Classification and Regression Tree (CART) and Random Forest

Juan Camilo Rivera. [1] Hugo Andres Dorado. [1]

[1]Big data and site-specific agriculture
Decision and Policy Analysis
Centro Internacional de Agricultura Tropical

Data analysis course, 2018

Outline

Overview of CART

Definition CART

The Classification And Regression Tree is a nonparametric technique that can select those variables and their interactions that are most important in determining an outcome or dependent variable.

- Ship structures from their radar - range
- Heart failure
- Distressed firm
- Technical aspects in crops

Overview of CART

Group 1	Group 2
AID	Discriminant analysis
THAID	Kernel density estimation
CHAID	K^{th} nearest neighbor
	Logistic regression
	Probit models

Table: Methods of classification

PRO

- CART makes no distributional assumptions.
- Mixture of categorical and continuous
- No affects missing values neither outliers
- Large data sets

CON

- No interval confidence for classify a new data set.

Basic Principles of CART Methodology

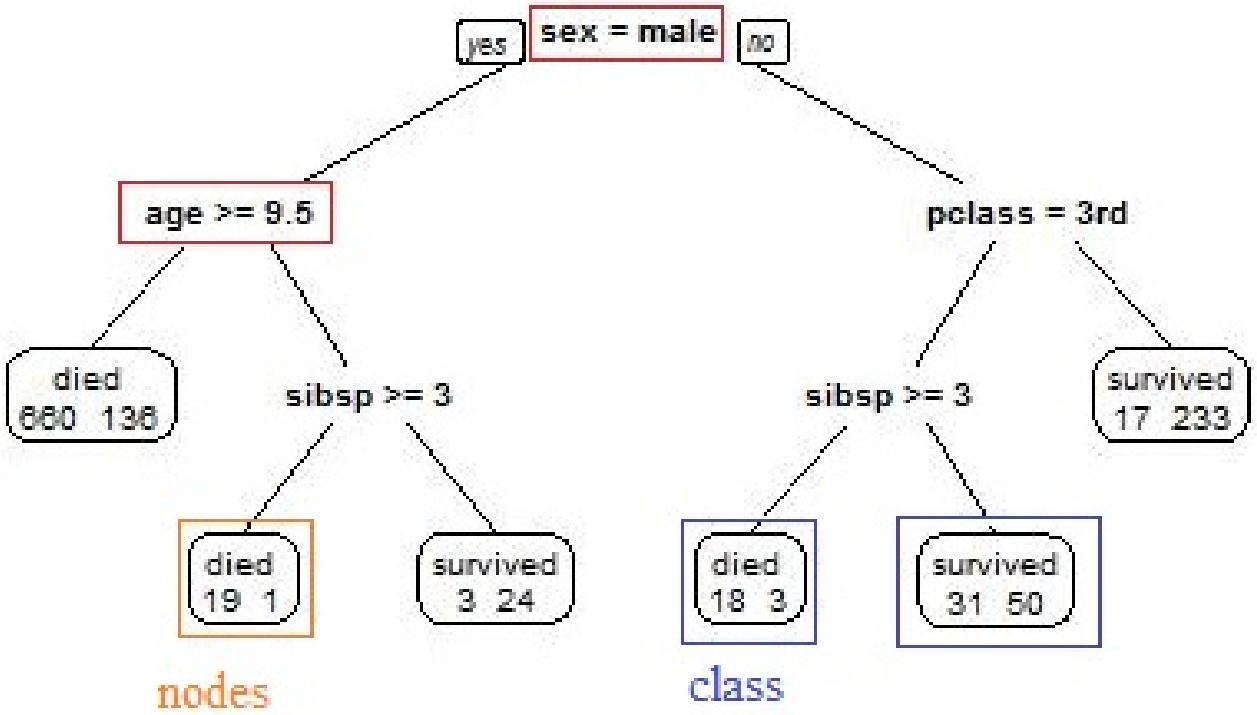


Figure: Titanic data base. sibsp = number of spouses aboard, pclass = passenger class.

Basic Principles of CART Methodology

Components for building a classification tree:

- A set of questions upon which to base a split. Type of questions:
  1. $X \geq d$
  2. $X = d$
- Splitting rules
  1. Gini criterion

$$i(t) = 1 - S \tag{1}$$

with *impurity function*.

$$S = \sum p^2(t|j) \quad for \quad j = 1, 2, ...k \tag{2}$$

where a fixed node t

- goodness-of-split criteria

$$\Delta i(s, t) = i(t) - p_l[i(t_l)] - p_r[i(t_r)] \tag{3}$$

Case of study

The complexity of a tree is measured by the number of its terminal nodes.
Example: **Consumer Report Auto Data**
The cu.summary data base is composed by:

- `Reliability`: an ordered factor (contains NAs). Much worse, worse, average, better, much better.
- `Price`: price in dollars
- `Country`: country where car manufactured
- `Mileage`: gas mileage in miles contains
- `Type`: Small, Sporty, Compact, Medium, Large, Van

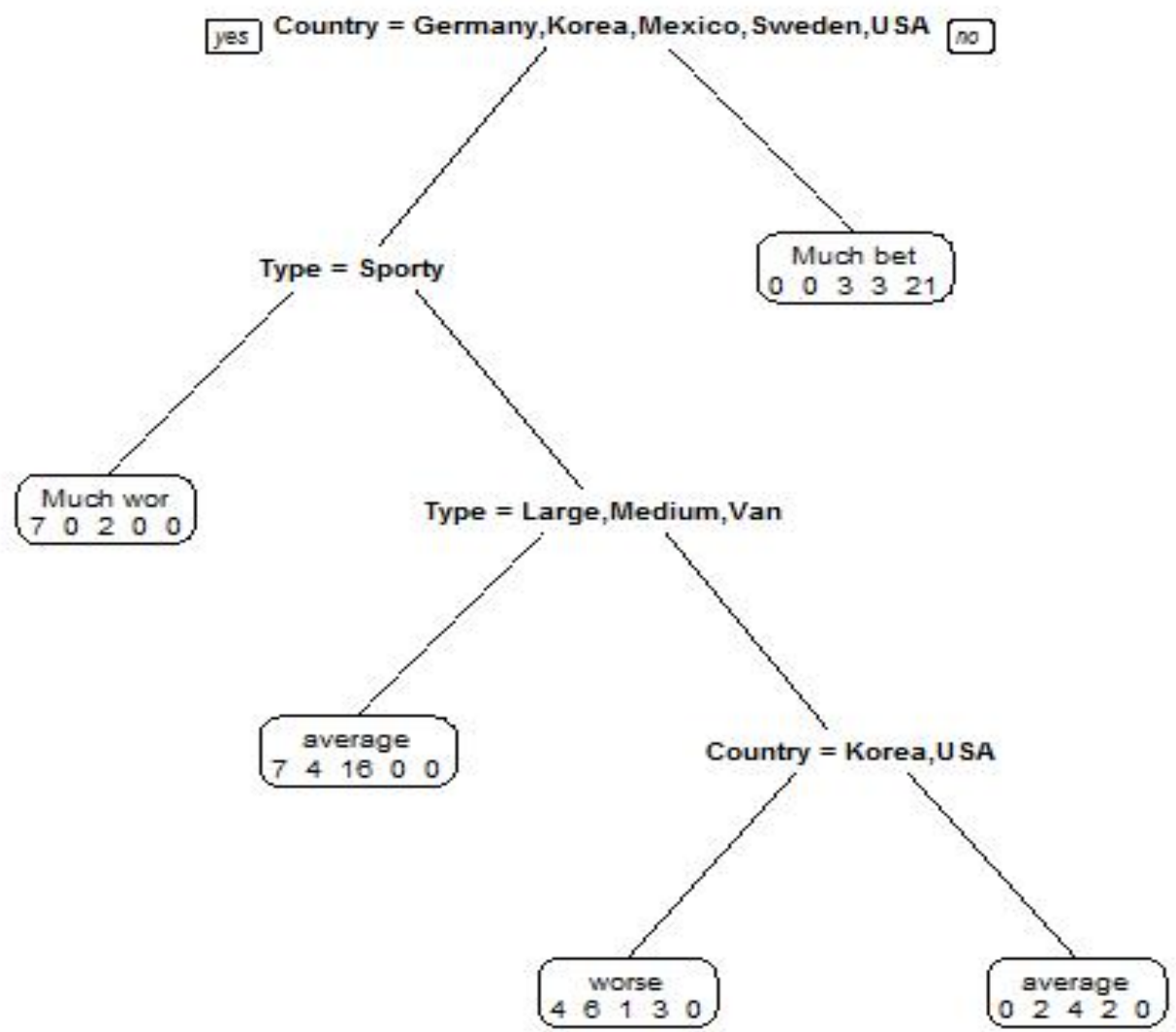
Libraries

rpart
partsm
rpart.plot

Functions

prune
prp
printcp

Case of study



Random Forest

Introduction

This algorithm is based on many random decision trees. It can be called as universal solution.

- regression and classification
- dimensional reduction methods
- outliers values
- treat missing values

How works?

Example:
Given a new passenger and knowing his or her personal information, we want to predict whether he or she will survive.

Building a machine learning model:
The algorithm builds (**ntree**) trees repeating the following steps:

- Generate the data to build the tree choosing a random row from data sampize times.
- Randomly select a number of features **mtry** .
- Build a decision tree based on the sampled data taking account of the selected featueres only.

Advantages and disadvantages

Advantages

- It can handle a large amount of data in high dimensionality.
- It can effectively estimate the missing data.
- Maintain accuracy with large amount data
- Boostrap sampling

Disadvantages

- It is not as good at regression as it is with classification.
- The data may become over - fit if the sample data is too noisy.
- It can act as black box approach for statistical modelers

For Further Reading I


Pierre Lafaye de Micheaux.
The R Software,
Fundamentals of programming and statistical analysis
Springer, 2013.


William Sullivan
Machine Learning for Beginners Guide Algorithms