### Introduction

Decision Trees (DTs) represent a supervised learning technique used for predicting values of responses. DTs learn decision rules from features. They can are used for:

* Regression analysis
* Classification

For this reason the algorithms are also known as Classification and Regression Trees (CART). DT/CART models use the machine learning technique called the *adaptive basis function models*. DT/CART models learn directly from the data as compared to classic statistics models where initial model has to be indicated.

### Decision Trees key features

* Are not linear in parameters as compared to linear regression
* Split the feature space into a number of simple rectangular regions divided by axis parallel splits
* The mean of the rectangle in feature space to which the observation belongs is used for predicting a particular observation
* The algorithm computes a locally optimal maximum likelihood estimate (MLE) for the parameters
* Produces *if-then-else* decision rulesets (graphical flowcharts)
* The algorithms can be used for predicting future asset prices and liquidity

### Mathematical Overview

Under a probabilistic adaptive basis function specification the model is the following:

– mean response in a particular region

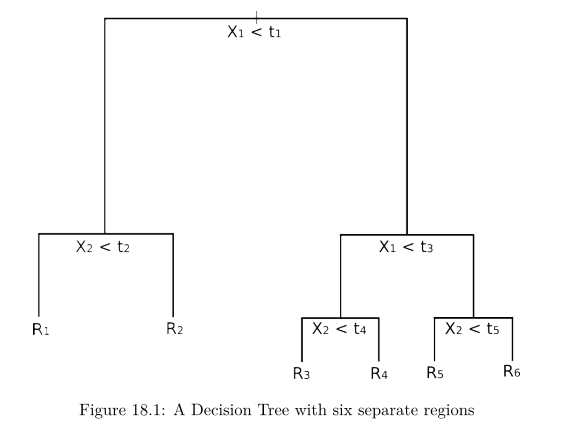
and - how each variable is split at a particular threshold value. These splits define how the feature space in is carved into separate ‘hypeblock’ regions.

### Decision Trees for Regression

We consider a regression problem with two feature variables .

– numerical response

This situation can be shown as follows:



Question 1:

Is (threshold value) ?

Yes: Is (threshold value) ?

Yes: The value is in region 1

No: The value is in region 2

Is (threshold value) ?

No: Is (threshold value) ?

Yes: Is (threshold value) ?

Yes: The value is in region 3

No: The value is in region 4

No: Is (threshold value) ?

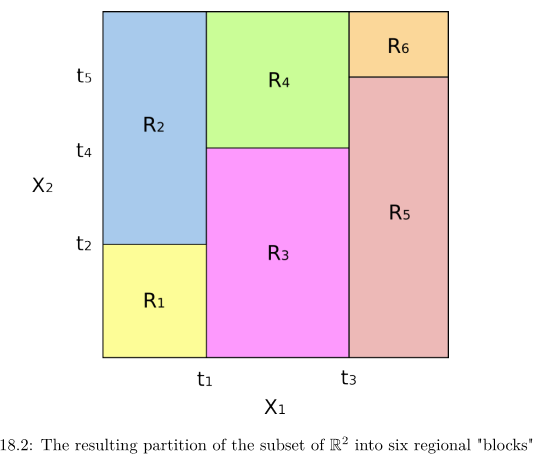
No: Is (threshold value) ?

Yes: The value ends in region 5

No: The value ends in region 4

Concepts in tree based regression:

* Axis parallel splitting – every split of the domain that is aligned with one of the feature axis.
* Axis parallel splitting can be used for higher dimensions data
* For a feature space of dimension (a subset of ) the space is divided into regions each denoted by
* The above chart shows how a subset of is trained using example data



### Pseudocode for CART Algorithm

### 

### The algorithm has to minimize some form of error criterion. Residual Sum of Squares (RSS) can be used. RSS in the case of a partitioned feature space with M partitions is the following:

Algorithm stopping criteria:

* Indicate maximum tree depth
* Sufficient training examples
* Regions are sufficiently homogeneous

### Logistic regression versus decision trees

|  |  |
| --- | --- |
| Logistic regression | Decision trees |
| Parametric model | Non-parametric model |
| Provides good results when there is linearity | Provides good results when variables are categorical |
| Assumptions are made regarding the response variable (binomial or bernoulli distribution) | No assumptions regarding data distribution |
| Logistic curve is predefined | No assumption regarding the shape of the model; results are based on the data |
| Outliers and missing data influence negatively model results | Outliers and missing data do not influence results |
| Does not handle non-linear relation between the data | Handles non-linear relation |

### Pruning the tree

* The three splitting process has to be adjusted in order to surpass the dangers of overfitting and the bias-variance tradeoff
* *Cost-complexity pruning* means the introduction of an additional tuning parameter which balances the depth of the tree with goodness of fit to the training data. This approach is described in detail by James et al (2013) and by Hastie et al (2009).

### Decision Trees for Classification

* DTs can be used in regression analysis and in classification
* A response value is predicted in regression analysis while a category is forecasted in classification process
* The difference is that the *mode* of the training region has to be used in the classification process instead of a *mean* value in regression
* Classification requires additional splitting criteria as the Residual Sum of Squares (RSS) does not work in this case
* Splitting criteria for classification: 1) Error Rate (Hit Rate) 2) Gini Index 3) Cross-Entropy

### Error Rate/Hit Rate

Hit Rate -> fraction of training observations in a particular region that does not belong to the most widely occurring class

-> fraction of training data in region that belong to class .

### Gini Index

The Gini Index -> shows how pure a region is (how much of the training data in a particular region belongs to a single class).

The Gini index is small if contains data that is mostly from a single class .

### 

### Cross-Entropy/Deviance