DSLA
DATA SCIENCE LEARNING ACADEMY

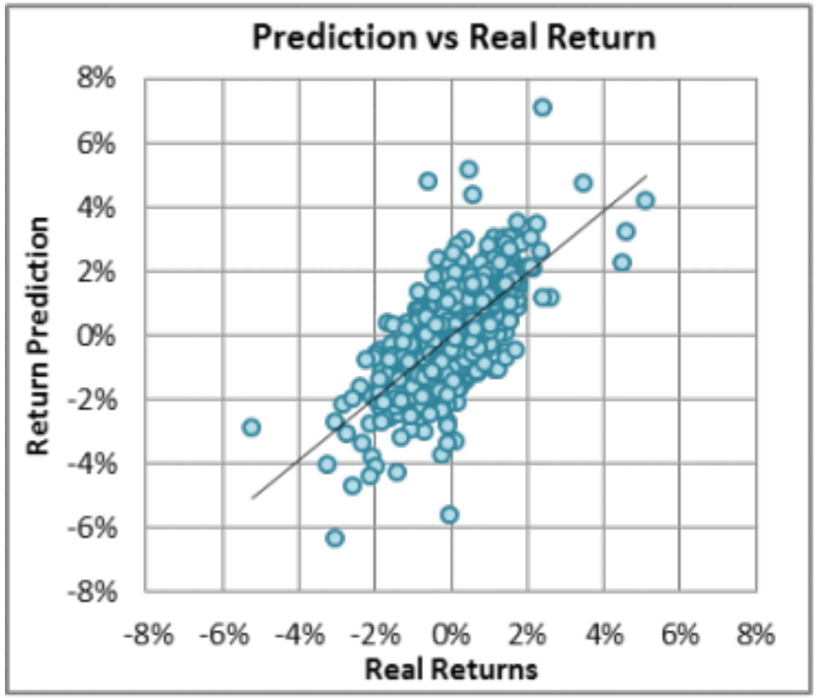
Introduction to Trees

DSLA COURSE

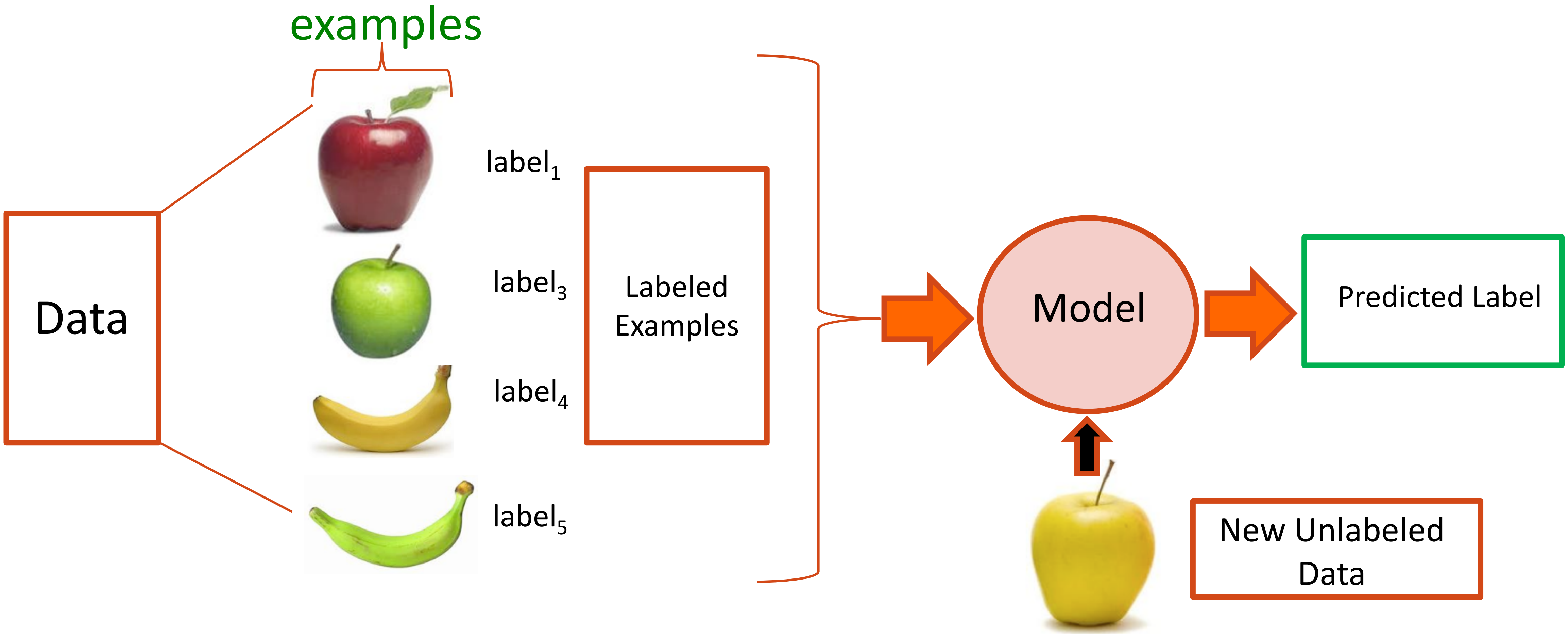
ROHIT PADEBETTU

Regression vs Classification

Regression vs Classification



Classification



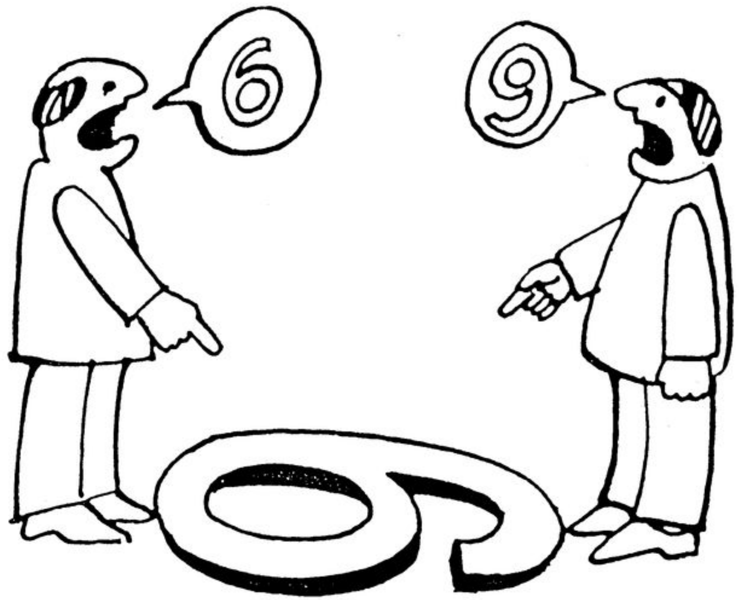
Classification – Real Life Examples

Hospital Emergency Room measures blood pressure, age, history of illness etc. of newly admitted patients. A decision needs to made regarding admitting the patients into ICU. They want to admit high risk patients. Problem is how to discriminate between high risk and low risk patients?

A Bank receives hundreds and thousands of loan applications with information about salary, age, marital status, other loans, credit history, employment status etc. The bank wants to make loans to those people who are most likely going to repay the loan. Problem is how to discriminate between those who are a good credit risk and those who are bad credit risk?

Classification – Other Examples

- Face/Speech Recognition
- Character Recognition
- Spam Detection
- Customer Segmentation
- Recommendation Systems

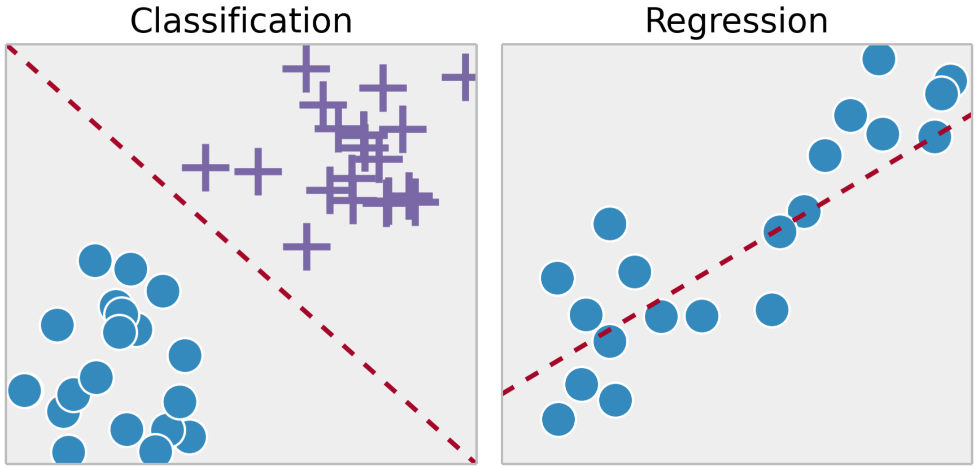


Classification – How?

Logistic, LDA, QDA, KNN for Classification – We got this!

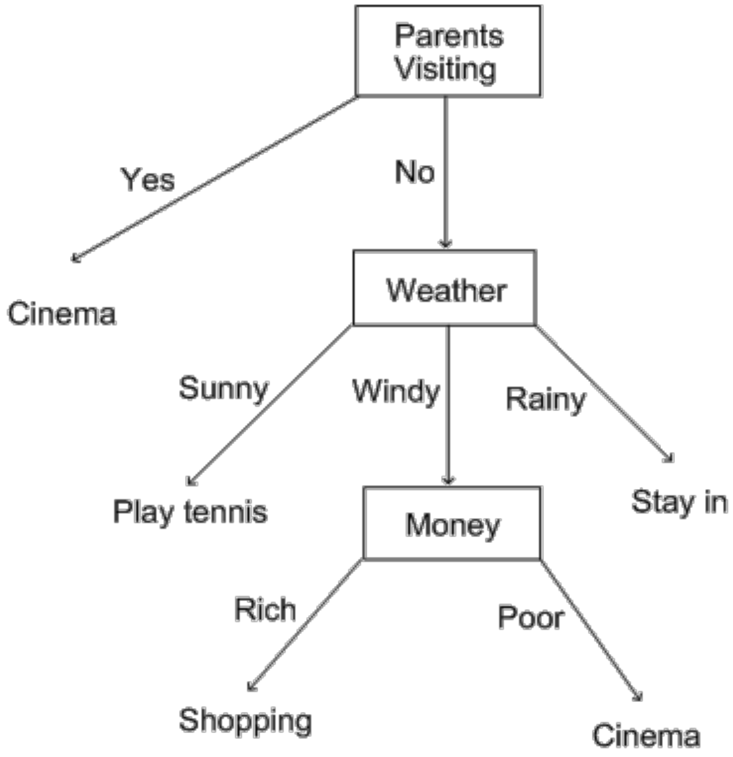
In this session we will learn about **tree based algorithms** which can be used for regression as well as classification

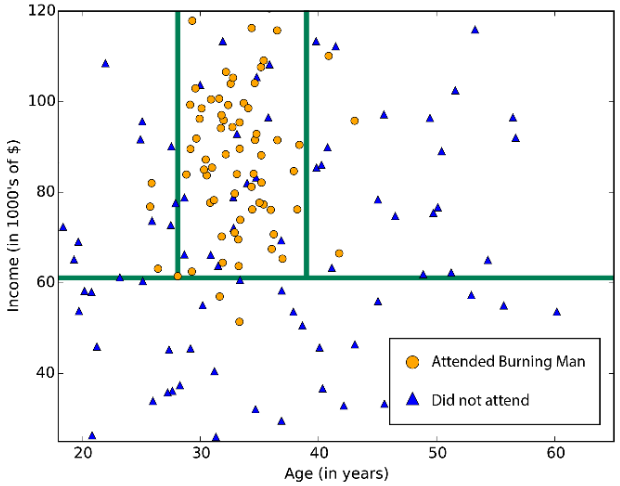
- Decision Trees
- Pruning and Bagging Trees
- Random Forests
- Gradient Boosting Method

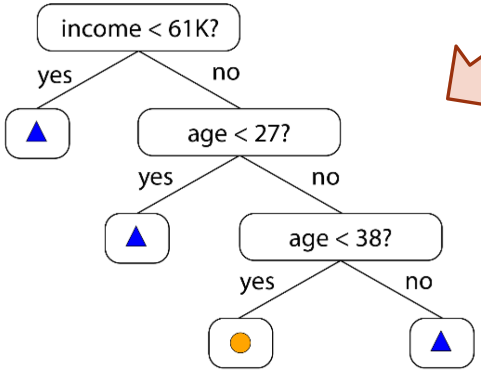


Decision Trees – Examples

How do you decide what to do on a Holiday ?





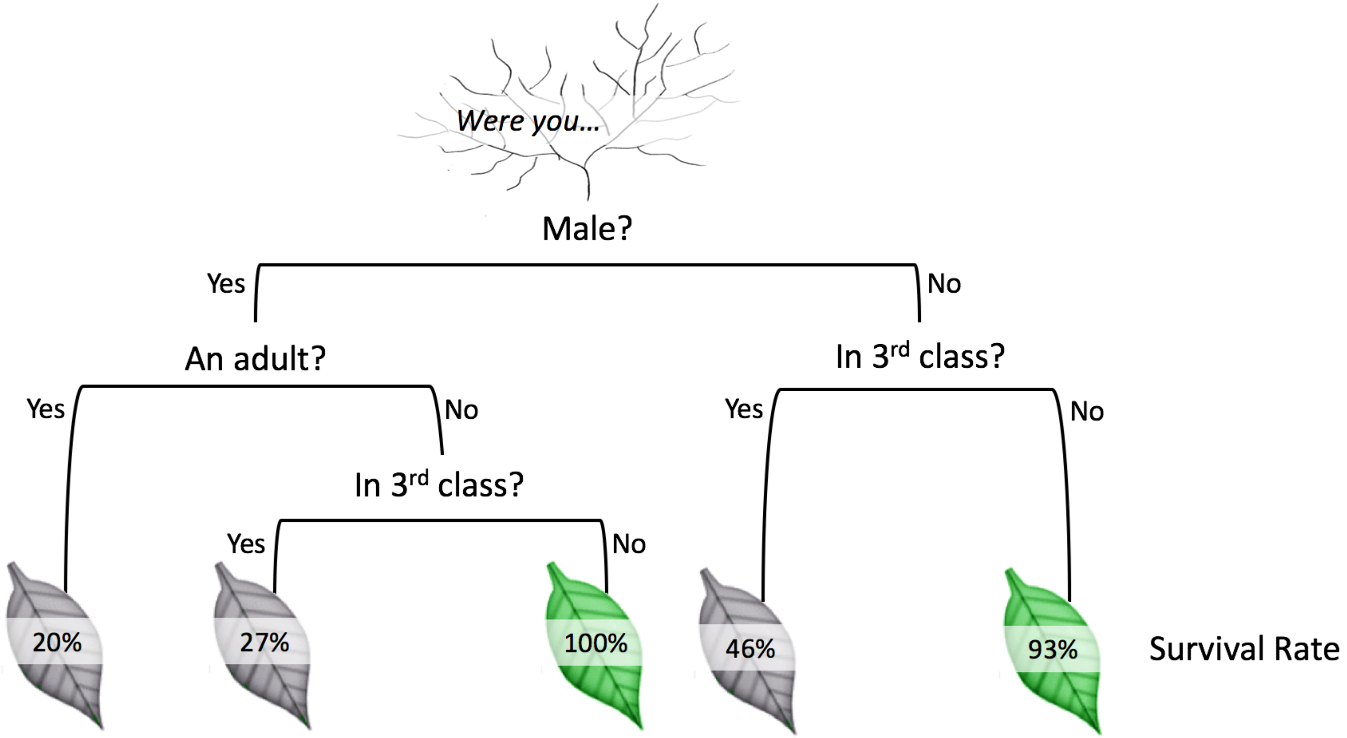


How do you know who attended Burning Man?

Decision Trees – Examples

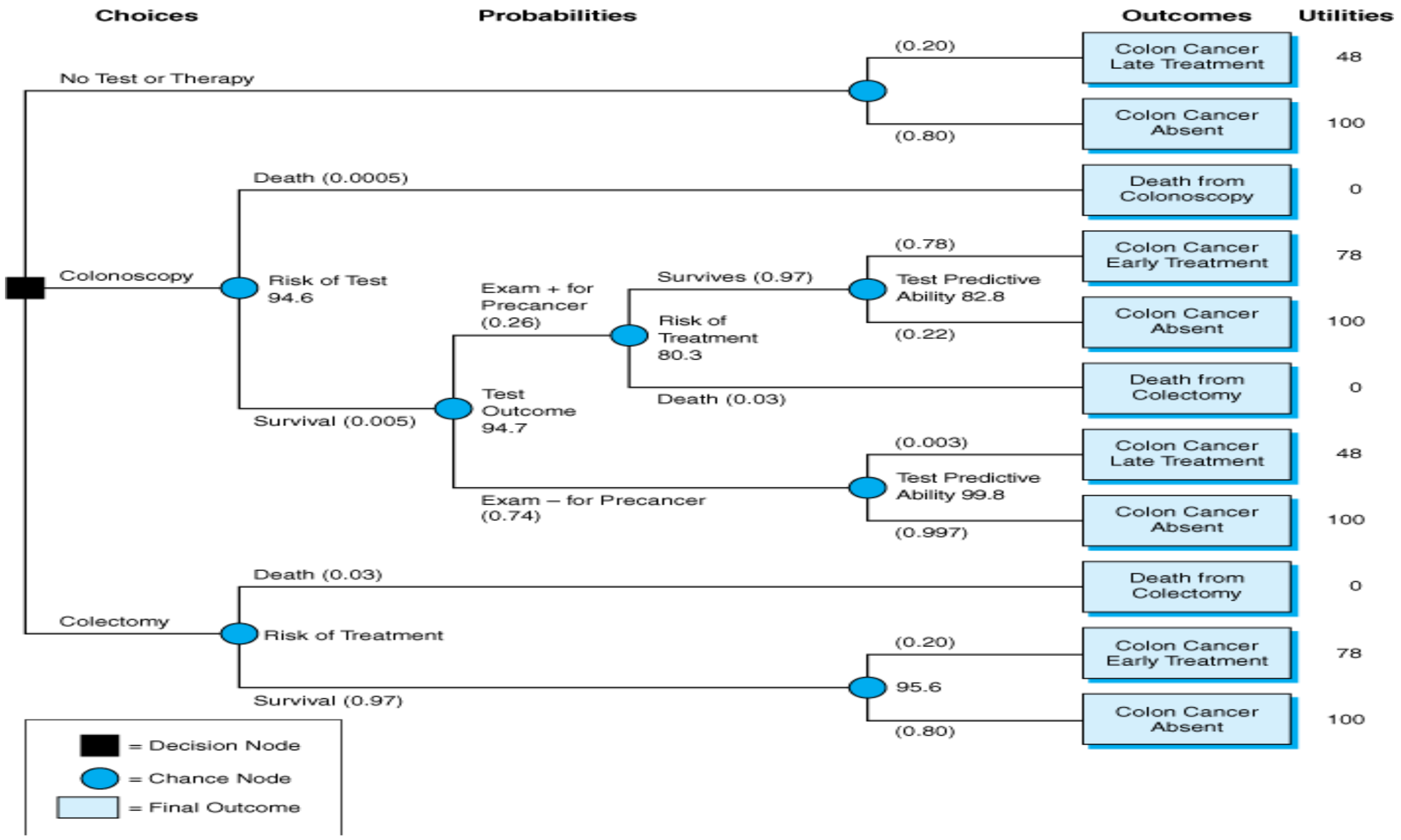
Would you survive if you were on the Titanic ?

Actual Kaggle Dataset



Decision Trees – Examples

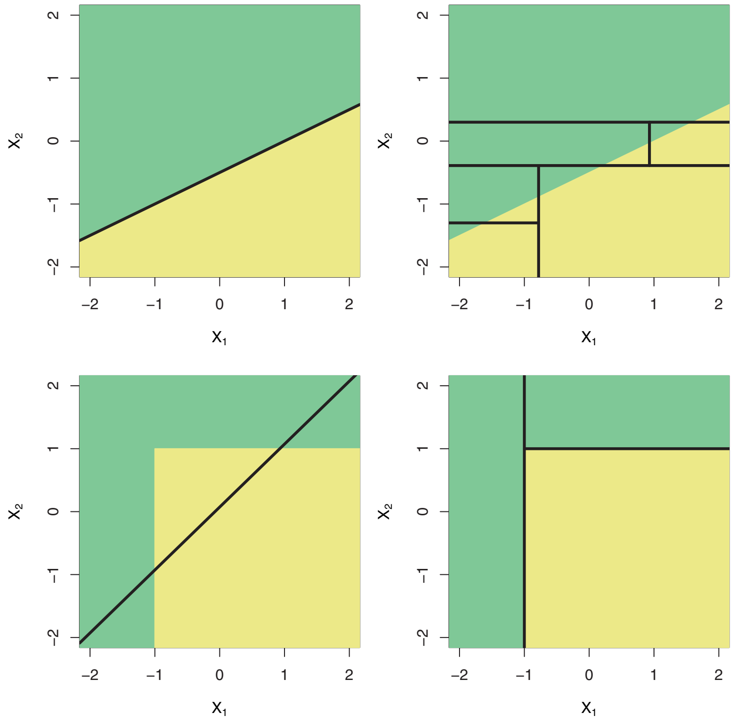
Even for serious Medical Diagnosis



Source: Dawson B, Trapp RG: *Basic & Clinical Biostatistics*, 4th Edition:
http://www.accessmedicine.com

Decision Trees – When?

When to use Decision Trees over the other Algorithms we have learnt ?

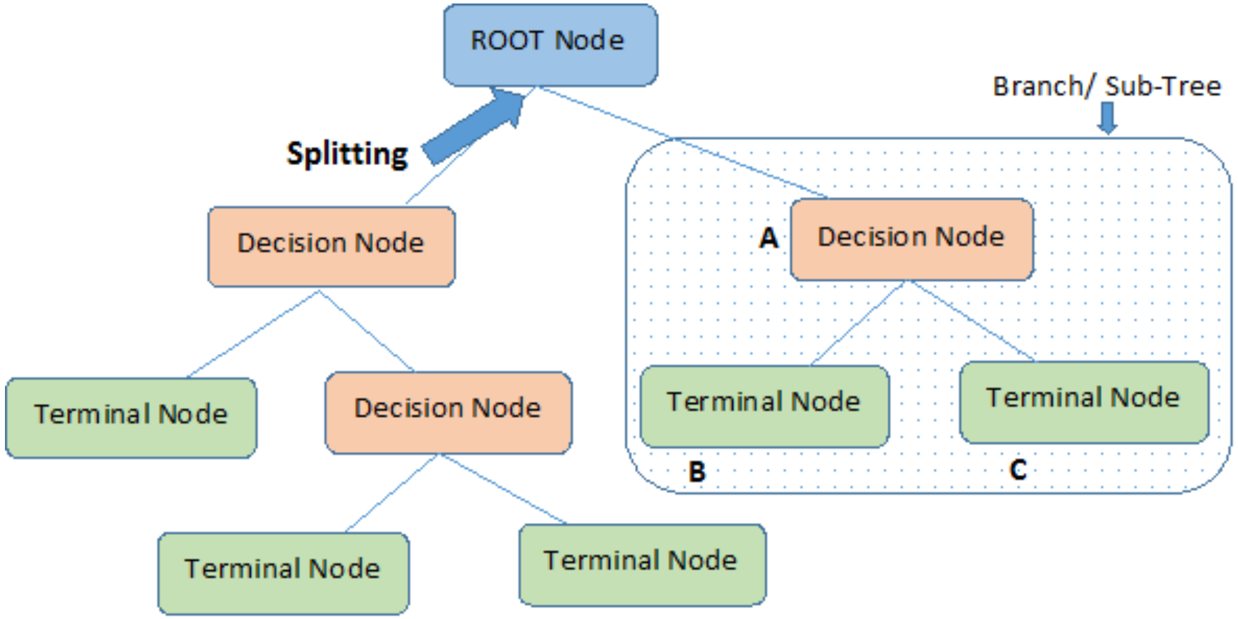


1. If the relationship between dependent & independent variable is well approximated by a linear model, linear regression will outperform tree based model.

2. If there is a high non-linearity & complex relationship between dependent & independent variables, a tree model will outperform a classical regression method.

3. If you need to build a model which is easy to explain to people, a decision tree model will always do better than a linear model. Decision tree models are even simpler to interpret than linear regression!

Decision Trees – How?

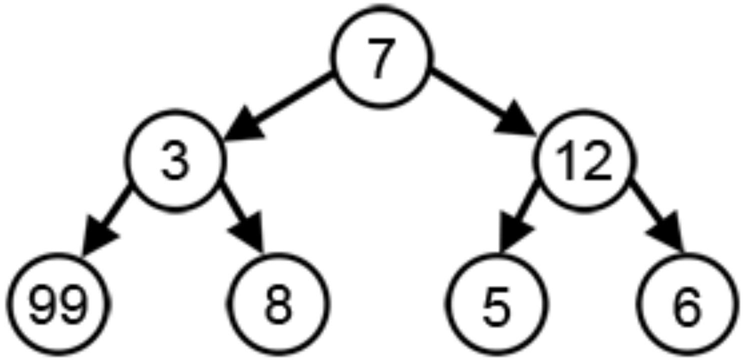


Note:- A is parent node of B and C.

1. Start at the Root Node with all the data
2. Split data based on a predictor using a decision criteria.
3. This produces two sub-nodes which are split further using a different attribute and only data assigned to that node.
4. Tree is grown until terminal or leaf nodes are produced which satisfy the criteria

Decision Trees – Greedy Search

A search method that makes the best locally optimum choices at each stage with the hope of finding a global optimum



Criteria here is to find the largest sum, but Greedy Algorithm makes the locally optimum choice.

Common criteria used in growing Trees

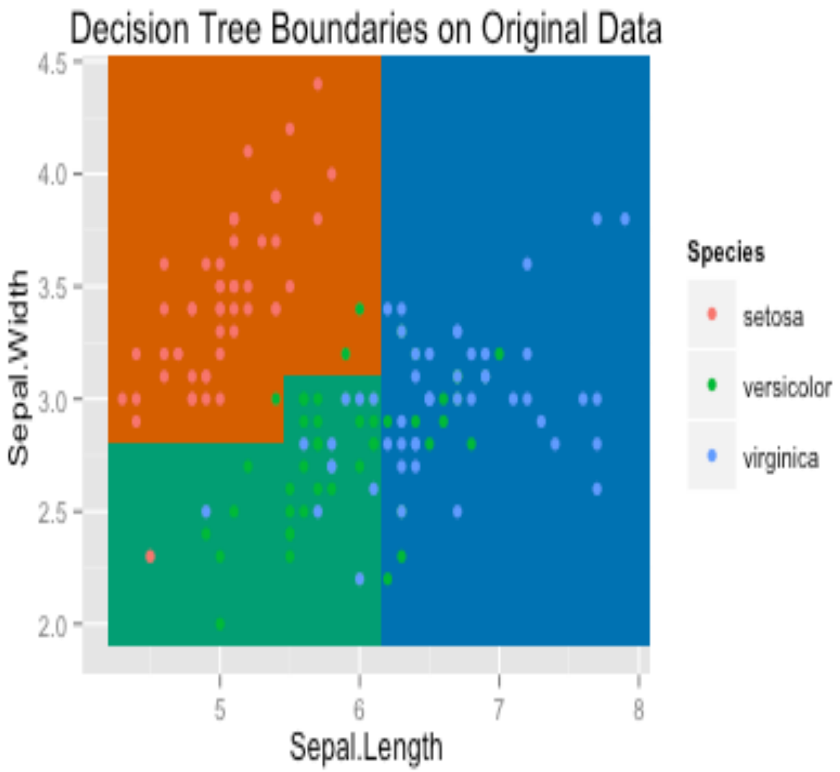
1. Minimum Classification Error
2. Maximum Information Gain
3. Least Entropy or Maximum Purity
4. Maximum Reduction in Variance
5. Depth of Tree
6. Minimum Observations at a node

Decision Trees – Advantages

1. Easy to Understand
2. Less Data Cleaning required compared to other models
3. Handles Quantitative as well as Categorical Variables
4. Non Parametric – No Mathematical assumptions about the data
5. Easily handle complex non-linear relationships in data
6. Easy to Visualize and Interpret
7. Rules based

Decision Trees – Disadvantages

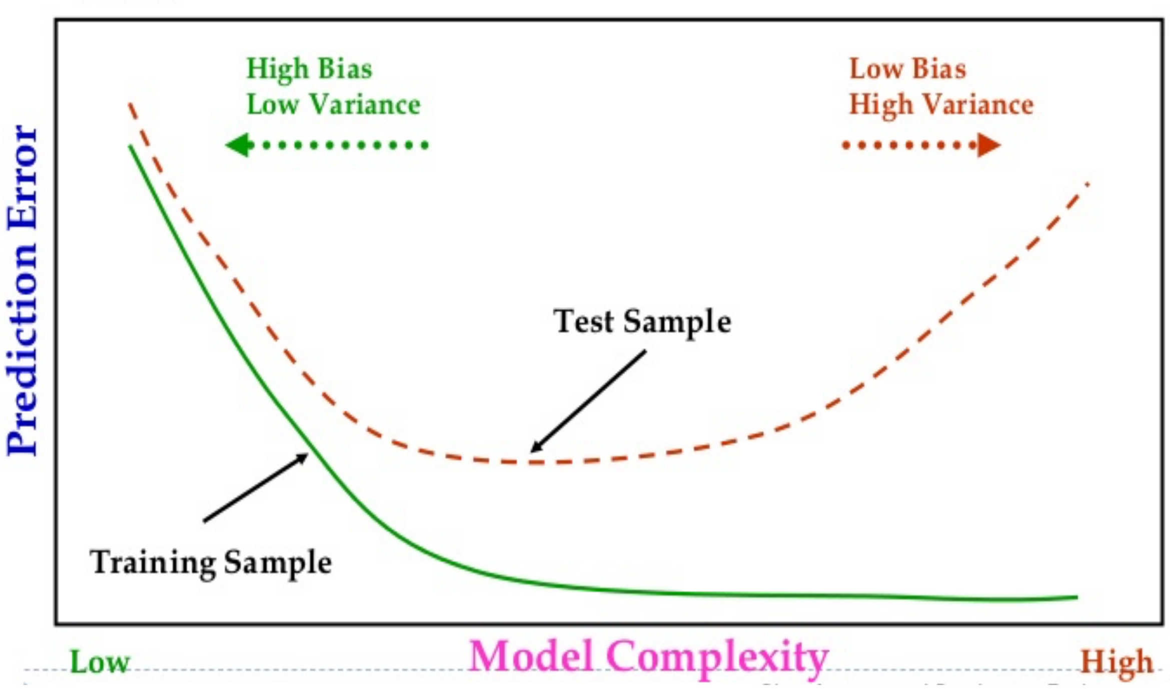
1. Large deep trees may suffer from over-fitting data
2. Once a mistake is made at higher level, sub trees can be wrong
3. Process of growing a tree is computationally expensive
4. Some information loss when handling continuous variables
5. Relies on rectangular approximation which might not be good for some datasets
6. Less Data and more classes can lead to over-fitting



Bias Variance Problem

Under fitting leads to overly simplified models which haven’t learned all the patterns in the data.

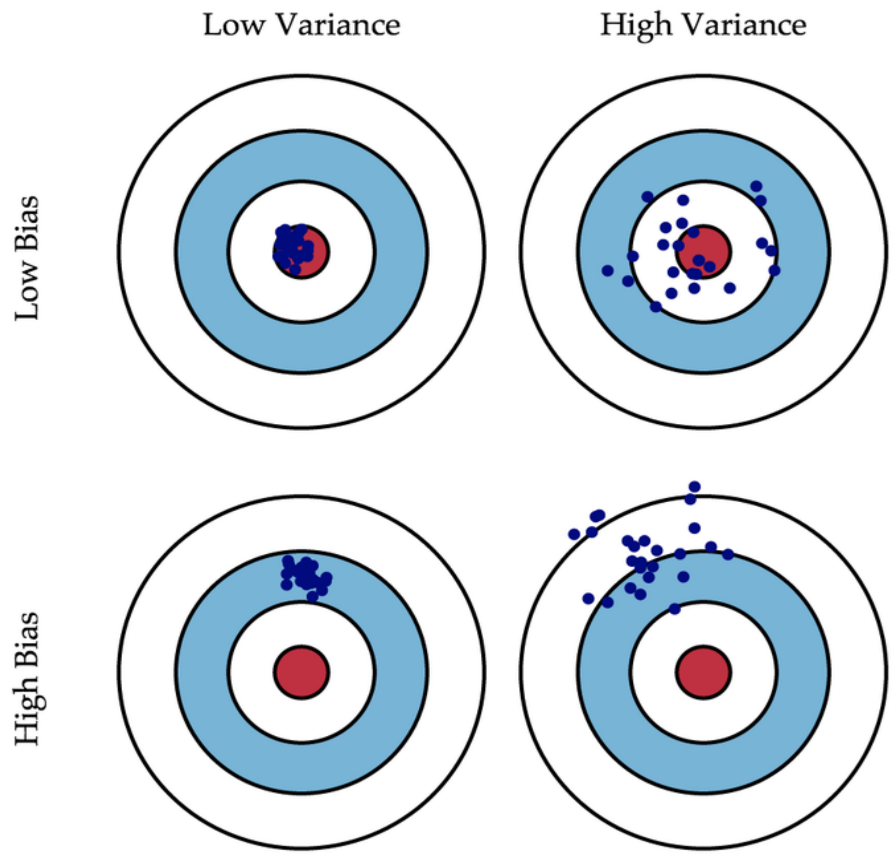
Stubs or one node trees are prone to this



Overfitting leads to increased model complexity which leads to a variance problem as the model learns the noise in the data.

Large trees are prone to this

Bias Variance Problem



To Reduce Variance

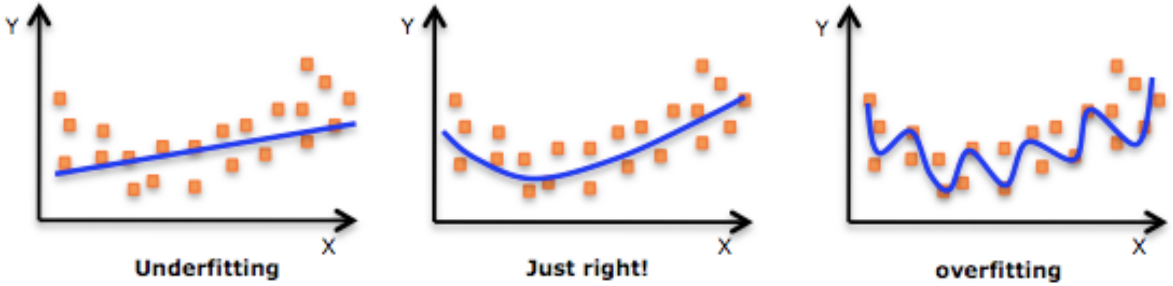
- Get more data
- Reduce Features
- Regularization
- Pruning
- Ensemble Models

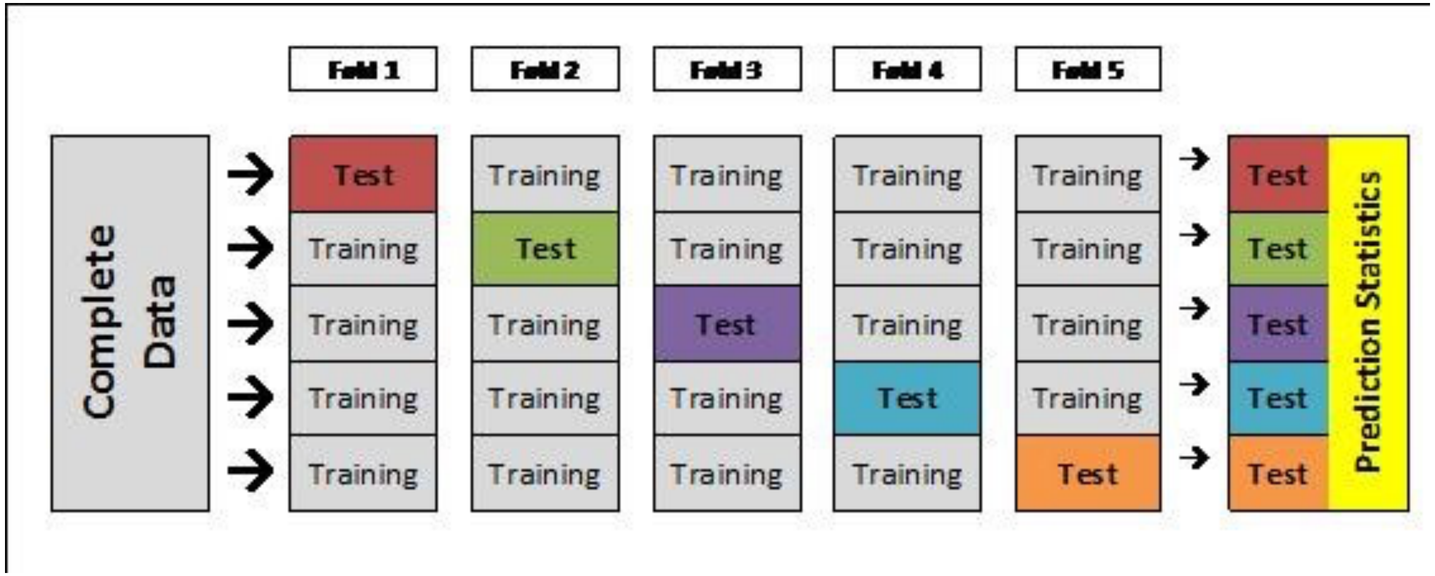
To Reduce Bias

- Obtain more features
- More data – but it doesn’t help beyond a point

The key is pay attention to difference between training set error and test/validation set error

Cross Validation





Complete Data	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Prediction Statistics
	Test	Training	Training	Training	Training	Test
	Training	Test	Training	Training	Training	Test
	Training	Training	Test	Training	Training	Test
	Training	Training	Training	Test	Training	Test
	Training	Training	Training	Training	Test	Test

Cross Validation is used for

1. Estimating Model error
2. Generalizing a model
3. Model Selection
4. Model Parameter Selection
5. A solution for over-fitting

Avoid the model learning the noise in the data by repeatedly testing it using “pretend test samples”

Pruning Trees

Pruning is replacing entire sub-trees with a leaf node if doing so improves validation error. This prevents over-fitting and helps generalize the model to unseen data.

Two approaches to Pruning

Pre-Pruning

Halt tree construction early and do not split a node if fit criteria is not met.

- Minimum observations in node
- Depth of Tree
- Threshold for information gain
- Threshold for classification error

Pre-Pruning is faster

Post-Pruning

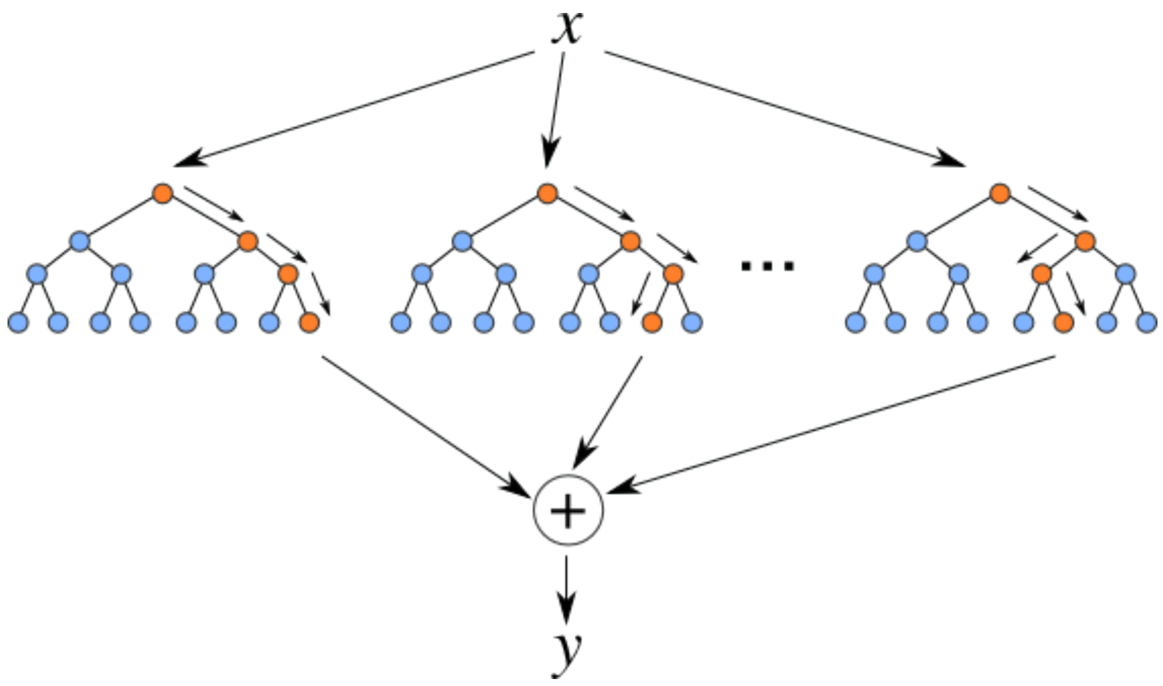
Grow the tree fully to **allow overfitting** and then prune

- Get a sequence of pruned trees by removing sub-trees/branches
- Use the validation set to decide which is “best pruned tree”

Post-Pruning is more accurate

Bagging Trees – Ensemble Learning

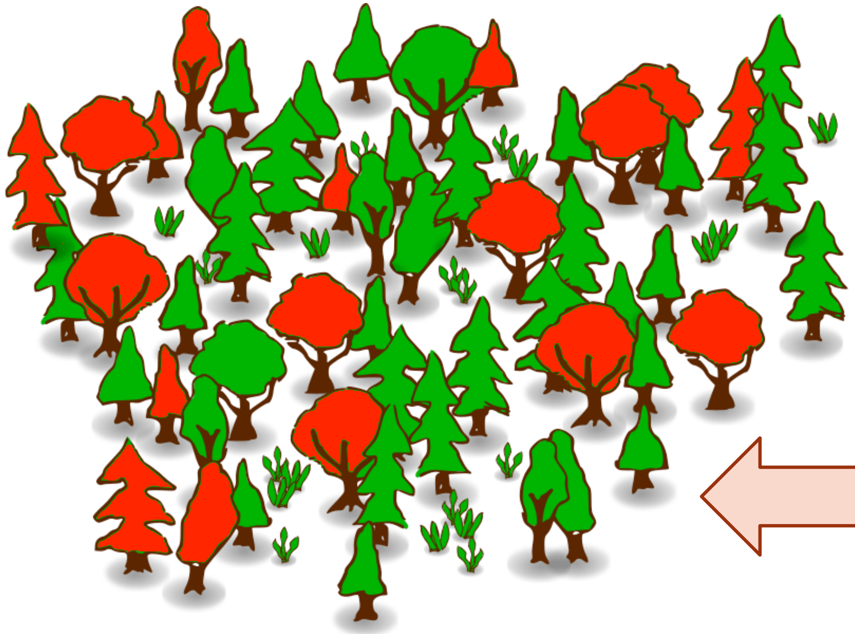
Another very popular cure for over-fitting/ high variance problem



- Randomly select samples from training data
- Fully grow trees for each sample
- Each such tree makes prediction on the test data
- Average the predictions (regression) or use the most popular vote (classification)

Random Forest

Random Forest is a more generalized version of the Bagging Algorithm



A diversified variety of trees are grown, which leads to less variance and more accuracy

- Randomly select samples from training data
- **Also randomly select number of predictors in each such sample**
- Fully grow trees for each sample
- Each such tree makes prediction on the test data
- Average the predictions (regression) or use the most popular vote (classification)

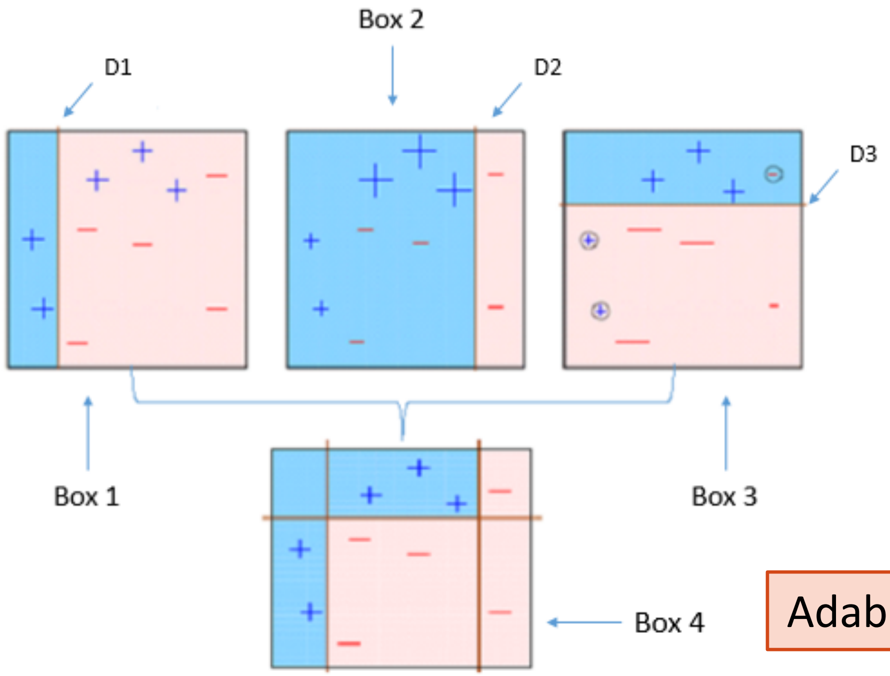
Random Forests

Advantages	Disadvantages
• Does Regression and Classification	• May over-fit particularly noisy data
• Power to handle large data set with high dimensionality	• In regression cannot predict beyond the limits of the data
• Handles missing data	• Is a bit of black box in what the model does
• Handles imbalanced classes data	
• Outputs importance of variables	

Boosting Algorithms

Boosting refers to this general problem of producing a very accurate prediction rule by combining rough and moderately inaccurate rules-of-thumb.

- Sequentially combines a bunch of weak learners to produce a strong learner.
- Instead of fully grown trees, decision stumps are produced
- Incorrectly classified observations by weak learners are given higher weight
- New weak learners are added focusing their ability on these misclassified observations



Adaboost

Gradient Boosting Machine

Very popular Machine Learning model. The XGBoost variant of this model wins quite a lot of Kaggle competitions

- Is a more generalized version of AdaBoost
- New weak learner focuses on the residual errors of the existing trees in the ensemble and tries to model them

Machine Learning leads to …

Leaf Classification Case

DSLA
DATA SCIENCE LEARNING ACADEMY

Team Introductions

Leaf Classification Case

Lunch

Leaf Classification Case

Demo Time

Course Assignments

Programming Assignments

Reading Assignments

Presentation Assignments

Technical Skills Assignments

Writing Assignments

Technical Assignment

Complete & Submit Code on GitHub for Mushroom Classification

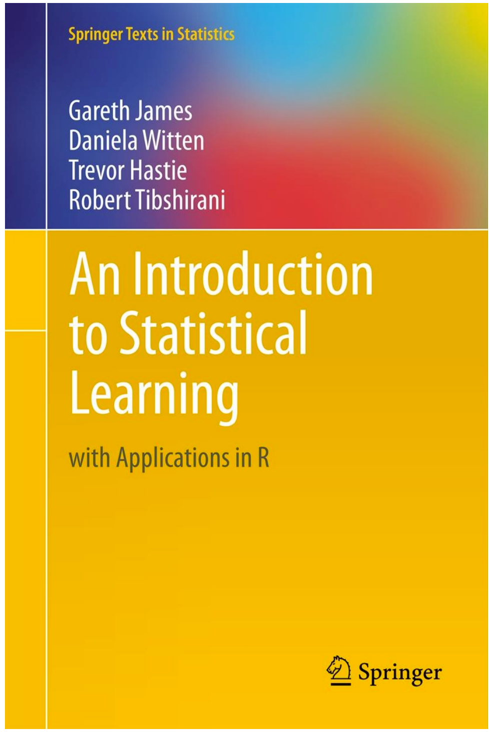
Programming Assignment

Install & Complete: Swirl – Regression Models

Install & Complete: Swirl – Getting & Cleaning Data

Reading Assignment

Read Chapter 4: Classification
Read Chapter 8: Tree Based Methods



Springer Texts in Statistics

Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani

An Introduction to Statistical Learning

with Applications in R

Springer

Writing Assignment

No Writing Assignment this week !!!

Presentation Assignment

By Saturday Submit

Your Presentations on Mushroom Classification Case

1. Technical Presentation
2. Business Presentation (Not to exceed 5 slides)