GENERAL ASSEMBLY

# DATA SCIENCE

## 10 WEEK PART TIME COURSE

## Week 6 - Decision Trees
## Tuesday 27th June 2017

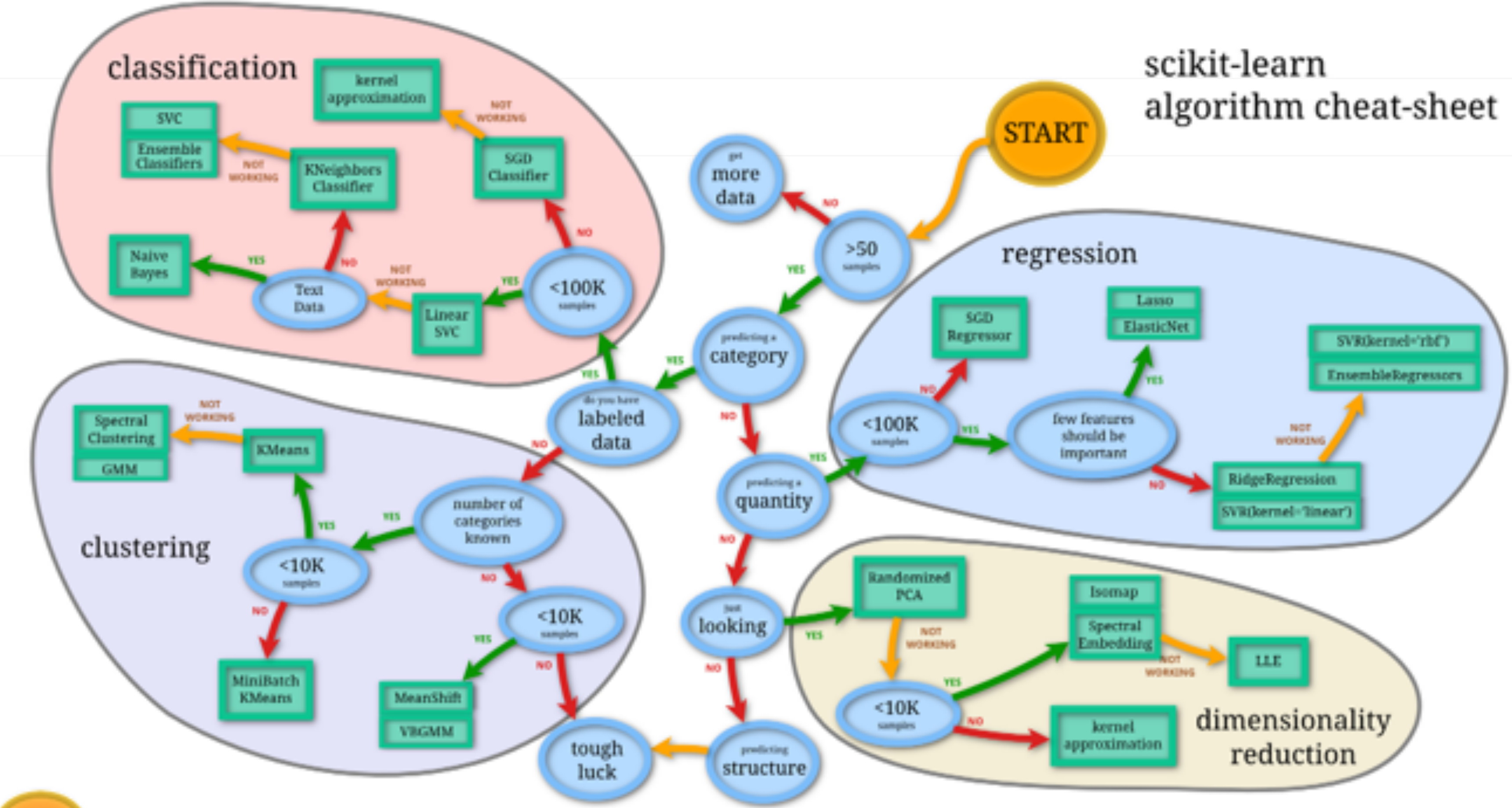
# AGENDA

1. What are decision trees?
2. How decision trees work
3. Visual example
4. Lab
5. Discussion

DATA SCIENCE PART TIME COURSE

# DECISION TREES

# scikit-learn algorithm cheat-sheet



Back

# WHAT ARE DECISION TREES?

- A supervised learning technique that can be used for classification or regression.

# WHAT ARE DECISION TREES?

- A supervised learning technique that can be used for classification or regression.
- Visually engaging and easy to interpret.

# WHAT ARE DECISION TREES?

- A supervised learning technique that can be used for classification or regression.
- Visually engaging and easy to interpret.
- Foundation for getting into very powerful techniques.

# WHAT ARE DECISION TREES?

- A supervised learning technique that can be used for classification or regression.
- Visually engaging and easy to interpret.
- Foundation for getting into very powerful techniques.
- Great for explaining to people!

## DECISION TREES - SHORTCOMINGS

- Prone to overfitting.

# DECISION TREES - SHORTCOMINGS

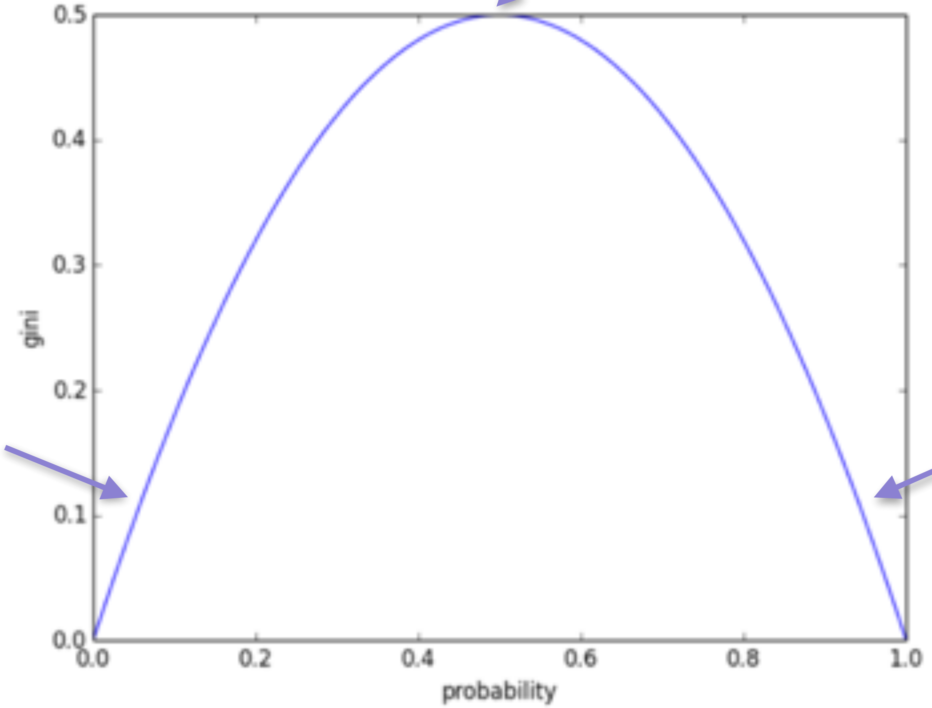
- Prone to overfitting.
- Predictive power is lower in comparison to many other modern techniques.

# DECISION TREES - HOW IT WORKS

- Scans for a feature to split on that results in the greatest separation between classes in the resulting nodes.

# DECISION TREES – HOW IT WORKS

## The Gini Index



Equal ratio of target classes 50:50

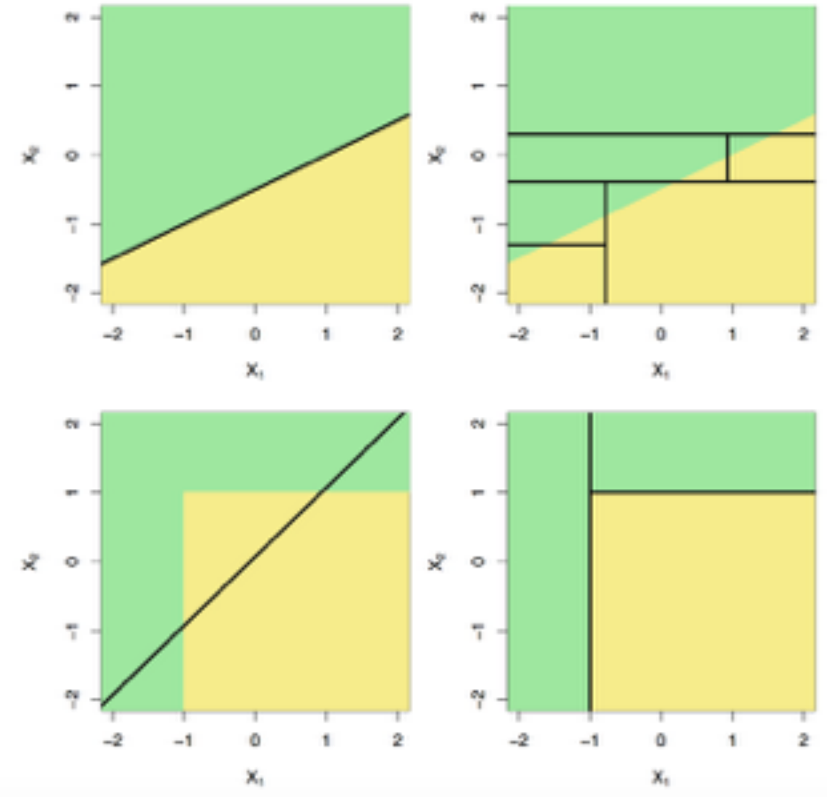
High purity of class 0

High purity of class 1

# DECISION TREES - HOW IT WORKS

- Scans for a feature to split on that results in the greatest separation between classes in the resulting nodes.
- Non-linear.

# DECISION TREES - HOW IT WORKS



Linear decision boundary

Non-linear decision boundary

# DECISION TREES - HOW IT WORKS

- Scans for a feature to split on that results in the greatest separation between classes in the resulting nodes.
- Non-linear
- Greedy process
- Splits within splits

# DECISION TREES - HOW IT WORKS

- Scans for a feature to split on that results in the greatest separation between classes in the resulting nodes.
- Non-linear
- Greedy process
- Splits within splits
- For a classification tree, we predict that each observation belongs to the most commonly occurring class of training observations in the region to which it belongs.

# DECISION TREES - HOW IT WORKS

- Scans for a feature to split on that results in the greatest separation between classes in the resulting nodes.
- Non-linear
- Greedy process
- Splits within splits
- For a classification tree, we predict that each observation belongs to the most commonly occurring class of training observations in the region to which it belongs.
- We naturally get combinations of features used for our prediction.

# VISUALISING A DECISION TREE FIT

http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

DATA SCIENCE PART TIME COURSE

# LAB

# SYNCHING YOUR FORK WITH THE COURSE REPO

1. re-name your labs with lab_name.<yourname>.ipynb  (to prevent a conflict)
2. cd <path to the root of your SYD_DAT_8 local repo>
3. commit your changes ahead of sync
   - git status
   - git add .
   - git commit -m “descriptive label for the commit”
   - git status
4. download new material from official course repo (upstream) and merge it
   - git checkout master  (ensures you are in the master branch)
   - git fetch upstream
   - git merge upstream/master

DATA SCIENCE - Week 6 Day 1

# DISCUSSION TIME

- Review of last week
- Further Reading for Decision Trees
- Check in with homework/course project
- Pre-Reading

WEEK 5 Review

# DISCUSSION TIME

- Recommendations
- SQL & Productivity Tools
- Linear Regression

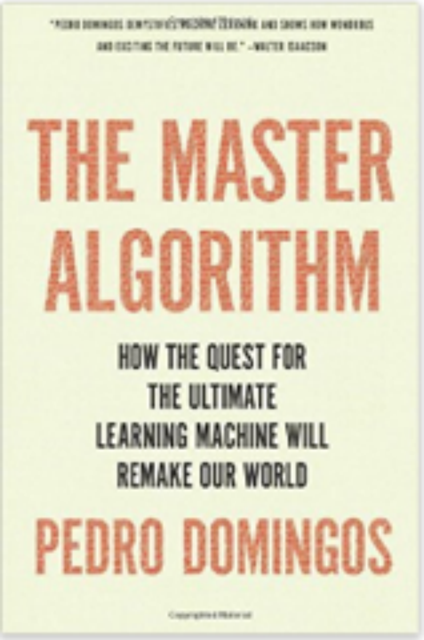
# DISCUSSION TIME

CIA using Decision Trees

- https://www.cia.gov/library/center-for-the-study-of-intelligence/kent-csi/vol18no4/html/v18i4a03p_0001.htm

Overview of difference approaches to building models

- https://www.amazon.com/Master-Algorithm-Ultimate-Learning-Machine/dp/0465065708


"PEDRO DOMINGOS DEMYSTIFIES MACHINE LEARNING AND SHOWS HOW WONDROUS AND EXCITING THE FUTURE WILL BE." —WALTER ISAACSON
THE MASTER ALGORITHM
HOW THE QUEST FOR THE ULTIMATE LEARNING MACHINE WILL REMAKE OUR WORLD
PEDRO DOMINGOS

# DISCUSSION TIME

Homework/Course Project

- How’s Homework 2 going ?
- How are the projects going?

DATA SCIENCE – Week 6 Day 1

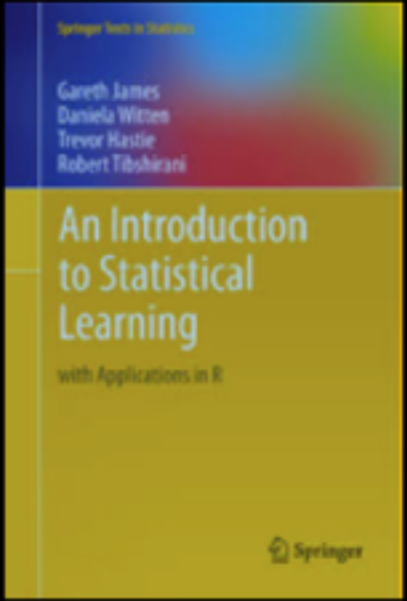
# PRE-READING

An Introduction to Statistical Learning

- Chapter 8 – Tree-Based Methods

Signup to Google Cloud Platform

- https://cloud.google.com/



Springer Texts in Statistics

Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani

An Introduction to Statistical Learning

with Applications in R

Springer