GENERAL ASSEMBLY

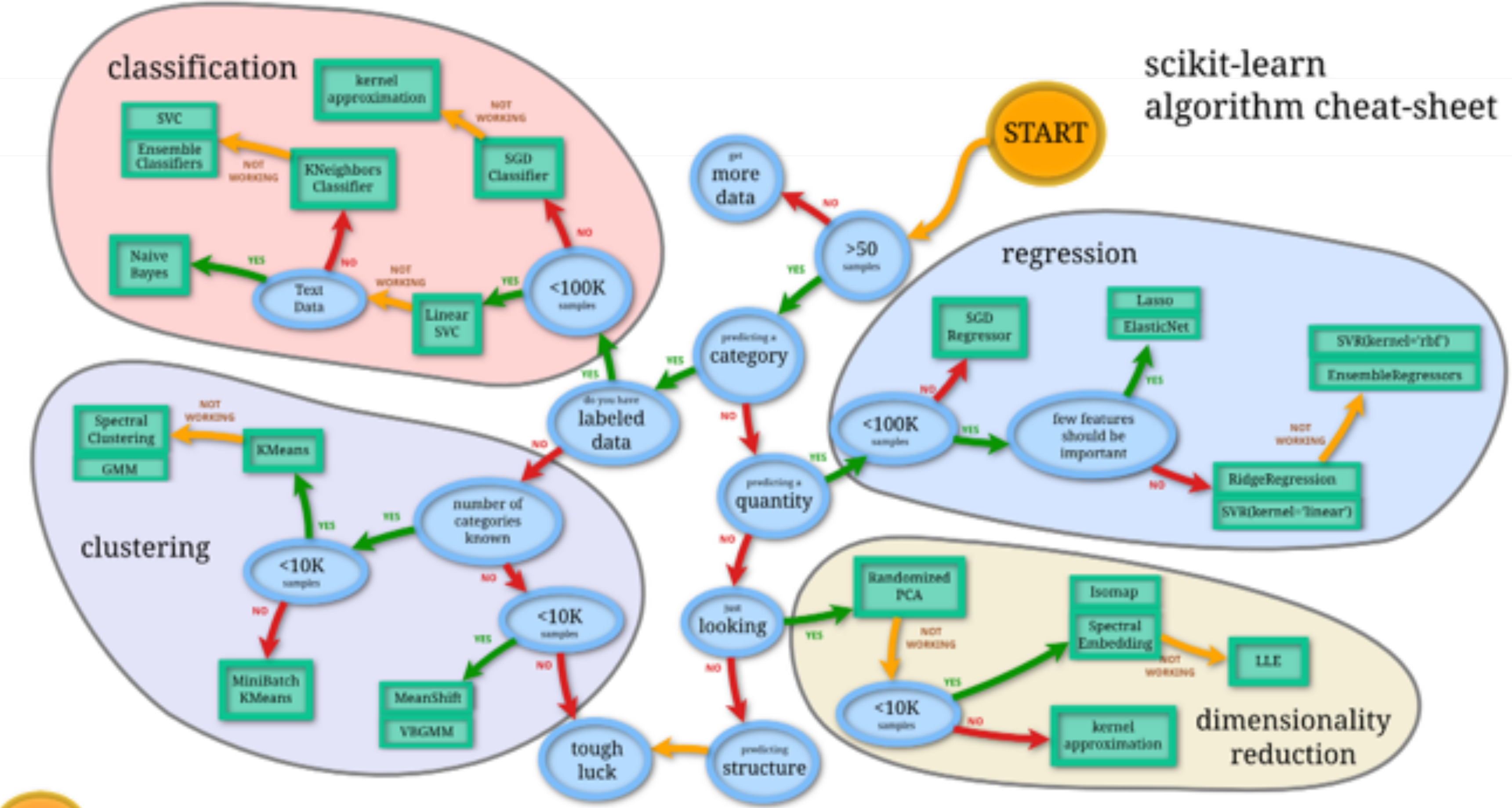
DATA SCIENCE

11 WEEK PART TIME COURSE

Week 3 - Logistic Regression
Monday 4th January 2016

AGENDA

1. Motivation
2. What is Logistic Regression?
3. Why use Logistic Regression
4. Lab
5. Homework Review



scikit-learn algorithm cheat-sheet

SUPERVISED LEARNING - REGRESSION & CLASSIFICATION

If the y variable is numeric then we have a regression problem - we are trying to predict a continuous number

If the y variable is a category (for example trying to predict a type of flower) the we have a classification problem - we are trying to classify what group that y belongs to.

DATA SCIENCE PART TIME COURSE

WHAT IS LOGISTIC REGRESSION?

We want to build a classifier that correctly identifies which class our target variable y belongs to given our input variable x.

Why not use the linear regression model?

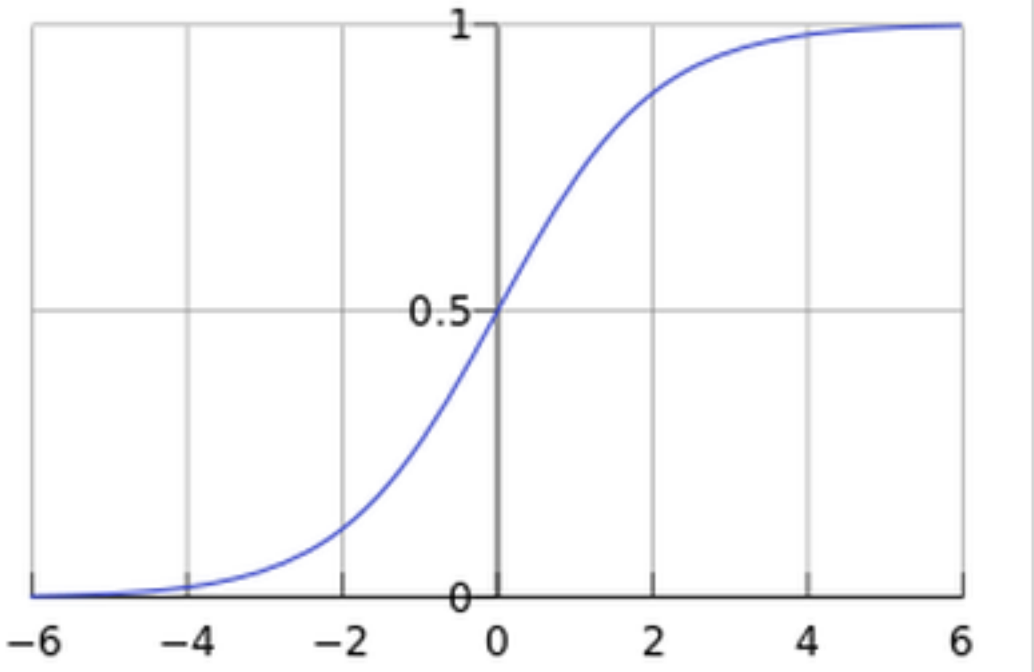
$$y = X\beta + \epsilon$$

- If we only have a binary response variable (0 or 1) it might make sense… BUT we can have our estimated value of $y > 1$ or $y < 0$ … which doesn't make sense.
- What of the case where we have more than one class? Linear regression cannot easily handle these cases.
- We want a classification method that can handle these cases and give us results we can easily interpret.

$$\mathbf{p(Y=1|X) = \beta_0 + \beta_1 X.}$$

- This is a good starting point but we still have the problem of p(Y) being outside the 0,1 range.
- We need to model $p(Y=1|X)$ using a function that gives outputs between 0 and 1.
- Basically we want something that looks like the following

LOGISTIC REGRESSION



Standard logistic sigmoid function

LOGISTIC REGRESSION

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

- This is the logit function,
- We can see that it this function is linear in X
- $\frac{p}{1-p}$ is called the ‘odds’ and can be any value from 0 to ∞
- $\log\left(\frac{p}{1-p}\right)$ is called the ‘log-odds’ or ‘logit’

LOGISTIC REGRESSION – WALKTHROUGH NOTEBOOK

- We will step through a notebook together and cover these concepts in a more tangible way.

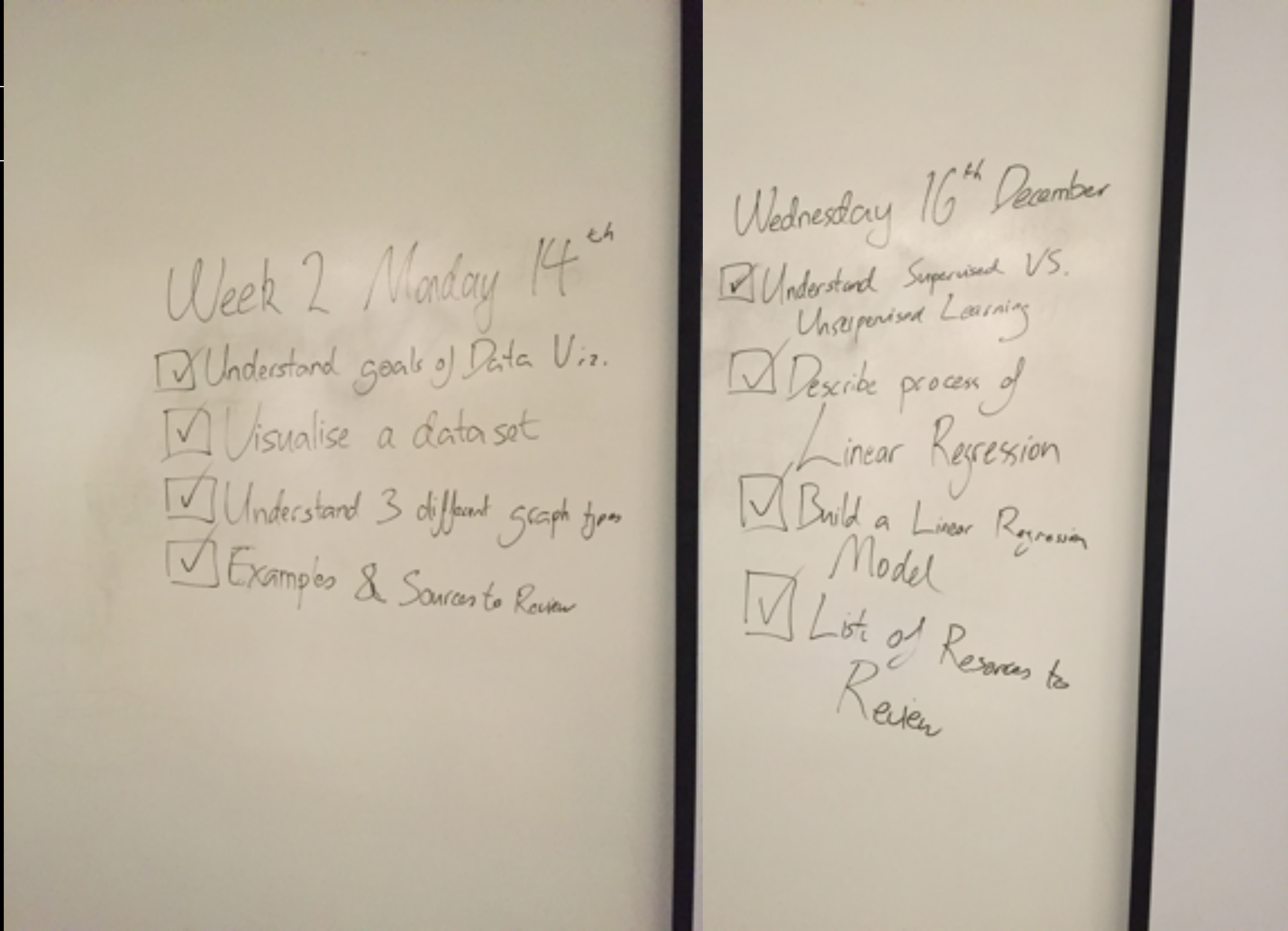
DATA SCIENCE PART TIME COURSE

LAB

DISCUSSION TIME

- Review of last week
- Further Reading for Logistic Regression
- Check in with homework/course project

WEEK 2



Week 2 Monday 14th

- ☑ Understand goals of Data Viz.
- ☑ Visualise a dataset
- ☑ Understand 3 different graph types
- ☑ Examples & Sources to Review

Wednesday 16th December

- ☑ Understand Supervised VS. Unsupervised Learning
- ☑ Describe process of Linear Regression
- ☑ Build a Linear Regression Model
- ☑ List of Resources to Review

DATA SCIENCE – Week 3 Day 1

DISCUSSION TIME

An Introduction to Statistical Learning

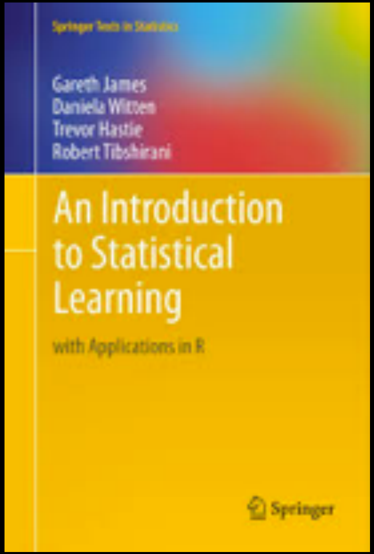
- Chapter 4 – Logistic Regression

Logistic Regression applied to loan applications

- https://github.com/nborwankar/LearnDataScience

Odds Ratio in Logistic Regression

- http://www.ats.ucla.edu/stat/mult_pkg/faq/general/odds_ratio.htm



Springer Texts in Statistics

Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani

An Introduction to Statistical Learning

with Applications in R

Springer

DATA SCIENCE – Week 3 Day 1

HOMEWORK 1

Highlights

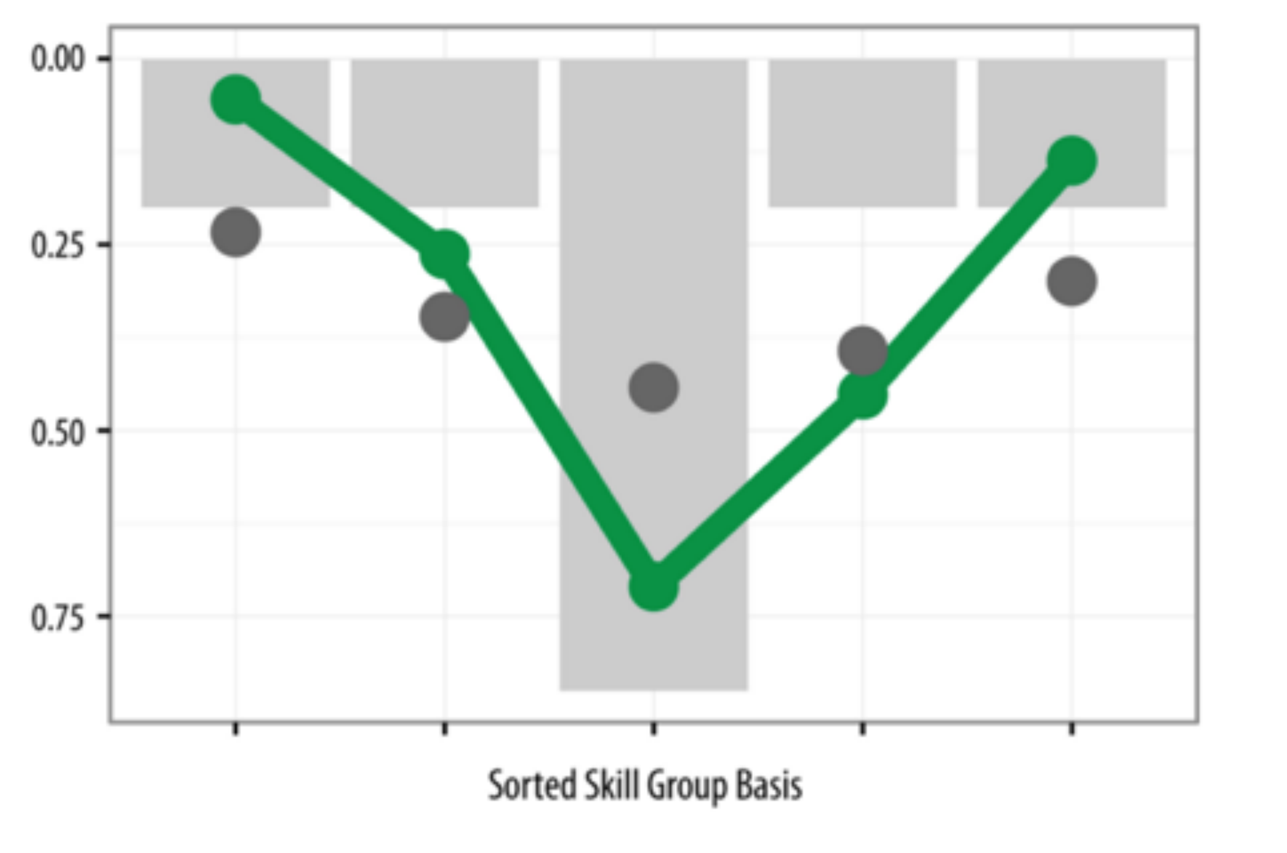
- *Experience is key in developing your skill set as a data scientist. You will become better with time, industry exposure, and proper mentoring* – Michael
- *He also advises taking extra time to learn and develop skills, not only by studying but also by engaging on side projects and working with people that know about the field* – Sofia
- *Design pathway for the career development, so they can work effectively* – Xueyuan

HOMEWORK 1

Highlights

- ***Data Scientist are generally T shaped employees with a range of skills and indepth knowledge as shown in the data.*** **- Adrian**
- ***This paper favours a T-shaped approach for building a career in data science. That is to say, a data scientist should have a breadth of skills as well as expertise in at least one aspect in data science.*** **- Hans**
- ***This would mean that to be truly successful you would need a combination of different T shaped data scientists*** **- Angus**

DATA SCIENCE – Week 3 Day 1



HOMEWORK 1

Notes

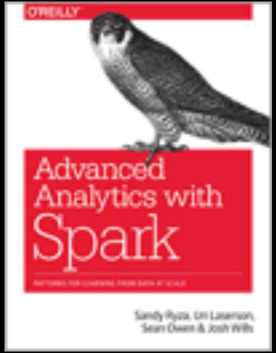
- Code Legibility is important. Anyone should be able to read your code at a later stage and make sense of it (this includes you in the future)
- Have a look at this style guide and steal some of the ideas https://google.github.io/styleguide/pyguide.html
- *Good comments don't repeat the code or explain it. They clarify its intent. Comments should explain, at a higher level of abstraction than the code, what you're trying to do.* - Code Complete, McConnell

DATA SCIENCE – Week 3 Day 1

HOMEWORK 1

Congratulations Michael !!!

O'REILLY
Becoming a Better Programmer
A HANDBOOK FOR PEOPLE WHO CARE ABOUT CODE
Pete Goodliffe


O'REILLY
Advanced Analytics with Spark
PATTERNS FOR LEARNING FROM DATA AT SCALE
Sandy Ryza, Uri Laserson, Sean Owen & Josh Wills

O'REILLY
Early Release
RAW & UNEDITED
Data Wrangling with Python
TIPS AND TOOLS TO MAKE YOUR LIFE EASIER
Jacqueline Kazil & Katharine Jarmul

DATA SCIENCE – Week 3 Day 1

DISCUSSION TIME

Homework/Course Project

- How’s Homework 2 going ?
- Did anyone make progress with their project over the break?
- After this week we will have the foundation for entering a kaggle competition, who’s setup an account?