FOR INSTRUCTOR PURPOSES ONLY

INSTRUCTOR NOTES

- Insert Text Here

FOR INSTRUCTOR PURPOSES ONLY

MATERIALS

- Insert Text Here

FOR INSTRUCTOR PURPOSES ONLY

PRE-WORK

- Insert Text Here

GENERAL ASSEMBLY

INTRODUCTION TO LOGISTIC REGRESSION

Chris Connell

LEARNING OBJECTIVES

- Build a Logistic regression classification model using the scikit learn library
- 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, and tune via grid search for regularization

COURSE

PRE-WORK

PRE-WORK REVIEW

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

INTRODUCTION TO LOGISTIC REGRESSION

EXERCISE

ANSWER THE FOLLOWING QUESTIONS

Read through the following questions and brainstorm answers for each:

1. What are the main differences between linear 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?

1. 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)?

DELIVERABLE

Answers to the above questions

OPENING

ODDS AND PROBABILITIES

PROBABILITIES

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

Fair coin flip

$$P(heads) = \frac{1}{2} = 0.5$$

Fair die roll

$$P(1\ or\ 2) = \frac{2}{6} = \frac{1}{3} = 0.333$$

Deck of playing cards

$$P(diamond\ card) = \frac{13}{52} = \frac{1}{4} = 0.25$$

ODDS

$$odds = \frac{P(occurring)}{P(not\ occurring)}$$

$$odds = \frac{p}{1-p}$$

Fair coin flip

$$odds(heads) = \frac{0.5}{0.5} = 1\ or\ 1:1$$

Fair die roll

$$odds(1\ or\ 2) = \frac{0.333}{0.666} = \frac{1}{2} = 0.5\ or\ 1:2$$

Deck of playing cards

$$odds(diamond\ card) = \frac{0.25}{0.75} = \frac{1}{3} = 0.333\ or\ 1:3$$

ODDS RATIO

The odds ratio is exactly what it says it is, a ratio of two odds

Fair coin flip

$$P(heads) = \frac{1}{2} = 0.5$$

$$odds(heads) = \frac{0.5}{0.5} = 1 \; or \; 1:1$$

Loaded coin flip

$$P(heads) = \frac{7}{10} = 0.7$$

$$odds(heads) = \frac{0.7}{0.3} = 2.333$$

$$Odds\ ratio = \frac{odds_1}{odds_0}$$

$$Odds\ ratio = \frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_0}}$$

$$Odds\ ratio = \frac{\frac{.7}{.3}}{\frac{.5}{.5}} = \frac{.7}{.3} \times \frac{.5}{.5} = \frac{.35}{.15} = 2.333$$

The odds of getting "heads" on the loaded coin are 2.333x greater than the fair coin.

OPENING

INTRODUCTION TO LOGISTIC REGRESSION

Logistic regression is a generalization of the linear regression model to classification problems

- The name is somewhat misleading
  - "Regression" comes from fact that we fit a linear model to the feature space
  - But it is really a technique for classification, not regression

- We use a linear model, similar to linear regression, in order to solve if an item *belongs* or *does not* belong to a class model
  - It is a binary classification technique: $y = \{0, 1\}$
  - Our goal is to classify correctly two types of examples:
    - Class 0, labeled as 0, e.g., *"belongs"*
    - Class 1, labeled as 1, e.g., *"does not belong"*

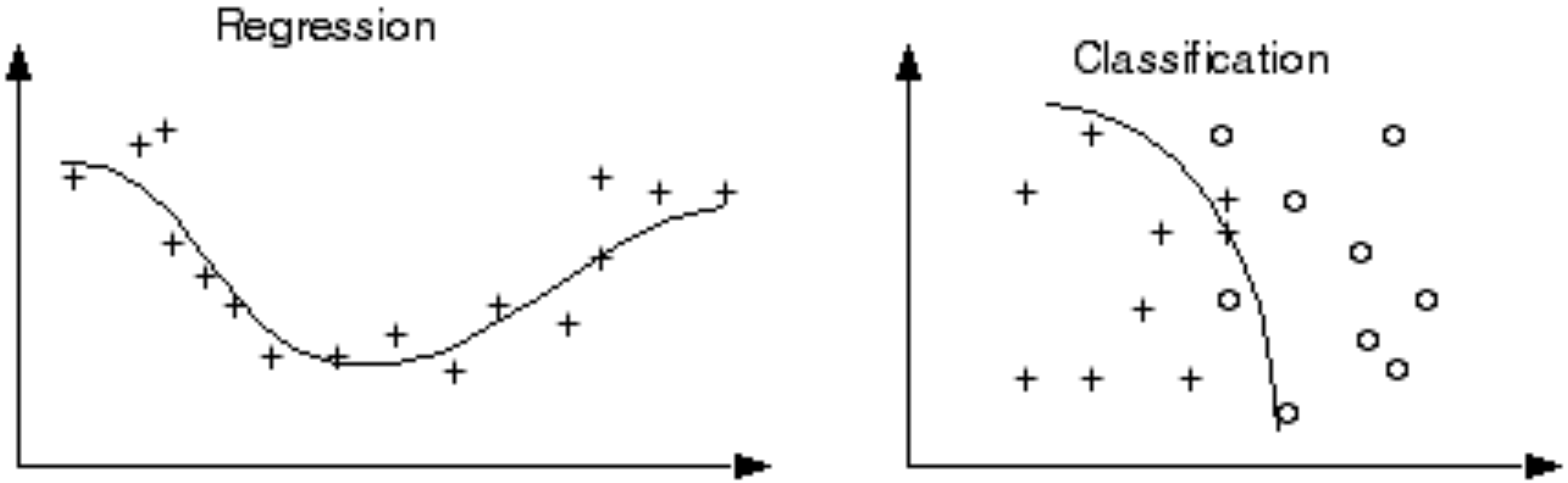
Why is logistic regression so valuable to know?

- It addresses many commercially valuable classification problems, such as:
- Fraud detection (e.g., payments, e-commerce)
- Churn prediction (marketing)
- Medical diagnoses (e.g., is the test positive or negative?)
- and many, many others…

CHALLENGE! LINEAR REGRESSION RESULTS FOR CLASSIFICATION

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

- Classification is used predict class labels by select a line that separates them



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?

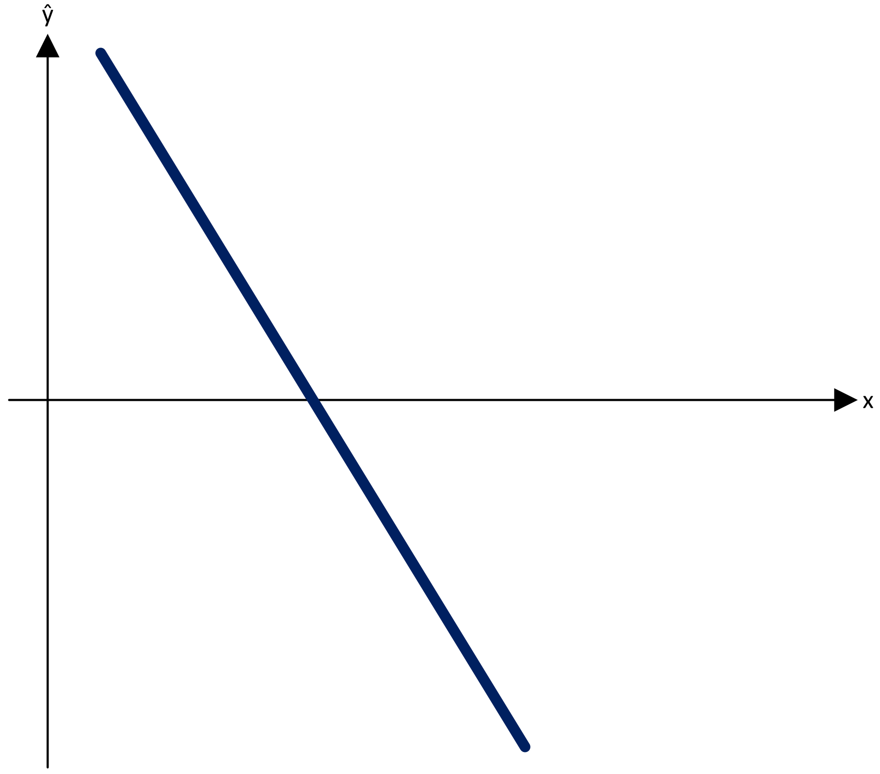
REGRESSION RESULTS FOR CLASSIFICATION

With linear regression, $\hat{y}$ is in $]-\infty;\ +\infty[$, not $[0;1]$. How do we fix this for logistic regression?

- The key variable in any regression problem is the outcome variable $\hat{y}$ given the covariate x

$$\hat{y} = \hat{\beta}x$$

- With linear regression, $\hat{y}$ takes values in $]-\infty;\ +\infty[$
- However, with logistic regression, $\hat{y}$ takes values in the unit interval $[0;1]$



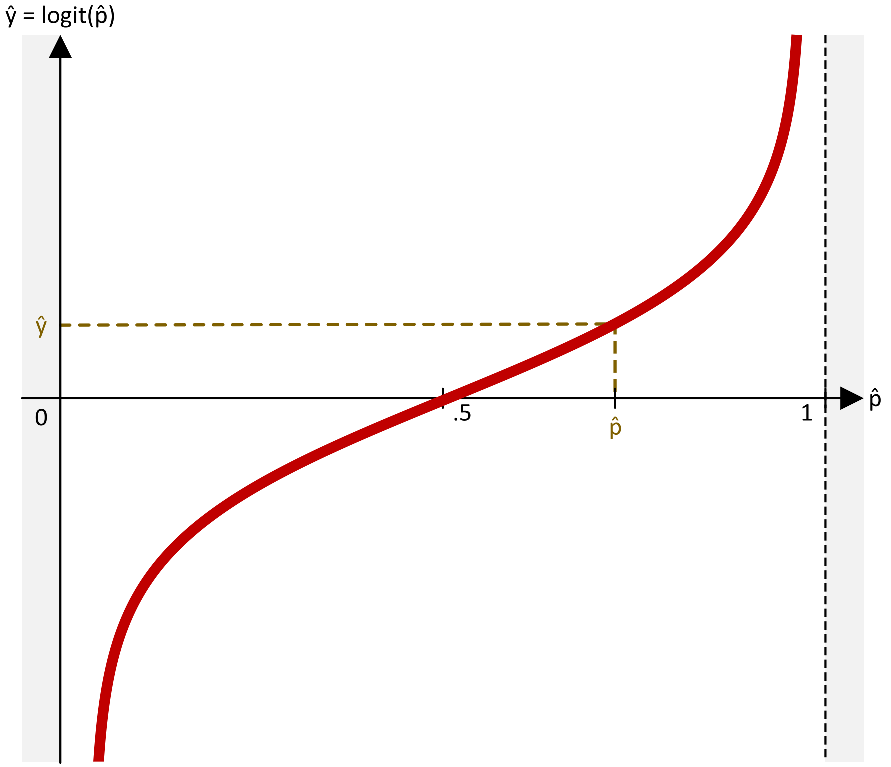
LINK FUNCTIONS AND THE SIGMOID FUNCTION

- 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.
- 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.

With transformations called the *logit* function (a.k.a., the *log-odds* function) and its inverse, the *logistic* function (a.k.a., *sigmoid* function)

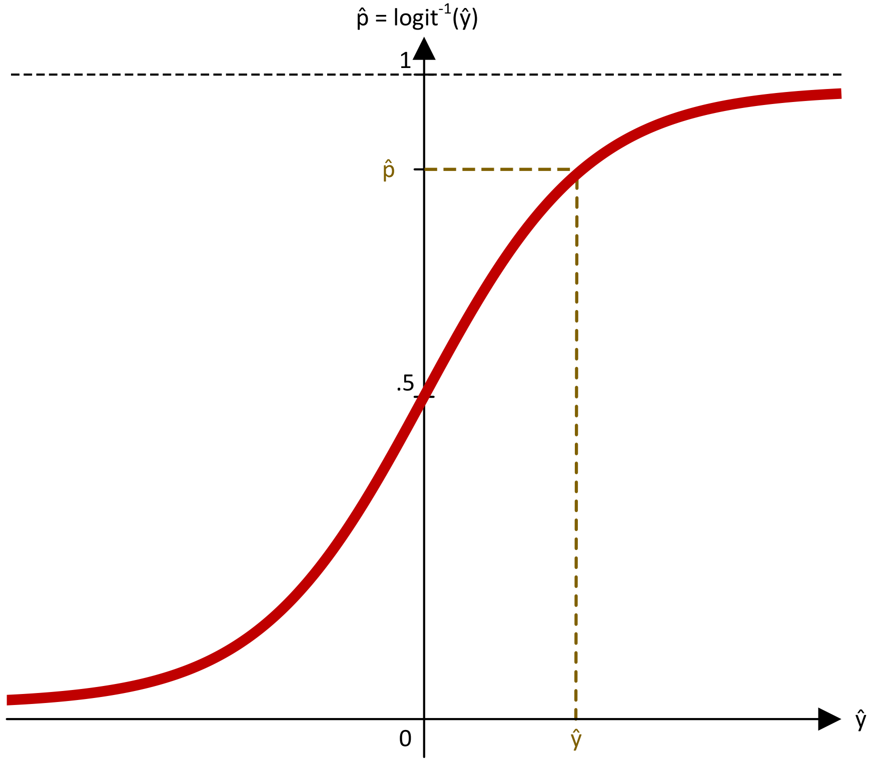
$logit$ maps $\hat{p}$ ($[0;1]$) to $\hat{y}$ ($]-\infty;\ +\infty[$)

$$logit(\hat{p}) = ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{y}$$



$\pi = logit^{-1}$ maps $\hat{y}$ ($]-\infty;\ +\infty[$) to $\hat{p}$ ($[0;1]$)

$$\pi(\hat{y}) = \frac{e^{\hat{y}}}{e^{\hat{y}}+1} = \hat{p}$$



ACTIVITY: KNOWLEDGE CHECK

EXERCISE

ANSWER THE FOLLOWING QUESTIONS

1. What is the difference between odds, odds ratio, and probability?

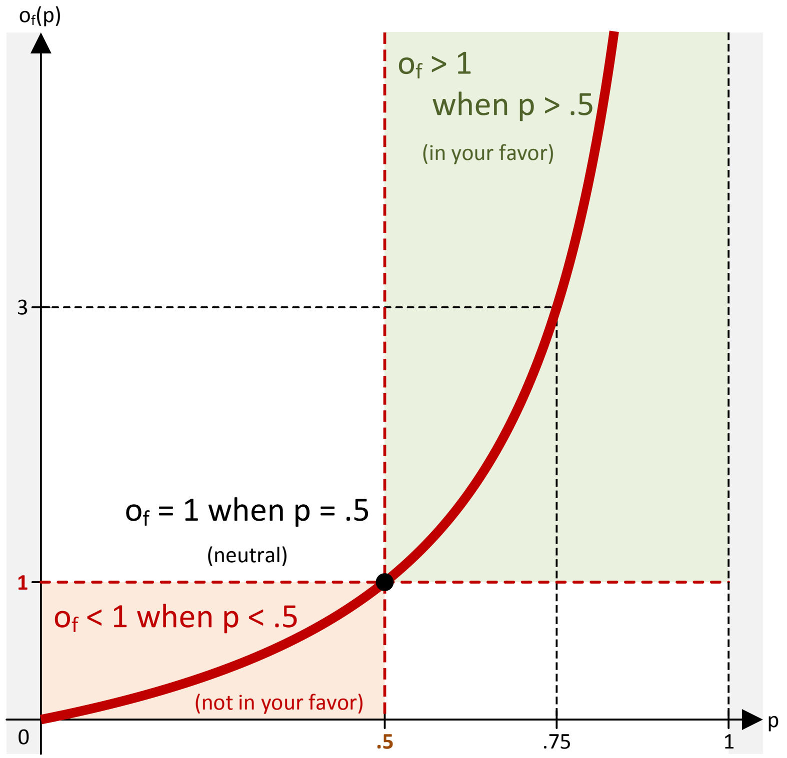
DELIVERABLE

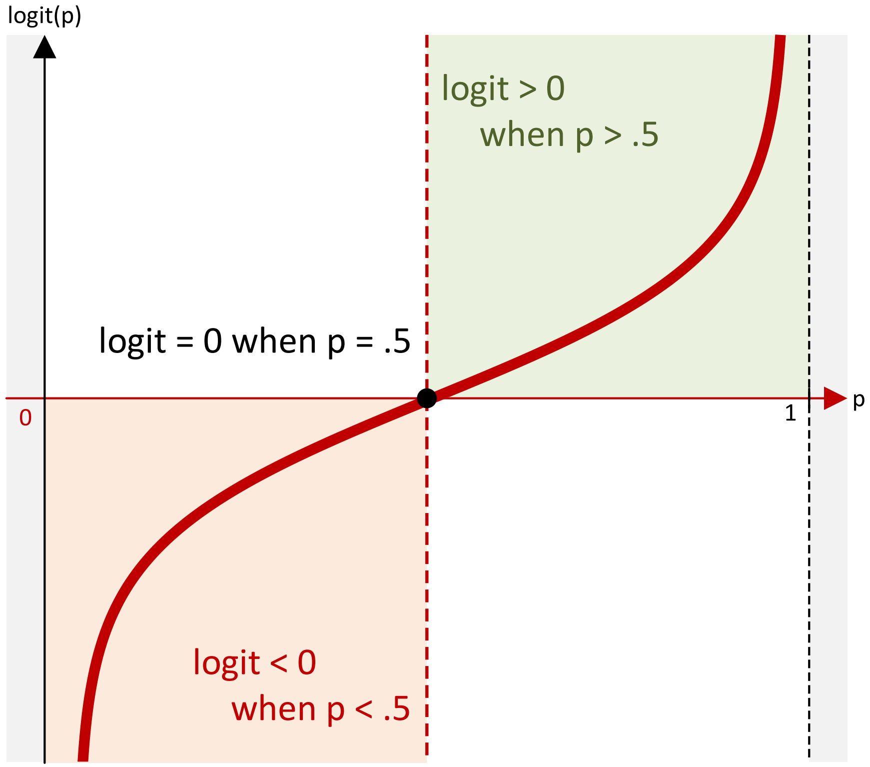
Answers to the above questions

Why is the *logit* function also called the *log-odds* function?

$$\underbrace{o_f}_{odds\ (in\ favor)} = \frac{\overbrace{p}^{\substack{probability\ that\ the\ event\ (with\ probability\ p)\\ happens}}}{\underbrace{1-p}_{\substack{probability\ that\ the\ event\\ does\ not\ happen}}}$$

$$logit(p) = ln(o_f) = ln\left(\frac{p}{1-p}\right)$$

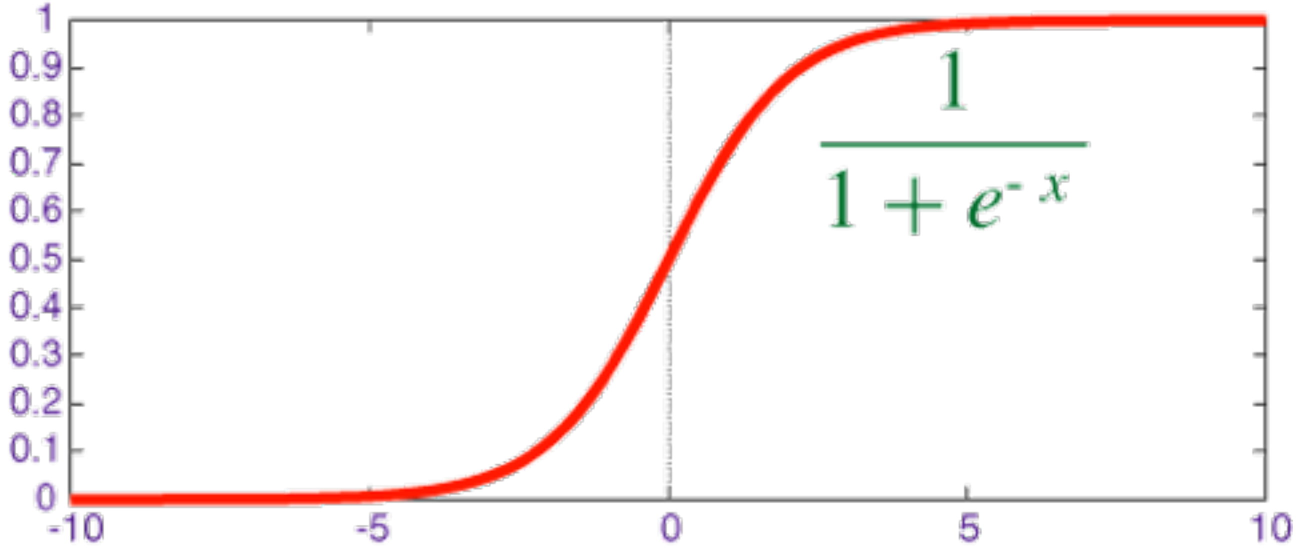




THE SIGMOID FUNCTION

- 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.



INTRODUCTION

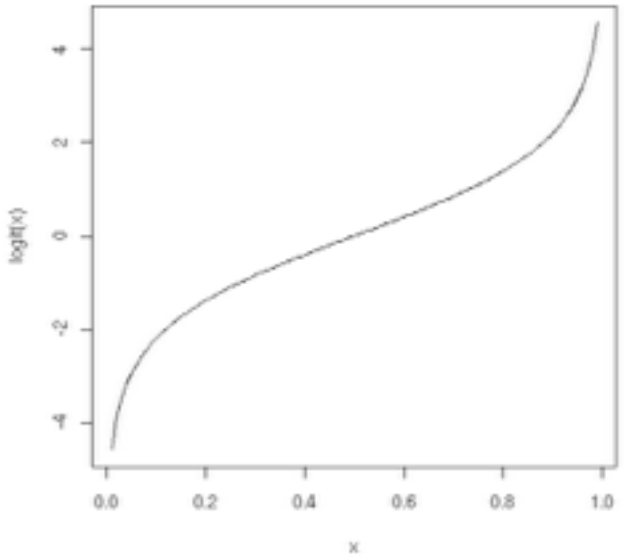
LOGISTIC REGRESSION

FIX 3: ODDS AND LOG-ODDS

- The *logit* function is the inverse of the *sigmoid* function.

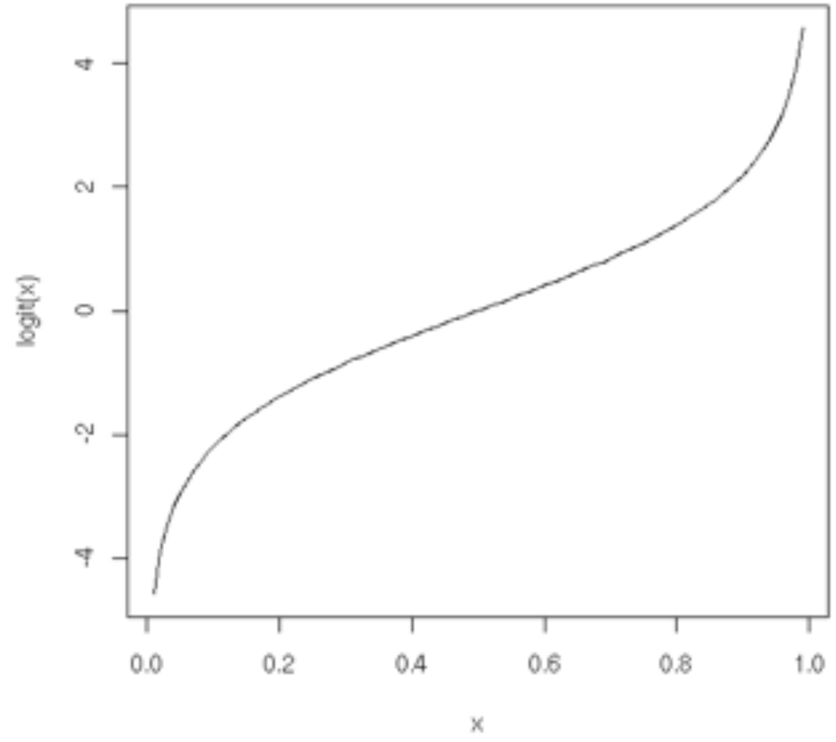
- This will act as our *link* function for logistic regression.

- Mathematically, the logit function is defined as $Ln\left(\frac{P}{1-P}\right)$



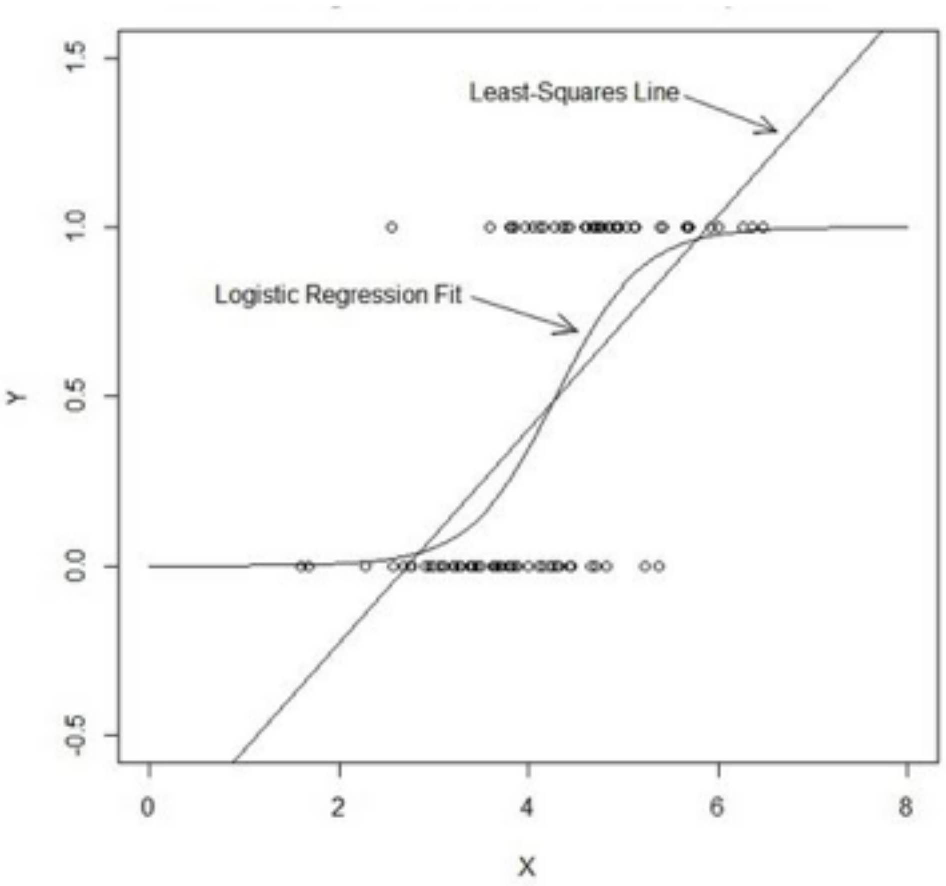
ODDS AND LOG-ODDS

- The value within the natural log, p / (1-p) represents the *odds*. Taking the natural log of odds generates *log odds*.



PUTTING IT ALL TOGETHER

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



ACTIVITY: KNOWLEDGE CHECK

EXERCISE

ANSWER THE FOLLOWING QUESTIONS

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

DELIVERABLE

Answers to the above questions

PUTTING IT ALL TOGETHER

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

$$0.2 = \ln(p / (1-p))$$

- Taking the inverse would give us a probability of ~0.55.

$$1 / (1 + e^{-0.2})$$

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

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

PUTTING IT ALL TOGETHER

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

$$Ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1$$

- The odds multiply by e^{B1} for every 1-unit increase in x.

$$\mathrm{OR} = \frac{\mathrm{odds}(x+1)}{\mathrm{odds}(x)} = \frac{\frac{F(x+1)}{1-F(x+1)}}{\frac{F(x)}{1-F(x)}} = \frac{e^{\beta_0+\beta_1(x+1)}}{e^{\beta_0+\beta_1 x}} = e^{\beta_1}$$

ACTIVITY

WAGER THOSE ODDS!

ACTIVITY: WAGER THOSE ODDS!

EXERCISE

DIRECTIONS (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. Stanford : Iowa, 5:1
    b. Alabama : Michigan State, 20:1
    c. Clemson : Oklahoma, 1.1:1
    d. Houston : Florida State, 1.8:1
    e. Ohio State : Notre Dame, 1.6:1

DELIVERABLE

The desired probabilities

ACTIVITY: WAGER THOSE ODDS!

EXERCISE

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
```

DELIVERABLE

The desired probabilities

CODE ALONG

LOGISTIC REGRESSION IMPLEMENTATION

The Iris dataset, Take 2

Iris Setosa | **Iris Versicolor** | **Iris Virginica**

Source: Flickr

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 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.

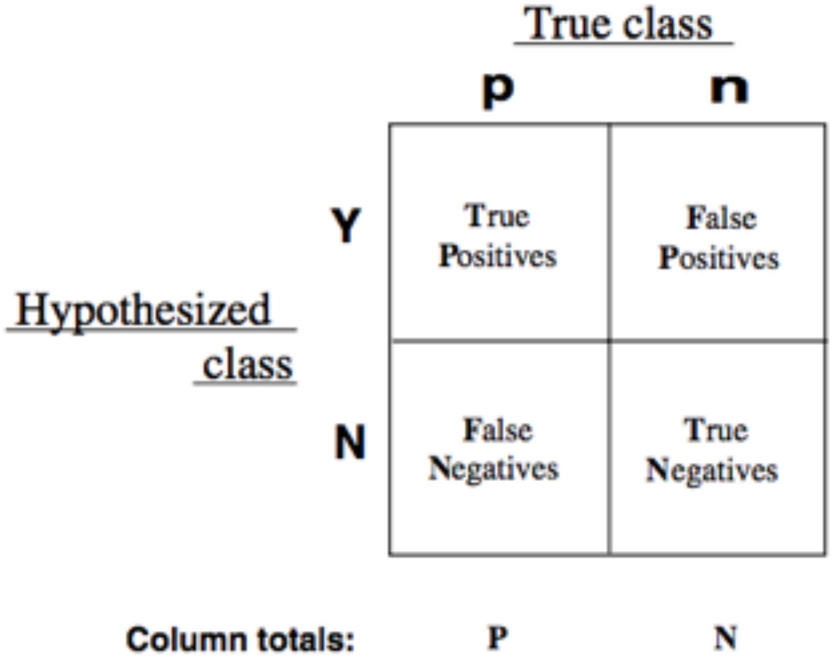
ADVANCED CLASSIFICATION METRICS

- 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

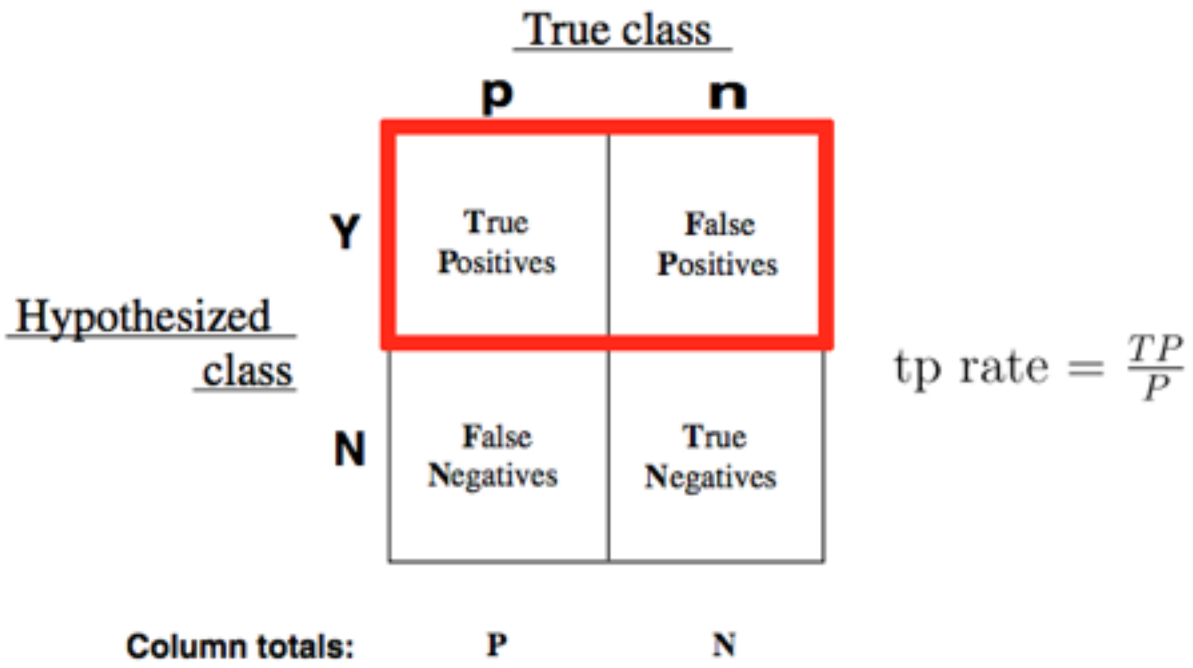
- We can split up the accuracy of each label by using the *true positive rate* and the *false positive rate*.

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



ADVANCED CLASSIFICATION METRICS

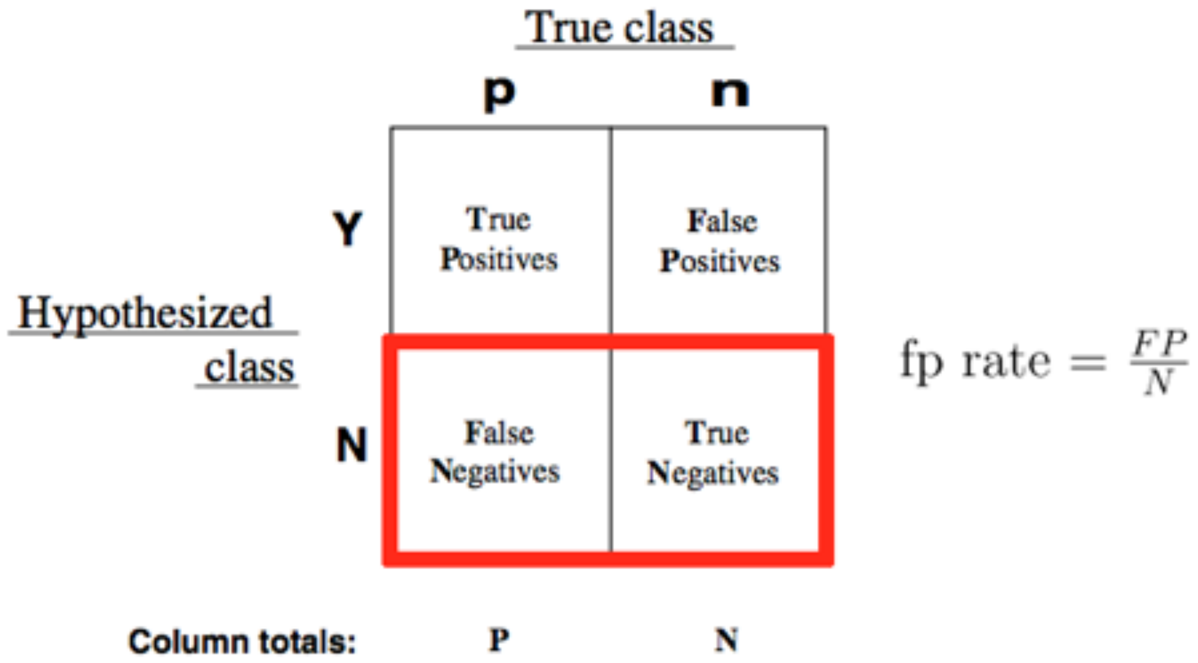
- True Positive Rate (TPR) asks, “Out of all of the target class labels, how many were accurately predicted to belong to that class?”
- For example, given a medical exam that tests for cancer, how often does it correctly identify patients with cancer?



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?"

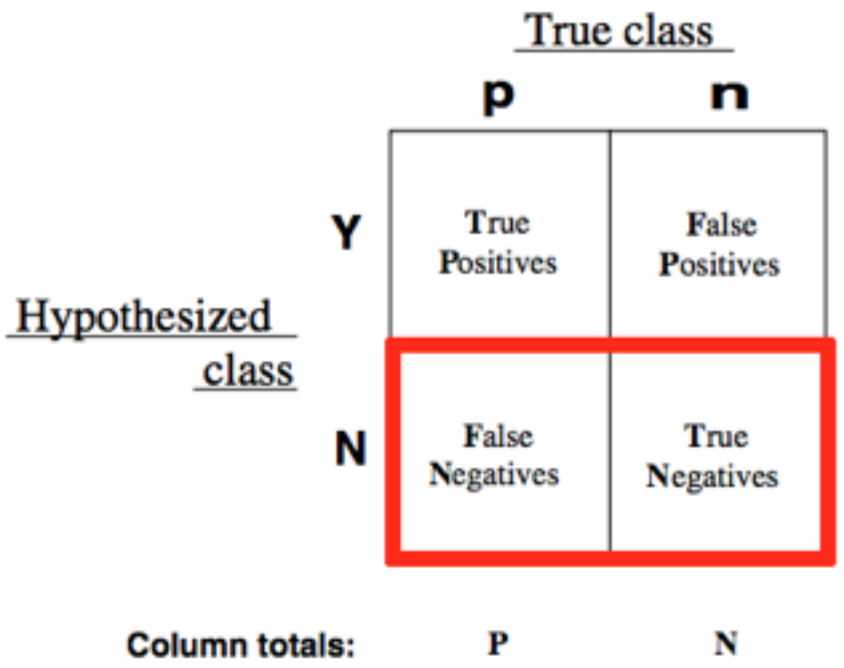
- For example, given a medical exam that tests for cancer, how often does it trigger a "false alarm" by incorrectly saying a patient has cancer?



ADVANCED CLASSIFICATION METRICS

- These can also be inverted.

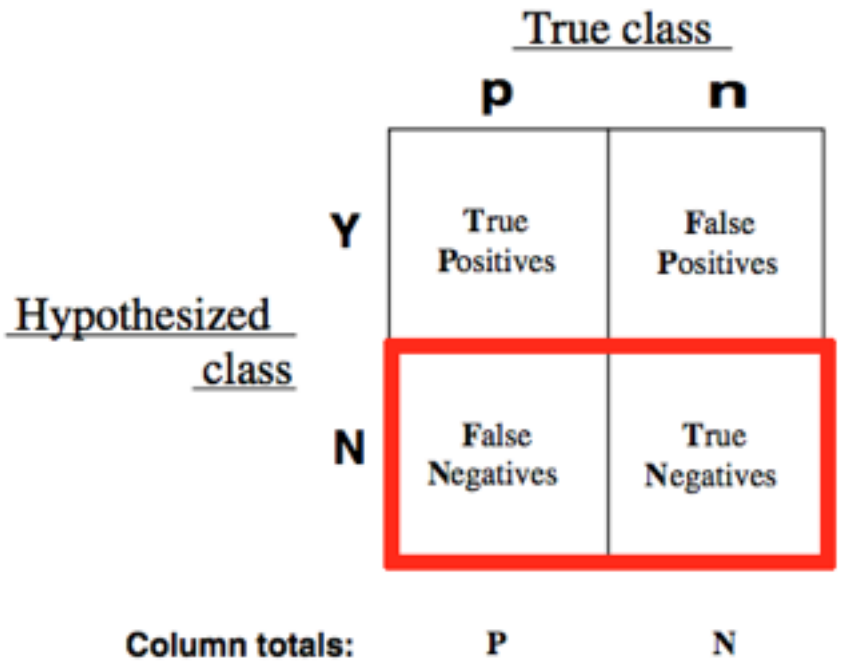
- How often does a test *correctly* identify patients without cancer?



$$\frac{TN}{TN+FN}$$

ADVANCED CLASSIFICATION METRICS

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



$$\frac{FN}{TN+FN}$$

ADVANCED CLASSIFICATION METRICS

- The true positive and false positive rates gives us a much clearer pictures 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 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 optimize two numbers at once.

- Logically, we like a single number for optimization.

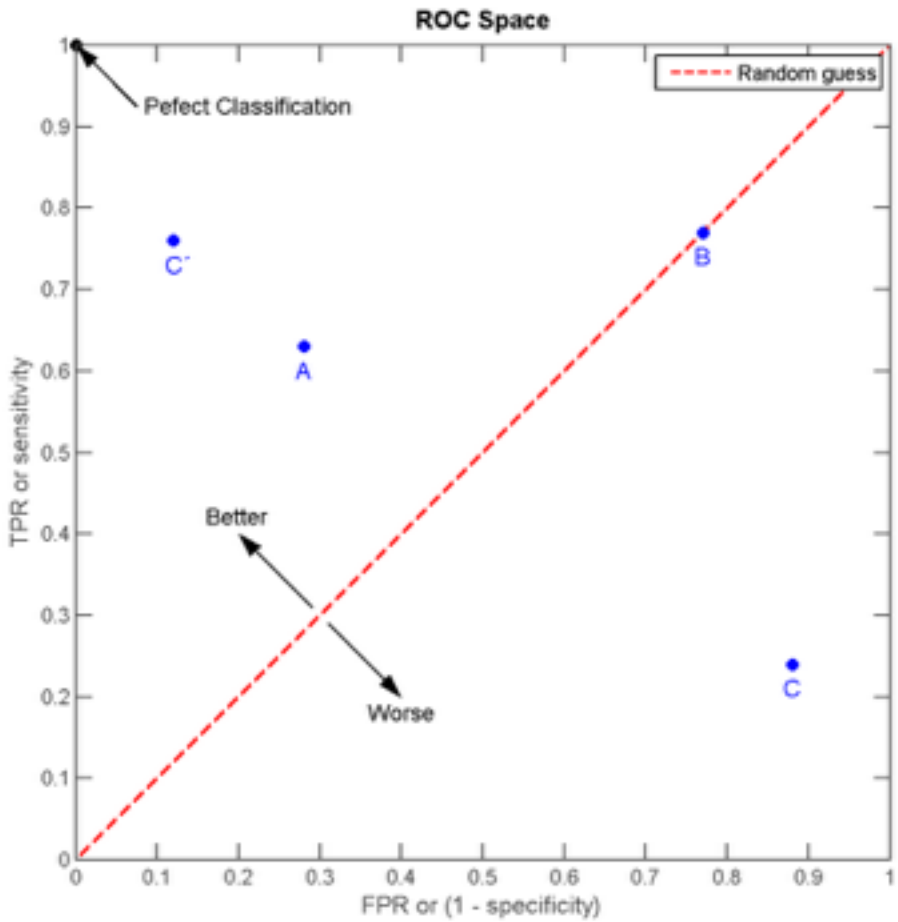
- Can you think of any ways to combine our two metrics?

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) summarizes the impact of TPR and FPR in one single value.

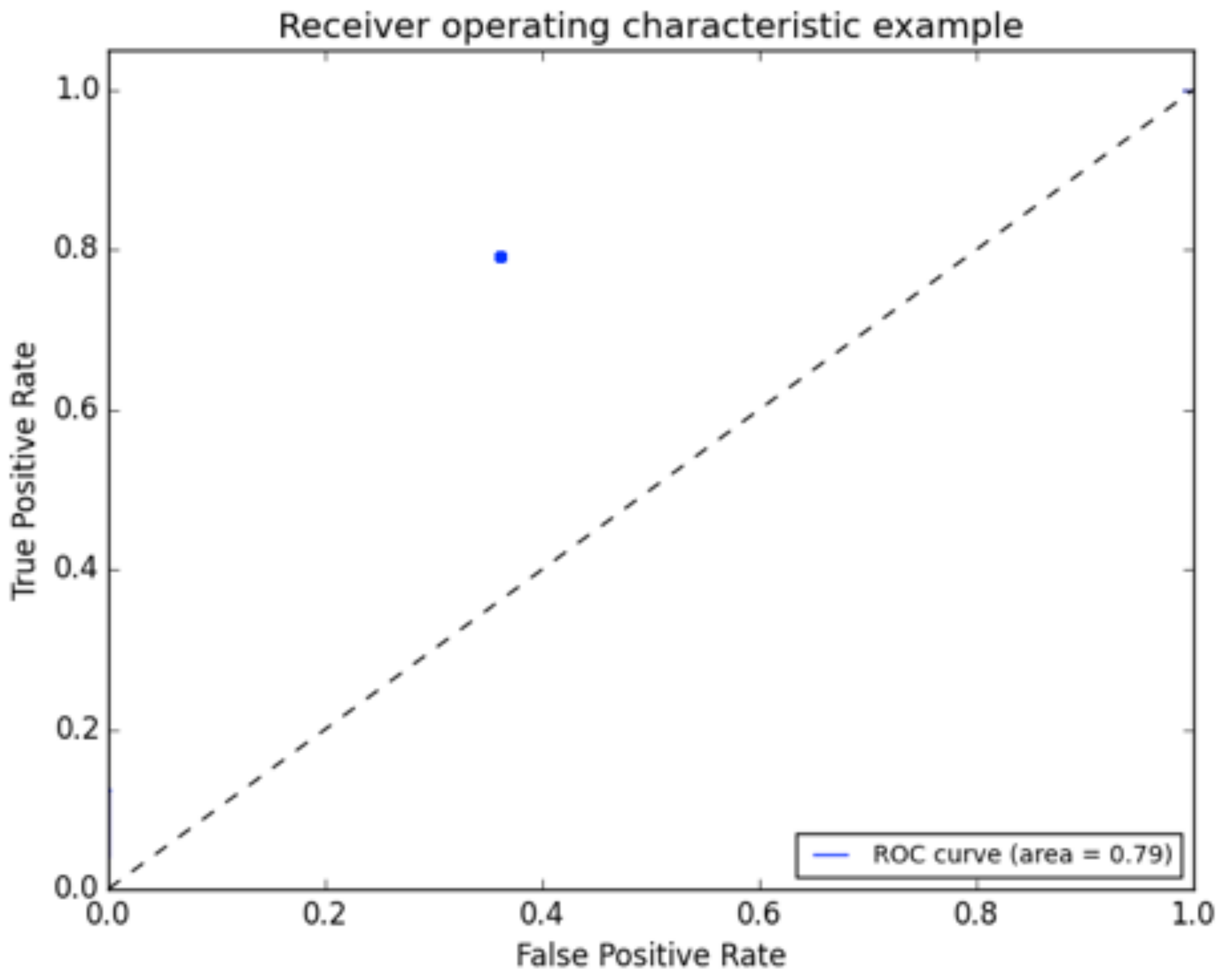
ADVANCED CLASSIFICATION METRICS

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



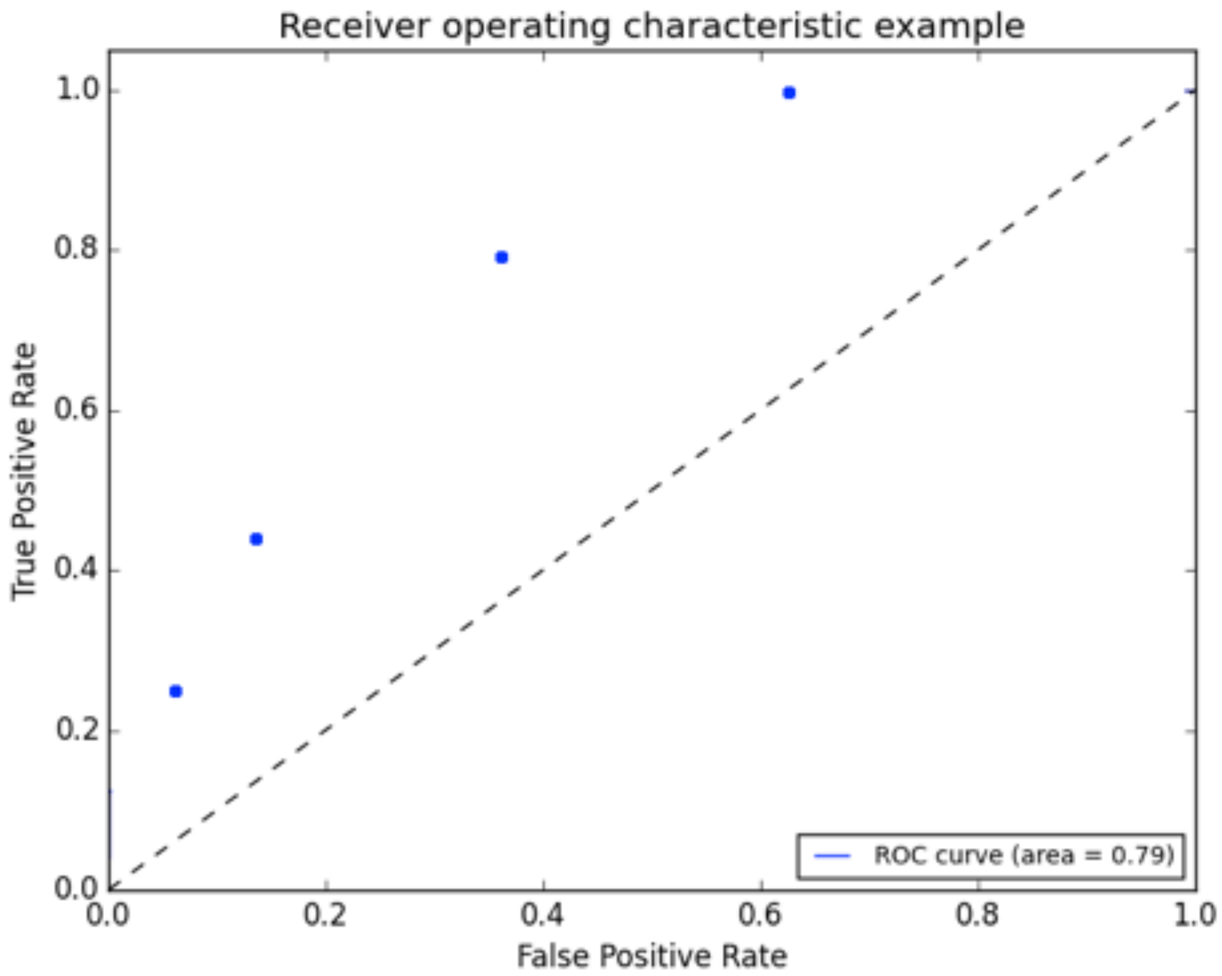
ADVANCED CLASSIFICATION METRICS

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



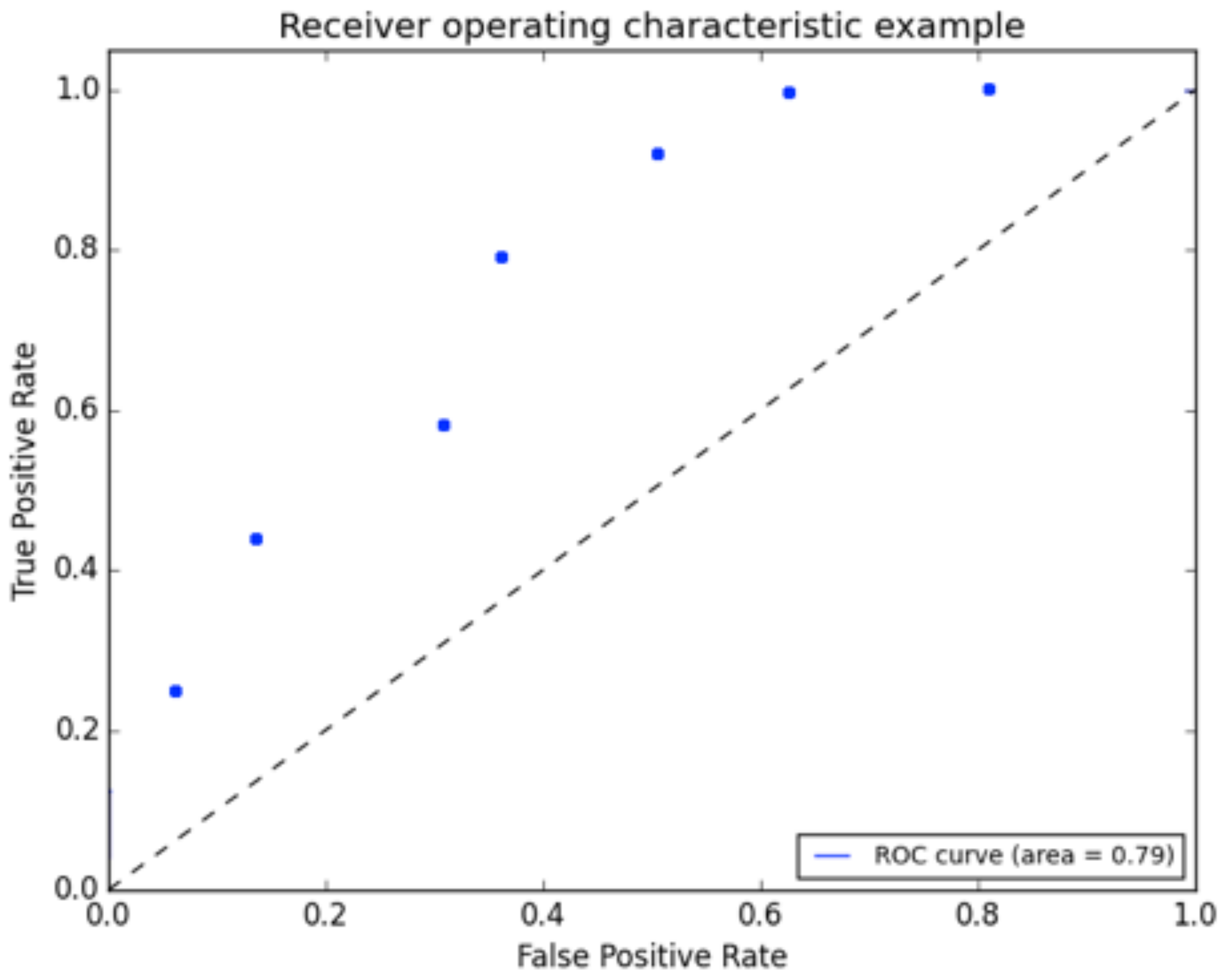
ADVANCED CLASSIFICATION METRICS

- We can continue adding pairs for different thresholds



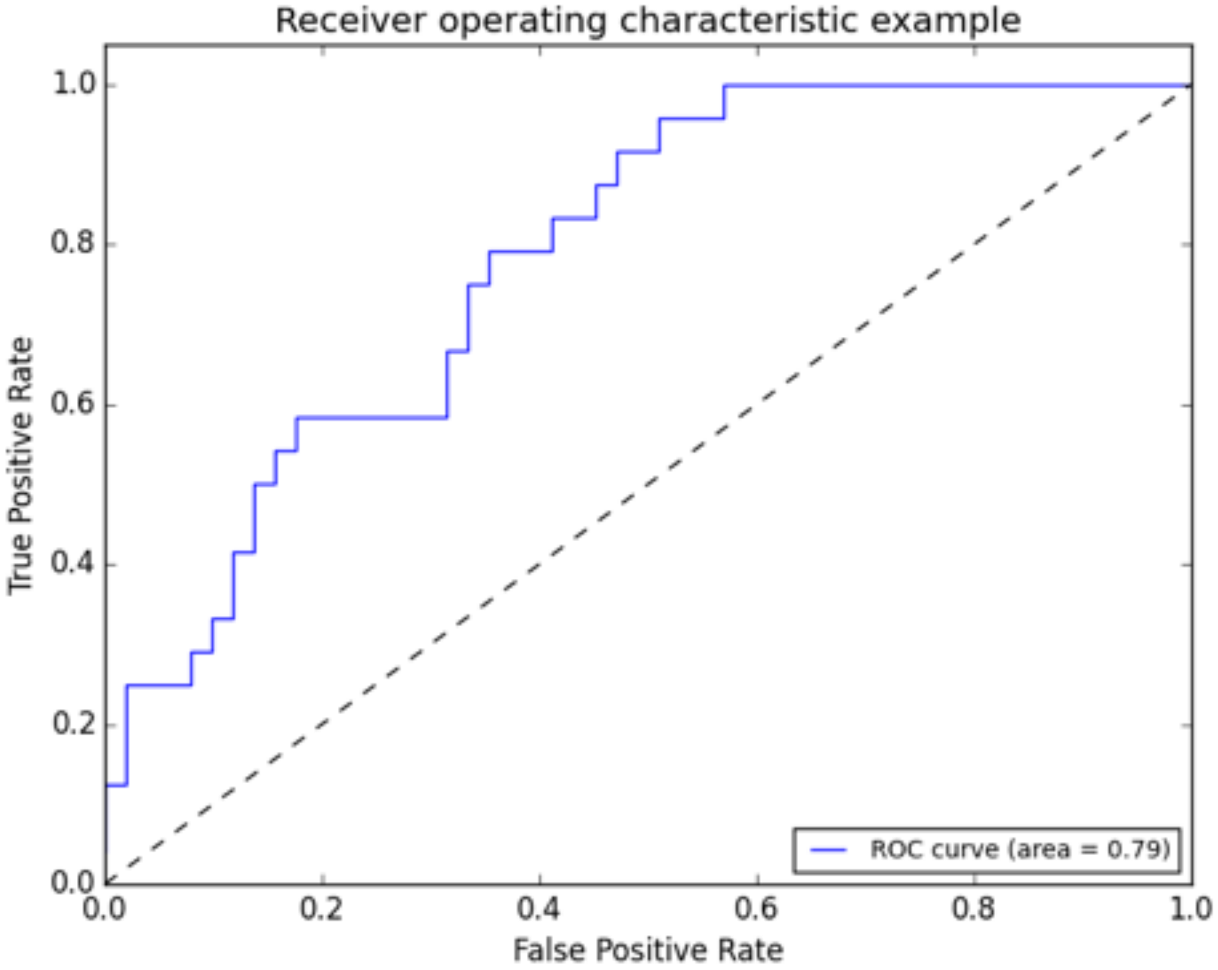
ADVANCED CLASSIFICATION METRICS

- We can continue adding pairs for different thresholds



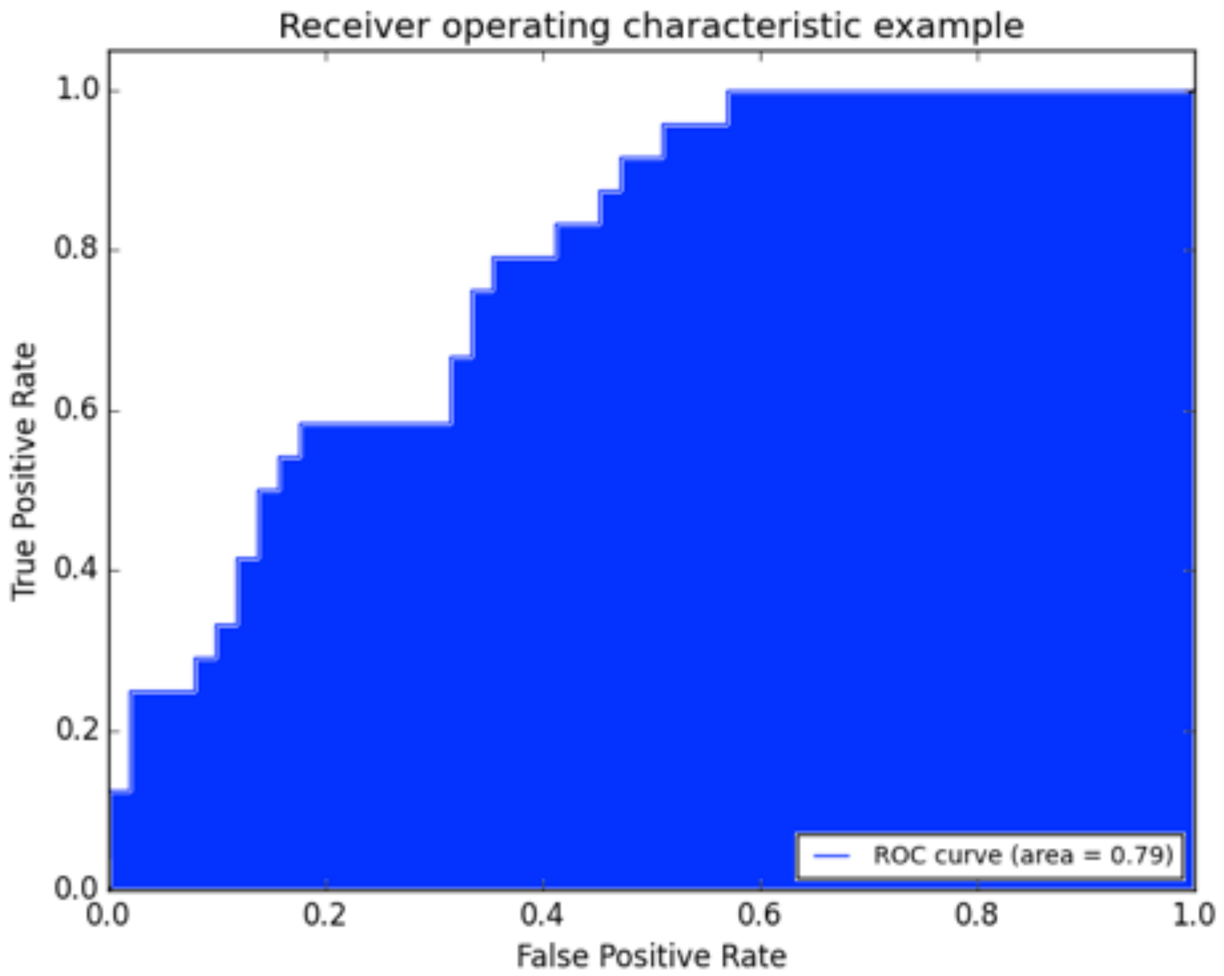
ADVANCED CLASSIFICATION METRICS

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



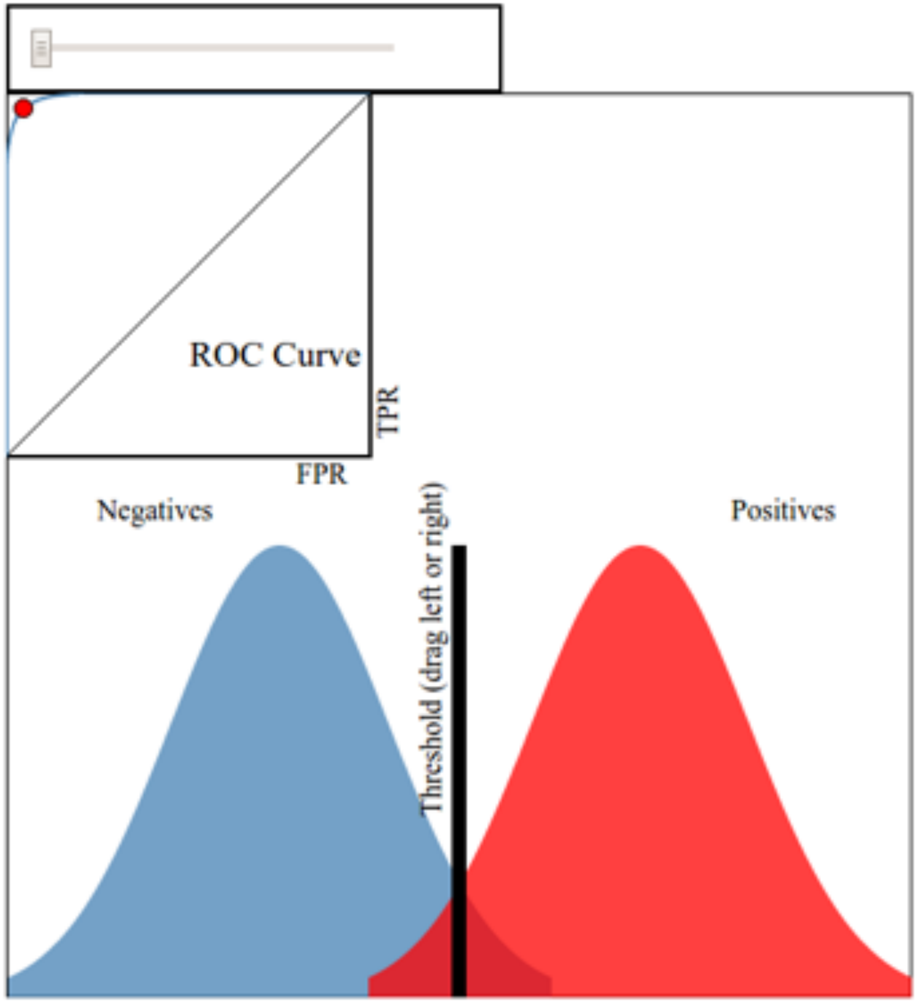
ADVANCED CLASSIFICATION METRICS

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



ADVANCED CLASSIFICATION METRICS

- This interactive visualization can help practice visualizing ROC curves.



ADVANCED CLASSIFICATION METRICS

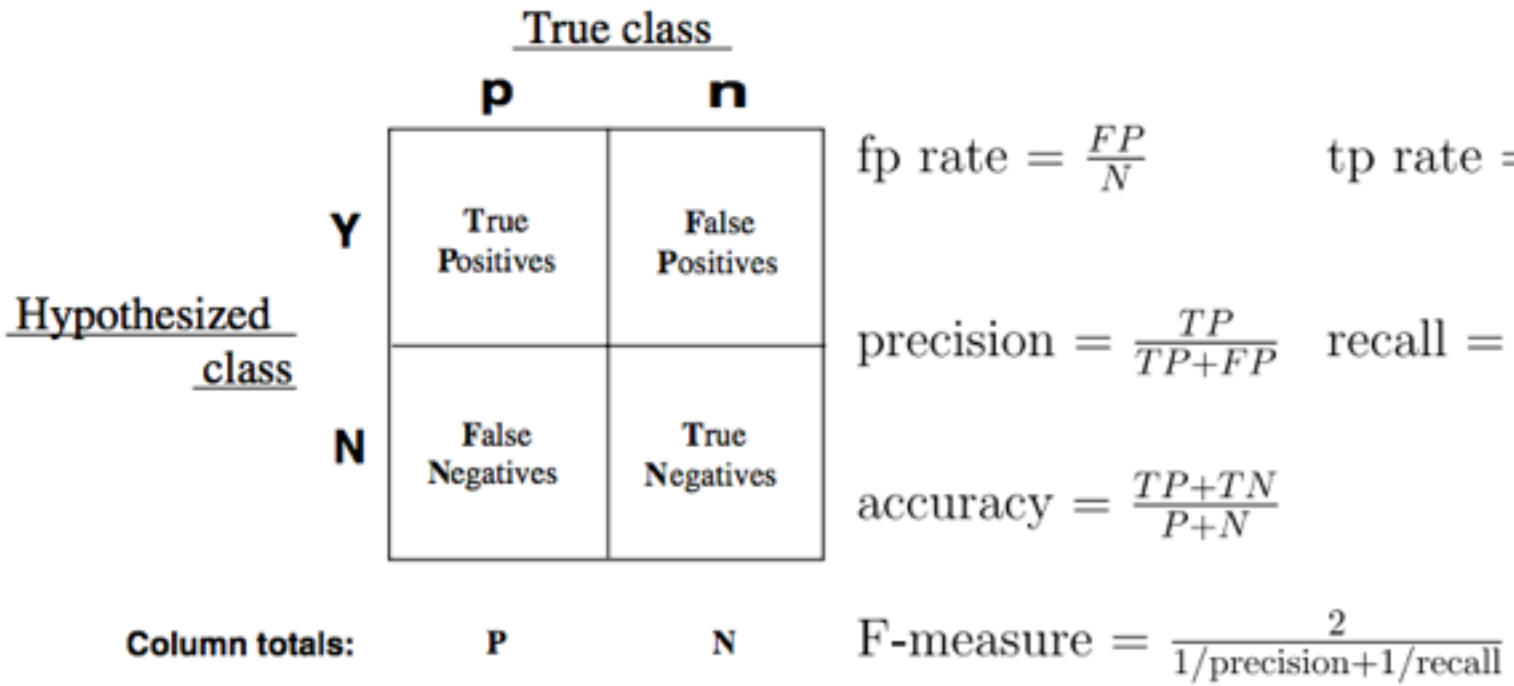
- If we have a TPR of 1 (all positives are marked positive) and FPR of 0 (all negatives are not marked positive), we'd have an AUC of 1. This means everything was accurately predicted.

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

GUIDED PRACTICE

WHICH METRIC SHOULD I USE?

ACTIVITY: WHICH METRIC SHOULD I USE?

EXERCISE

DIRECTIONS (15 minutes)

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 optimize TPR, FPR, and AUC.

DELIVERABLE

Answers for each example

ACTIVITY: WHICH METRIC SHOULD I USE?

EXERCISE

DIRECTIONS (15 minutes)

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.

DELIVERABLE

Answers for each example

INDEPENDENT PRACTICE

EVALUATING LOGISTIC REGRESSION WITH ALTERNATIVE METRICS

ACTIVITY: EVALUATING LOGISTIC REGRESSION

EXERCISE

DIRECTIONS (35 minutes)

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 sklearn. For a worst case scenario, identify one or two strong features that would be useful to include in this model.

DELIVERABLE

Answers to the above question and a Logistic model on the Titanic data

ACTIVITY: EVALUATING LOGISTIC REGRESSION

EXERCISE

DIRECTIONS (35 minutes)

1. Spend 1-2 minutes considering which *metric* makes the most sense to optimize. Accuracy? FPR or TPR? AUC? Given the business problem of understanding survival rate aboard the Titanic, why should you use this metric?

1. Build a tuned Logistic model. Be prepared to explain your design (including regularization), metric, and feature set in predicting survival using any tools necessary (such as a fit chart). Use the starter code to get you going.

DELIVERABLE

Answers to the above question and a Logistic model on the Titanic data

CONCLUSION

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.

COURSE

BEFORE NEXT CLASS

BEFORE NEXT CLASS

DUE DATE

- Project:

LESSON

CREDITS

CITATIONS

- Title, Author: link
- Title, Author: link
- Title, Author: link

LESSON

Q & A

LESSON

EXIT TICKET

DON'T FORGET TO FILL OUT YOUR EXIT TICKET

THANKS!

NAME

- Optional Information:
- Email?
- Website?
- Twitter?