GENERAL ASSEMBLY

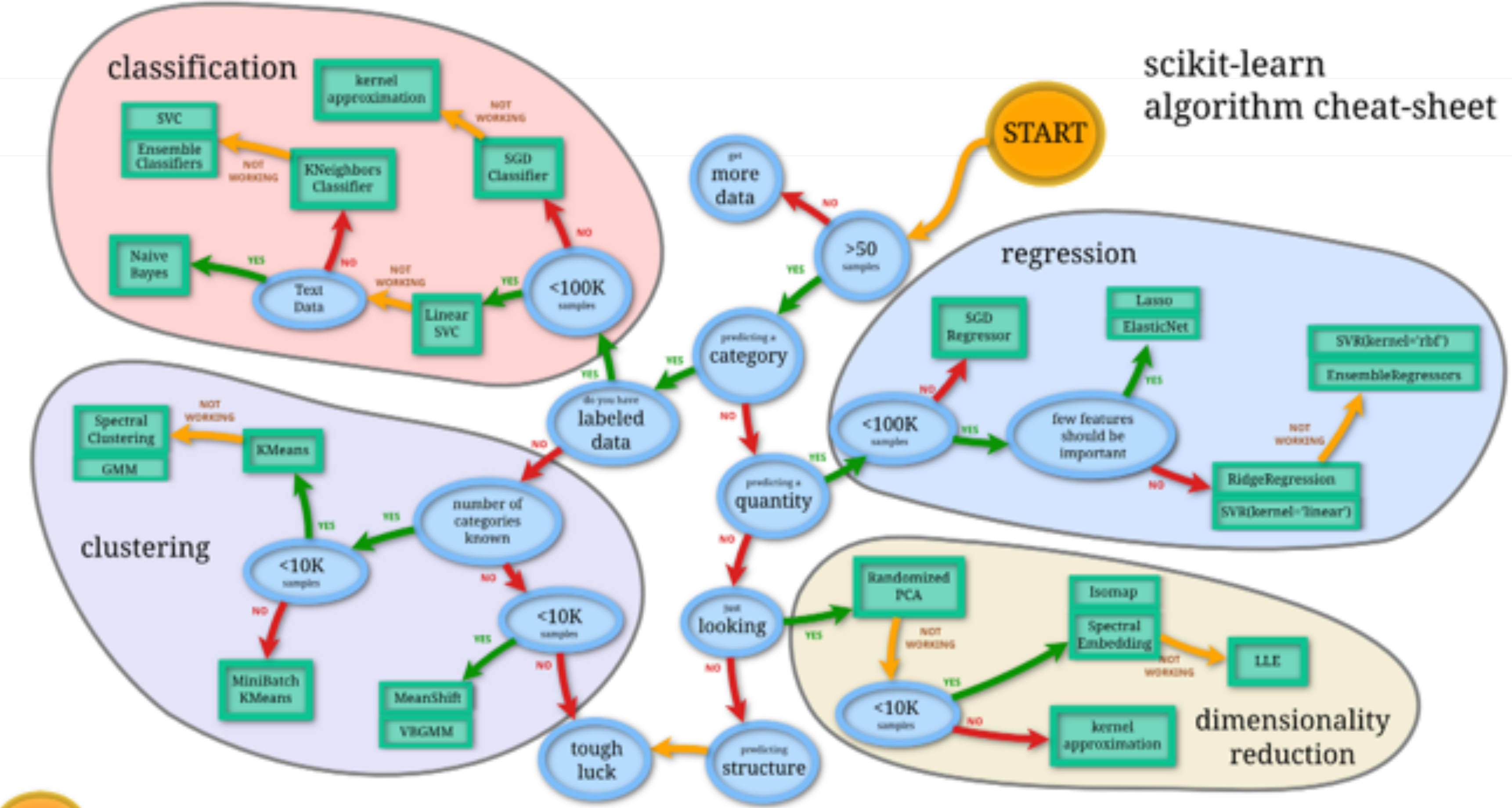
DATA SCIENCE

10 WEEK PART TIME COURSE

Week 3 - Logistic Regression
Tuesday 6th June 2017

AGENDA

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



Back

scikit learn

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?

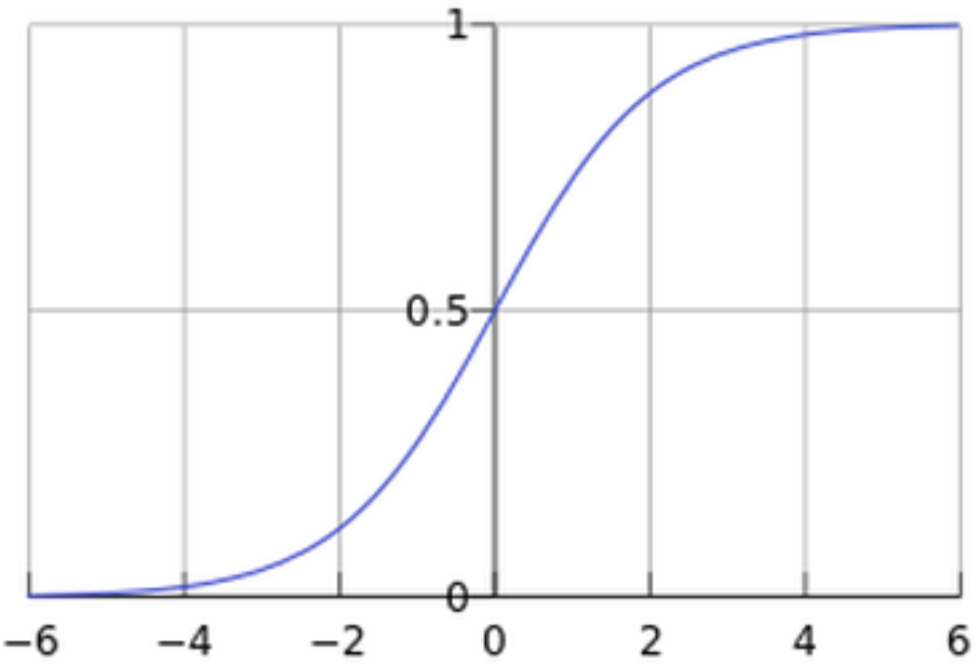
$$y = X\beta + \epsilon$$

- 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

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

- 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’

DATA SCIENCE PART TIME COURSE

EVALUATING LOGISTIC REGRESSION

ACCURACY SCORE

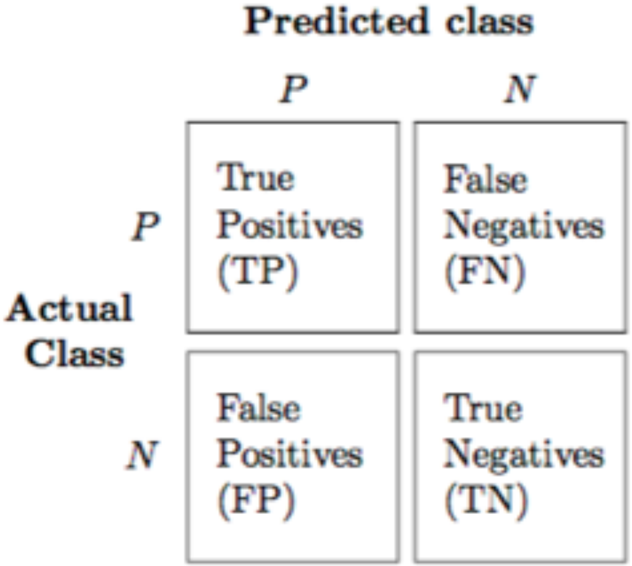
This is simply the fraction of correct predictions from the model.

So it is the number of correct predictions divided by the number of observations in our dataset.

CONFUSION MATRIX

A confusion matrix shows us what the predicted class was against what the actual class was. The true class makes up the rows or the vertical axes and the predicted class makes up the columns or the horizontal axis.

Any entries in the diagonal of the matrix are those that are correctly classified.

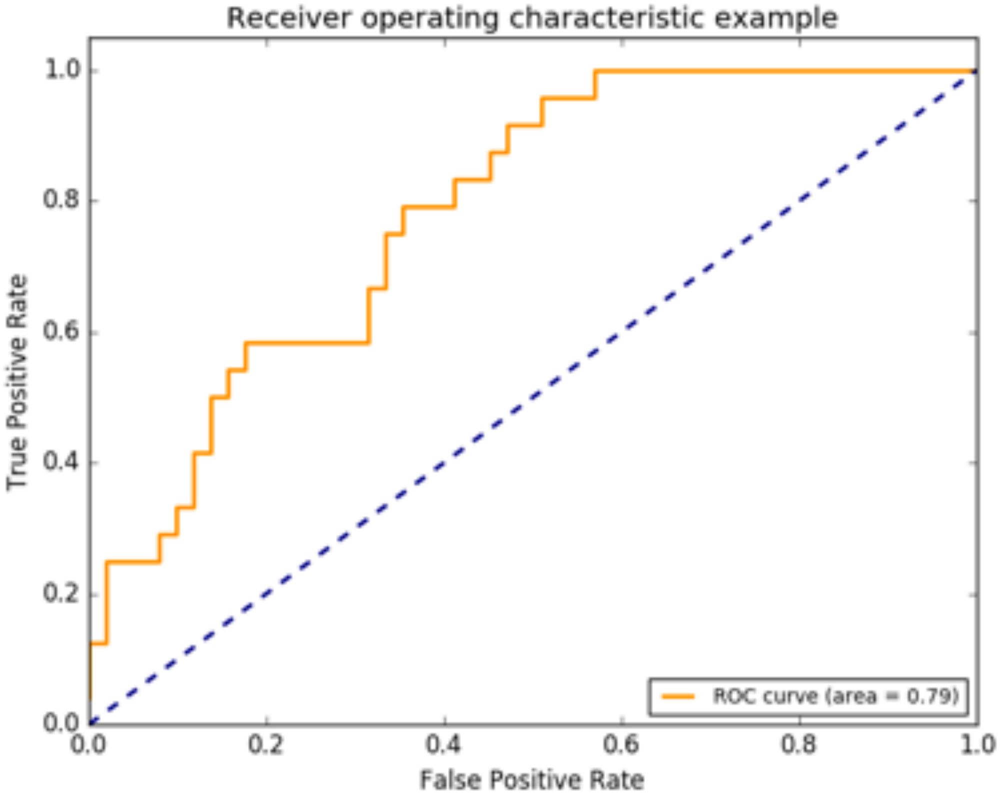


		Predicted class	
		P	*N*
Actual Class	*P*	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

RECEIVER OPERATING CHARACTERISTIC (ROC)

The Receiver Operating Characteristic or ROC curve shows the performance of a binary classifier system as its discrimination threshold is varied. It is created by plotting the fraction of true positives out of the positives (TPR = true positive rate) vs. the fraction of false positives out of the negatives (FPR = false positive rate), at various threshold settings.

RECEIVER OPERATING CHARACTERISTIC (ROC)



RECEIVER OPERATING CHARACTERISTIC (ROC) - AREA UNDER THE CURVE(AUC)

By computing the Area Under the Curve of the ROC curve we get a single number summary of accuracy.

The closer that number is to 1 the more accurate our model is.

LOGISTIC REGRESSION - WALKTHROUGH NOTEBOOK

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

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

DISCUSSION TIME

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

WEEK 2 Review

DISCUSSION TIME

- Visualisation
- Supervised vs Unsupervised Learning
- Linear Regression

DISCUSSION TIME

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

DATA SCIENCE – Week 3 Day 1

DISCUSSION TIME

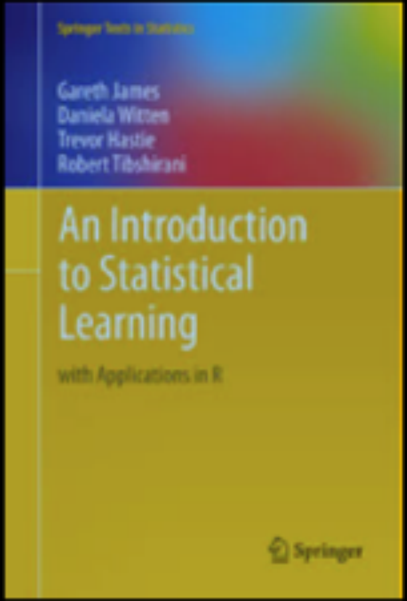
Homework/Course Project

- How's Homework 1 going ?
- How are the projects going?

PRE-READING

An Introduction to Statistical Learning

- Chapter 5 - Resampling Methods



Springer Texts in Statistics

Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani

An Introduction to Statistical Learning

with Applications in R

Springer