Rensselaer

why not change the world?®

Regression: Linear Regression, Decision Trees, Random Forest

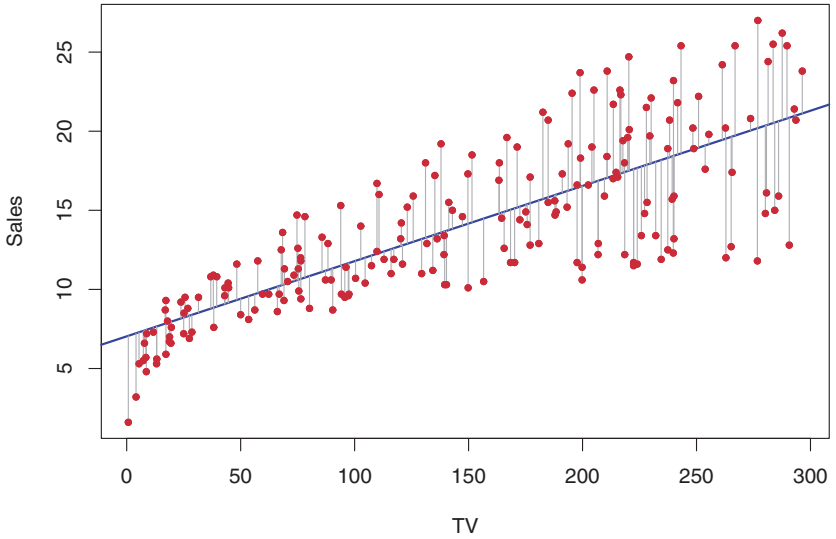
Ahmed Eleish

Data Analytics ITWS-4600/ITWS-6600/MATP-4450/CSCI-4960

February 4th, 2025

Tetherless World Constellation
Rensselaer Polytechnic Institute

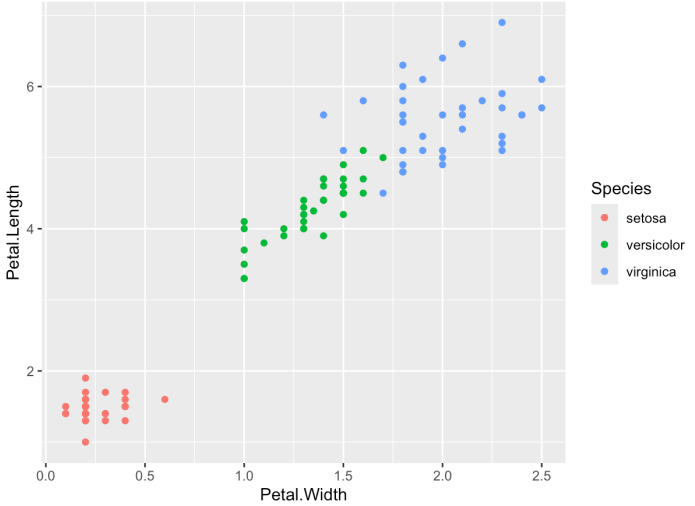
Regression



x-axis: independent numeric variable
y-axis: dependent numeric variable

Look for:
- trend? direction?
- are points tightly grouped?

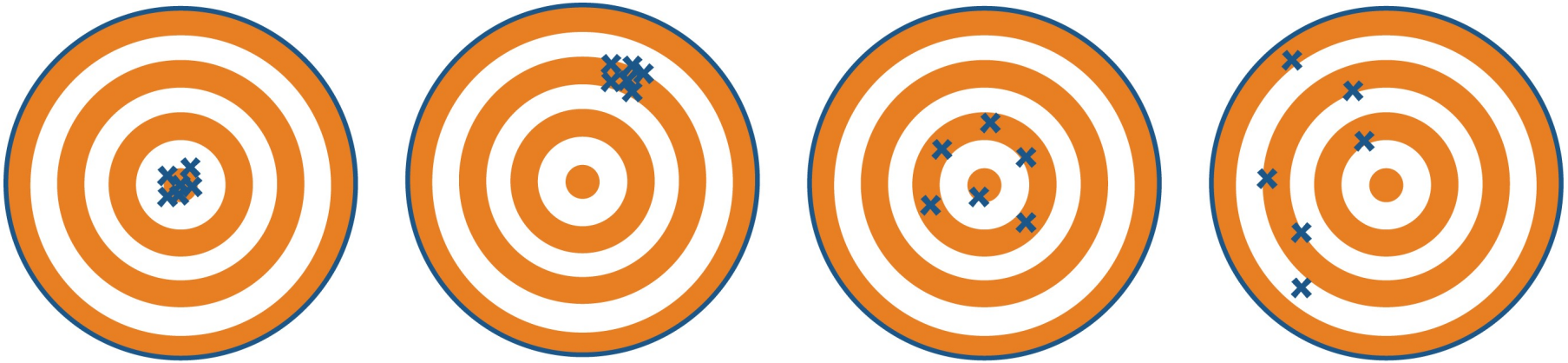
Classification/Clustering



x-axis: numeric variable
y-axis: numeric variable

Look for:
- structure: groups? group separation?

Accurate vs. Precise



High Accuracy
High Precision

Low Accuracy
High Precision

High Accuracy
Low Precision

Low Accuracy
Low Precision

http://climatica.org.uk/climate-science-information/uncertainty

Linear Regression

Regression

Linear Regression: In regression, fitting covariate and response data to a line is referred to as linear regression.

Covariate: A variable that is possibly predictive of the outcome under study control variable, ***explanatory variable, independent variable, predictor***

Response: dependent variable

Intercept: The expected value of the response variable when the value of the predictor variable is 0.

Slope: the average increase in Y associated with a one-unit increase in X

Reference/Resources:
The Elements of Statistical Learning. Hastie • Tibshirani • Friedman, 2nd Edition.
Introduction to Probability and Statistics, 4th Edition by Beaver.
Introduction to Statistical Learning with R, 7th Edition (ISLR).

Simple Linear Regression

- Let's take a look at the Least Squares Method for a single covariate (single regression).

- Utilizing the statistical notion of estimating parameters from data points, we find the estimates (coefficients) using the least squares method.

- We will look at evaluating linear models.

Least Squares Method

Equation of line: $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x$

Let *n* be a positive integer. For a given data *(x1,y1), ..., (x_n,y_n)* $\in \mathbb{R}\times\mathbb{R}$,

- we obtain the intercept β_0 and slope β_1 using the least squares method.
- Residual Sum of Squares (RSS), the *i*th residual $e_i = y_i - \hat{y}_i$

$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$$

Or

$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

More precisely, we minimize RSS

$$\text{RSS} = \sum_{i=1}^{n}(y_i - \widehat{\beta_0} - \widehat{\beta_1}x_i)^2$$

Sum of squared distances between (x_i, y_i) and $(x_i, \widehat{\beta_0}+\widehat{\beta_1}x_i)$ over i = 1,...,n

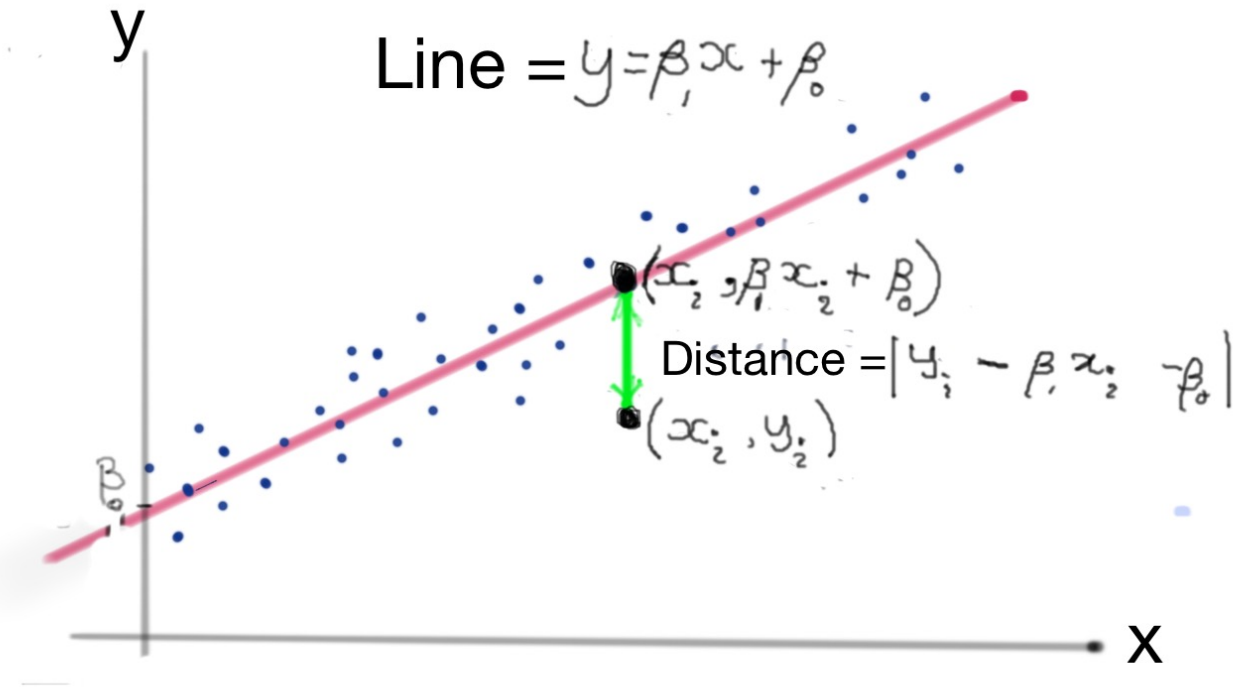


Figure: obtain $\widehat{\beta_0}$ and $\widehat{\beta_1}$ that minimize $\sum_{i=1}^{n}(y_i - \widehat{\beta_0} - \widehat{\beta_1}x_i)$ via least squares method

• We partially differentiate *L* by β_0 and β_1 and let them be equal to zero, we obtain the following equations:

$$\frac{\partial L}{\partial \widehat{\beta_0}} = -2\left(\sum_{i=1}^{n}\left(y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i\right)\right) = 0 \qquad \text{Eq(1)}$$

$$\frac{\partial L}{\partial \widehat{\beta_1}} = -2\left(\sum_{i=1}^{n} x_i\left(y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i\right)\right) = 0 \qquad \text{Eq(2)}$$

Where the partial derivative is calculated by differentiating each variable and regarding the other variables as constants. In this case, β_0 and β_1 are regarded as constants when differentiating *L* by β_0 and β_1 respectively.

- By solving Eq (1) and Eq (2) when:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 \neq 0 \qquad \text{Eq(3)}$$

i.e., $x_1 = x_2 = \cdots = x_N$ is not true.

Where:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{N}\sum_{i=1}^{n} y_i$$

- We can obtain:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})} \qquad \text{Eq(4)}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \qquad \text{Eq(5)}$$

Assessing the Coefficient Estimates

True relationship between X and Y:

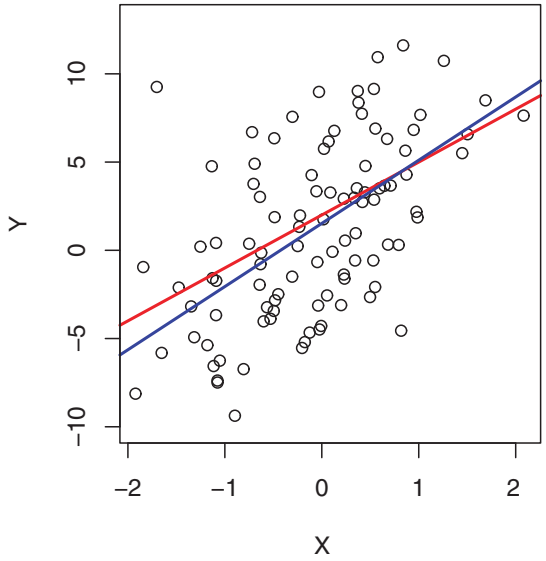
$$Y = \beta_0 + \beta_1 X + \epsilon.$$

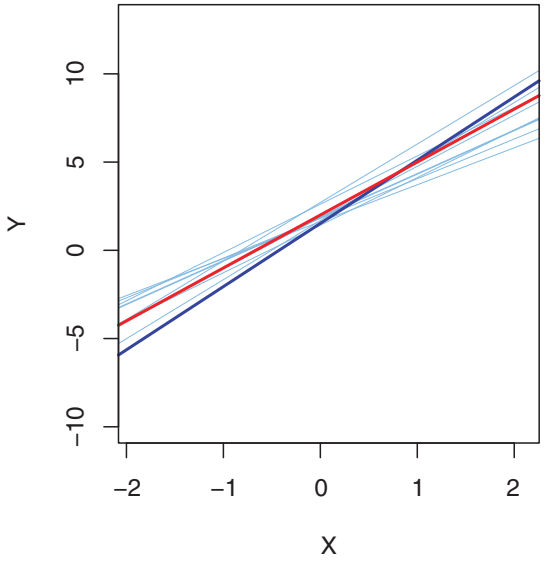
- Where ϵ is a mean-zero random error

Red: true relationship

Dark Blue: least squares regression line

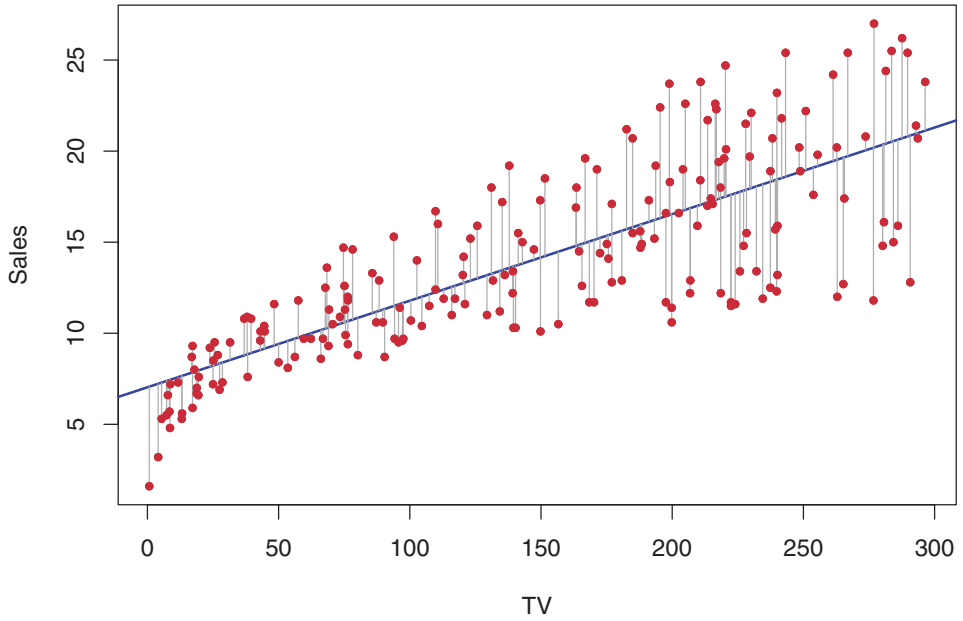
Light Blue: least squares regression lines for multiple random subsets





Evaluating Linear Models

- Sales vs. TV ad spending
- Sales in 1000s of units
- TV ad spending in 1000s of $



Evaluating Linear Models

Values of coefficients >> their Std. errors

High t-statistic

Very low p-value

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

Hypothesis (more TV ads → more sales)

H0 : There is no relationship between X and Y

Ha : There is some relationship between X and Y

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)},$$

Reject the null hypothesis!

Residual Standard Error

- Mean sales ≈ 14,000 units

RSE = 3.26 = 3,260 units
good/bad?

R^2
- measures the proportion of the variability in *Y* that can be explained using *X*
- has a value between 0,1

Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

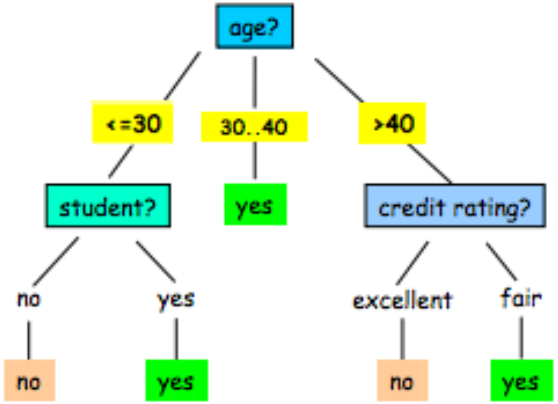
$$\mathrm{TSS} = \sum(y_i - \bar{y})^2$$

Decision Trees

Decision tree classifier

Classification by Decision Tree Induction

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31..40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31..40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
>40	medium	no	excellent	no



buys_computer ?

More on this later in Group 2 ...

Decision Trees

Decision trees can be applied to both regression and classification problems

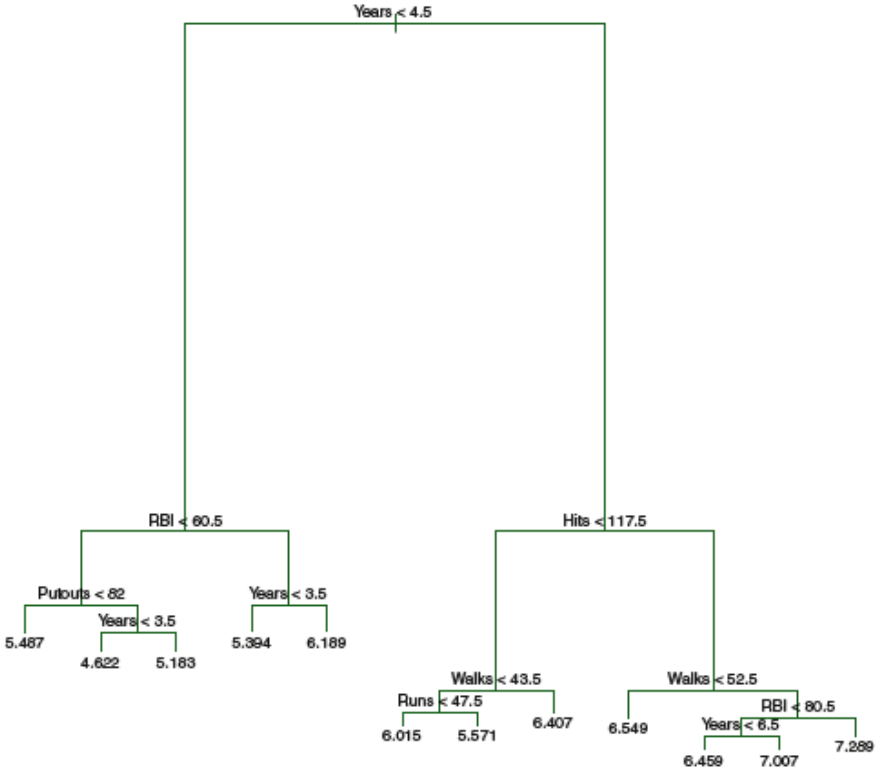
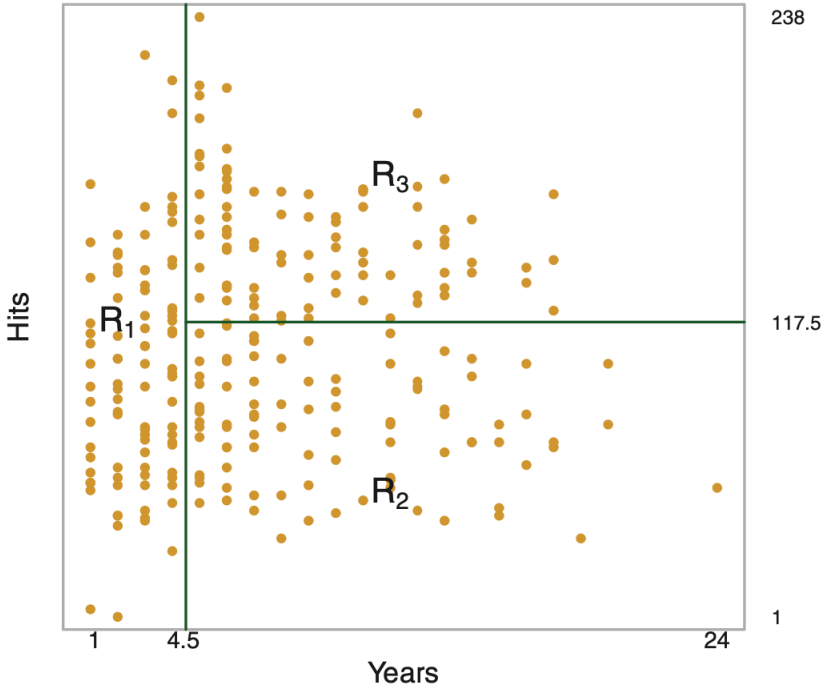


FIGURE 8.4. *Regression tree analysis for the* `Hitters` *data. The unpruned tree that results from top-down greedy splitting on the training data is shown.*

Reference/Resources: Introduction to Statistical Learning with R -7 Edition: Chapter 8

Decision Trees



Reference/Resources: Introduction to Statistical Learning with R -7 Edition: Chapter 8

Advantages and Disadvantages of Decision Trees

Advantages:

- Trees are very easy to explain to people, in fact, they are even easier to explain than linear regression.

- Some people believe that decision trees are more closely mirror human decision-making than other regression and classification techniques.

- Trees can be displayed graphically, and easily interpreted even non-experts (especially if the tree is small) can understand.

Reference/Resources: Introduction to Statistical Learning with R -7 Edition: Chapter 8

Advantages and Disadvantages of Decision Trees

Disadvantages:

• Unfortunately, trees generally do not have the same level of predictive accuracy as some of the other regression and classification approaches

• Additionally, trees can be very non-robust. In other words, a small changes in the data can cause a large change in the final estimated tree

• However, by aggregating many decision trees, using methods like *bagging*, *random forest* and *boosting*, the predictive performance of trees can be substantially improved.

Decision Tree – Classification

• When we implement decision trees for classification, the idea is to split the data into subsets. So that each subset belongs to one particular class.

• In other words, splitting the data into regions, that are separated by decision boundaries, where each region's samples have only one class.

Decision Tree

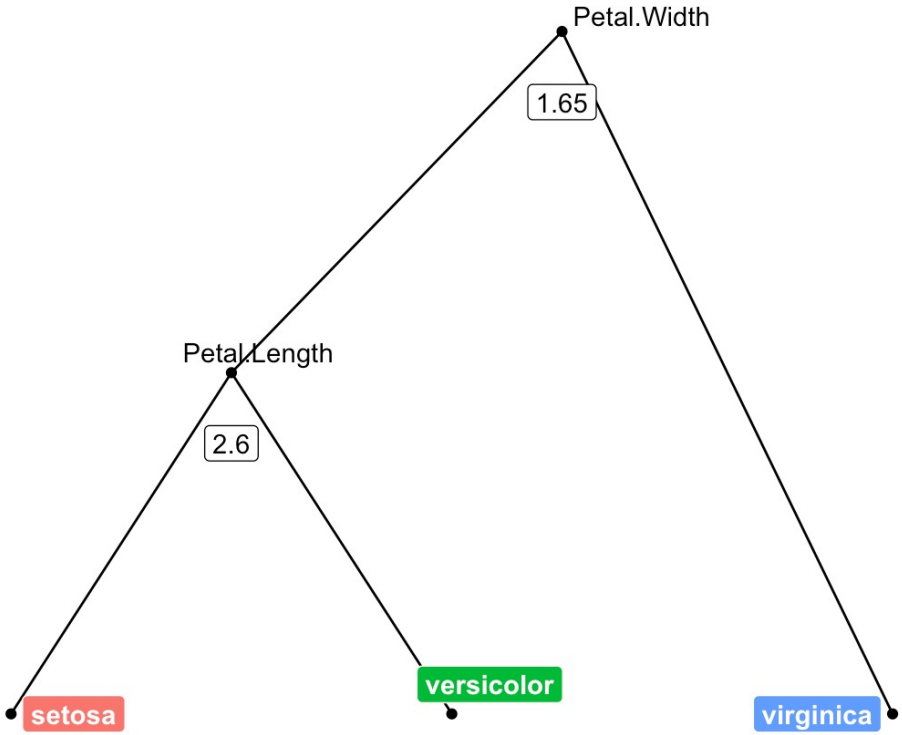
• Decision Tree has a hierarchical structure with “nodes” and “directed edges”

• The top node is defined as the “root node” and the nodes at the bottom are called as “leaf nodes”

• Nodes that are neither root node nor the leaf nodes are identified as “internal nodes” in the decision tree.

• There is a “class label” associated with each leaf node

Decision Tree – Classification



Decision Tree

- Classification decisions are made by traversing the decision tree

- Traversing starts from the root node (from the top of the tree).

- The root node and the internal nodes have test conditions. Those test conditions determine which path to traverse on the tree.

Decision Tree

- **When a leaf node is reached through traversing, the category of the leaf node determines the classification**.

- The **depth is measured from the root node** and the **depth at the root node is zero**.

- **The depth of the decision tree**: Tree Depth is calculated by counting the number of edges in the longest path from the root node to a leaf node.

Decision Tree

• Number of nodes in the decision tree determine the size of the tree.

• **The decision tree constructing algorithm is referred to as a tree induction algorithm**.

Impurity Measure

• The goal is to have the resulting subsets to be homogeneous as possible and minimize the impurity.

• In practice, we don’t get pure homogeneous subsets, there are impurities.

• A **common impurity measure to determining the best split is “Gini-index”**

• The **lower the Gini-index value, the higher the purity of the split**.

Impurity Measure

• The decision tree will select the split that minimize the Gini-index.

• Besides the Gini-index, there are other impurity measures available such as:

- entropy or information gain
- misclassification rate

• The decision tree will test all variables to determine the best way to split a node using a purity measure such as Gini-index to compare different possibilities

• **Tree induction algorithms repeatedly split nodes to get more and more homogeneous subsets**.

• When does this process stop? When does the algorithm stop growing the tree?

When to Stop splitting the nodes?

• There are several criteria that can be used to determine the when a node shouldn't be split into subsets.

• The **induction algorithm can stop expanding a node when all samples in the node have the same class label**.

• Since getting pure subsets is difficult to archive with real world data, **the stopping criteria can be modified to a certain percentage of the samples in the node. i.e 95% of have the same class label.**

Stopping criteria

• The algorithm can stop expanding a node when the number of samples in the node falls below a certain minimum number.

• The induction algorithm can stop expanding a node when the improvement in impurity measure is way too small to measure (too small to make a much difference in classification result).

• The algorithm can also stop expanding when it reaches maximum tree-depth.

Random Forest(s)

Random Forest

- Random Forest is based on decision trees.
- In Random Forest we build large number of trees, where each tree is based on a bootstrap sample.
- Then, what we do is we average those predictions together in order to get the predictive probabilities of each class across all the different trees.

Random Forest

Cons:

- Speed (it can be slow; it has to build large numbers of trees)

- Interpretability (it can be hard to interpret in the sense that you have large number of trees that are averaged together and those trees represent the bootstrap samples and are complicated to understand)

Random Forest

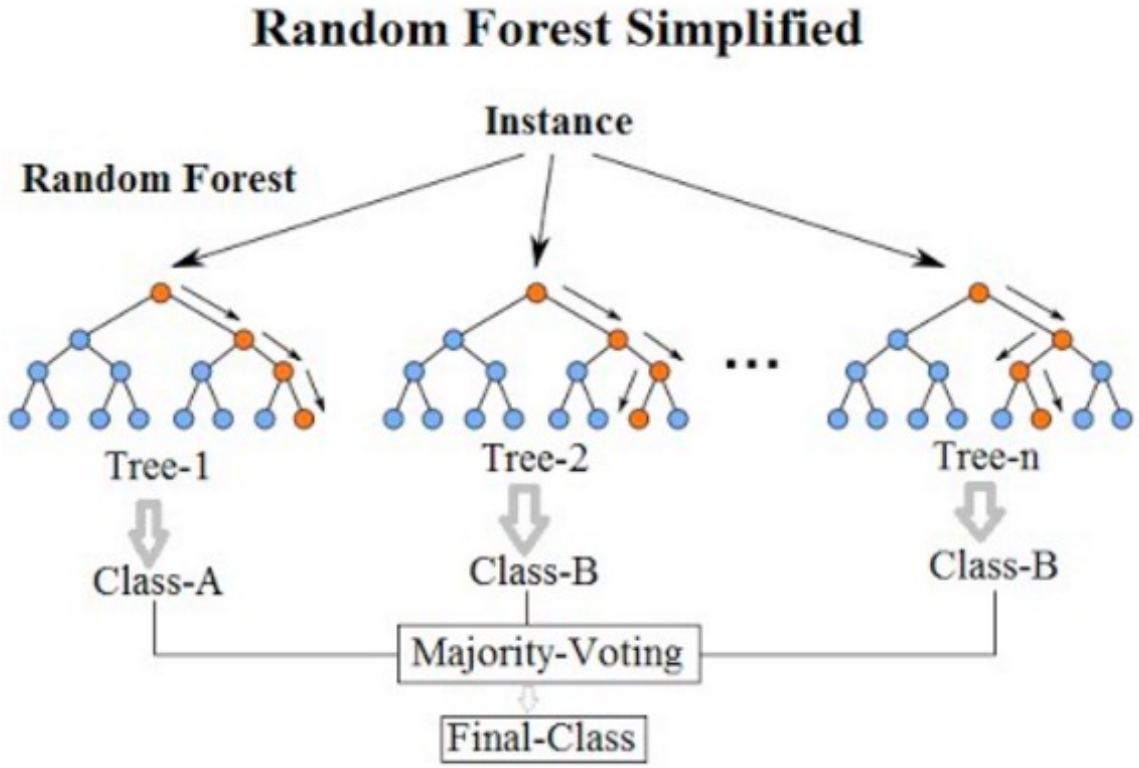


Image Resource: https://commons.wikimedia.org/wiki/File:Random_forest_diagram_complete.png

Random Forest

The original algorithm was created in 1995 by Tin Kam Ho.

An extension of the algorithm was developed by Leo Breiman and Adele Cutler, who registered "Random Forests" as a trademark in 2006

- http://www.stat.berkeley.edu/~breiman/RandomForests/

Random Forest Algorithm

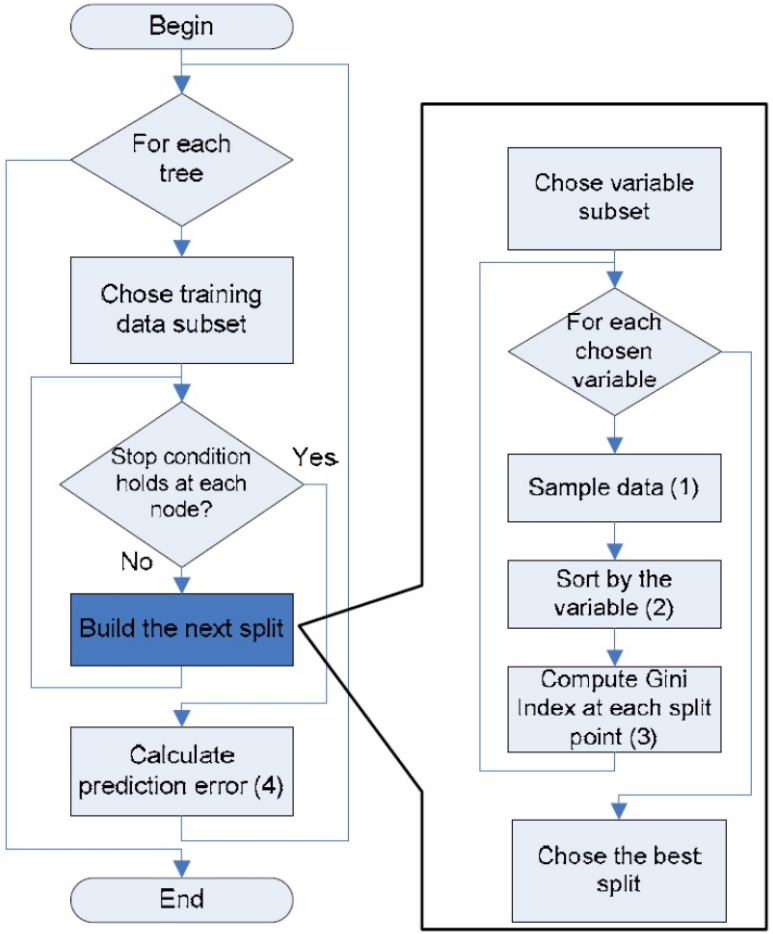
- Let N_{trees} be the number of trees to build
for each of N_{trees} iterations:

1. Select a new bootstrap sample from training set
2. Grow an un-pruned tree on this bootstrap.
3. At each internal node, randomly select m_{try} predictors and determine the best split using only these predictors.
4. Do not perform cost complexity pruning. Save tree as is, along side those built thus far.

Output overall prediction as the average response (regression) or majority vote (classification) from all individually trained trees

Ref: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=5f31bcc21ab2155c084527648d436b036126b30d

Random Forest



Image/ Photo Credit: Albert A. Montillo

Thanks!