GENERAL ASSEMBLY

INTRODUCTION TO LOGISTIC REGRESSION

Angelo Klin
Katra Analytics

LEARNING OBJECTIVES

- Build a Logistic Regression Classification Model
- Describe a sigmoid function, odds and the odds ratio as well as how they relate to logistic regression
- Evaluate a model using metrics such as classification accuracy/error, confusion matrix, ROC/AUC curves and loss functions

PRE-WORK

PRE-WORK REVIEW

- Implement a linear model (LinearRegression) with scikit-learn
- Understand what a coefficient is
- Recall metrics such as accuracy and misclassification
- Recall the differences between L1 and L2 regularisation

OPENING

INTRODUCTION TO LOGISTIC REGRESSION

DEFINE THE PROBLEM

DEFINE THE PROBLEM

- Buying a Property
  - Getting credit approval for a loan based on Credit Score
  - Credit Scores range from 0 to 1200
  - Current Credit Score is 590
  - Data from a financial institution’s previous analysis
    - 500 cases were credit was approved or not

DEFINE THE PROBLEM

1. Model reality based on probability and odds of getting approval based on Credit Score

2. Figure the probability of 50%

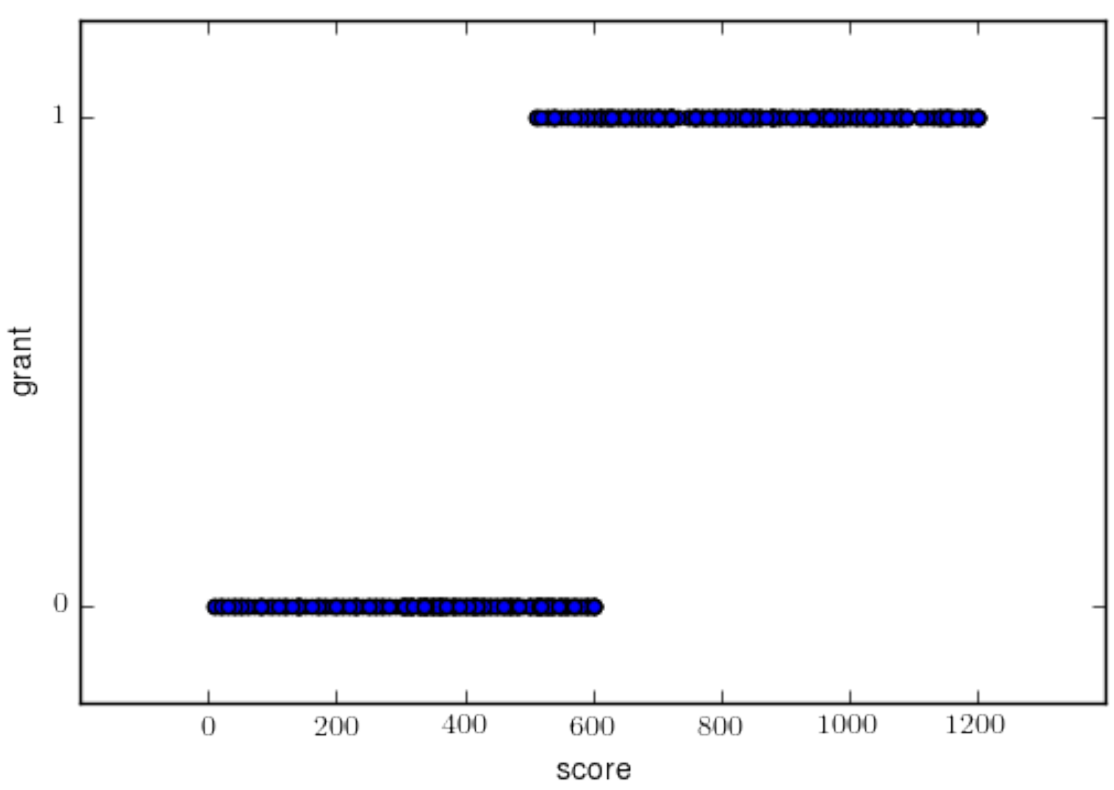
3. Use current Credit Score to check the probability and odds of getting approval

4. Check how the impact of improving the current Credit Score

DEFINE THE PROBLEM

- The actual cases (reality) are a compilation of people with a credit score that got results of credit being approved or not (yes / no)

approved	grant	score
no	0	330
no	0	340
no	0	510
yes	1	1020
yes	1	510
…	…	…
yes	1	810
no	0	500
yes	1	670
no	0	350



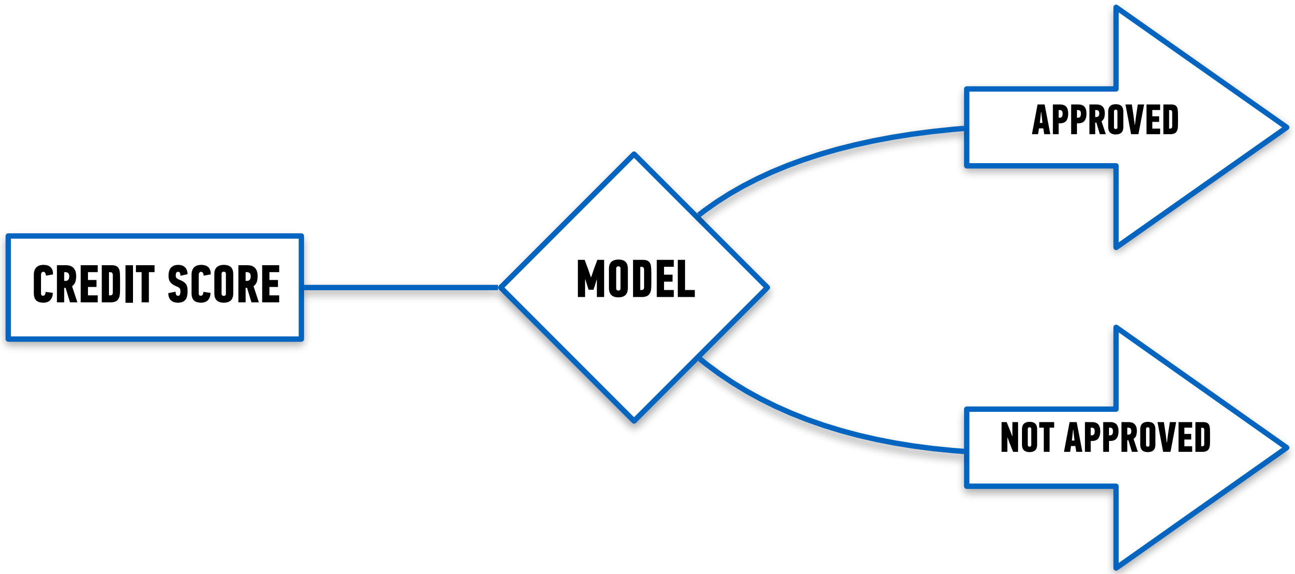
- $n = 500$

LOGISTIC REGRESSION

- Model the probability of an event occurring based on the independent variables
- Estimate the probability that an event occurs vs the probability that an event does not occur
- Predict the effect of the independent variables on the output variable
- Classify observation by estimating the probability that an observation is of some particular category

PROCESS

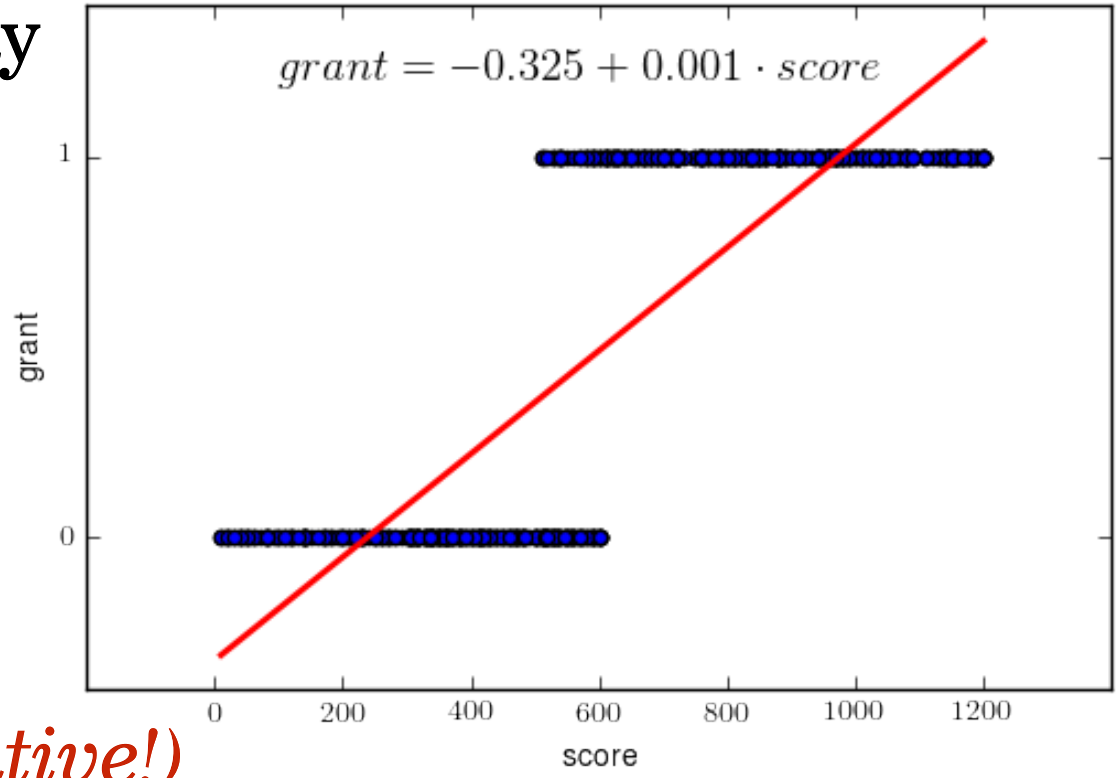
- Define reality by creating a model to predict the outcomes



- As per the Model, what is the probability of getting approval given a particular Credit Score?

TRYING LINEAR REGRESSION

- Option: Assume grant as a probability
  $P(grant = 1) = p$, where $0 \leq p \leq 1$
  - $p = -0.325 + 0.001 \cdot 590 = 0.476$
    - Increase of 1 unit of score adds 0.001 probability of a grant
    - Result is fine, although bad news



- $p = -0.325 + 0.001 \cdot 100 = -0.189$ *(negative!)*
- $p = -0.325 + 0.001 \cdot 990 = 1.019$ *(over one!)*
- Linear Regression does **not** make a good model on extremes

CHALLENGE! LINEAR REGRESSION RESULTS FOR CLASSIFICATION

- But, since most classification problems are binary (0 or 1) and 1 is greater than 0, does it make sense to apply the concept of regression to solve classification?
- How might we contain those bounds?
- Let's review some approaches to make classification with regression feasible

IMPROVE ON LINEAR REGRESSION

- Make it positive
  - $f(x) = |x|$ or $f(x) = x^2$ or $f(x) = e^x$
- Make it less or equal to one

$$f(x) = \frac{x}{1+x}$$

PROBABILITY

$$P = \frac{outcomes\ of\ interest}{all\ outcomes}$$

$$P(die = even) = \frac{3}{6} = \frac{1}{2} = 0.5$$

ODDS

$$Odds = \frac{P(occurring)}{P(not\ occurring)} = \frac{p}{1-p}$$

$$Odds = \frac{P(die = even)}{P(die = not\ even)} = \frac{\frac{3}{6}}{1 - \frac{3}{6}} = \frac{\frac{1}{2}}{\frac{1}{2}} = \frac{1}{1} = 1:1$$

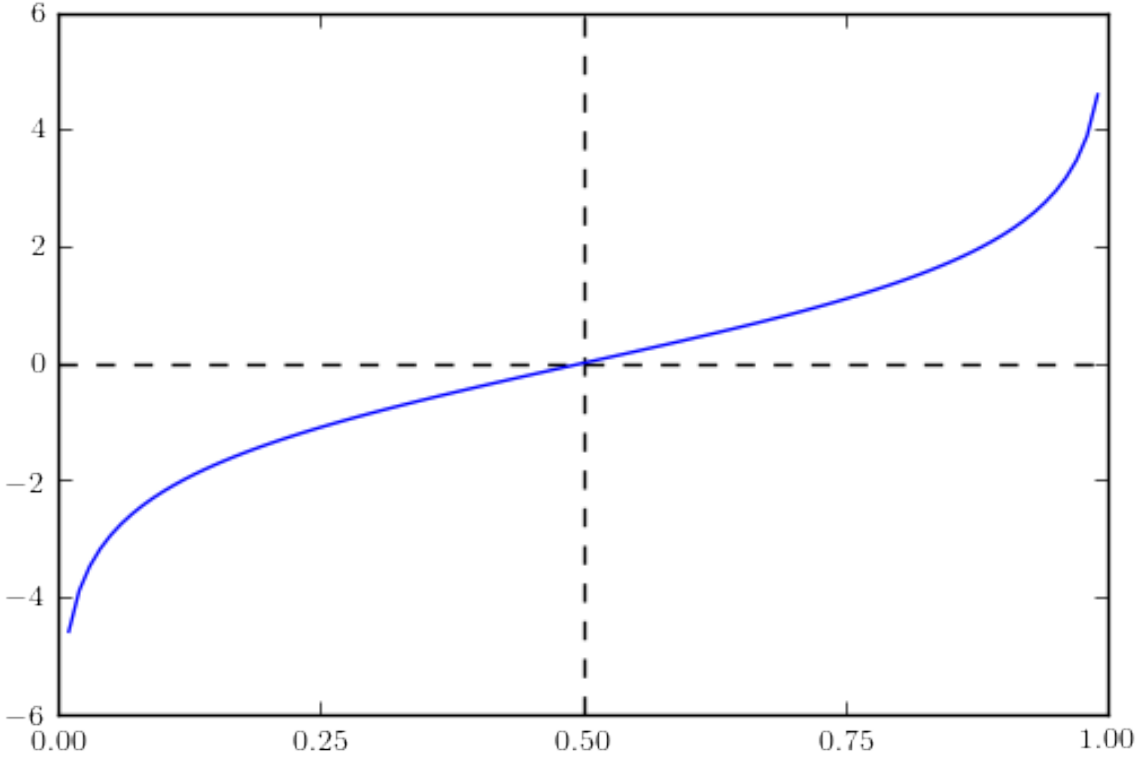
ODDS RATIO

$$Odds\ Ratio = \frac{Odds_a}{Odds_b} = \frac{\frac{p_a}{1-p_a}}{\frac{p_b}{1-p_b}}$$

$$Odds\ Ratio = \frac{Odds_{die=even}}{Odds_{die=1,2}} = \frac{\frac{p(die=even)}{1-p(die=even)}}{\frac{p(die=1,2)}{1-p(die=1,2)}} = \frac{\frac{\frac{3}{6}}{1-\frac{3}{6}}}{\frac{\frac{2}{6}}{1-\frac{2}{6}}} = \frac{\frac{\frac{1}{2}}{\frac{1}{2}}}{\frac{\frac{1}{3}}{\frac{2}{3}}} = \frac{1}{\frac{1}{2}} = 2$$

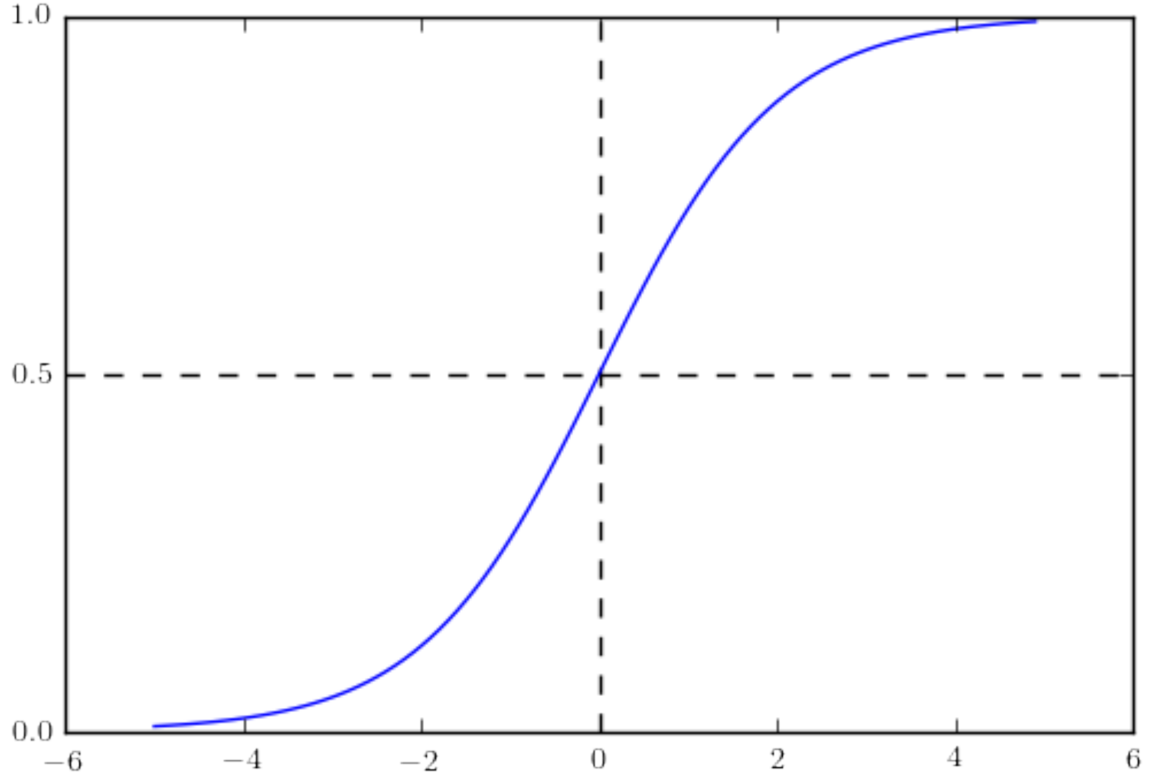
LOG ODDS OR LOGIT

$$f(x) = logit(p) = log(odds) = log\left(\frac{p}{1-p}\right); 0 \leq p \leq 1$$



SIGMOID

$$f(x) = sigmoid(x) = logit^{-1}(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{1+e^x}; -\infty \leq x \leq \infty$$



DOING THE MATH

$$logit(p) = ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_1$$

$$\frac{p}{1-p} = e^{\alpha+\beta_1 x_1}$$

$$p = (1-p)e^{\alpha+\beta_1 x_1}$$

$$p = e^{\alpha+\beta_1 x_1} - p \cdot e^{\alpha+\beta_1 x_1}$$

$$p + p \cdot e^{\alpha+\beta_1 x_1} = e^{\alpha+\beta_1 x_1}$$

$$p(1 + e^{\alpha+\beta_1 x_1}) = e^{\alpha+\beta_1 x_1}$$

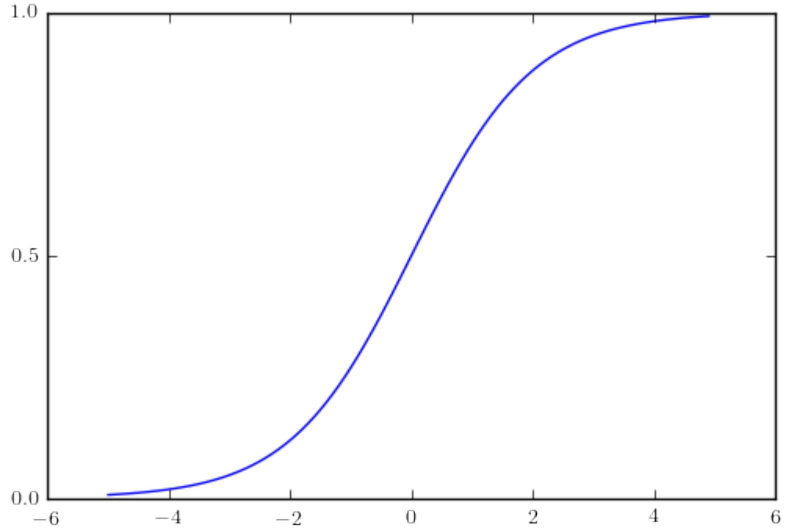
$$\hat{p} = \frac{e^{\alpha+\beta_1 x_1}}{1 + e^{\alpha+\beta_1 x_1}}$$

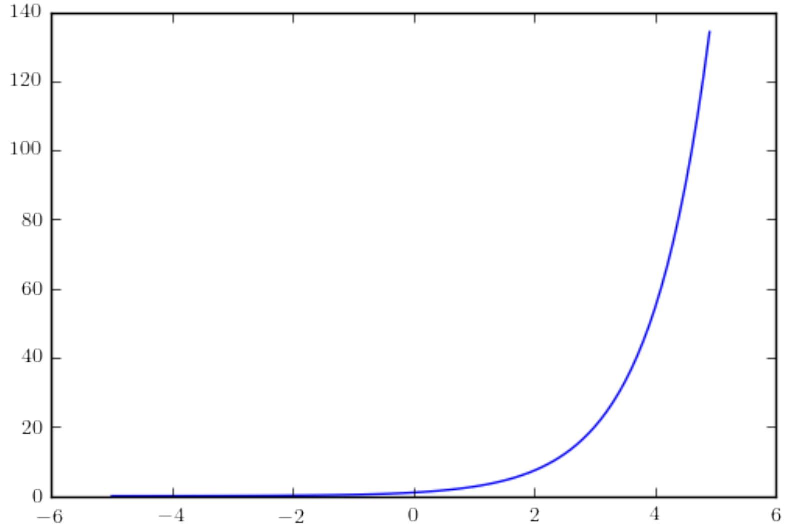
PROBABILITY «–» ODDS «–» LOG ODDS

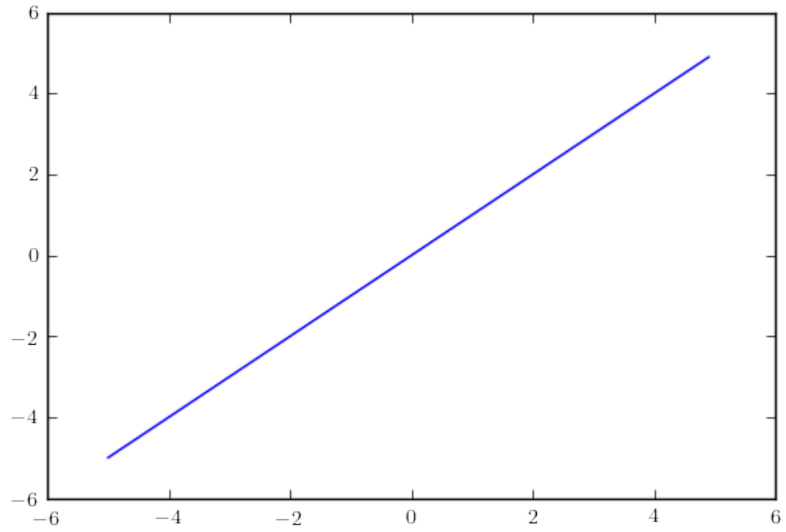
$$p(y) = \frac{e^{\alpha+\beta_1 x_1}}{1 + e^{\alpha+\beta_1 x_1}}$$

$$odds(p(y)) = \frac{p(y)}{1-p(y)} = \frac{\frac{e^{\alpha+\beta_1 x_1}}{1+e^{\alpha+\beta_1 x_1}}}{1 - \left(\frac{e^{\alpha+\beta_1 x_1}}{1+e^{\alpha+\beta_1 x_1}}\right)}$$

$$log(odds(p(y))) = \alpha + \beta_1 x_1$$







CHALLENGE! LINEAR REGRESSION RESULTS FOR CLASSIFICATION

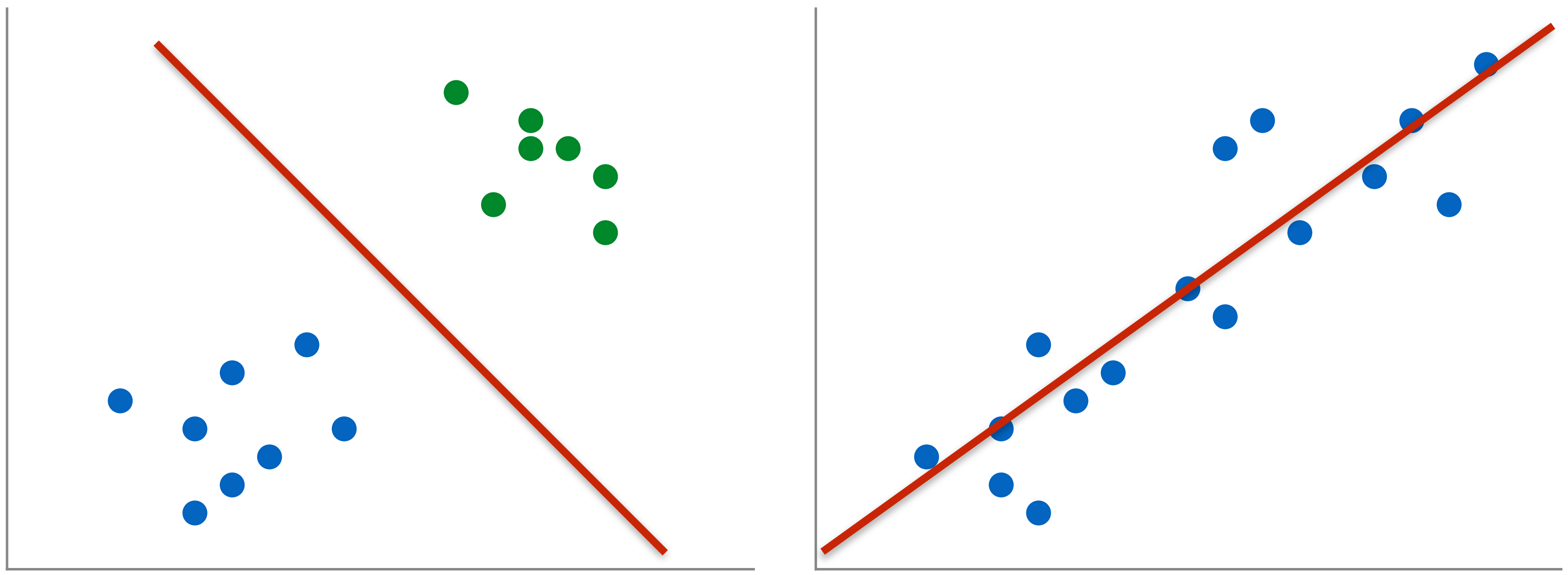
- In Logistic Regression
  - How much the Odds change for one unit change in Odds Ratio

CHALLENGE! LINEAR REGRESSION RESULTS FOR CLASSIFICATION

- Regression results can have a value range from -∞ to ∞

- Classification is used when predicted values (i.e. Class Labels) are not greater than or less than each other

Classification   Regression



FIX 1: PROBABILITY

- One approach is predicting the probability that an observation belongs to a certain class

- We could assume the prior probability (the bias) of a class is the class distribution

FIX 1: PROBABILITY

- For example, suppose we know that roughly 700 of 2,200 people from the Titanic survived
  - Without knowing anything about the passengers or crew, the probability of survival would be ~0.32 (32%)

- However, we still need a way to use a linear function to either increase or decrease the probability of an observation given the data about it

FIX 2: LINK FUNCTIONS AND THE SIGMOID FUNCTION

- Another advantage to OLS (Least Squared) is that it allows for generalised models using a link function
- Link functions allows us to build a relationship between a linear function and the mean of a distribution
- We can now form a specific relationship between our linear predictors and the response variable

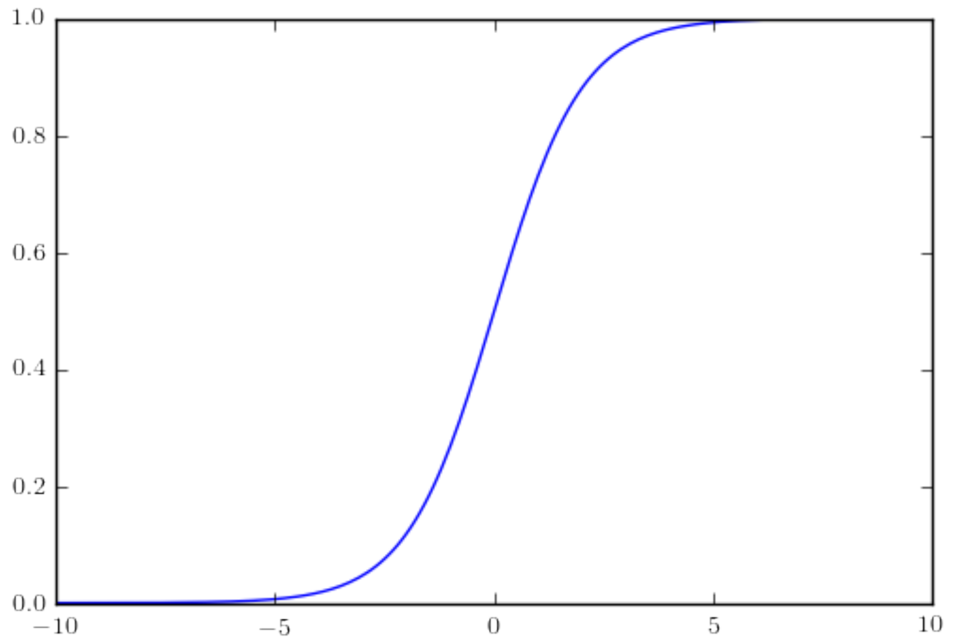
FIX 2: LINK FUNCTIONS AND THE SIGMOID FUNCTION

- For classification, we need a distribution associated with categories
  - Given all events, what is the probability of a given event?

- The link function that best allows for this is the logit function, which is the inverse of the sigmoid function

FIX 2: LINK FUNCTIONS AND THE SIGMOID FUNCTION

- A sigmoid function is a function that visually looks like an **S**



- Mathematically, it is defined as

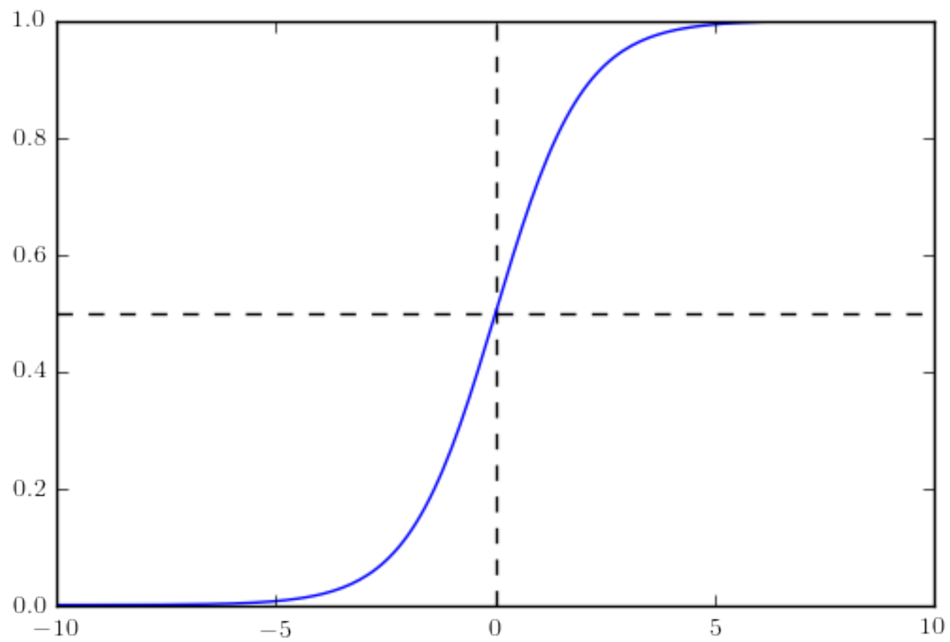
$$f(x) = \frac{1}{1 + e^{-x}}$$

FIX 2: LINK FUNCTIONS AND THE SIGMOID FUNCTION

- Recall that *e* is the inverse of the natural log
  - As x increases, the results is closer to 1
  - As x decreases, the result is closer to 0

- When x = 0, the result is 0.5

$$f(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{1+e^x}$$



FIX 2: LINK FUNCTIONS AND THE SIGMOID FUNCTION

- Since x decides how much to increase or decrease the value away from 0.5, x can be interpreted as something like a coefficient

- However, we still need to change its form to make it more useful

DEMONSTRATION

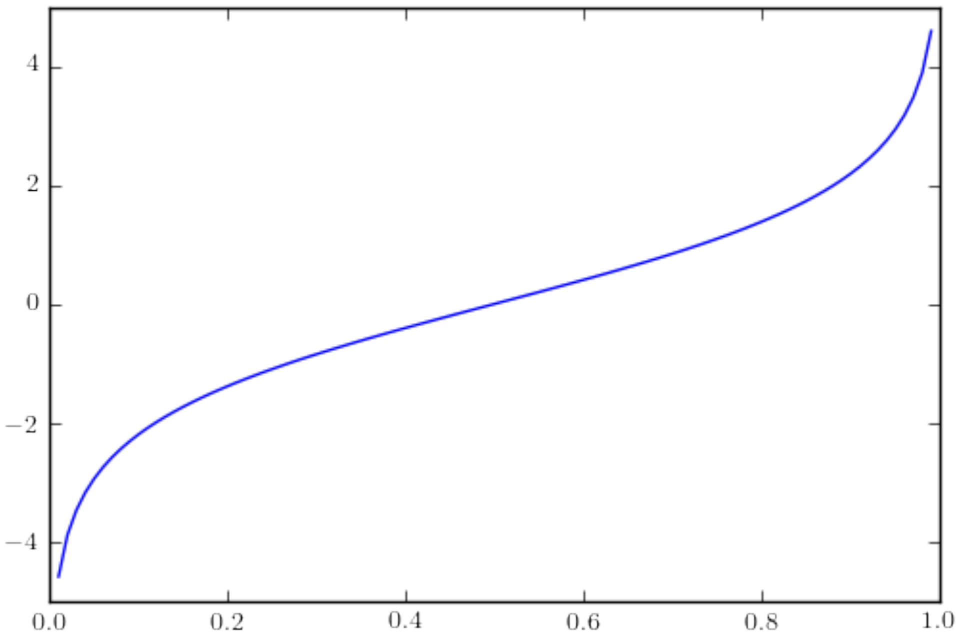
PLOTTING A SIGMOID FUNCTION

PLOTTING A SIGMOID FUNCTION

- Use the sigmoid function definition with values of x between -6 and 6 to plot it on a graph
- Do this by hand or write Python code to evaluate it
- Recall that $e = 2.71$
- Do we get an the “S” shape we expect?

FIX 3: ODDS AND LOG-ODDS

- The logit function is the inverse of the sigmoid function
- This will act as our link function for logistic regression



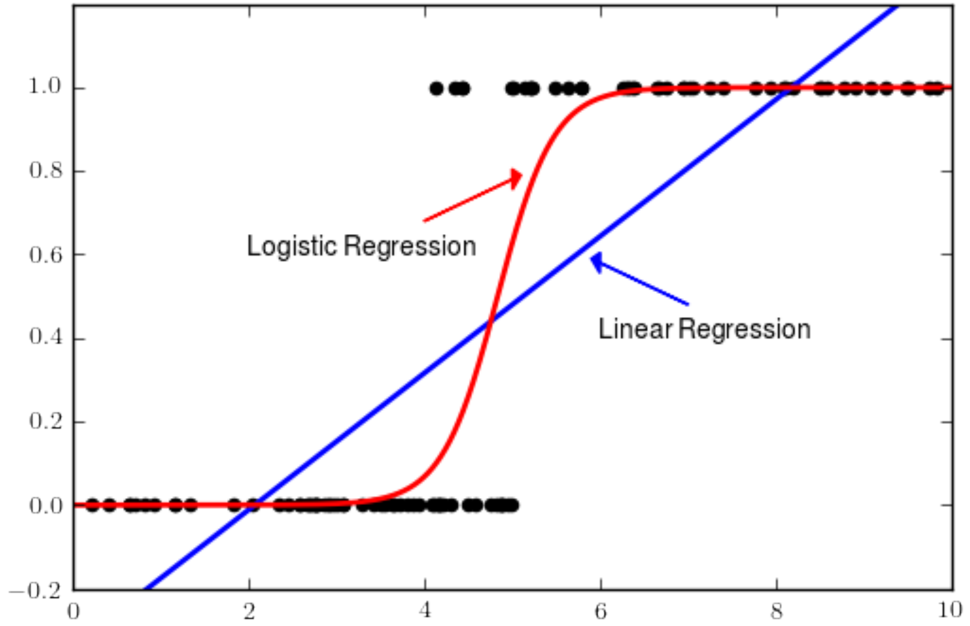
- Mathematically, the logit function is defined as

$$f(x) = ln\left(\frac{P}{1-P}\right)$$

- The value within the natural log represents the odds
- Taking the natural log of odds generates log odds

FIX 3: ODDS AND LOG-ODDS

- The logit function allows for values between -∞ and ∞, but provides us probabilities between 0 and 1



ACTIVITY: KNOWLEDGE CHECK

EXERCISE

DIRECTIONS: ANSWER THE FOLLOWING QUESTIONS

1. Why is it important to take values between $-\infty$ and ∞, but provide probabilities between 0 and 1?
2. What does this remind us of?

FIX 3: ODDS AND LOG-ODDS

- For example, the logit value (log odds) of 0.2 (or odds of ~1.2:1)

$$ln\left(\frac{P}{1-P}\right) = 0.2$$

- With a mean probability of 0.5, the adjusted probability would be ~0.55

$$\frac{1}{1+e^{-0.2}} \approx 0.55$$

- To calculate this in python, we could use the following

```
1 / (1 + numpy.exp(-0.2))
```

FIX 3: ODDS AND LOG-ODDS

- While the logit value represents the coefficients in the logistic function, we can convert them into odds ratios that make them more easily interpretable

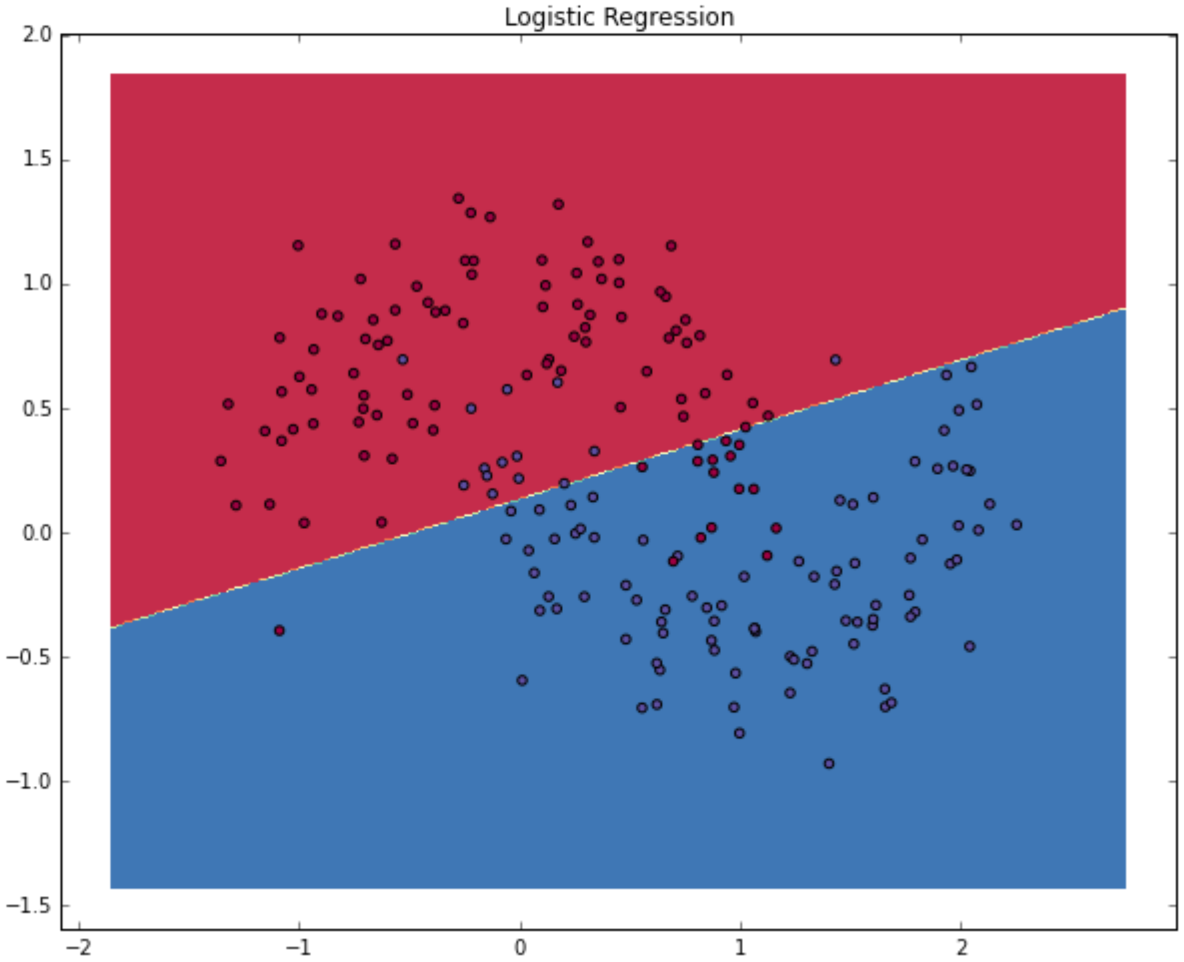
$$ln\left(\frac{P}{1-P}\right) = \alpha + \beta_1 X_1$$

- For every 1 unit increase in x the odds multiply by e^{β_1}

$$\frac{odds(x+1)}{odds(x)} = \frac{\frac{F(x+1)}{1-F(x+1)}}{\frac{F(x)}{1-F(x)}} = \frac{e^{\alpha+\beta_1(x+1)}}{e^{\alpha+\beta_1 x}} = e^{\beta_1}$$

FIX 3: ODDS AND LOG-ODDS

- With these coefficients, we get our overall probability
  - The logistic regression draws a linear decision line which divides the classes



GUIDED PRACTICE

WAGER THOSE ODDS!

ACTIVITY: WAGER THOSE ODDS!

EXERCISE

DIRECTIONS: ANSWER THE FOLLOWING QUESTIONS (15 MINUTES)

1. Given the odds below for some football games, use the *logit* function and the *sigmoid* function to solve for the *probability* that the "better" team would win.

a. Collingwood	: Carlton	5.0 : 1
b. Essendon	: West Coast	20.0 : 1
c. Geelong	: Adelaide	1.1 : 1
d. Port Adelaide	: Richmond	1.8 : 1
e. Hawthorn	: Sydney	1.6 : 1

```
def logit_func(odds):
  # uses a float (odds) and returns back the log odds (logit)
  return None

def sigmoid_func(logit):
  # uses a float (logit) and returns back the probability
  return None
```

INDEPENDENT PRACTICE

LOGISTIC REGRESSION IMPLEMENTATION

ACTIVITY: LOGISTIC REGRESSION IMPLEMENTATION

EXERCISE

DIRECTIONS: ANSWER THE FOLLOWING QUESTIONS (15 MINUTES)

1. Use the data collegeadmissions.csv and the LogisticRegression estimator in scikit-learn to predict the target variable admit
   a. What is the bias or prior probability of the dataset?
   b. Build a simple model with one feature and explore the coef_ value. Does this represent the odds or logit (log odds)?
   c. Build a more complicated model using multiple features. Interpreting the odds, which features have the most impact on admission rate? Which features have the least?
   d. What is the accuracy of your model?

INTRODUCTION

ADVANCED CLASSIFICATION METRICS

ADVANCED CLASSIFICATION METRICS

- Accuracy is only one of several metrics used when solving a classification problem
- Accuracy = total **predicted** **correct** / total observations
- Accuracy alone does not always give us a full picture
- If we know a model is 75% accurate, it does not provide **any** insight into why the 25% was wrong

ADVANCED CLASSIFICATION METRICS

- Was it wrong across all labels?
- Did it just guess one class label for all predictions?
- It is important to look at other metrics to fully understand the problem

ADVANCED CLASSIFICATION METRICS

- Imagine a study evaluating a new test that screens people for a disease
- Each person taking the test either has or does not have the disease
- The test outcome can be positive (classifying the person as having the disease) or negative (classifying the person as not having the disease)
  - In general, Positive = identified and Negative = rejected
- The test results for each subject may or may not match the subject's actual status

Classification	Conclusion	Interpretation
True Positive	Correctly Identified	Sick people correctly identified as sick
False Positive	Incorrectly Identified	Healthy people incorrectly identified as sick
True Negative	Correctly Rejected	Healthy people correctly identified as healthy
False Negative	Incorrectly Rejected	Sick people incorrectly identified as healthy

ADVANCED CLASSIFICATION METRICS

- We can split up the accuracy of each label by using the true positive rate and the false positive rate

- For each label, we can put it into the category of a true positive, false positive, true negative or false negative

		Actual Class	
		positive	negative
Predicted Class	True	True Positives	False Positives
	False	False Negatives	True Negatives
	Total	Positive	Negative

ADVANCED CLASSIFICATION METRICS

- Sensitivity or True Positive Rate (TPR) asks
  - "Out of all of the target class labels, how many were accurately predicted to belong to that class?"
- For example, given a medical exam that tests for cancer, how often does it correctly identify patients with cancer?

		Actual Class	
		positive	negative
Predicted Class	True	TP	FP
	False	FN	TN
	Total	Positive	Negative

$$TPR = \frac{TP}{P} = \frac{TP}{(TP + FN)}$$

ADVANCED CLASSIFICATION METRICS

- Fall-out or False Positive Rate (FPR) asks
  - "Out of all items not belonging to a class label, how many were predicted as belonging to that target class label?"
- For example, given a medical exam that tests for cancer, how often does it trigger a **false alarm** by incorrectly saying a patient has cancer?

		Actual Class	
		positive	negative
Predicted Class	True	TP	FP
	False	FN	TN
	Total	Positive	Negative

$$FPR = \frac{FP}{N} = \frac{FP}{(FP + TN)}$$

ADVANCED CLASSIFICATION METRICS

- These can also be inverted
- Negative Predictive Value (NPV)
- How often does a test *correctly* identify patients without cancer?

		Actual Class	
		positive	negative
Predicted Class	True	TP	FP
	False	FN	TN
	Total	Positive	Negative

$$NPV = \frac{TN}{(TN + FN)}$$

ADVANCED CLASSIFICATION METRICS

- How often does a test *incorrectly* identify patient as cancer-free?

		Actual Class	
		positive	negative
Predicted Class	True	TP	FP
	False	FN	TN
	Total	Positive	Negative

$$FNR = \frac{FN}{(TP + FN)}$$

ADVANCED CLASSIFICATION METRICS

- The True Positive and False Positive Rates give us a much clearer pictures of where predictions begin to fall apart
- This allows us to adjust our models accordingly
- A good classifier would have a True Positive Rate approaching 1 and a False Positive Rate approaching 0
- In our smoking problem, this model would accurately predict all of the smokers as smokers and not accidentally predict any of the nonsmokers as smokers

ADVANCED CLASSIFICATION METRICS

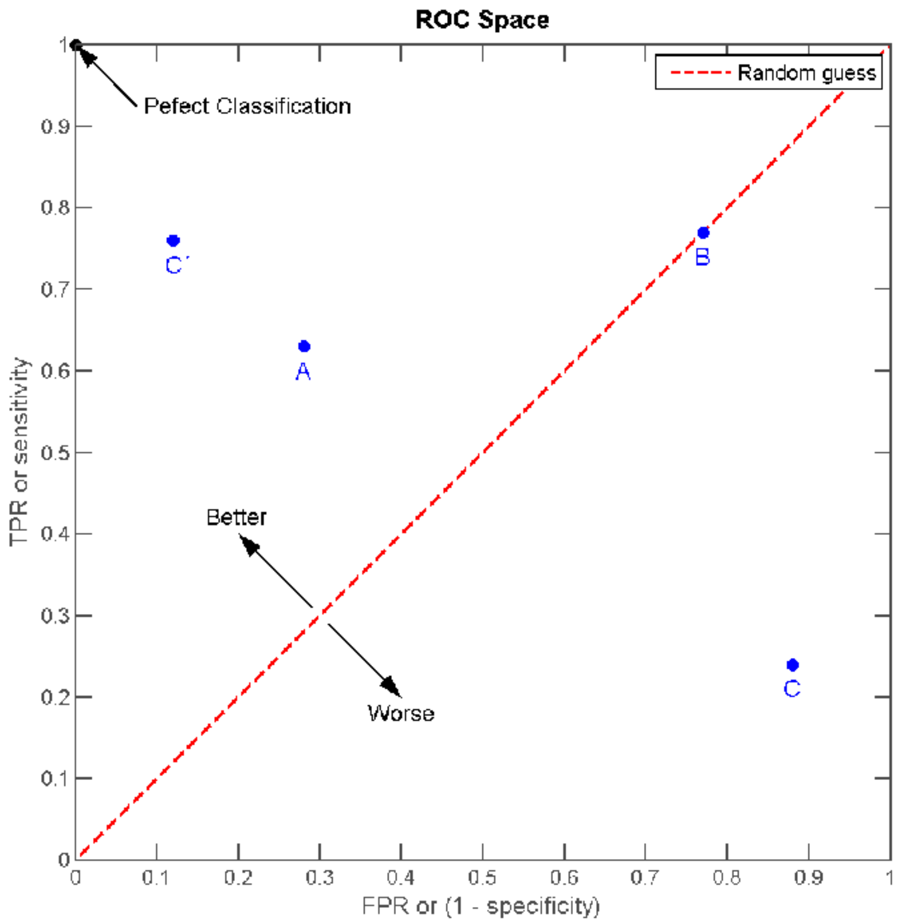
- We can vary the classification threshold for our model to get different predictions
  - But how do we know if a model is better overall than other model?
- We can compare the FPR and TPR of the models, but it can often be difficult to optimise two numbers at once
- Logically, we like a single number for optimisation
- Can you think of any ways to combine our two metrics?

ADVANCED CLASSIFICATION METRICS

- This is where the Receiver Operation Characteristic (ROC) curve comes in handy
- The curve is created by plotting the True Positive Rate against the False Positive Rate at various model threshold settings
- Area Under the Curve (AUC) summarises the impact of TPR and FPR in one single value

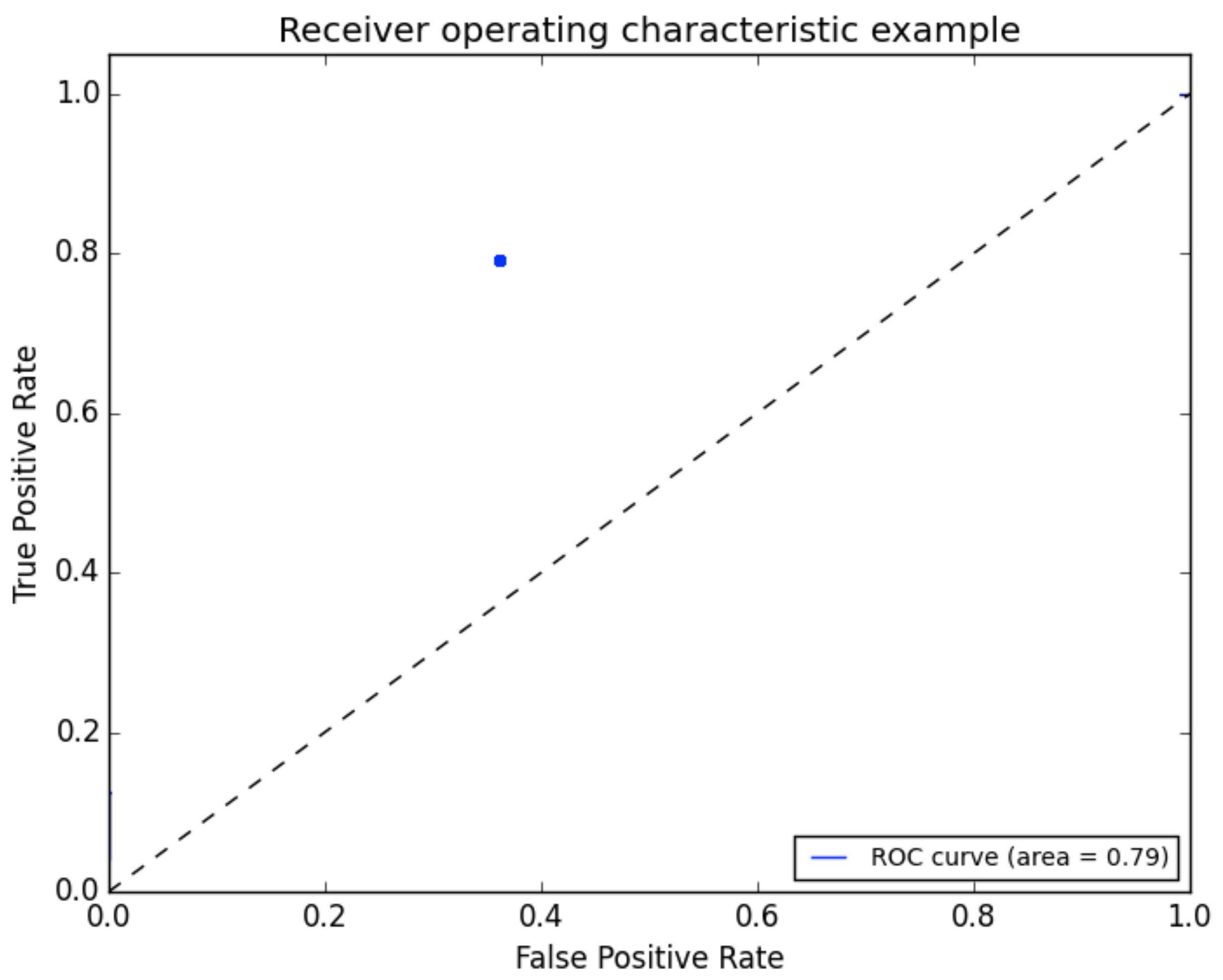
ADVANCED CLASSIFICATION METRICS

- There can be a variety of points on an ROC curve



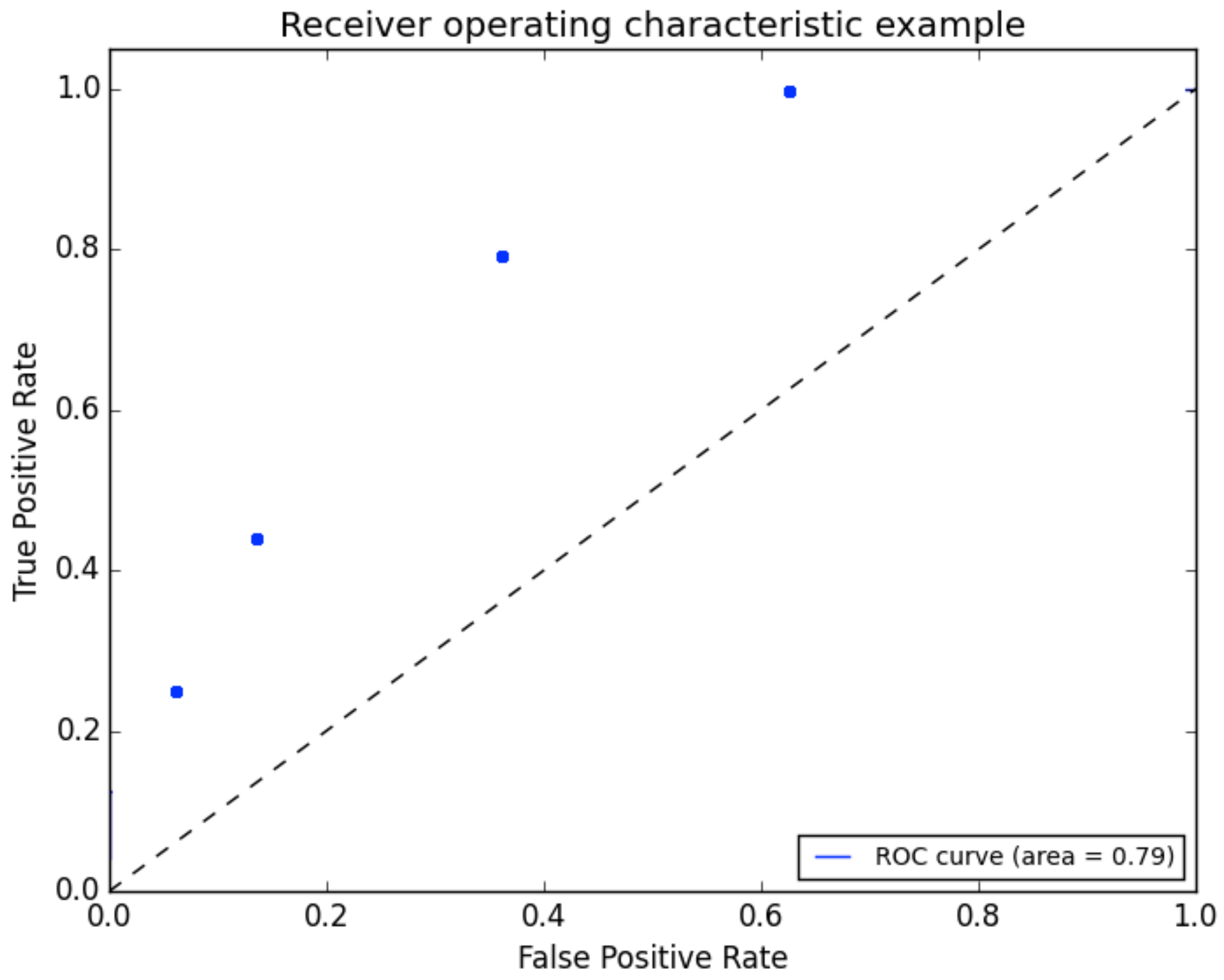
ADVANCED CLASSIFICATION METRICS

- We can begin by plotting an individual TPR/FPR pair for one threshold



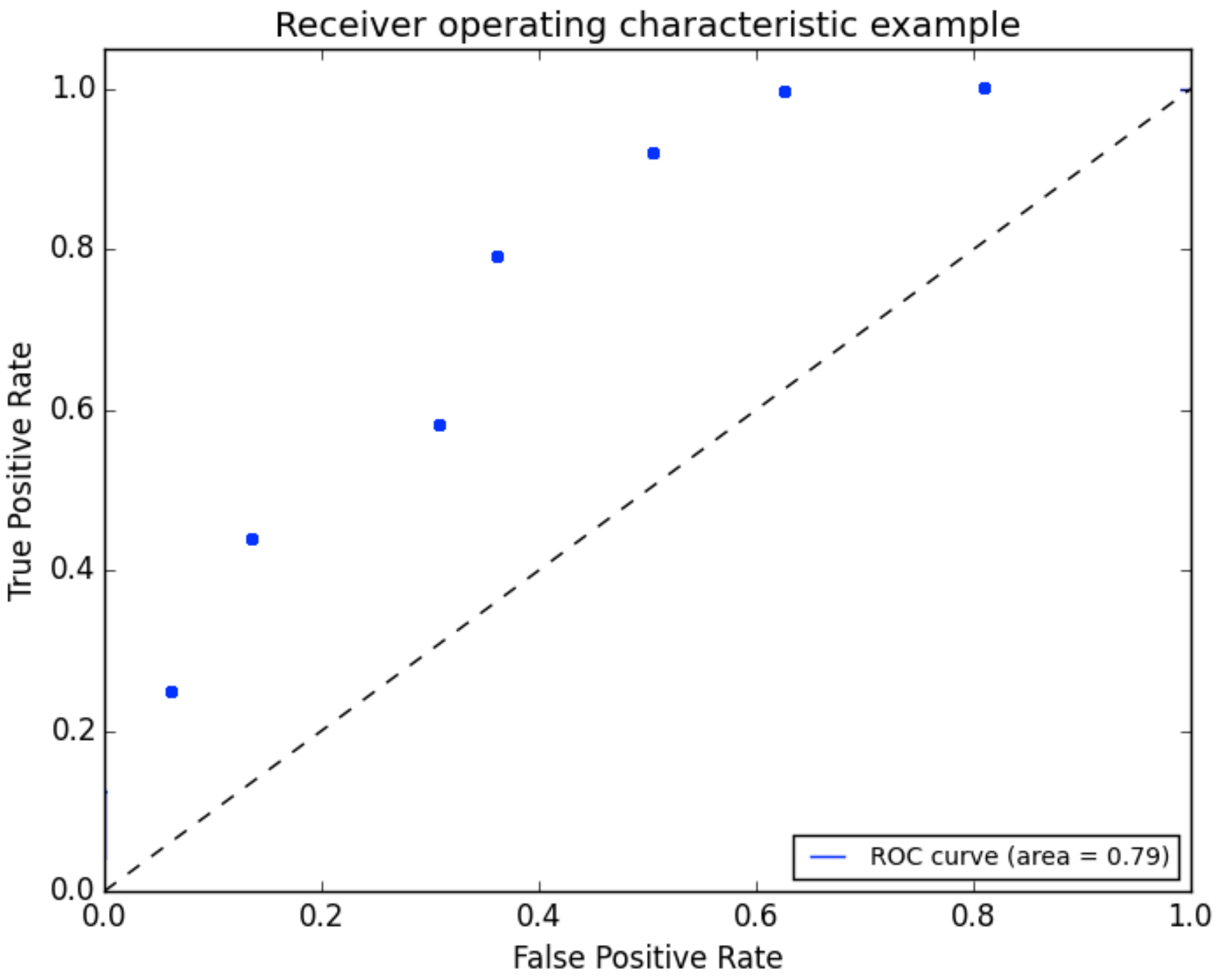
ADVANCED CLASSIFICATION METRICS

- We can add pairs for different thresholds



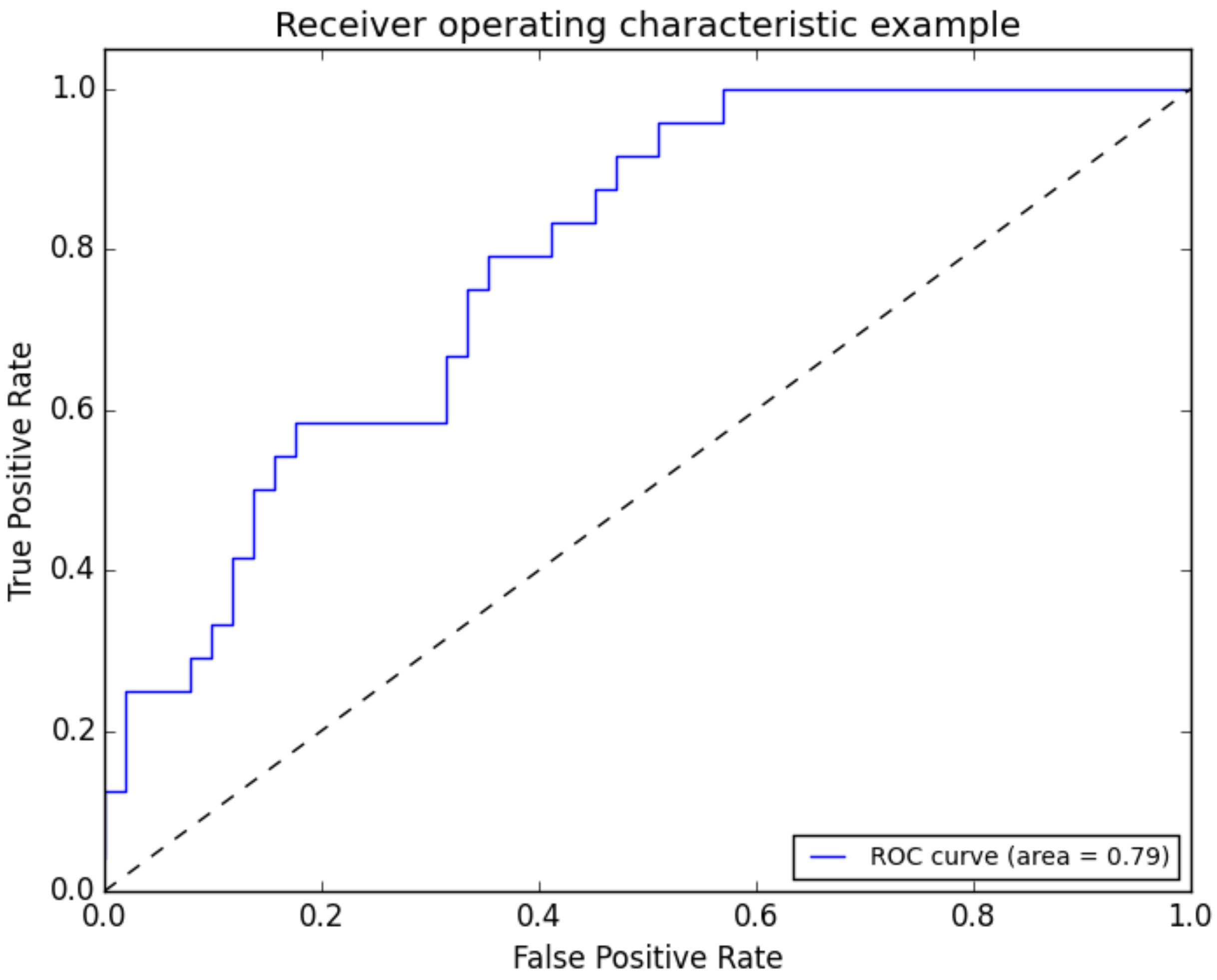
ADVANCED CLASSIFICATION METRICS

- We can continue adding pairs for different thresholds



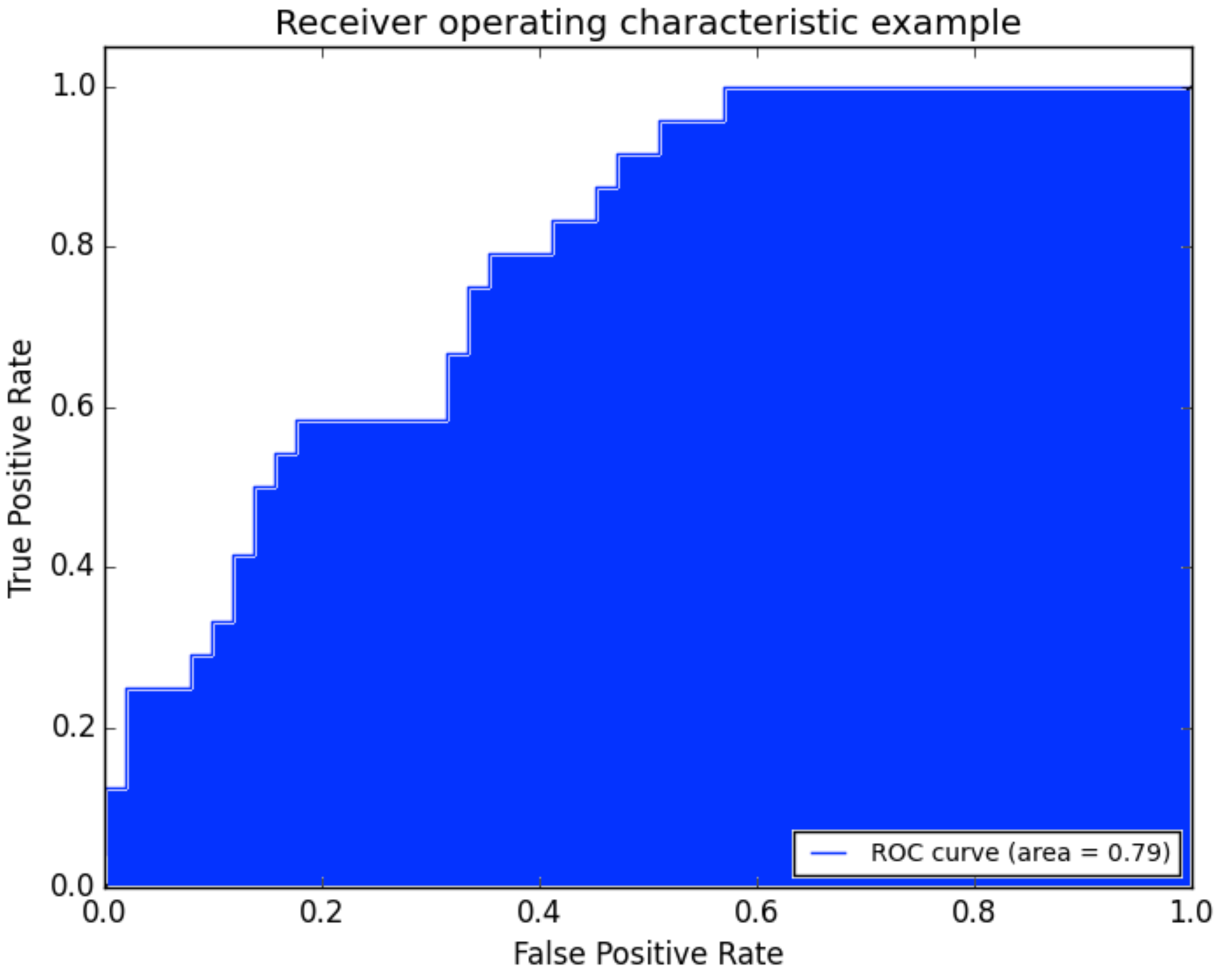
ADVANCED CLASSIFICATION METRICS

- Finally, we create a full curve that is described by TPR and FPR



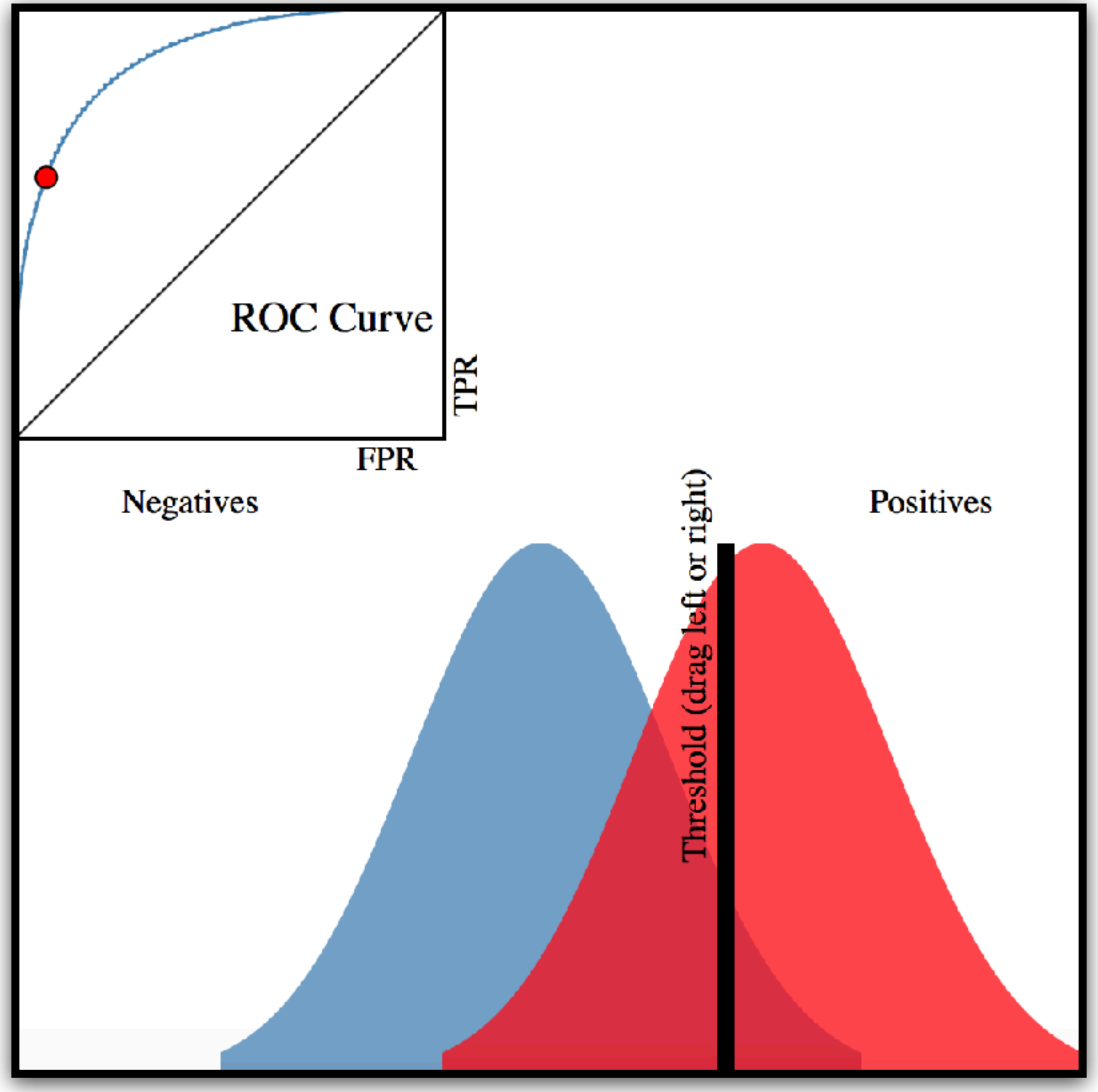
ADVANCED CLASSIFICATION METRICS

- With this curve, we can find the Area Under the Curve (AUC)



ADVANCED CLASSIFICATION METRICS

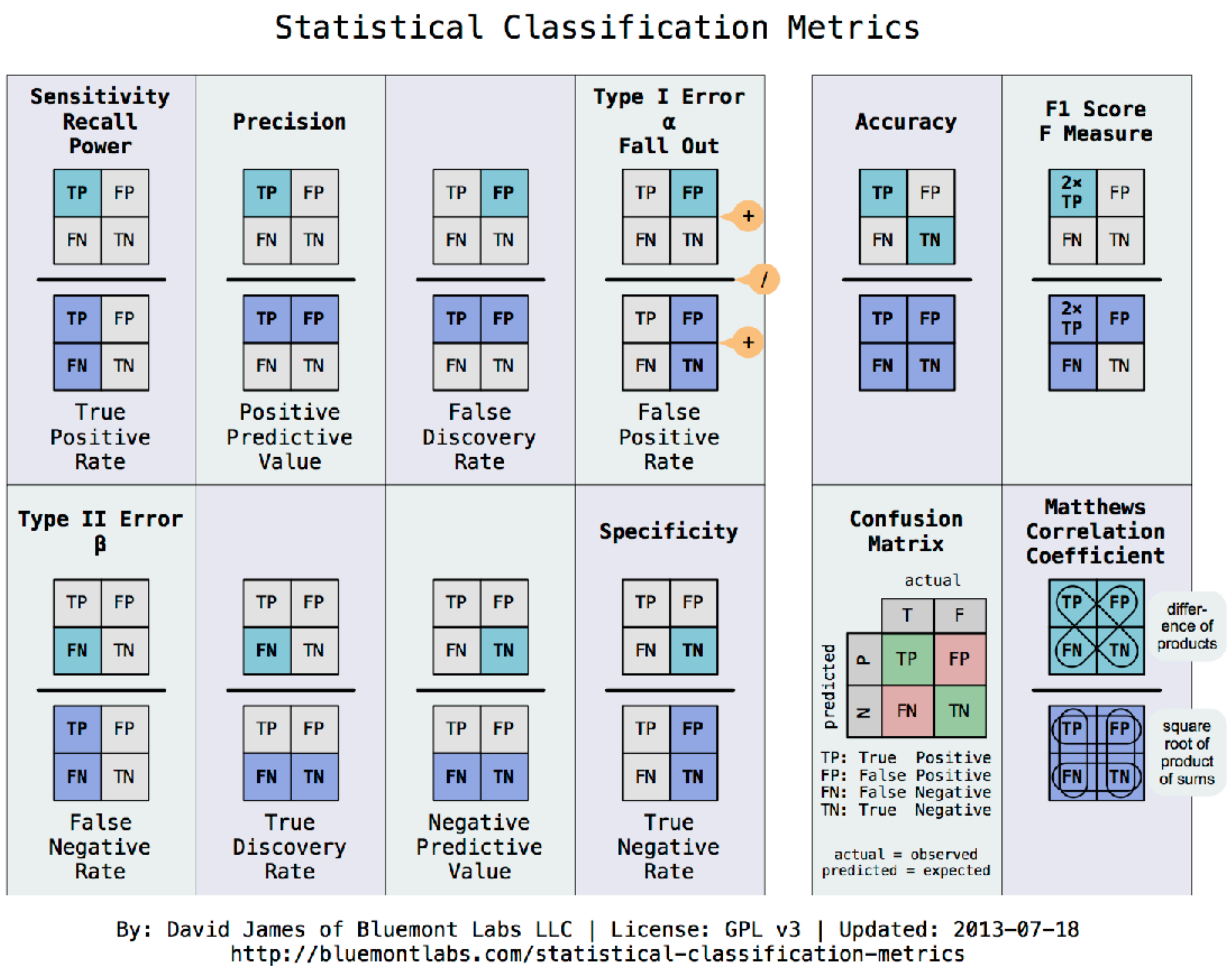
- This interactive visualisation can help practice visualising ROC curves



ADVANCED CLASSIFICATION METRICS

- Sensitivity or True Positive Rate (TPR) $$TPR = \frac{TP}{P} = \frac{TP}{(TP + FN)}$$
- Specificity (SPC) or True Negative Rate (TNR) $$SPC = \frac{TN}{N} = \frac{TN}{(FP + TN)}$$
- Precision or Positive Predictive Value (PPV) $$PPV = \frac{TP}{(TP + FP)}$$
- Negative Predictive Value (NPV) $$NPV = \frac{TN}{(TN + FN)}$$
- Fall-out or False Positive Rate (FPR) $$FPR = \frac{FP}{N} = \frac{FP}{(FP + TN)}$$
- False Discovery Rate (FDR) $$FDR = \frac{FP}{(FP + TP)} = 1 - PPV$$
- Miss Rate or False Negative Rate (FNR) $$FNR = \frac{FN}{(FN + TP)}$$

ADVANCED CLASSIFICATION METRICS



- There are several other common metrics that are similar to TPR and FPR.
- scikit-learn has all of the metrics located on one page

ADVANCED CLASSIFICATION METRICS

- If we have a TPR of 1 (all positives are marked positive) and FPR of 0 (all negatives are not marked positive), we would have an AUC of 1
  - This means everything was accurately predicted
- If we have a TPR of 0 (all positives are not marked positive) and an FPR of 1 (all negatives are marked positive), we would have an AUC of 0
  - This means nothing was predicted accurately
- An AUC of 0.5 would suggest randomness (somewhat) and is an excellent benchmark to use for comparing predictions (i.e. is my AUC above 0.5?)

GUIDED PRACTICE

WHICH METRIC SHOULD I USE?

ACTIVITY: WHICH METRIC SHOULD I USE?

EXERCISE

DIRECTIONS: (15 MINUTE)

While AUC seems like a “golden standard”, it could be further improved depending upon your problem. There will be instances where error in positive or negative matches will be very important. For each of the following examples:

1. Write a confusion matrix: true positive, false positive, true negative, false negative. Then decide what each square represents for that specific example
2. Define the benefit of a true positive and true negative
3. Define the cost of a false positive and false negative
4. Determine at what point does the cost of a failure outweigh the benefit of a success? This would help you decide how to optimise TPR, FPR and AUC

Examples:

1. A test is developed for determining if a patient has cancer or not
2. A newspaper company is targeting a marketing campaign for “at risk” users that may stop paying for the product soon
3. You build a spam classifier for your email system

INDEPENDENT PRACTICE

EVALUATING LOGISTIC REGRESSION WITH ALTERNATIVE METRICS

ACTIVITY: WHICH METRIC SHOULD I USE?

EXERCISE

DIRECTIONS: (35 MINUTE)

Kaggle’s common online exercise is exploring survival data from the Titanic

1. Spend a few minutes determining which data would be most important to use in the prediction problem. You may need to create new features based on the data available. Consider using a feature selection aide in scikit-learn. For a worst case scenario, identify one or two strong features that would be useful to include in this model.
2. Spend 1-2 minutes considering which metric makes the most sense to optimise. Accuracy? FPR or TPR? AUC? Given the business problem of understanding survival rate aboard the Titanic, why should you use this metric?
3. Build a tuned Logistic model. Be prepared to explain your design (including regularisation), metric and feature set in predicting survival using any tools necessary (such as a fit chart). Use the starter code to get you going.

TOPIC REVIEW

TOPIC REVIEW

- What is the link function used in logistic regression?
- What kind of machine learning problems does logistic regression address?
- What do the coefficients in a logistic regression represent?
  - How does the interpretation differ from ordinary least squares?
  - How is it similar?
- How does True Positive Rate and False Positive Rate help explain accuracy?
- What would an AUC of 0.5 represent for a model?
  - What about an AUC of 0.9?
- Why might one classification metric be more important to tune than another? Give an example of a business problem or project where this would be the case

BEFORE NEXT CLASS

DUE DATE

- Project
  - Unit Project 3

Q & A