Welcome to General Assembly

- **WiFi** GA Guest
- **Password** yellowpencil

GENERAL ASSEMBLY

DATA SCIENCE

10 WEEK PART TIME COURSE

Week 3 - Logistic Regression

Course Plan

UNITS

UNIT 1: FOUNDATIONS OF DATA MODELING	Introduction to Data Science	Lesson 1
	Elements of Data Science	Lesson 2
	Data Visualisation	Lesson 3
	Linear Regression	Lesson 4
	Logistic Regression	Lesson 5
	Model Evaluation	Lesson 6
	Regularisation	Lesson 7
	Clustering	Lesson 8
UNIT 2: DATA SCIENCE IN THE REAL WORLD	Recommendations	Lesson 9
	SQL + Productivity	Lesson 10
	Decision Trees	Lesson 11
	Ensembles	Lesson 12
	Natural Language Programming	Lesson 13
	Cloud Computing	Lesson 14
	Time Series	Lesson 15
	Soft Skills	Lesson 16
	Network Analysis	Lesson 17
	Neural Networks	Lesson 18
	Final Projects Presentations	Lesson 19
	Final Projects Presentations	Lesson 20

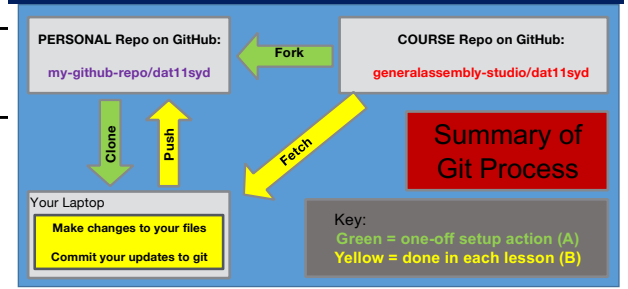
Paul & James review final project ideas

GA

Git & GitHub – 1 Pager Guide!

- Squash the GIT confusion from last class!
- 1-Pager follows through the steps clearly
- Run through this once together ➔ CRYSTAL CLEAR

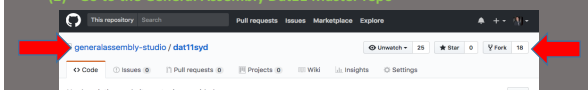
1-Pager - Git and GitHub



(Part A) - One-off setup:

(1) Login to GitHub as you

(2) Go to the General Assembly Dat11 Master repo



(3) Click "Fork" to send a copy of the COURSE repo to your own GitHub

(4) Clone your PERSONAL repo to your own workspace (laptop or server):

```
git clone https://github.com/yourgithubname/yourgithubrepo
```

(5) Set your upstream to the COURSE repo:

```
git remote add upstream https://github.com/generalassembly-studio/dat11syd
```

(Part B) EVERY CLASS:

At the START of the class, you'll need to sync the latest materials from the COURSE repo:

(1) Make sure you are in the dat11syd directory:

```
cd ~/workspace/dat11syd
```

(2) Make sure to select the "master" branch of your repo:

```
git checkout master
```

(3) Fetch the latest changes from the UPSTREAM repo (i.e the course repo)

```
git fetch upstream
```

(4) Merge the changes from the upstream repo to your master branch:

```
git merge upstream/master
```

DURING the class:

(5) Before editing, either copy files to your "students/" folder, or rename them

At the END of every class:

(6) Make sure you are in the dat11syd directory:

```
cd ~/workspace/dat11syd
```

(7) Add any files that you've updated to your git registry:

```
git add -A
```

(8) Commit the changes with a sensible comment:

```
git commit -m "my updates for lesson 7"
```

(9) Push your changes to your PERSONAL repo:

```
git push origin master
```

DONE!!!!!

Git & GitHub – 1 Pager Guide!

(Part B) EVERY CLASS:

At the START of the class, you'll need to sync the latest materials from the COURSE repo:

(1) **Make sure you are in the dat11syd directory:**

```
cd ~/workspace/dat11syd
```

(2) **Make sure to select the "master" branch of your repo:**

```
git checkout master
```

(3) **Fetch the latest changes from the UPSTREAM repo (i.e the course repo)**

```
git fetch upstream
```

(4) **Merge the changes from the upstream repo to your master branch:**

```
git merge upstream/master
```

DURING the class:

(5) **Before editing, either copy files to your "students/" folder, or rename them**

At the END of every class:

(6) **Make sure you are in the dat11syd directory:**

```
cd ~/workspace/dat11syd
```

(7) **Add any files that you've updated to your git registry:**

```
git add -A
```

(8) **Commit the changes with a sensible comment:**

```
git commit -m "my updates for lesson 7"
```

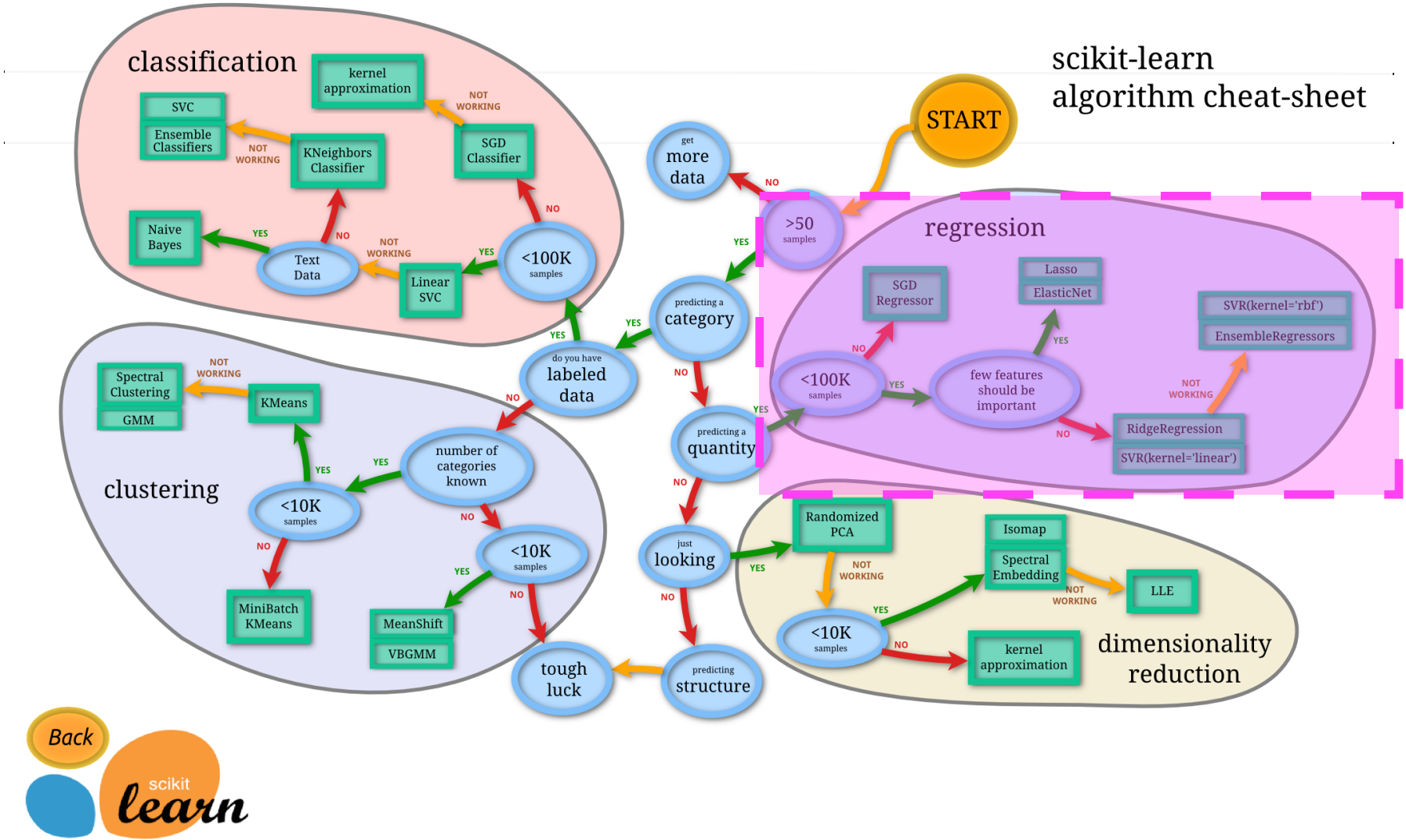
(9) **Push your changes to your PERSONAL repo:**

```
git push origin master
```

DONE!!!!!

AGENDA

1. Motivation
2. What is Logistic Regression?
3. Evaluating Logistic Regression
4. Lab4a - Logistic Regression using Titanic data
5. Lab4b - Jupyter Notebook Practice
6. Lab5 - Statistical Inference
7. Homework Review


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

DATA SCIENCE PART TIME COURSE

WHAT IS LOGISTIC REGRESSION?

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

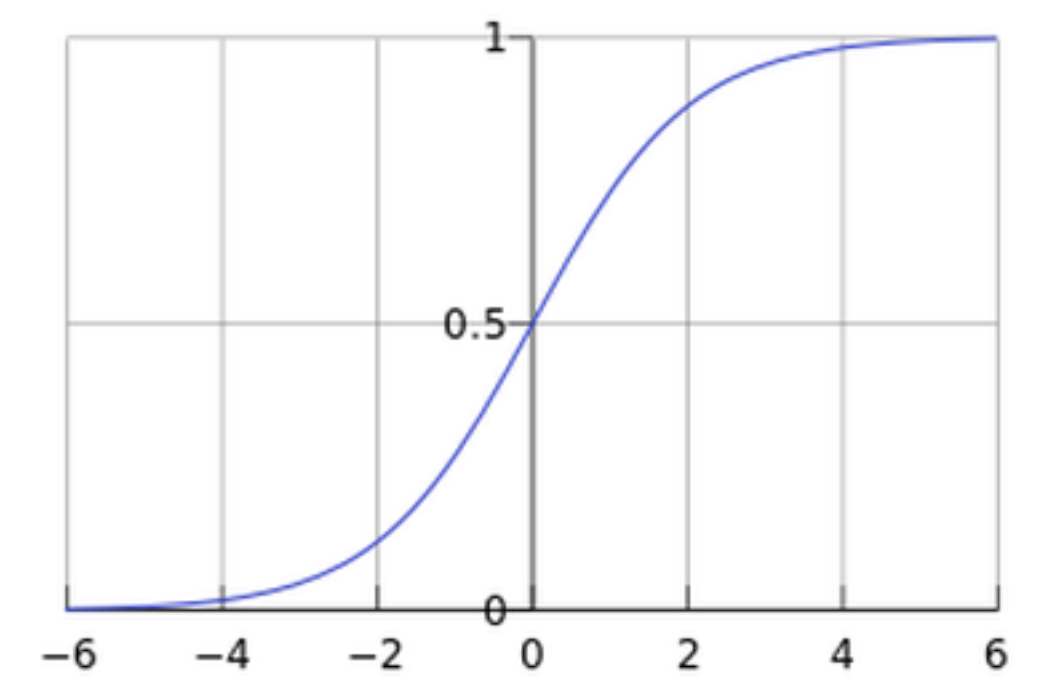
Why not use the linear regression model?

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

- Predicting
- Regression is limited to (0,1) prediction could be > 1 or < 0 …
- 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’

DATA SCIENCE PART TIME COURSE

EVALUATING LOGISTIC REGRESSION

How to Classify Using Probability?

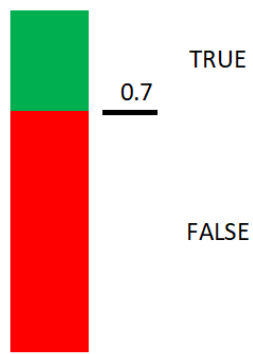
- We transformed a discrete prediction problem into a continuous prediction problem. How do we go back and answer the original problem?

  Q: How do we convert a probability to a prediction?

➢A: Specify a threshold for classification:

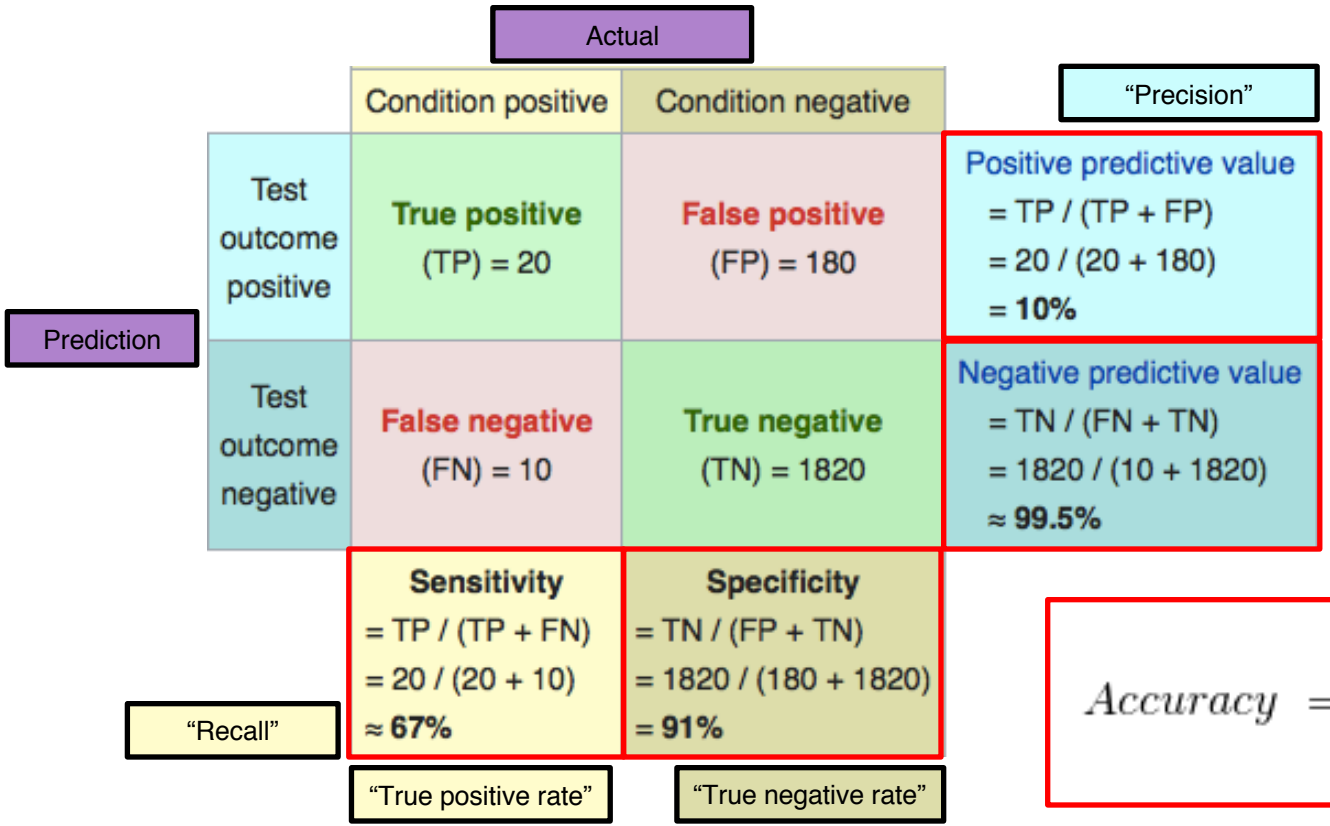
e.g. 'True' = $\mathbf{p}(y = 1 \mid X) > 0.7$

'False' = $\mathbf{p}(y = 1 \mid X) \leq 0.7$




DEFINITIONS!!!
DO NOT PANIC!!!!

Accuracy of a Classification Model



$$True\ Positive\ Rate$$
$$TPR = \frac{TP}{Actual\ P}$$

$$False\ Positive\ Rate$$
$$FPR = \frac{FP}{Actual\ N}$$

$$True\ Negative\ Rate$$
$$TNR = \frac{TN}{Actual\ N}$$

$$False\ Negative\ Rate$$
$$FNR = \frac{FN}{Actual\ P}$$

$$Accuracy = \frac{TP + TN}{Total\ Population}$$

$$Accuracy = \frac{20 + 1820}{20 + 1820 + 180 + 10} = \frac{1840}{2030} = 0.91 = 91\%$$

Accuracy of a Classification Model

accuracy score:

$$\text{accuracy\_score} = (TP + TN) / (TP + FP + TN + FN)$$

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.

Knowledge Check

- **true positive rate?**
  - also called ...
    - _____?
    - _____?
- **true negative rate:**
  - also called ...
    - _____?

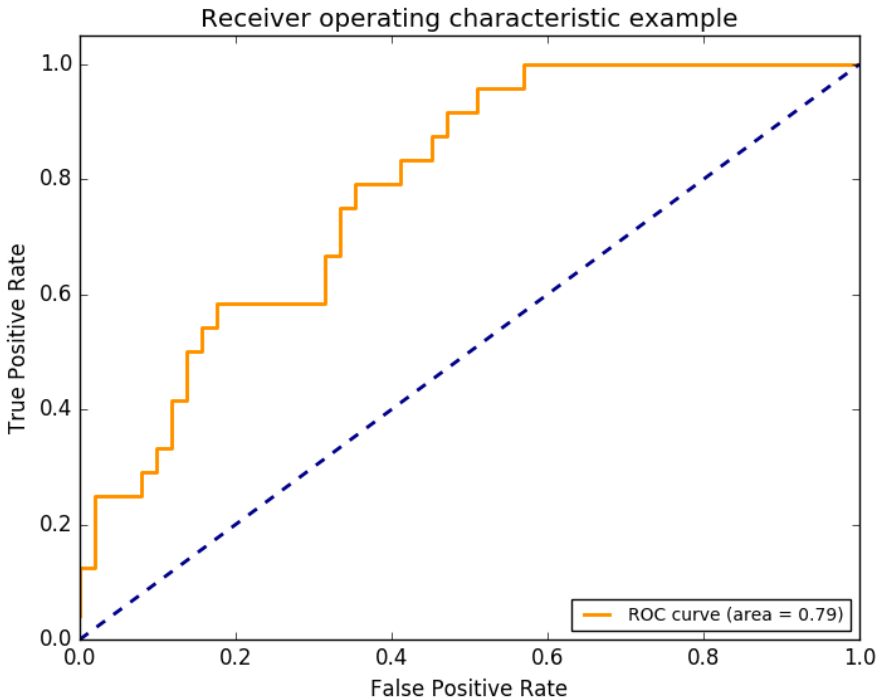
A confusion matrix shows us what the predicted class was against what the actual class was. The true class makes up the rows or the vertical axes and the predicted class makes up the columns or the horizontal axis.

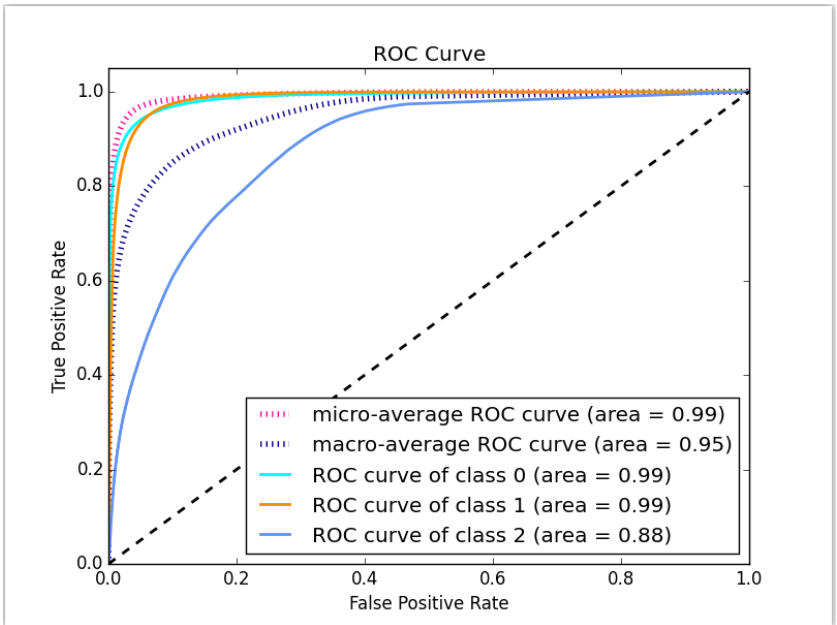
Any entries in the diagonal of the matrix are those that are correctly classified.

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

- Receiver Operating Characteristic or ROC curve
- Shows the performance of a binary classifier system as its discrimination threshold is varied
- Plot 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)





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.

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

LAB: dat11syd/lessons/lesson-04/lab/

- **Lab4a_Logistic_Regression_Titanic**

Dummy Variables

- *Used for representing categorical variables in a linear regression model*
  - Sex: male = {0, 1}
  - Pclass: {1, 2, 3}
  - Age group: child = {0, 1}

Data Set:
Titanic Survival

Data Dictionary:
('titanic-dic.csv')

Notebook:
'Lab 4 - Logistic Regression.ipynb'

Variable	Definition	Key
Survived	survival	0 = No, 1 = Yes
Pclass	ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
Name	name	
Sex	sex	
Age	age in years	
Sibsp	# of siblings, spouses aboard	
Parch	# of parents, children aboard	
Ticket	ticket number	
Fare	passenger fare	
Cabin	cabin number	
Embarked	port of embarkation	

Logistic Regression Modelling

sklearn.linear_model.**LogisticRegression**(*penalty='l2', dual=False, tol=0.0001, C=1.0, fit_intercept=True, intercept_scaling=1, class_weight=None, random_state=None, solver='liblinear', max_iter=100, multi_class='ovr', verbose=0, warm_start=False, n_jobs=1*)

http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Method:

1. split the dataset into a training subset and a test subset
2. train the model on the training data
3. test the model on the test data

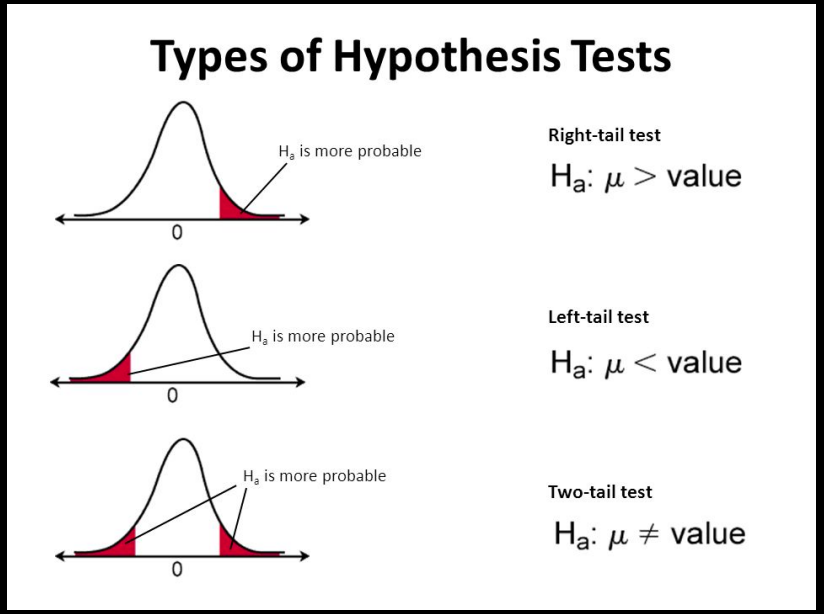
➢ *Issues with this approach?*

➢ *Resolutions?*

LAB: dat11syd/lessons/lesson-04/lab/

- **Lab4b_Jupyter_Practice**

Hypothesis Testing



LAB

LAB: dat11syd/lessons/lesson-04/lab/

- **Lab5_Statistical_Inference**
  - Probability Density Functions
  - The Normal Distribution
  - The t-Distribution
  - Student's t-Test

WEEK 2 Review

DISCUSSION TIME

- Visualisation
- Supervised vs Unsupervised Learning
- Linear Regression

DATA SCIENCE - Week 3 Day 1

DISCUSSION TIME

Logistic Regression applied to loan applications
(https://github.com/nborwankar/LearnDataScience)

Logistic Regression video (watch at 1.5x speed)
https://www.youtube.com/watch?v=-la3q9d7AKQ

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?

DATA SCIENCE - Week 3 Day 1

PRE-READING

An Introduction to Statistical Learning
Chapter 5 - Resampling Methods

OR

Watch:
Cross validation https://www.youtube.com/watch?v=CmEgvD_ov2o
Bootstrapping https://www.youtube.com/watch?v=PmKOOGVRsmc