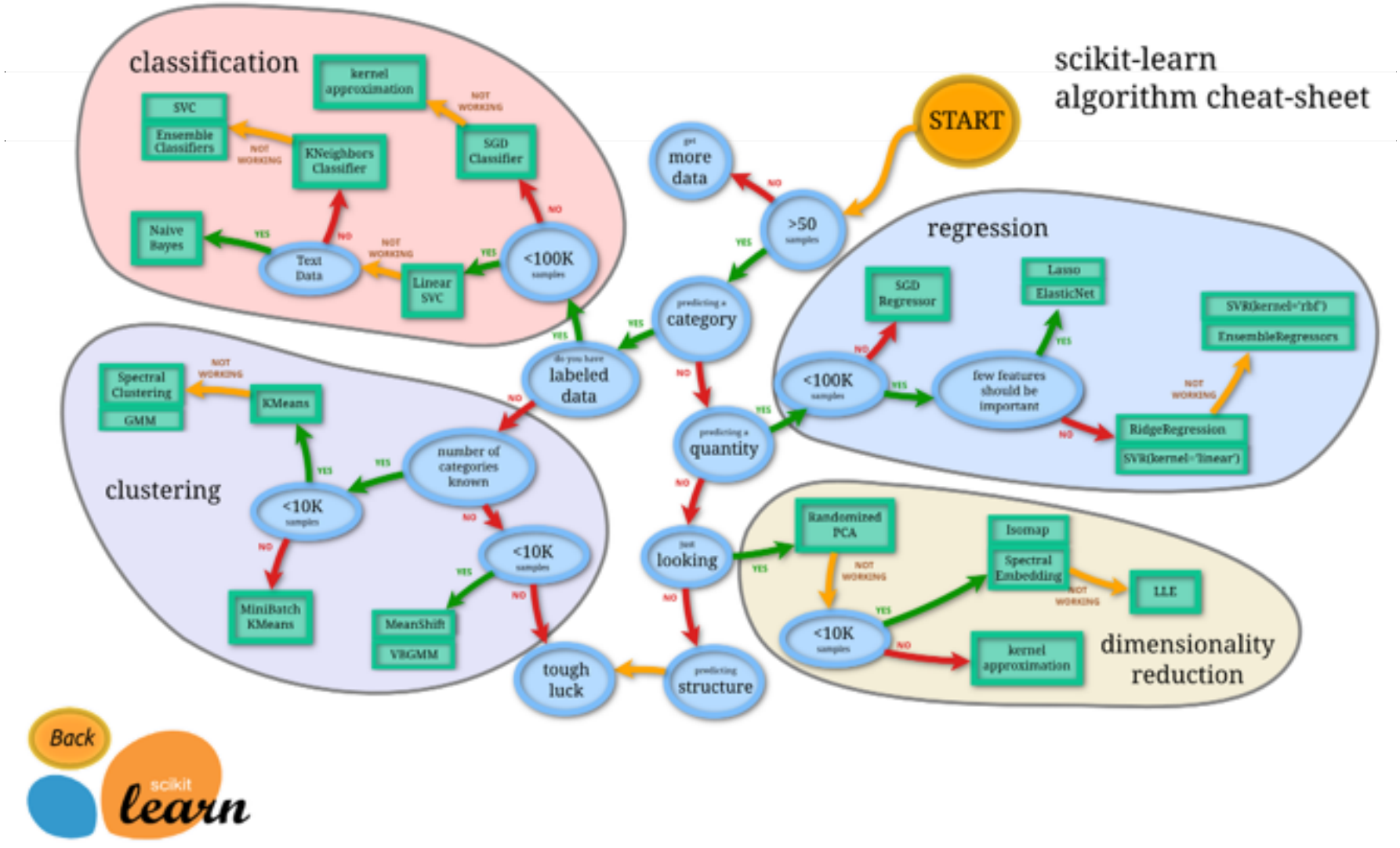
GENERAL ASSEMBLY

DATA SCIENCE
11 WEEK PART TIME COURSE

Week 3 – Logistic Regression
Wednesday 6th April 2016

AGENDA

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


scikit-learn
algorithm cheat-sheet
START
get more data
NO
>50 samples
YES
classification
kernel approximation
NOT WORKING
SVC
Ensemble Classifiers
NOT WORKING
KNeighbors Classifier
SGD Classifier
NO
Naive Bayes
YES
Text Data
NO
NOT WORKING
Linear SVC
YES
<100K samples
NO
regression
SGD Regressor
Lasso
ElasticNet
SVR(kernel='rbf')
EnsembleRegressors
<100K samples
NO
YES
few features should be important
YES
NOT WORKING
RidgeRegression
NO
SVR(kernel='linear')
YES
predicting a category
YES
do you have labeled data
NO
predicting a quantity
YES
NO
clustering
Spectral Clustering
GMM
NOT WORKING
KMeans
YES
number of categories known
YES
<10K samples
NO
NO
<10K samples
YES
MiniBatch KMeans
MeanShift
VBGMM
NO
just looking
YES
Randomized PCA
NOT WORKING
Isomap
Spectral Embedding
NOT WORKING
LLE
YES
<10K samples
NO
kernel approximation
dimensionality reduction
NO
predicting structure
tough luck
Back
scikit learn

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.

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.

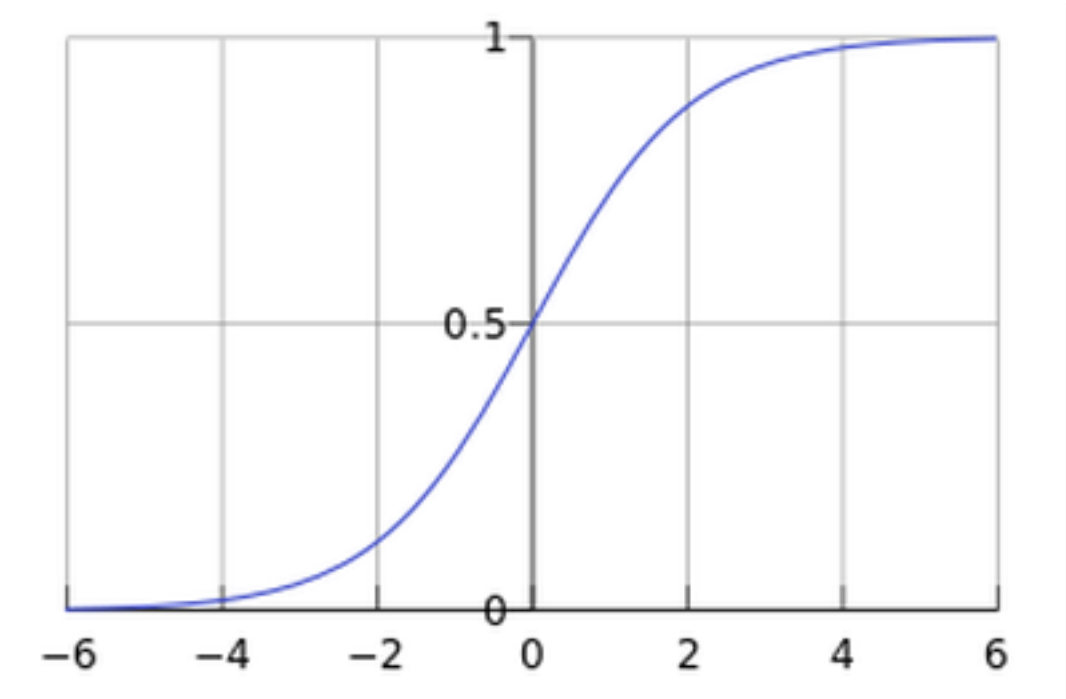
Why not use the linear regression model?

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

- 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.

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

- 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



Standard logistic sigmoid function

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

- 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'

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

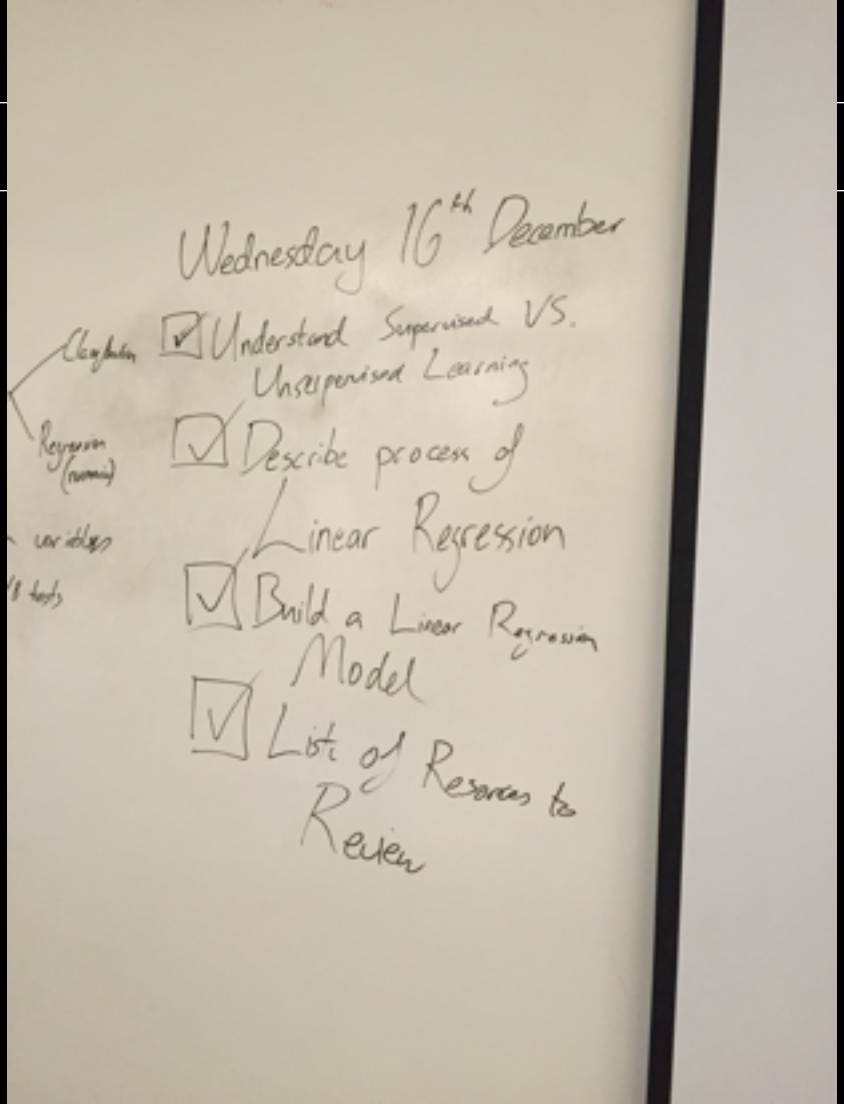

LAB

DATA SCIENCE – Week 3 Day 2

DISCUSSION TIME

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

WEEK 3


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 2

DISCUSSION TIME

An Introduction to Statistical Learning

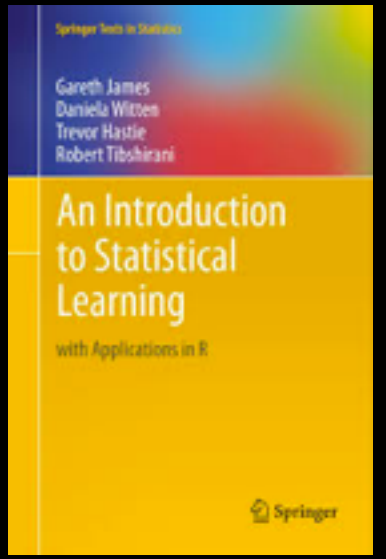
- Chapter 4 – Logistic Regression

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



DISCUSSION TIME

Homework/Course Project

- How’s Homework 1 going ?