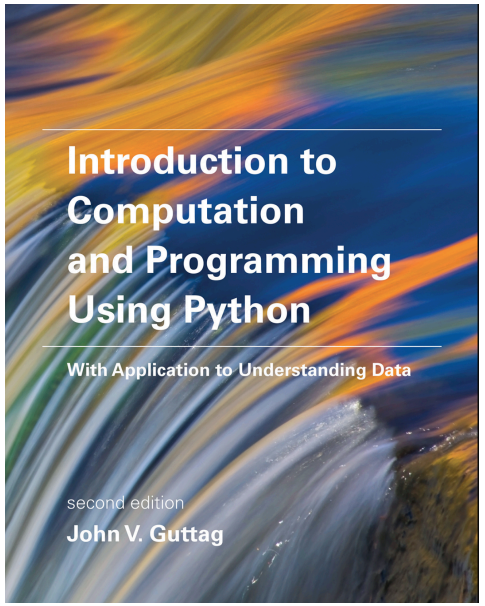
Decision Tree Models Machine Learning &Operations Research:

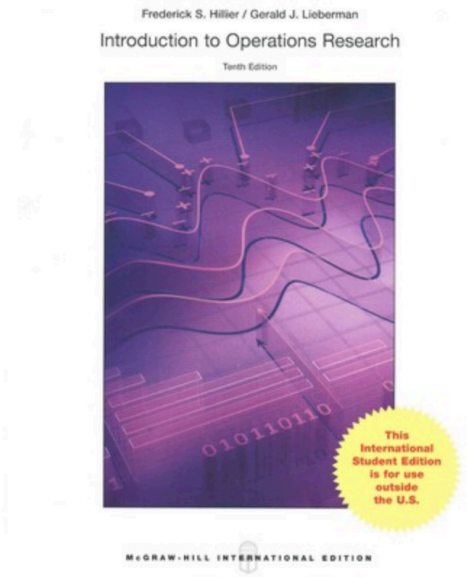
WILLIAM STEIMEL
スタイメルウィリアム

Table of Contents

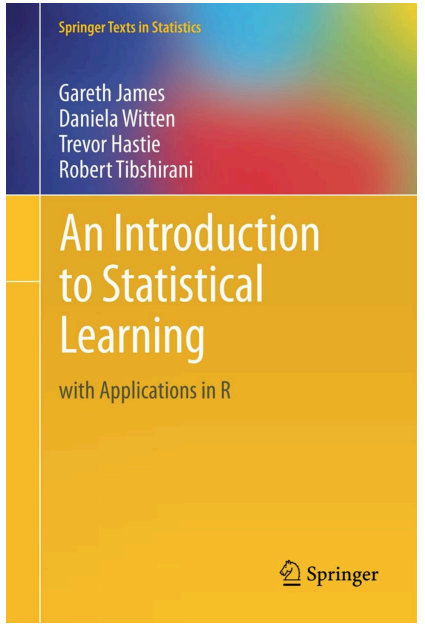
Sources



For Python



Chapter 16
Decision Analysis



Chapter 8 Tree based methods

What is a Decision Tree?

- Supervised Learning - Algorithm
- Predictive Model- Segmentation/splitting of the predictor space into a number of simple regions in order to make a prediction.
  - Non-Linear data structure (Hierarchical)
- Two types of decision trees (CART)
  - Classification Tree
    - Categorical Data (Passenger survived? Y or N)
  - Regression Tree
    - Continuous Data (Prediction of Housing Price)
- Widely used in Machine Learning/Operations Research
  - Application of Decision tree differs for Machine Learning and Operations Research

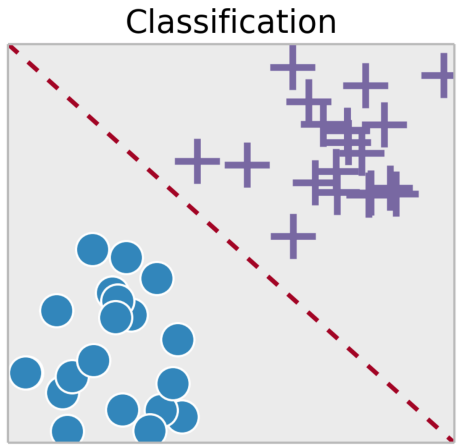
Machine Learning: Decision Tree

Categorical (Classification) vs Continuous (Regression) Decision tree Examples

Classification Tree

- Passenger Died (Y/N)
- Type of Flower(Rose,Lily)
- Loan Prediction (Y/N)

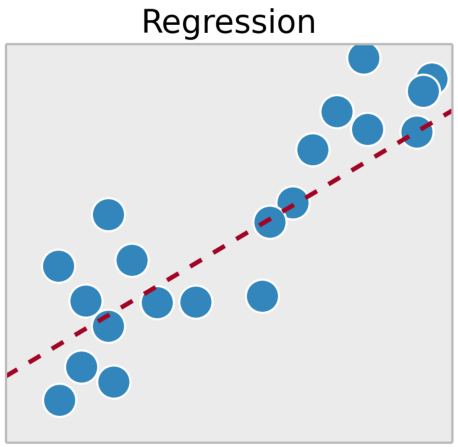
Pattern Classification



Regression Tree

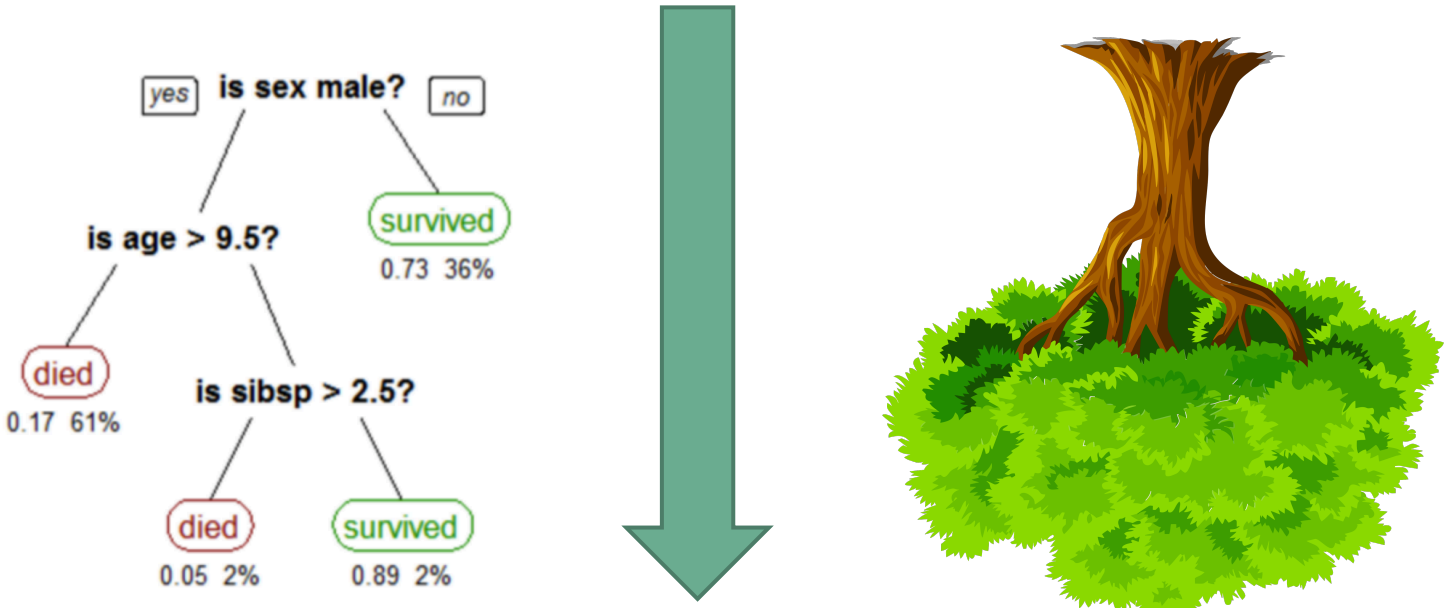
- Housing Price Prediction
- Stock Prediction
- Sales Prediction

Value Prediction



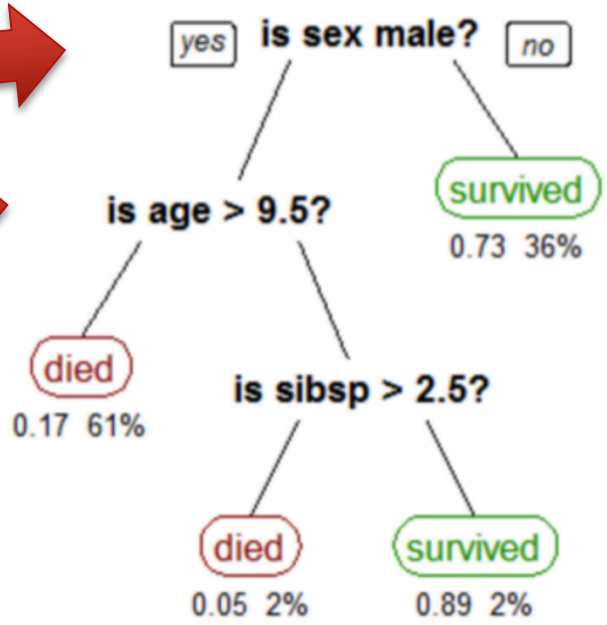
Classification Decision Tree Simple Example

- Like a reverse tree (Starts from the top and branches down)
- Based on popular Titanic Machine Learning Dataset



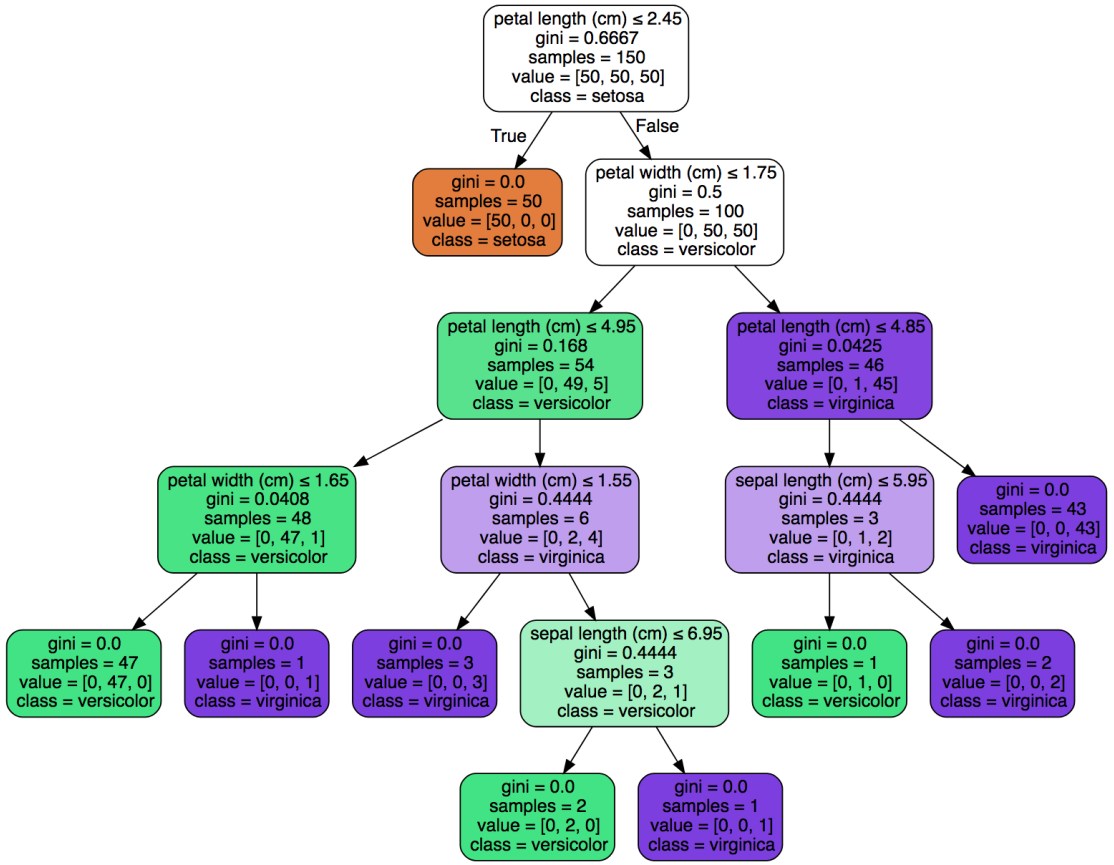
Classification Decision Tree Terminology

- Attributes – Sex, Age, sibsp (number of spouses or children)
- Internal node (condition)
- Edges (Branches in tree)
- Terminal Nodes – Leaves
  - (Decision) Died / Survived



Decision Tree Example from Scikit Learn Visualization

Example of Decision Tree Logic:



Classification Tree Evaluation- Best Split Measures

- Unlike Regression Tree's - Reduction of RSS cannot be used to make the binary splits in Classification Trees
- One method used is the classification error rate

$$E = 1 - \max_k(\hat{p}_{mk}).$$

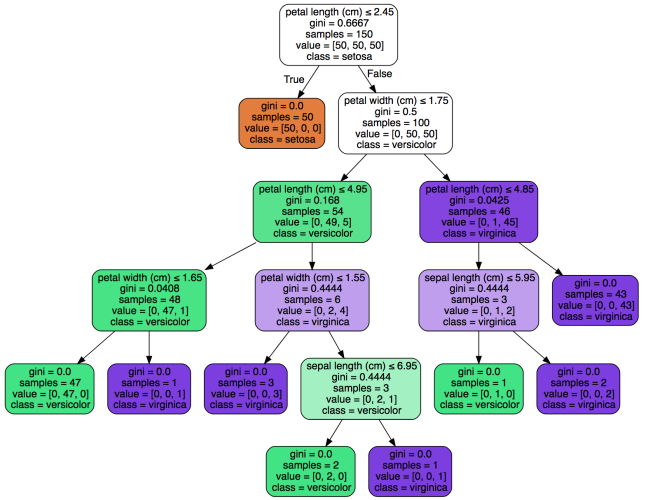
- Typically the Gini Index or Cross Entrophy formula are used to evaluate tree split quality.

Gini Index: Impurity

- Application to Decision Trees - Measures Impurity of the classes at each node
- Smaller value indicates that more observations come from a single class (More pure)
- Min = 0, Max = 1

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

Formula- Corrado Gini



Graph from Scikit Learn Example

Cross-Entrophy

- Application to Decision Trees - Measures Impurity of the classes at each node (similar to Gini Index)
- Smaller value indicates that more observations come from a single class (More pure)
- Min = 0, Max = 1

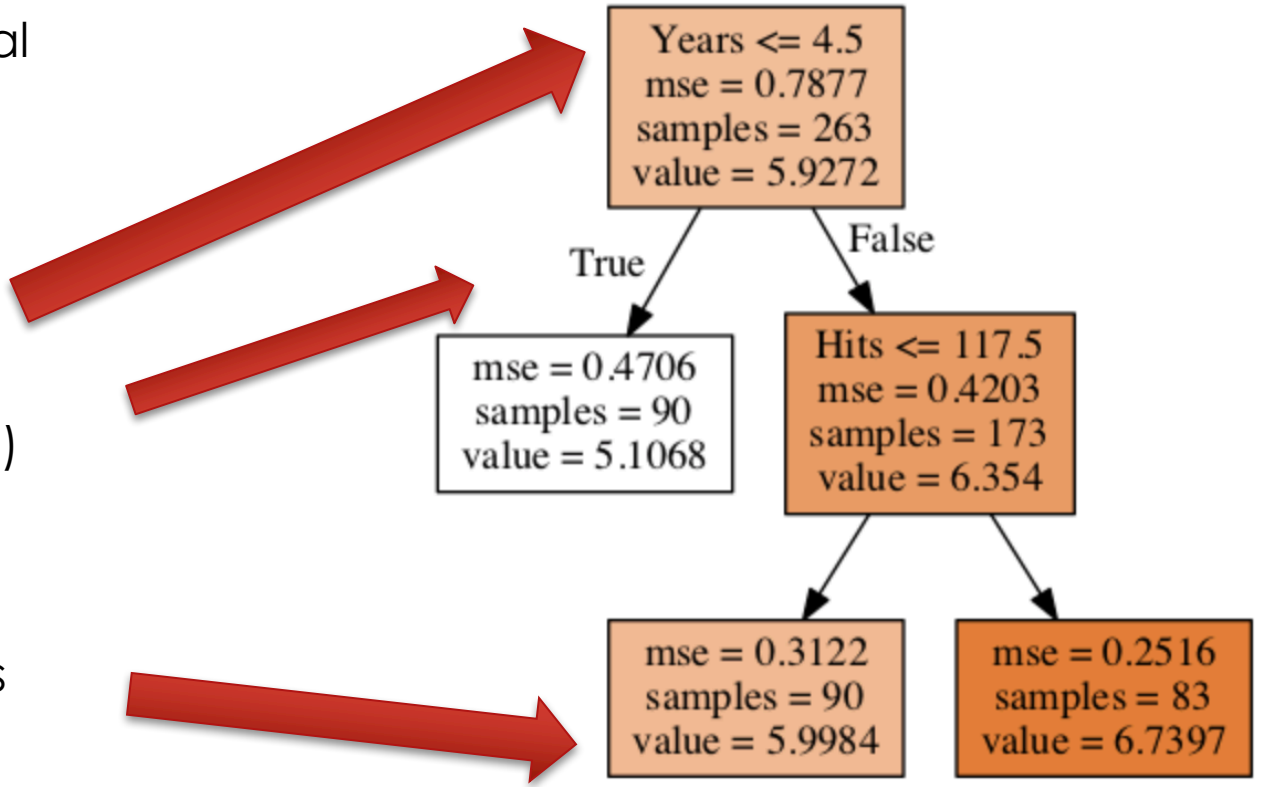
$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}.$$

What is Regression Tree?

Example from:
An Introduction to Statistical Learning

Hitters Dataset (Baseball)

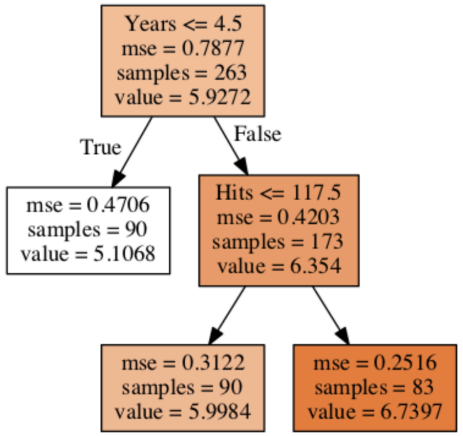
- Attributes – Years, Hits
- Internal node (condition)
- Edges (Branches in tree)
- Terminal Nodes – Leaves **(Predicted Value)**



Graph from Introduction to Statistical Learning

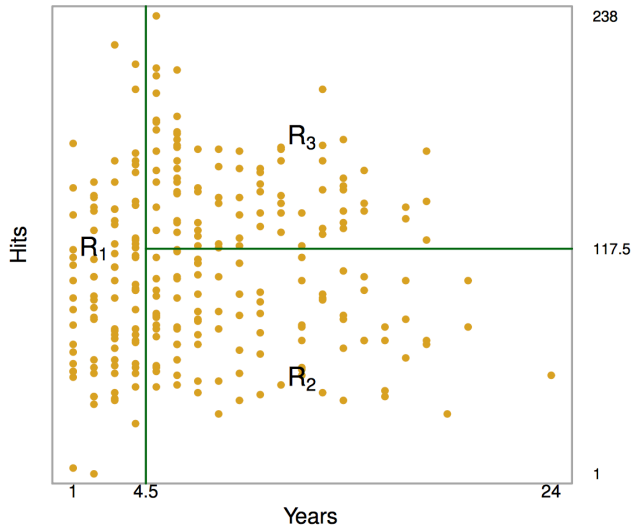
Segmentation of Decision Space

- Predicting baseball player's salary based on Years in the major leagues and hits made in the previous year.
- Players are segmented into three spaces within the predictor space



Segments

R1: Years <= 4.5 in major leagues
R2: Hits <= 117.5 & Years > 4.5
R3: Hits >= 117.5 & Years > 4.5



Graphs from Introduction to Statistical Learning

Split determination for Decision Tree Regression

- How are segments created within the predictive space?

The goal is to divide the predictor space that minimizes the RSS (Residual sum of squares)

$$RSS = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

- Step 1 : Divide the predictor space:

X1,X2,….,Xp – Into J Distinct and non-overlapping regions, R1, R2,…,RJ

- Step 2: For every observation that falls in Region Rj, we make the same prediction, (The mean of the training observations in Rj)

- This Method is however computationally infeasible

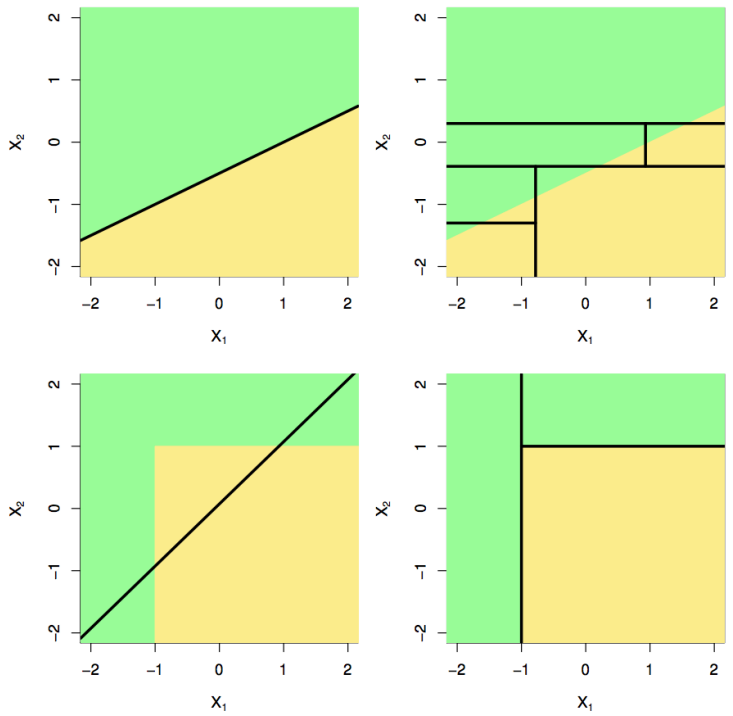
Decision Tree Regression (Greedy Approach)

- Recursive Binary Splitting (Greedy Approach)
  - Top-down approach as each step is dependent on the previous step
  - Greedy- Greedy because the decision tree makes the best decision at each step instead of looking ahead at future steps. (Optimal decision at current step)

$$R_1(j,s) = \{X|X_j < s\}, \quad R_2(j,s) = \{X|X_j \geq s\}$$

$$RSS = \sum_{x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2$$

Decision Tree vs Linear Models



Graph from Introduction to Statistical Learning

- Effectiveness depends on model and situation.
- If the relationship between the features is well approximated by linear model then linear model will be more effective. (Top Left)
- If the relationship is more complex or more non-linear between the features and response then decision trees may outperform traditional methods. (Bottom Right)

Pros/Cons of Decision Tree

Positive

- Easy to Read/Easy to visualize logic
- Easy to explain to anyone (Even non-technical stakeholder)
- More closely resemble human decision making compared to other regression/classification approaches (Hierarchical)

Negative

- Predictive accuracy not as powerful as some other new Machine Learning models

Decision Tree Performance Improvement Methods (Modern Approaches)

Bagging

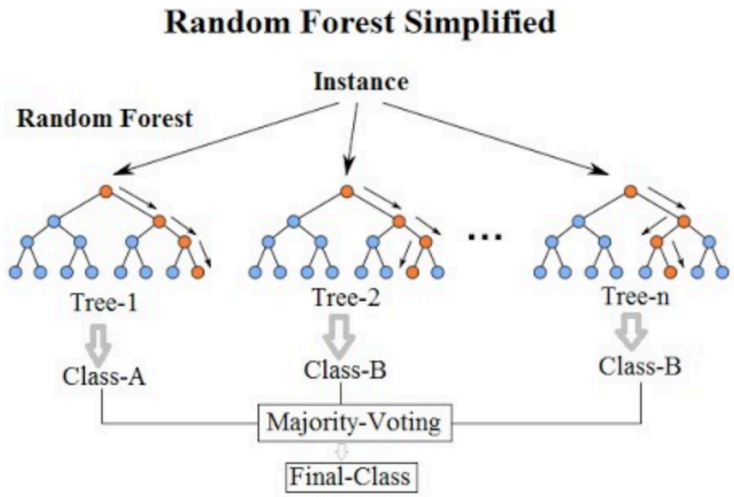
- Can be applied to many other statistical methods for regression/classification
- General purpose procedure for reducing the variance of a statistical method
  - In practice – Create multiple prediction models using separate training sets and average them together in order to obtain a single low-variance statistical model.
- (B) = Total Bootstrapped training sets
  - Quantitative Data- Uses Average Prediction across amount of training samples (B)
  - Qualitative Data – Uses majority vote across amount of training samples (B)
- Stronger performance compared to the use of a single decision tree
- Improves prediction accuracy at the expense of interpretability

Formula:

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x)$$

Random Forest

- Ensemble Learning (Combination of many decision trees)
- Improvements in prediction accuracy in comparison to bagging
- Improvements due to "de-correlation" of the trees – Random forests force each split to consider only a subset of the predictors



Boosting

- Like bagging- Can be applied to many other statistical methods for regression/classification
- Each tree is grown using information from previous trees (sequential growth)
- Boosting approach instead learns slowly

Figure 8.2 from Introduction to Statistical Learning

Algorithm 8.2 *Boosting for Regression Trees*

1. Set $\hat{f}(x) = 0$ and $r_i = y_i$ for all i in the training set.
2. For $b = 1, 2, \ldots, B$, repeat:
   (a) Fit a tree $\hat{f}^b$ with d splits ($d+1$ terminal nodes) to the training data (X, r).
   (b) Update $\hat{f}$ by adding in a shrunken version of the new tree:
   $$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x). \quad (8.10)$$
   (c) Update the residuals,
   $$r_i \leftarrow r_i - \lambda \hat{f}^b(x_i). \quad (8.11)$$
3. Output the boosted model,
   $$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x). \quad (8.12)$$

What is the Iris Dataset?

- 150 records – Practice Dataset for Classification of Iris Flower

- Attribute Information:
  - 1. Sepal length in cm
  - 2. Sepal width in cm
  - 3. Petal length in cm
  - 4. Petal width in cm
  - 5. Class: Iris Setosa, Iris Versicolour, Iris Virginica

- Source: https://archive.ics.uci.edu/ml/datasets/iris

What is the Boston Housing Dataset?

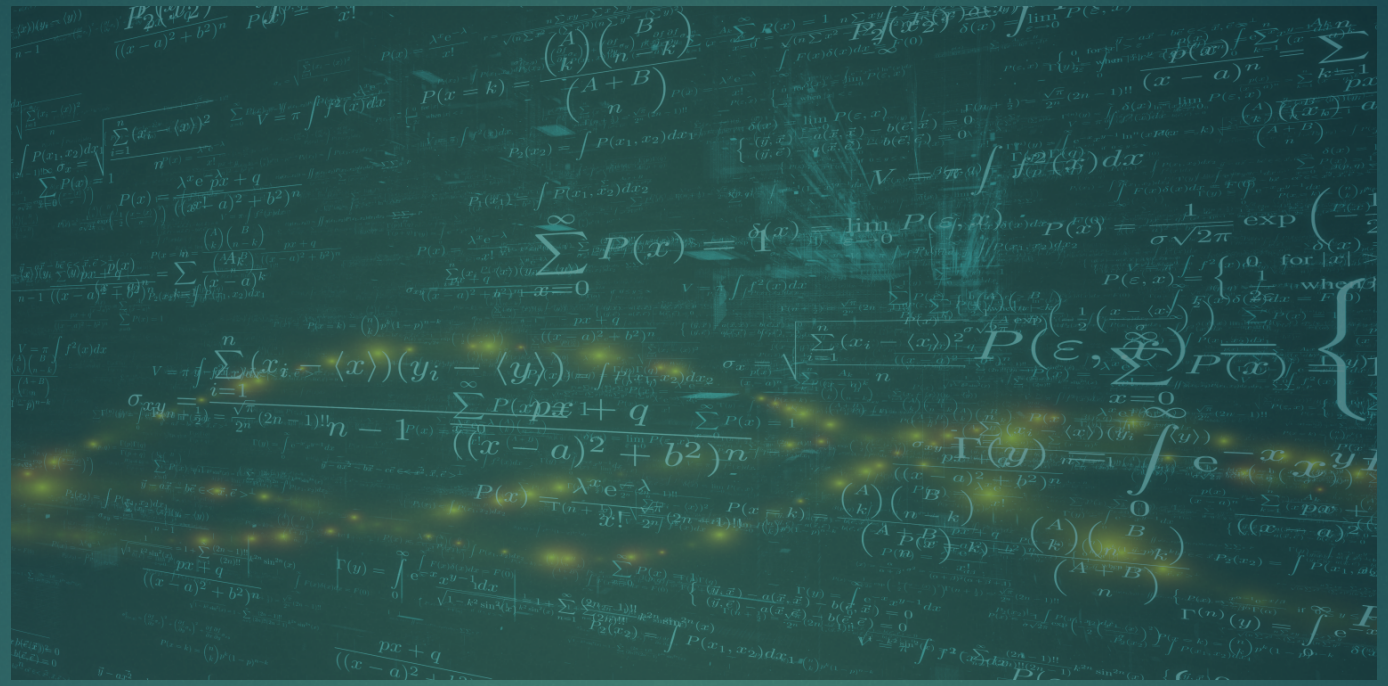
506 records – Practice Dataset for Regression and prediction of Boston Home Value
Attribute Information:

- CRIM
- ZN
- INDUS
- CHAS
- NOX
- RM
- AGE
- DIS
- RAD
- TAX
- PTRATIO
- B
- LSTAT
- MEDV

Source: https://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.names

Python Implementation

- Classification: ML-Iris-Dataset Jupyter Notebook
  - https://github.com/steimel64/steimel64.github.io/blob/master/Notebooks/Iris%20Notebook.ipynb
- Regression: ML-Boston-Housing Jupyter Notebook
  - https://github.com/steimel64/steimel64.github.io/blob/master/Notebooks/Boston%20Housing%20Notebook.ipynb

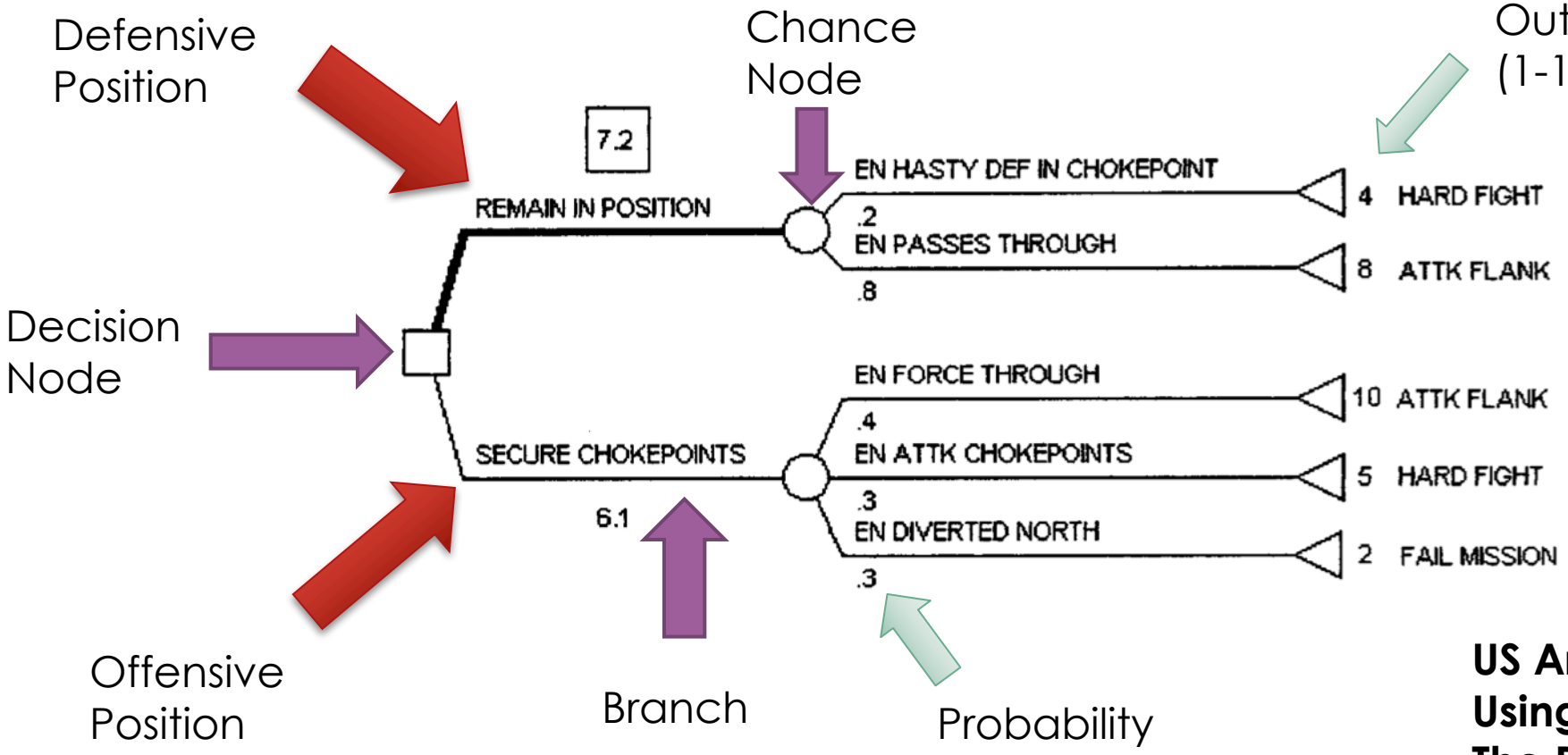


Operations Research

Operations Research Application of Decision Trees

- Used in Decision Analysis – Decisions are incredibly important for all organizations
  - Smart decisions contribute to Strategy
- Decision theory- Mathematical models that utilize probability theory and diagrams to create, visualize, and optimize decisions
  - The most logical decision based on data

Operations Research Decision Tree Example

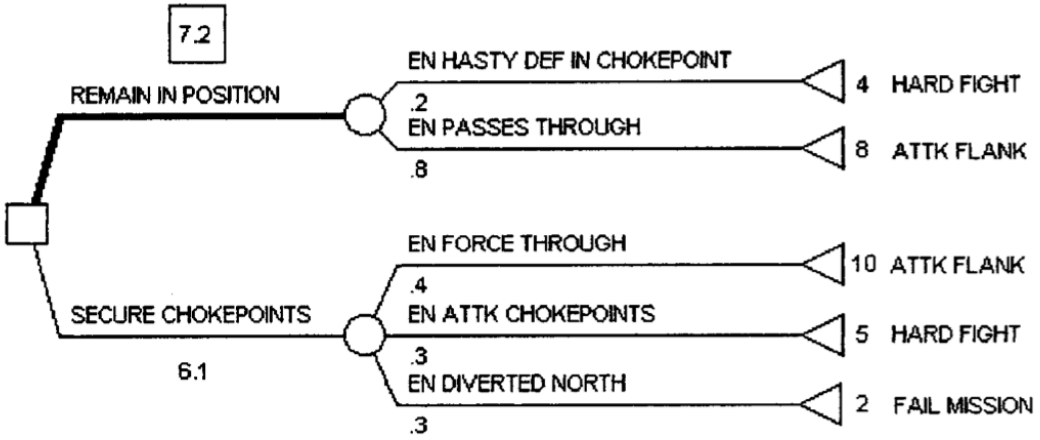


US Army Example:
Using Decision Trees To Direct The Planning Thought-Process:

Operations Research Decision Tree Example

- First Calculate Expected Payoff = Payoff Value x Probability
- For each Decision Node- Pick the branch with the largest expected payoff

US Army Example:



Remain in Position

4 * .2 = .8

8 * .8 = 6.4

Total = 7.2

Secure Chokepoints

10 * .4 = 4

5 * .3 = 1.5

2 * .3 = .6

Total = 6.1

In this case remain in position (Defensive) has the best payoff in comparison to Secure Chokepoints (Offensive)