GENERAL ASSEMBLY

INTRODUCTION TO LOGISTIC REGRESSION

Juan Ginzo

PRE-WORK REVIEW

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

TODAY

INTRODUCTION TO LOGISTIC REGRESSION

LEARNING OBJECTIVES

- Build a Logistic regression classification model using the scikit-learn library
- Describe the 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

INTRODUCTION TO LOGISTIC REGRESSION

ANSWER THE FOLLOWING QUESTIONS

Read through the following questions and brainstorm answers for each:

1. What are the main differences between linear regression models (ordinary least squares, OLS) and KNN models? What is different about how they approach solving the problem?
    a. For example, what is *interpretable* about OLS compared to what's *interpretable* in KNN?

2. What would be the advantage of using a linear model like OLS to solve a classification problem, compared to KNN?
    a. What are some challenges for using OLS to solve a classification problem (say, if the values were either 1 or 0)?

LOGISTIC REGRESSION

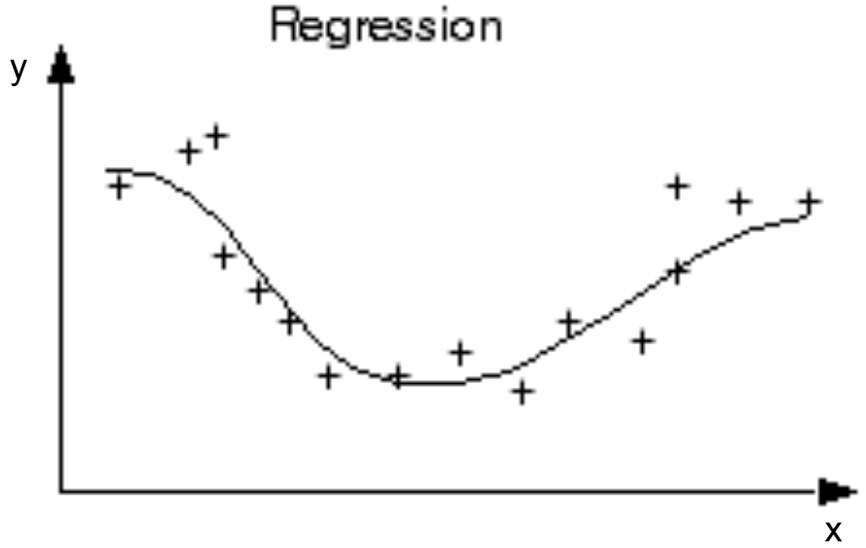
LOGISTIC REGRESSION

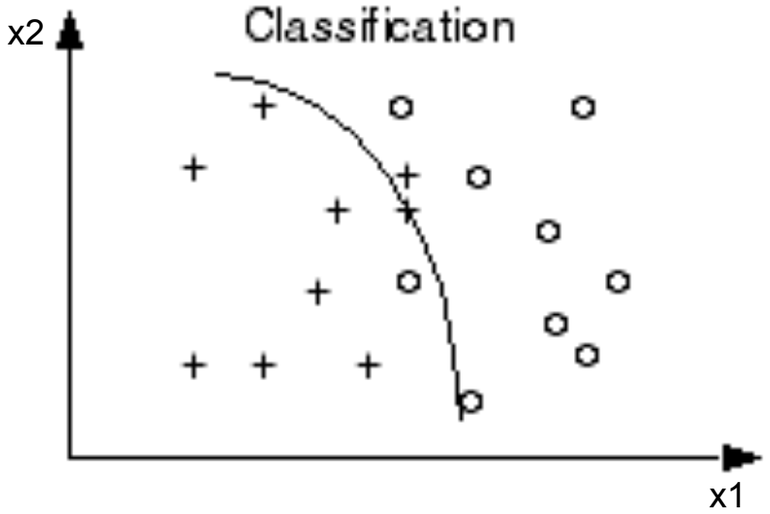
- Logistic regression is a *linear* approach to solving a *classification* problem.

- That is, we can use a linear model, similar to linear regression, in order to solve if an item *belongs* or *does not belong* to a class label.

WHY NOT 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.





WHY NOT 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? NO!

PROBABILITY

- One approach is predicting the probability that an observation belongs to a certain class, instead of the actual outcome.

- For example, suppose we know that roughly 700 of 2200 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 more information about passengers.

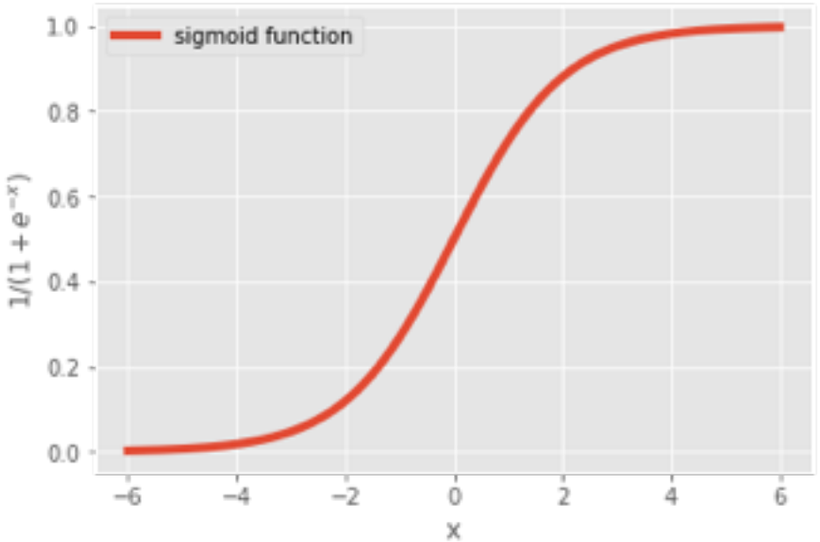
ACTIVITY: KNOWLEDGE CHECK

ANSWER THE FOLLOWING QUESTIONS BASED ON THE TITANIC EXAMPLE

1. Recall the relation between predictor and response variables in linear regression.
2. What's our response variable now?
3. What would be the analogue of the intercept, if we were trying to predict probability of survival? (In the context of logistic regression, it is often referred to as the bias or prior probability.)

THE SIGMOID FUNCTION

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

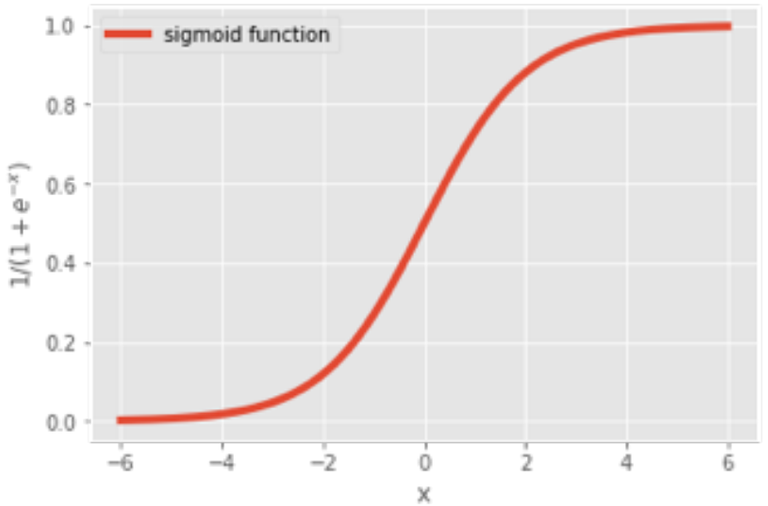


- Mathematically, it is defined as

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

THE SIGMOID FUNCTION

- Recall that the exponential function is the *inverse* of the natural log.
- As x increases, the result is closer to 1. As x decreases, the result is closer to 0.
- When x = 0, the result is 0.5.



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.

LOGISTIC FUNCTION

- For classification, we need a distribution associated with categories: given all events, what is the probability of a given event?
- In Logistic Regression, we use the ***logistic function,***

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}.$$

- This is a special case of a sigmoid function.

LOGISTIC FUNCTION

- The coefficients are found by fitting the function with a method called **maximum likelihood.**

- We seek estimates of the coefficients such that the predicted probability for each observation corresponds as closely as possible to the observed class (0 or 1).

- We **won't** get into the mathematical details of calculating the maximum likelihood estimates.

DEMO

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 the Euler constant e ~ 2.71.
- Do we get the “S” shape we expect?

ODDS AND LOGIT

ODDS

- With a bit of arithmetics applied on the logistic function, we can get to the following formula:

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}.$$

- The quantity on the right side is called an odd, which can take any value between 0 and infinite.

- Values for the odds close to 0 indicate a very low probability of a positive outcome.

- For example, if we were trying to predict probability of default, having odds of 9, would imply a probability of 0.9 of default, since 0.9 / (1 - 0.9) = 9

LOGIT FUNCTION

- One advantage of Logistic Regression to OLS is that it allows us to model this relationship between explanatory variables and probability of outcome with a *link function* called the **logit function or log odds**.

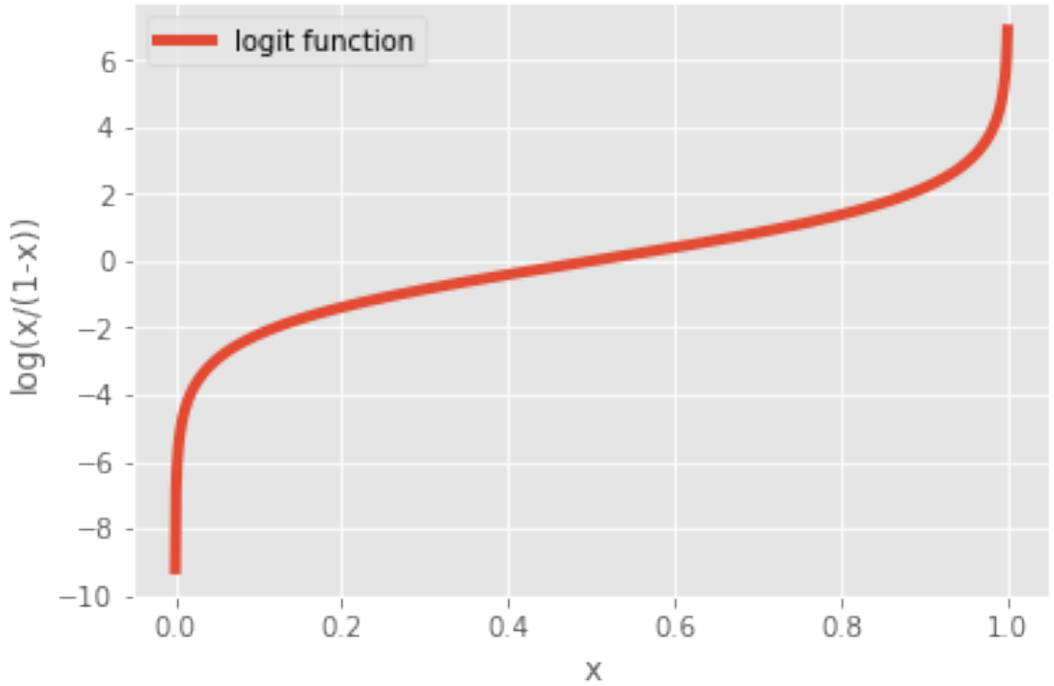
$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X.$$

- The logit function allow us to obtain probabilities from a linear function of the predictors.

- We can now form a specific relationship between our linear predictors and the response variable.

ODDS AND LOG-ODDS

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



ODDS AND LOG-ODDS

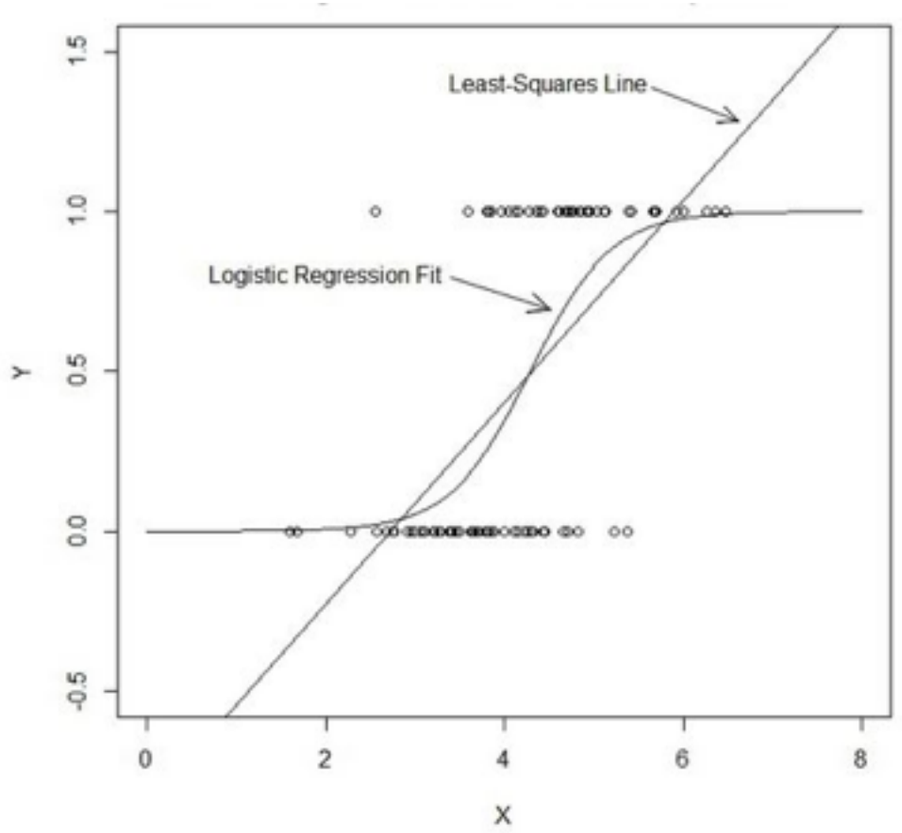
- Check that replacing the log odds into our sigmoid function will give us back the probabilities for the outcomes (defined by the logistic function).

$$f(x) = \frac{1}{1 + \mathrm{e}^{-x}} \qquad \frac{p(X)}{1 - p(X)} = e^{\beta_0 + \beta_1 X}.$$

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}.$$

ODDS AND LOG-ODDS

- The logit function takes odds between 0 and ∞ and returns values between 0 and 1.



ACTIVITY: KNOWLEDGE CHECK

EXERCISE

ANSWER THE FOLLOWING QUESTIONS

1. Why is it important to take values between -∞ and ∞, but provide probabilities between 0 and 1?

ODDS AND LOG-ODDS

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

- Solving for the probability gives

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

- To calculate this in python, we could use

$$1 \ / \ (1 + \text{numpy.exp}(-0.2))$$

ODDS AND LOG-ODDS

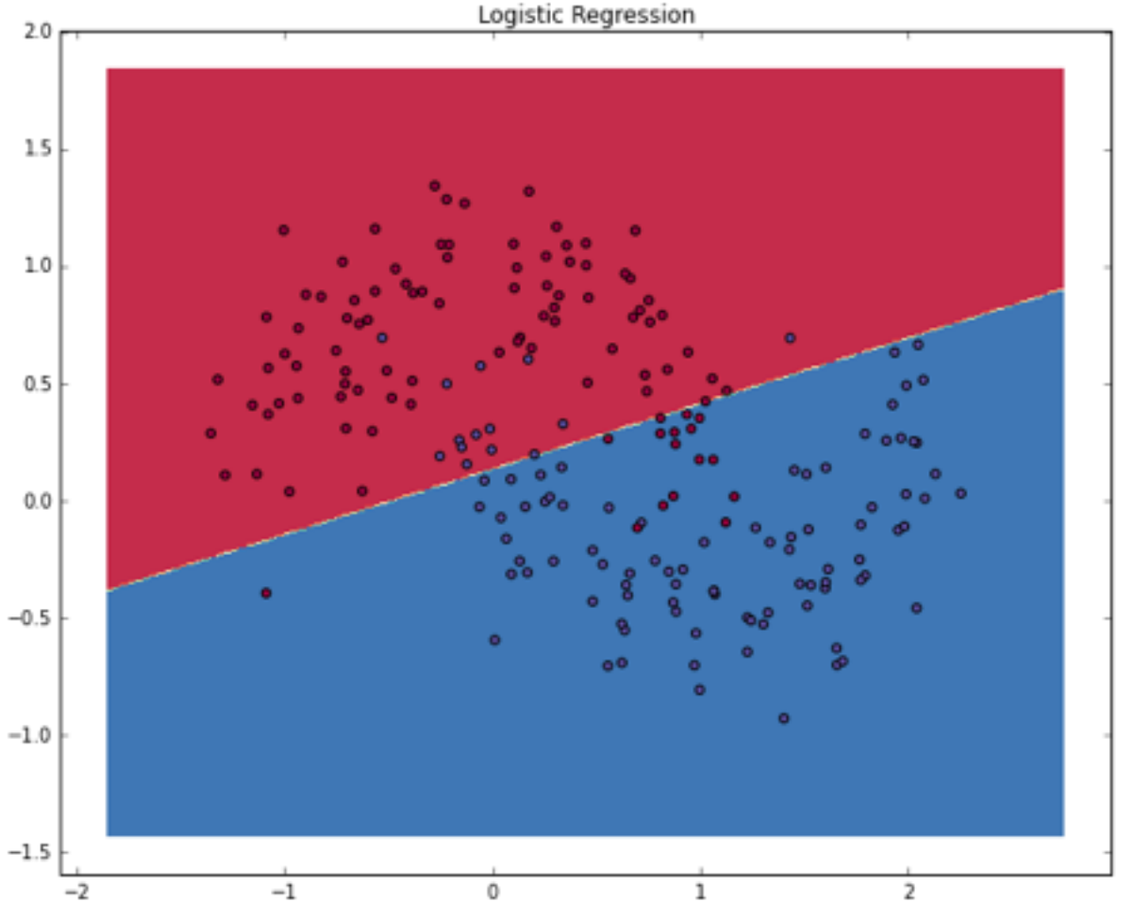
- While the logit value represents the combination of *coefficients and feature values* in the logistic function, we can convert them into odds ratios that make them more easily interpretable.

- The odds multiply by e^{β_1} for every 1-unit increase in x.

$$\text{oddsratio} = \frac{\text{odds}(\text{x}+1)}{\text{odds}(\text{x})} = \frac{e^{\beta_0+\beta_1(x+1)}}{e^{\beta_0+\beta_1 x}} = e^{\beta_1}$$

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

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. AlphaGo : Seedol, 4:1

    b. Chelsea : Leicester City, 20:1

    c. England : Wales, 1.1:1

    d. Brexit : Remain, 7:4

    e. President Trump : Not President Trump, 4:11

ACTIVITY: WAGER THOSE ODDS!

STARTER CODE

```
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

DIRECTIONS

Use the dataset `titanic.csv` and the `LogisticRegression` estimator in sklearn to predict the target variable Survived.

1. What is the bias, or prior probability, of the dataset?
2. Build a simple model with one feature and explore the `coef_` value. Does this represent the odds or logit (log odds)?
3. Build a more complicated model using multiple features. Interpreting the odds, which features have the most impact on Survival? Which features have the least?
4. What is the accuracy of your model?

ADVANCED CLASSIFICATION METRICS

ADVANCED CLASSIFICATION METRICS

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

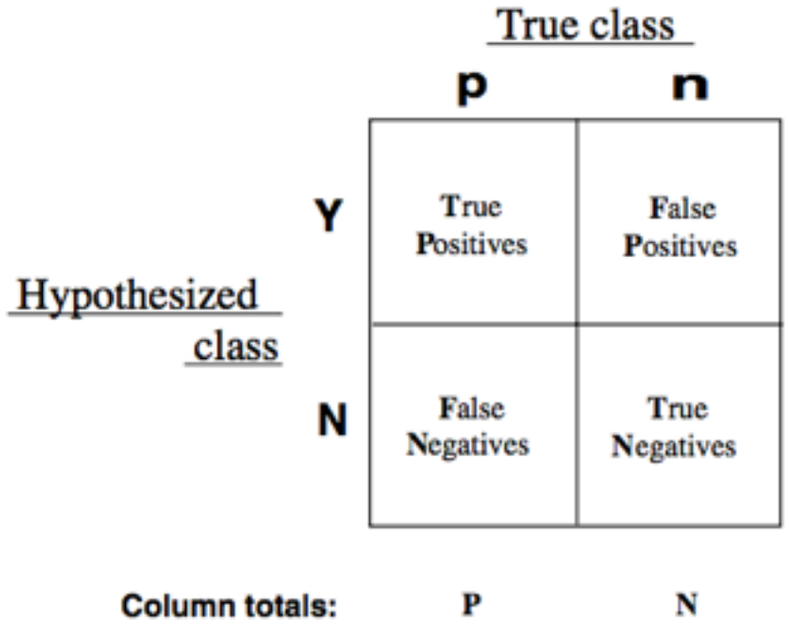
ADVANCED CLASSIFICATION METRICS

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

ADVANCED CLASSIFICATION METRICS

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

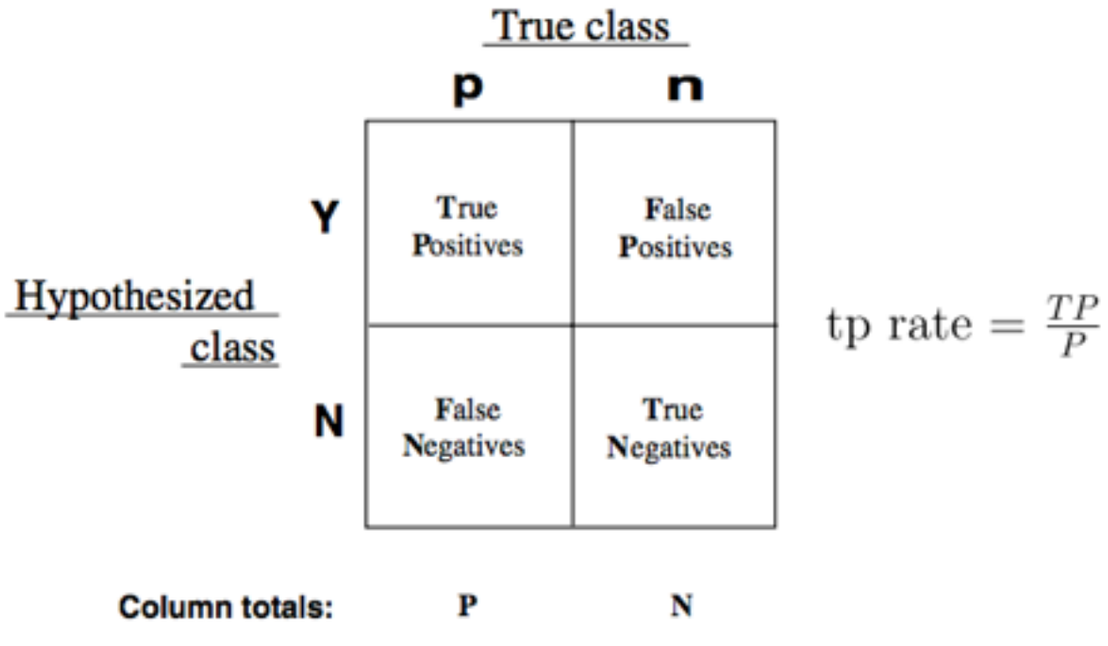
- For each label, we can put it into the category of a true positive, false positive, true negative, or false negative.



ADVANCED CLASSIFICATION METRICS

- 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?

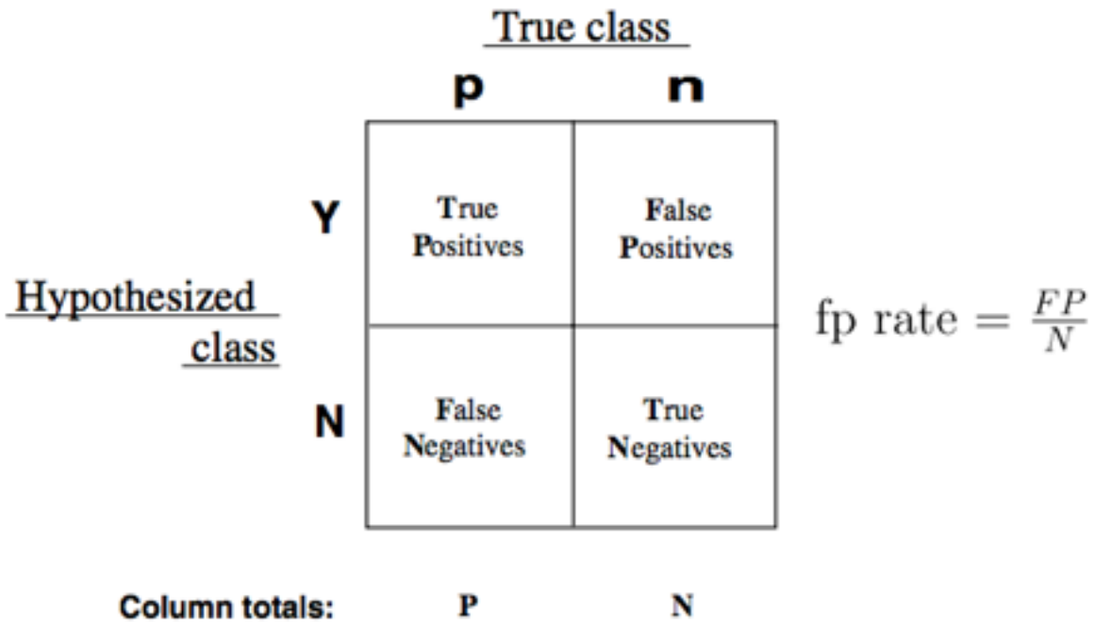


$$\text{tp rate} = \frac{TP}{P}$$

ADVANCED CLASSIFICATION METRICS

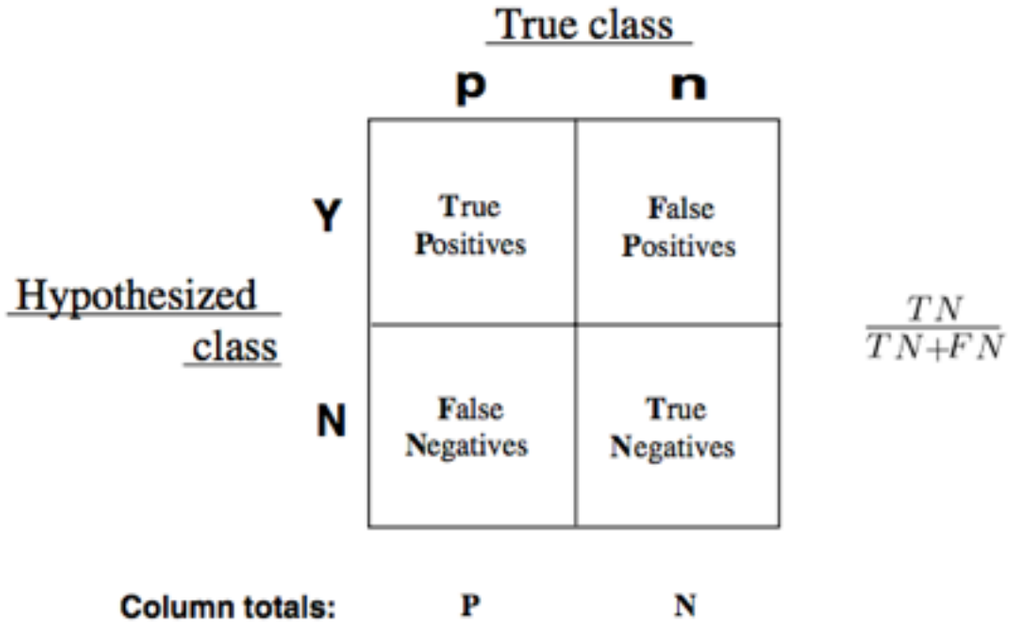
- False Positive Rate (FPR) asks, “Out of all items not belonging to a class label, how many were predicted as belonging to that target class label?”

- 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?



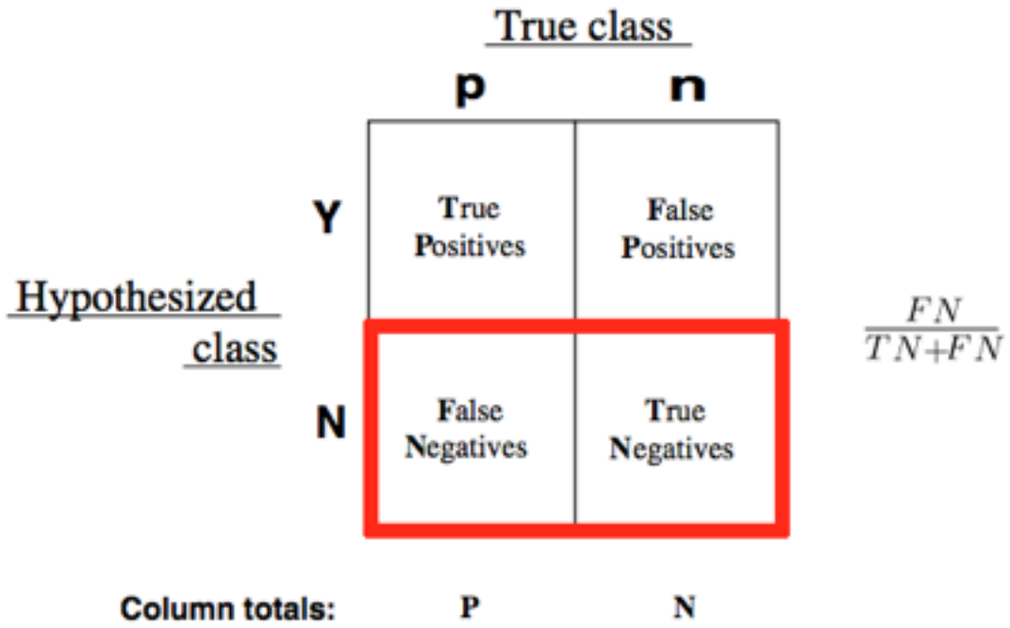
ADVANCED CLASSIFICATION METRICS

- These can also be inverted.
- How often does a test *correctly* identify patients without cancer?



ADVANCED CLASSIFICATION METRICS

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



ADVANCED CLASSIFICATION METRICS

- The true positive and false positive rates gives us a much clearer picture of where predictions begin to fall apart.

- This allows us to adjust our models accordingly.

ADVANCED CLASSIFICATION METRICS

- A good classifier would have a true positive rate approaching 1 and a false positive rate approaching 0.

- In our cancer problem, this model would accurately predict *all* of the patients with cancer as having cancer and not accidentally predict any of the patients not having cancer as having cancer.

ADVANCED CLASSIFICATION METRICS

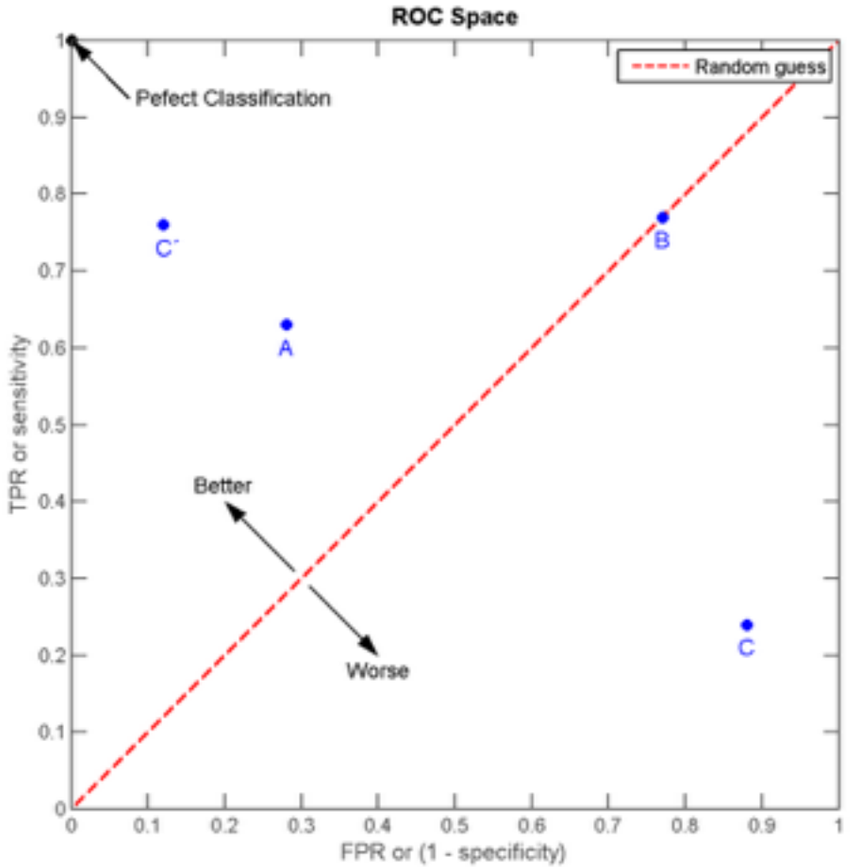
- We can vary the classification threshold for our model to get different predictions. But how do we know if a model is better overall than another model?
- We can compare the FPR and TPR of the models, but it can often be difficult to optimise two numbers at once.
- Logically, we like a single number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

ADVANCED CLASSIFICATION METRICS

- This is where the Receiver Operating 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.

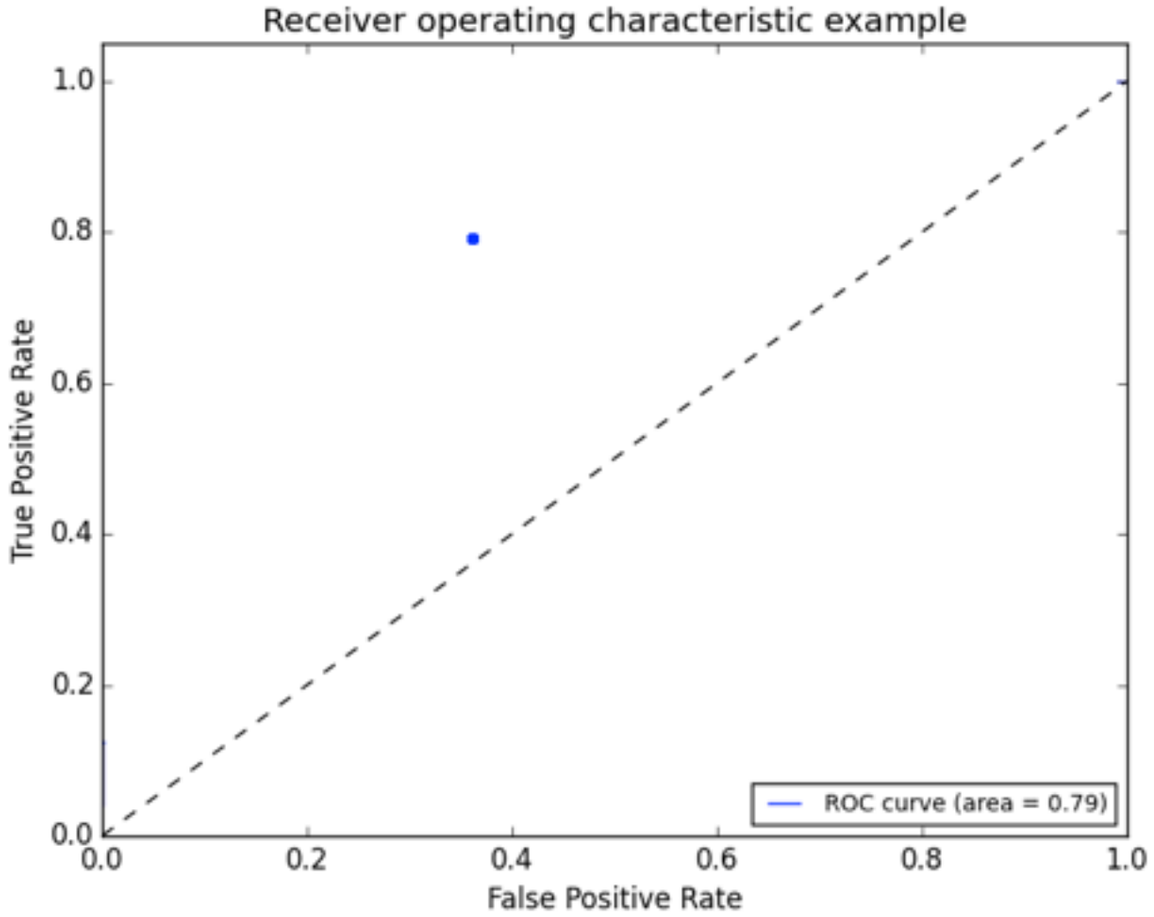
ADVANCED CLASSIFICATION METRICS

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



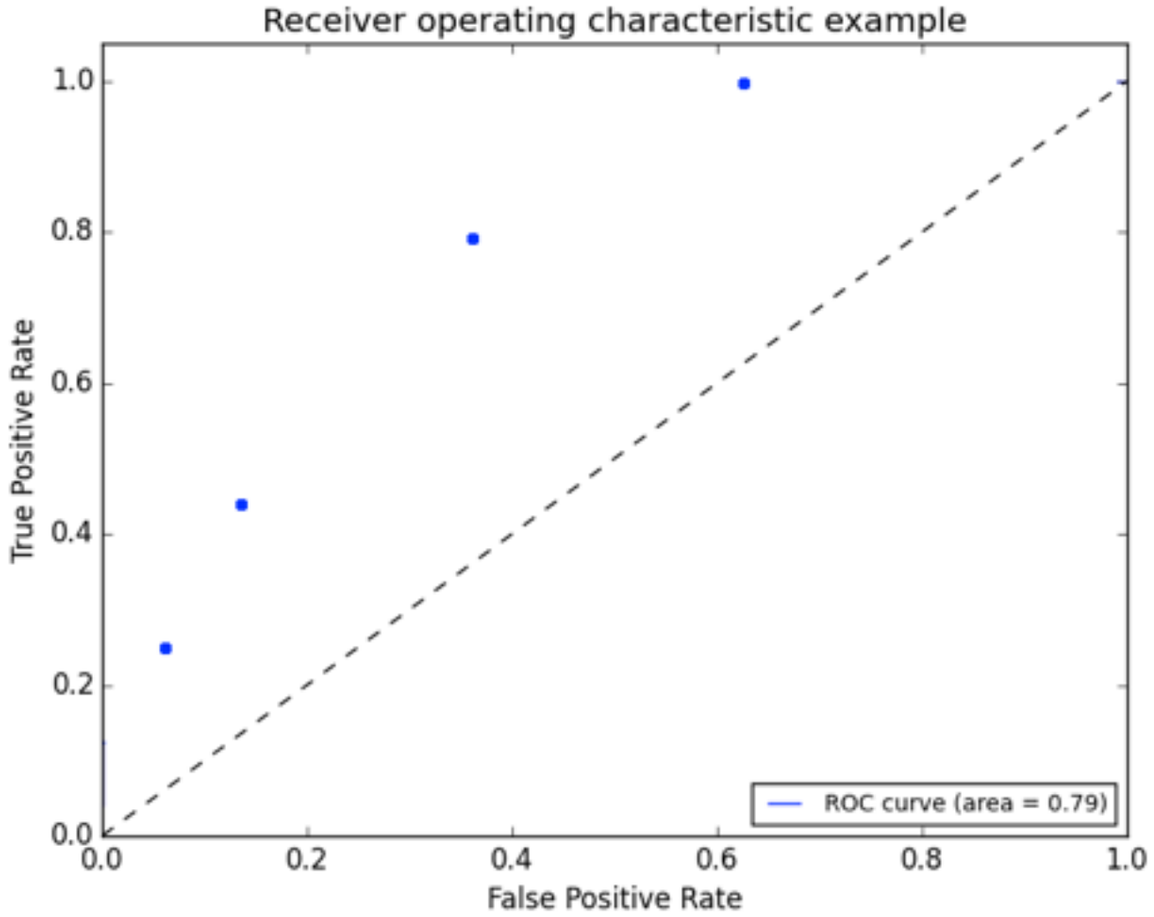
ADVANCED CLASSIFICATION METRICS

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



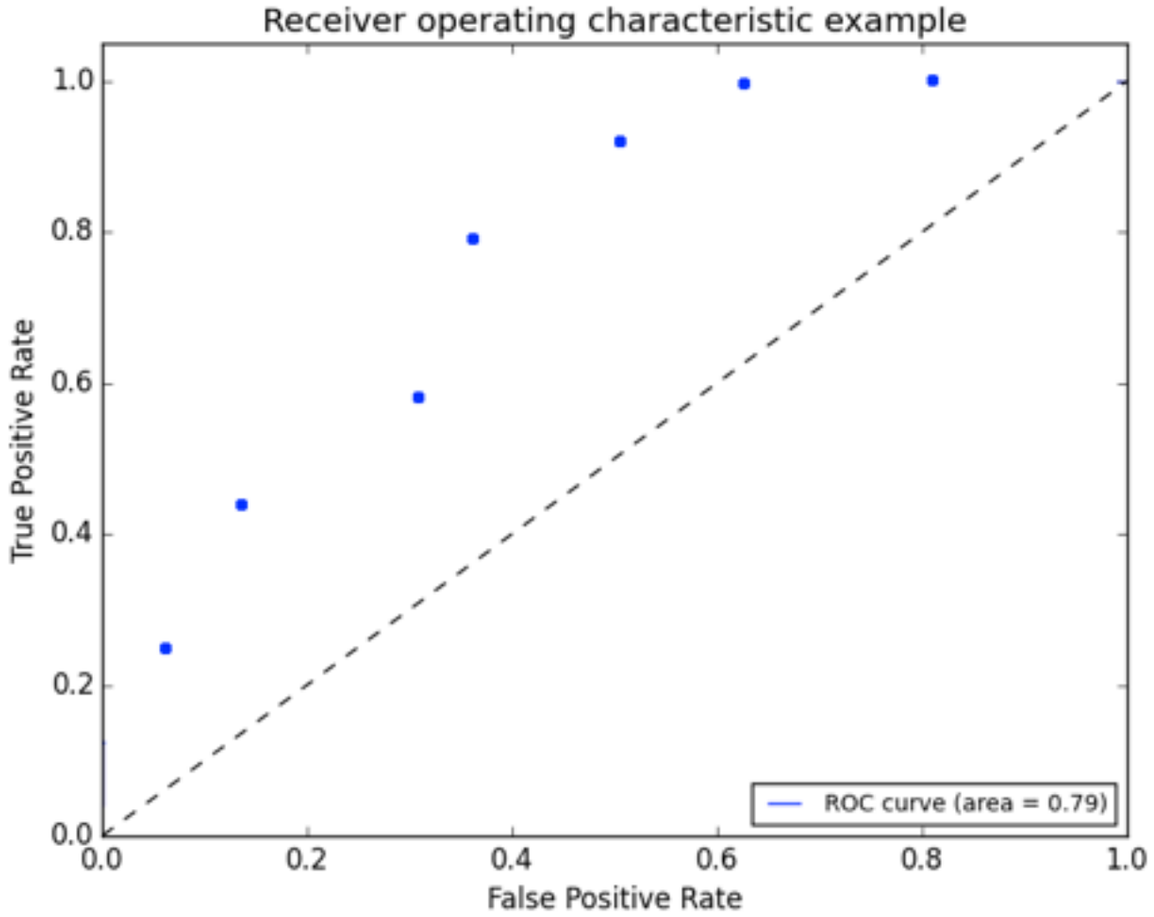
ADVANCED CLASSIFICATION METRICS

- We can continue adding pairs for different thresholds



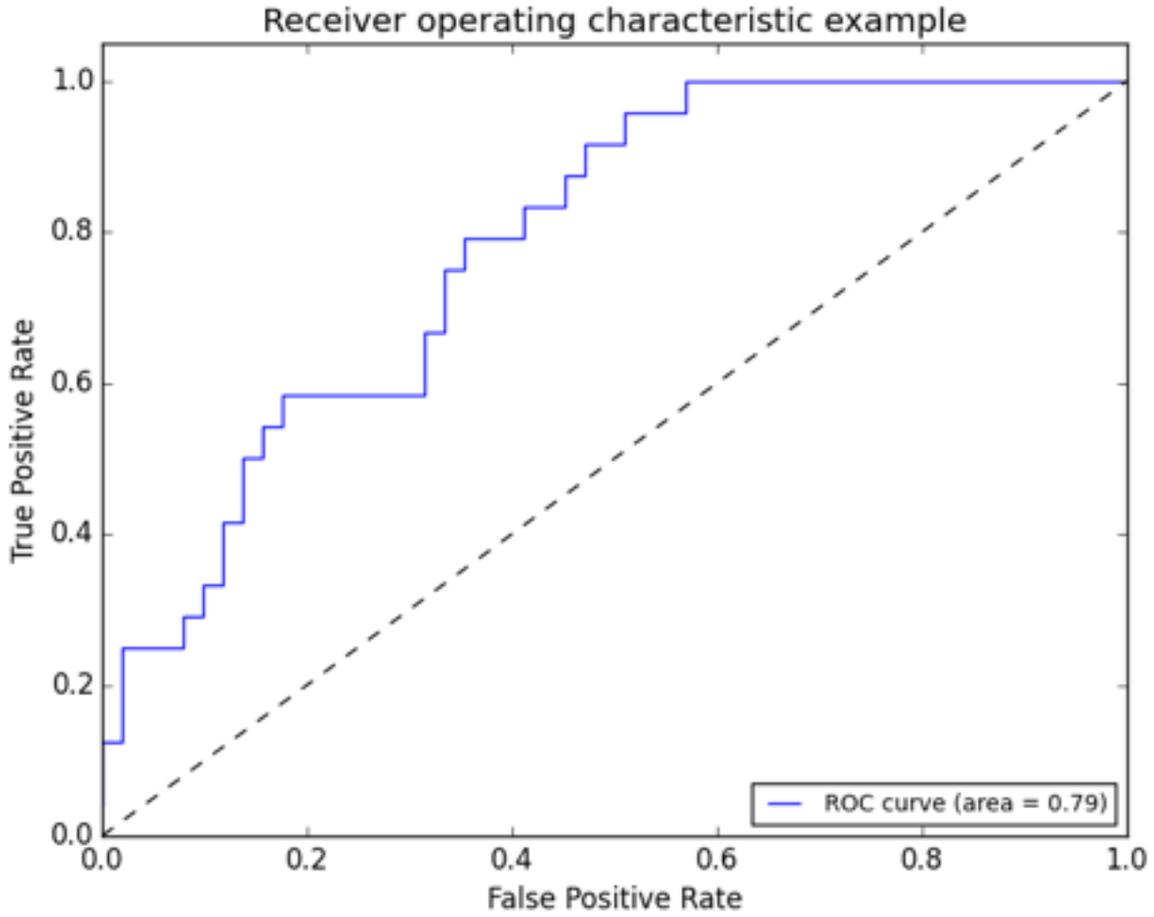
ADVANCED CLASSIFICATION METRICS

- We can continue adding pairs for different thresholds



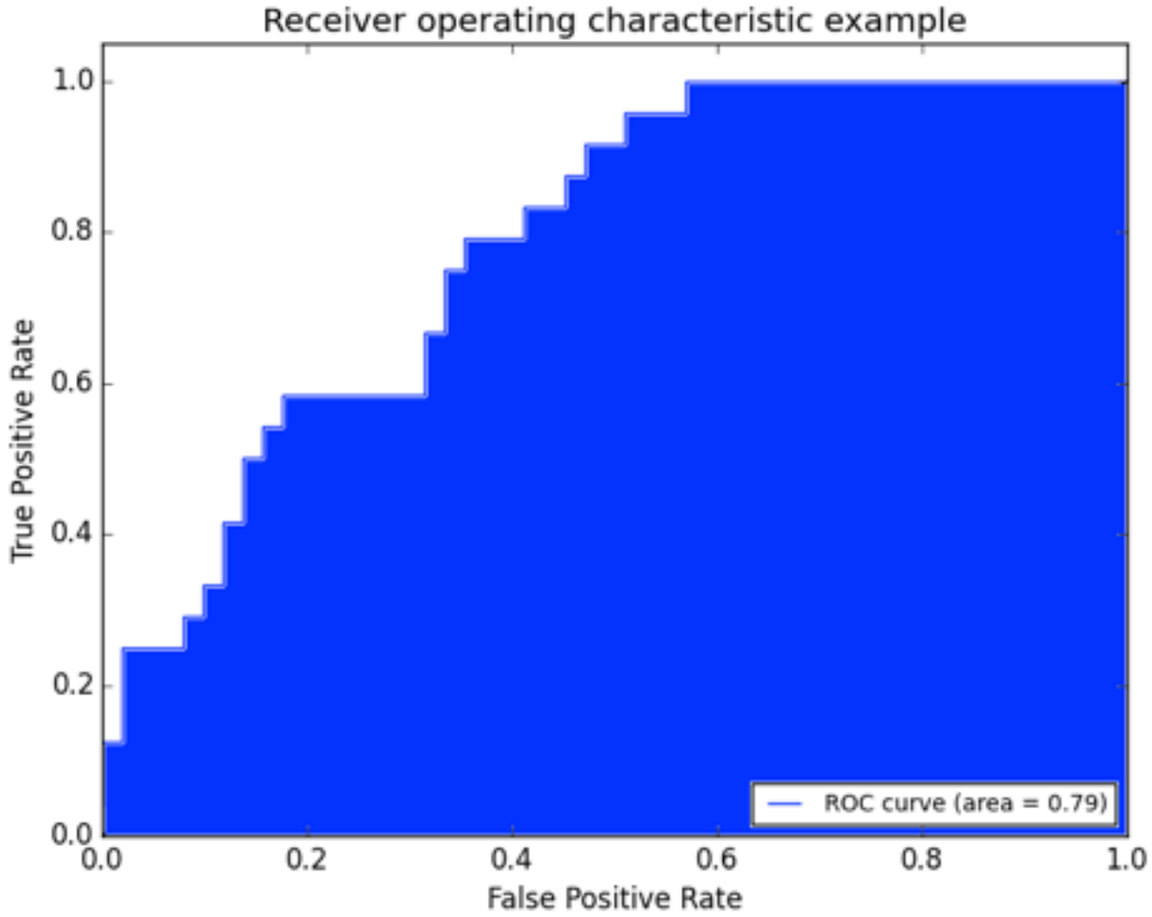
ADVANCED CLASSIFICATION METRICS

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



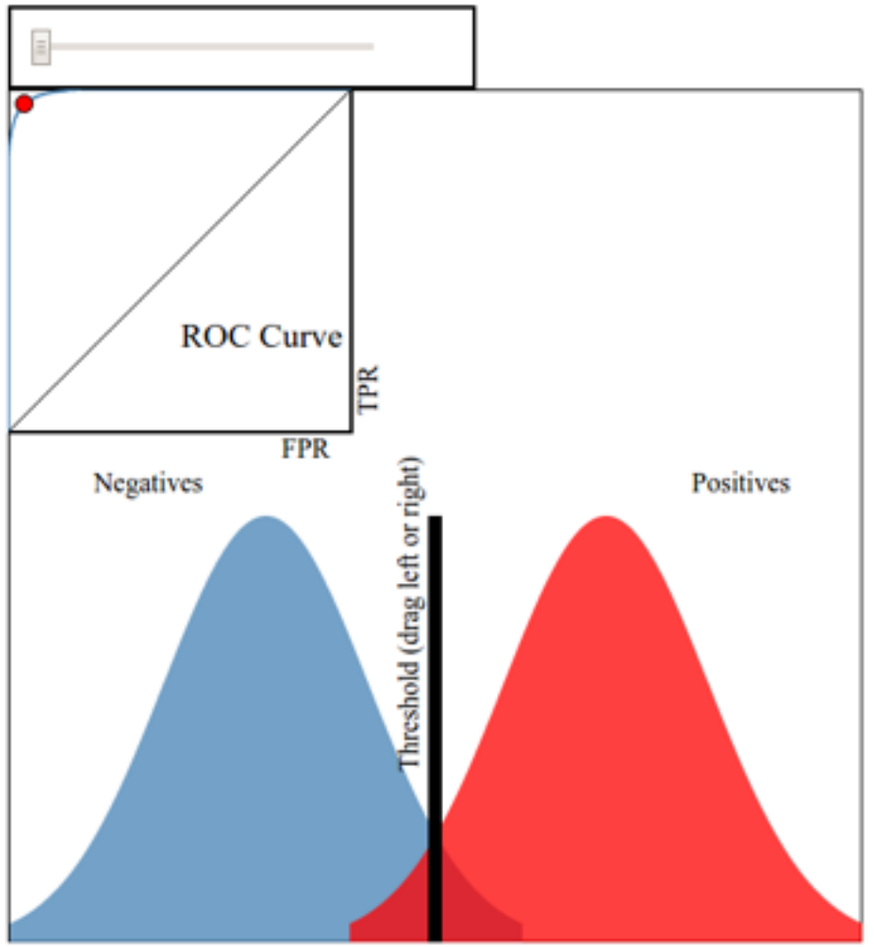
ADVANCED CLASSIFICATION METRICS

- 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

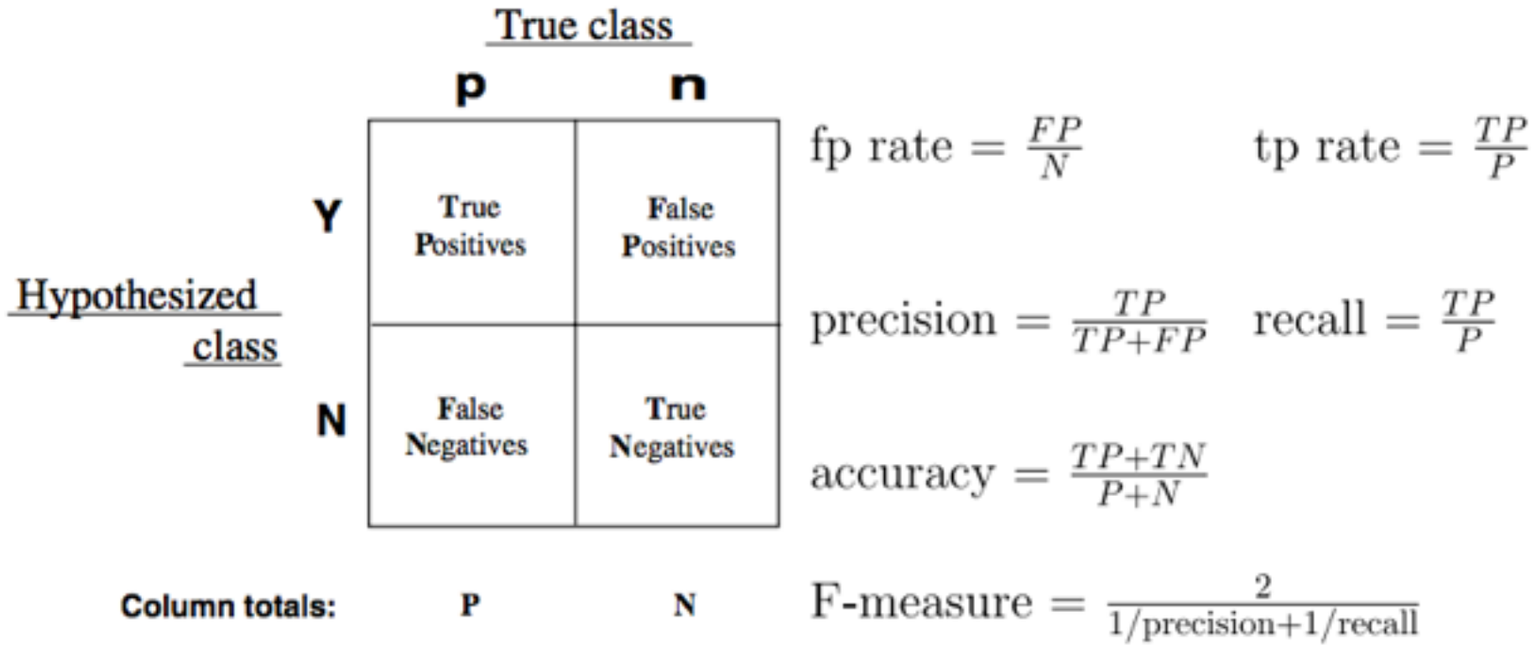
- If we have a TPR of 1 (all positives are marked positive) and FPR of 0 (all negatives are not marked positive), we have an AUC of 1. This means everything was accurately predicted.

- If we have a TPR of 0 (all positives are not marked positive) and an FPR of 1 (all negatives are marked positive), we 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?).

ADVANCED CLASSIFICATION METRICS

- There are several other common metrics that are similar to TPR and FPR.



- Sklearn has all of the metrics located on one convenient page.

GUIDED PRACTICE

WHICH METRIC SHOULD I USE?

ACTIVITY: WHICH METRIC SHOULD I USE?

EXERCISE

DIRECTIONS

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.

ACTIVITY: WHICH METRIC SHOULD I USE?

DIRECTIONS

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.

TOPIC REVIEW

REVIEW QUESTIONS

- What's 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?

REVIEW QUESTIONS

- 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.