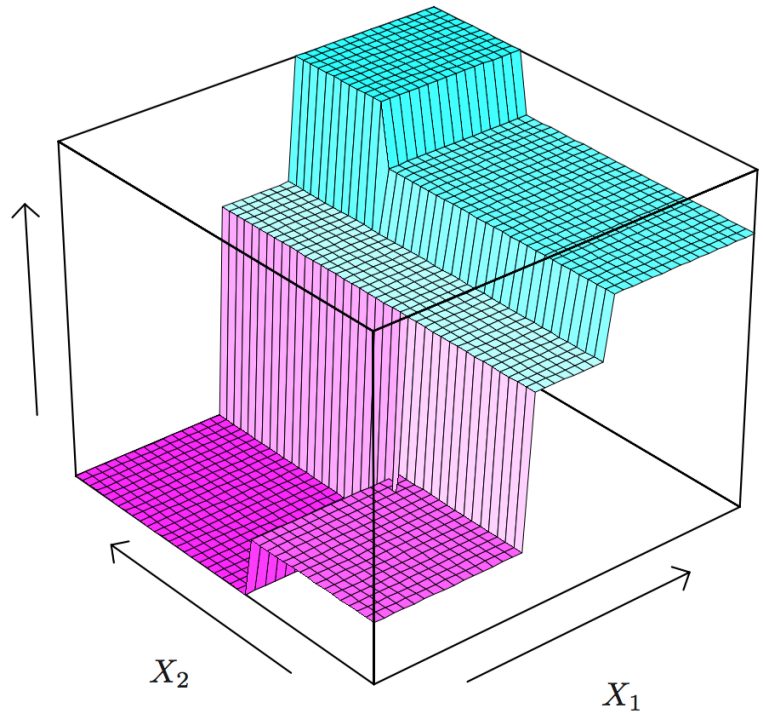




When we create partitions in this way, we can represent them in a tree structure. This provides a very simple way to explain the model to a non-expert (i.e. your boss or client)

Partitions in 3-dim space and associated tree



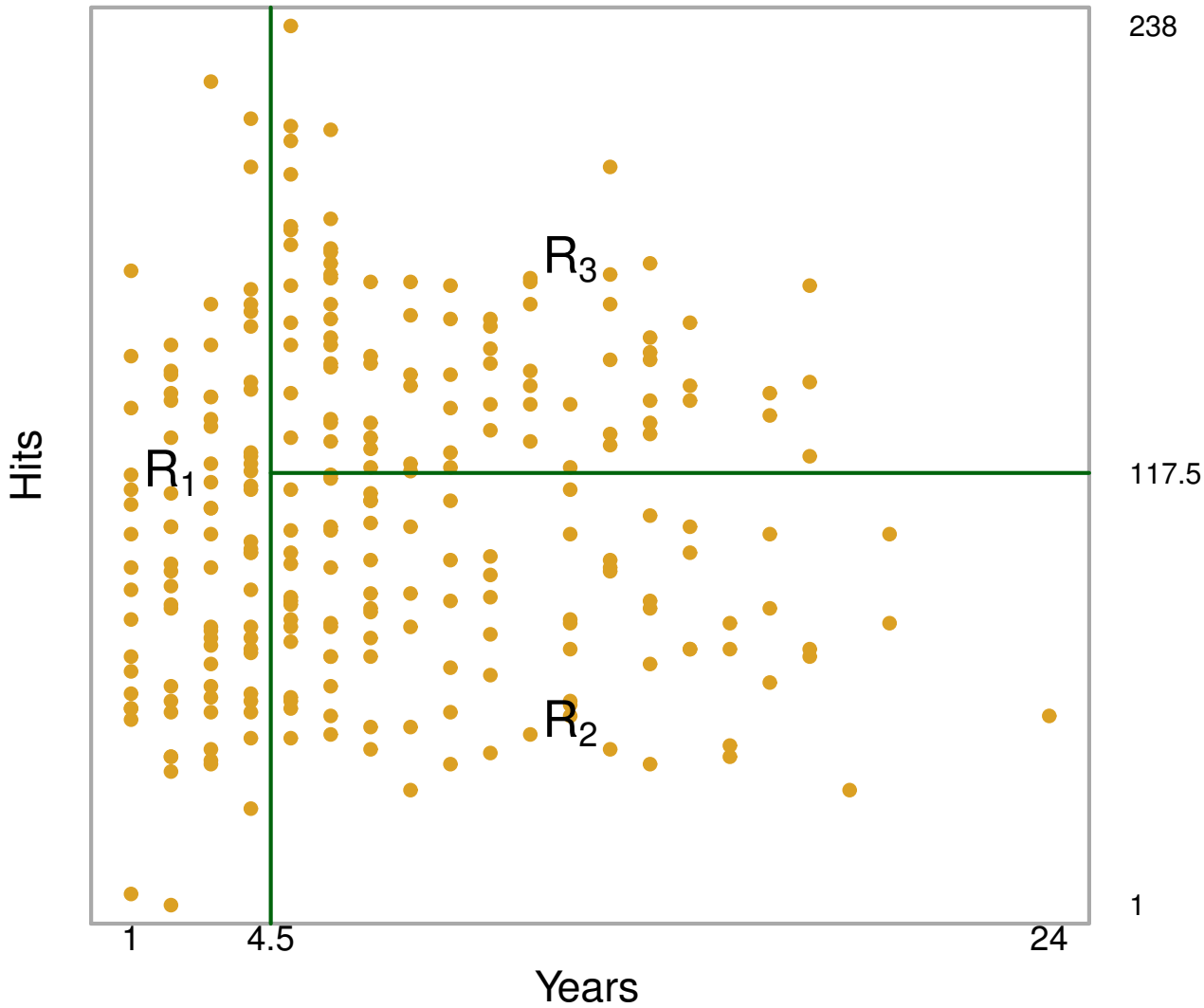


Example: Baseball Players' Salaries

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 it's log-transformed
- The predicted Salary for a player who played more than 4.5 years and had less than 117.5 hits last year is $\$1000 \times e^{6.00} = \$402,834$

Example: Baseball Players' Salaries



Some Natural Questions

- Where to split?
- How do we decide the adequate regions?
- How many regions?
- What values should we use for $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$?

We'll answer these questions in the next slides

Classification Trees
Appetizer

Classification Trees

- A classification tree is very similar to a regression tree
- The main difference is that we now have a categorical predictor
- The tree is grown on a similar way as with a regression tree
- There are various criteria to decide how to split the space of predictors

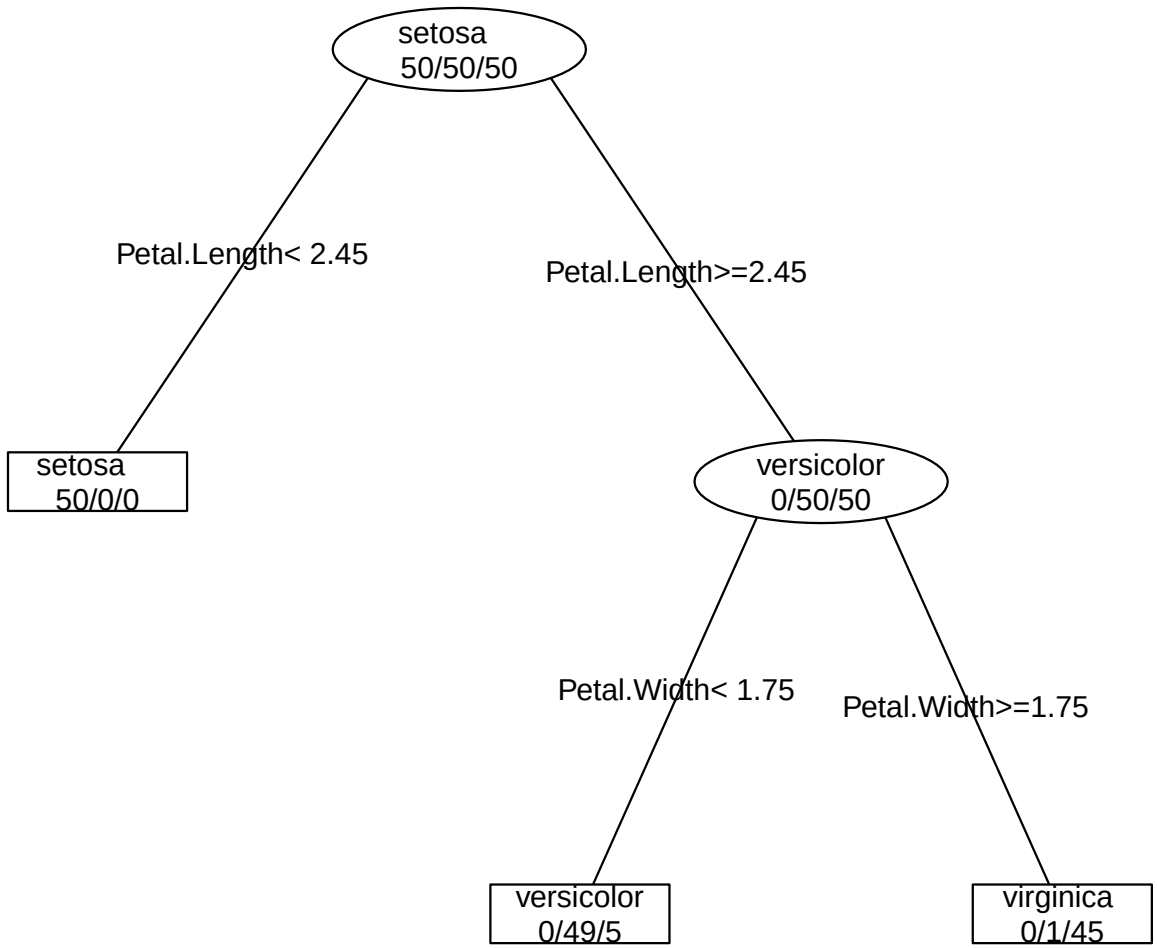
Iris Data Example

Tree for `iris` data

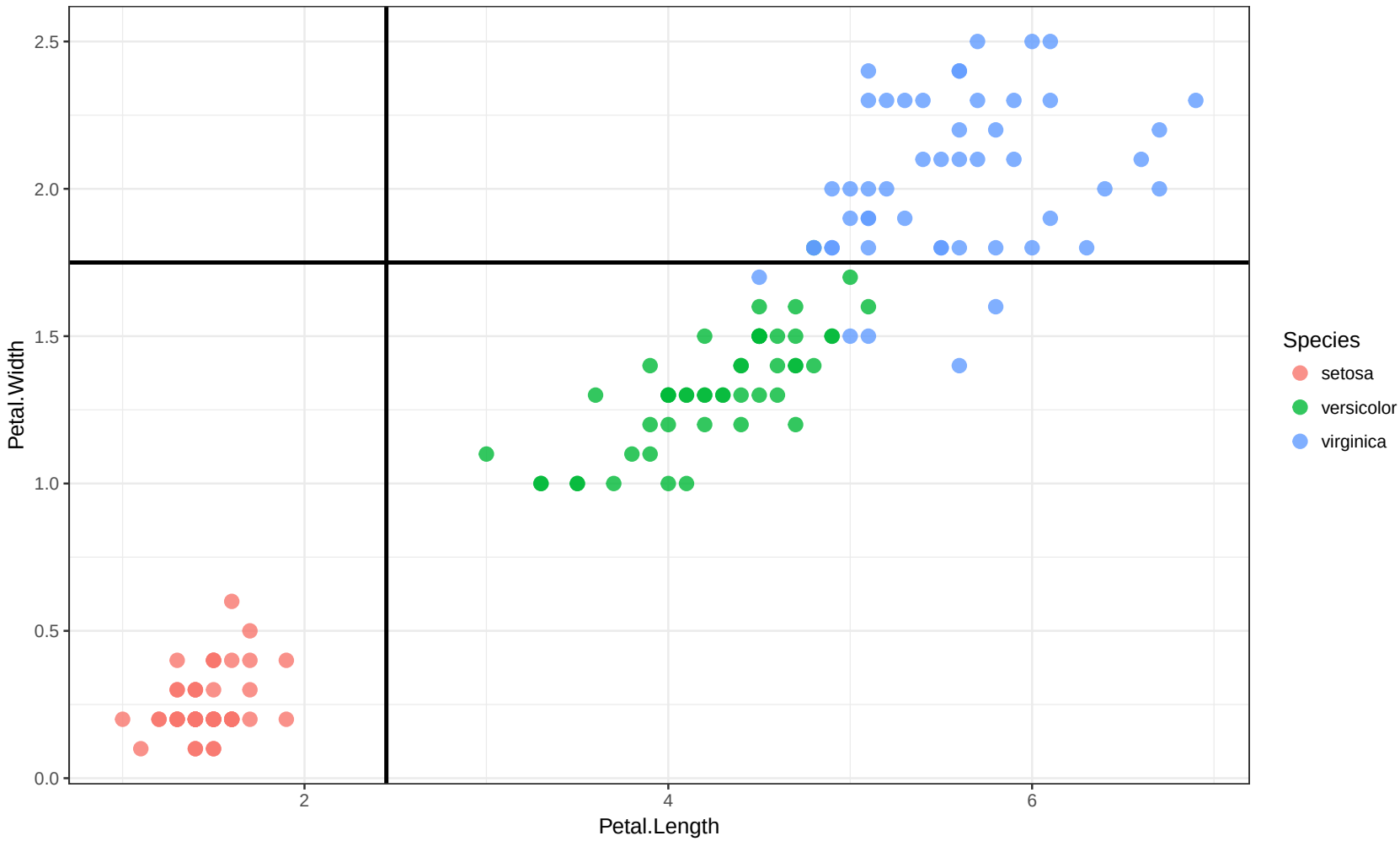
```
# fit tree
iris_tree <- rpart(Species ~ ., data = iris)

iris_tree
```

```
## n= 150
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
## 1) root 150 100 setosa (0.33333333 0.33333333 0.33333333)
##   2) Petal.Length< 2.45 50   0 setosa (1.00000000 0.00000000 0.00000000) *
##   3) Petal.Length>=2.45 100  50 versicolor (0.00000000 0.50000000 0.50000000)
##     6) Petal.Width< 1.75 54   5 versicolor (0.00000000 0.90740741 0.09259259) *
##     7) Petal.Width>=1.75 46   1 virginica (0.00000000 0.02173913 0.97826087) *
```

Tree for `iris` data



Space partitions for `iris` data



References

- **Fifty Years of Classification and Regression Trees** by Wei-Yin Loh (2014). *International Statistical Review, 82, 3, 329-348*.
- **History and Potential of Binary Segmentation for Exploratory Data Analysis** by James Morgan (2005). *Journal of Data Science*, 3, 123-136.
- **Classification and Regression Trees** by Breiman et al (1984).
- **Data Mining and Statistics for Decision Making** by Stephane Tuffery (2011). *Chapter 11: Classification and prediction methods*.
- **Fundamentals of Machine Learning for Predictive Data Analytics** by Kelleher et al (2015). *Chapter 4: Information-based Learning*.

References (French literature)

- **Analyse discriminante** by Mireille Bardos (2001). *Chapter 4: Arbres de partitionnement*. Dunod, Paris.
- **Statistique Exploratoire Multidimensionnelle** by Lebart et al (2004). *Chapter 3, section 3.5: Segmentation*. Dunod, Paris.
- **Statistique explicative appliquee** by Nakache and Confais (2003). *Chapter 8: Methode de segmentation CART*. Editions Technip, Paris.
- **Statistique: Methodes pour decrire, expliquer et prevoir** by Michel Tenenhaus (2008). *Chapter 13: Methodes de segmentation*. Editions Technip, Paris.