GENERAL ASSEMBLY

INTRODUCTION TO LOGISTIC REGRESSION

Jonathan Balaban

DAT2

LEARNING OBJECTIVES

- Build a Logistic regression classification model using the sklearn 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, confusion matrix, ROC/AUC curves, and loss functions

PRE-WORK REVIEW

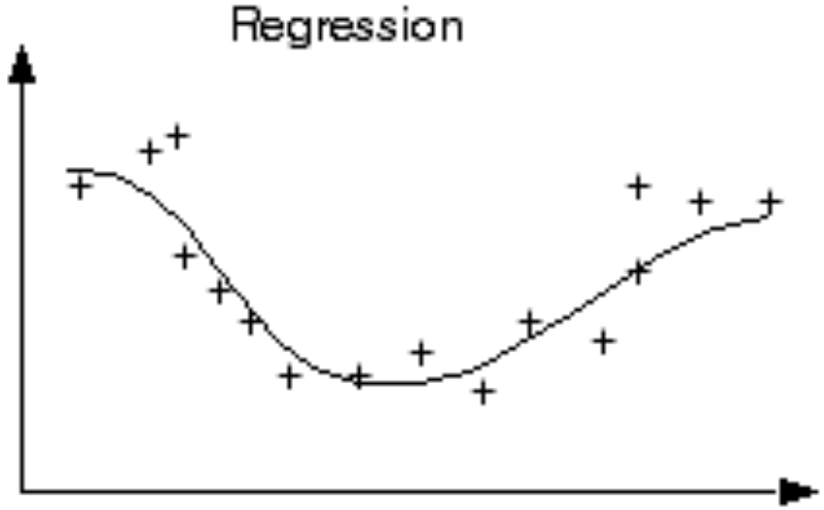
- Implement a linear model (LinearRegression) with sklearn
- Understand what a coefficient is
- Recall metrics such as accuracy and misclassification
- Recall the differences between L1 and L2 regularization

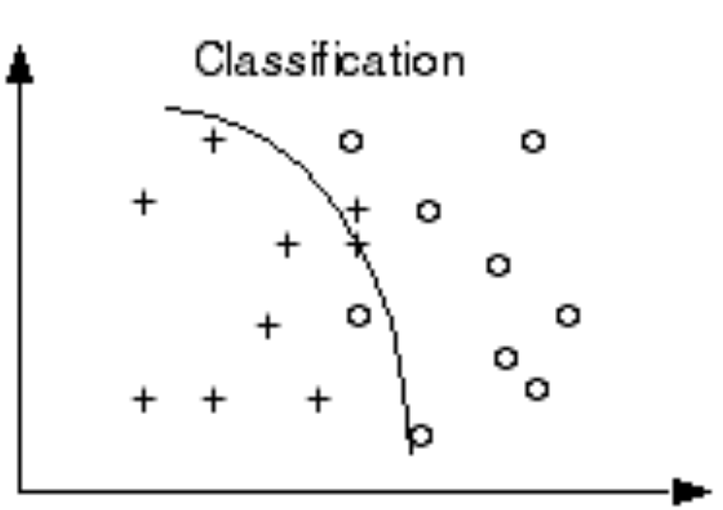
OPENING

INTRODUCTION TO LOGISTIC REGRESSION

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.





LINEAR REGRESSION RESULTS FOR CLASSIFICATION

- But, since many 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?
  - Output isn't normal
  - Predictions can be outside of [0,1], which violates laws of probability
  - Probability can also be U-shaped: Flu by Age
- How might we contain those bounds?
- Let's review some "fixes" to make classification with regression feasible.

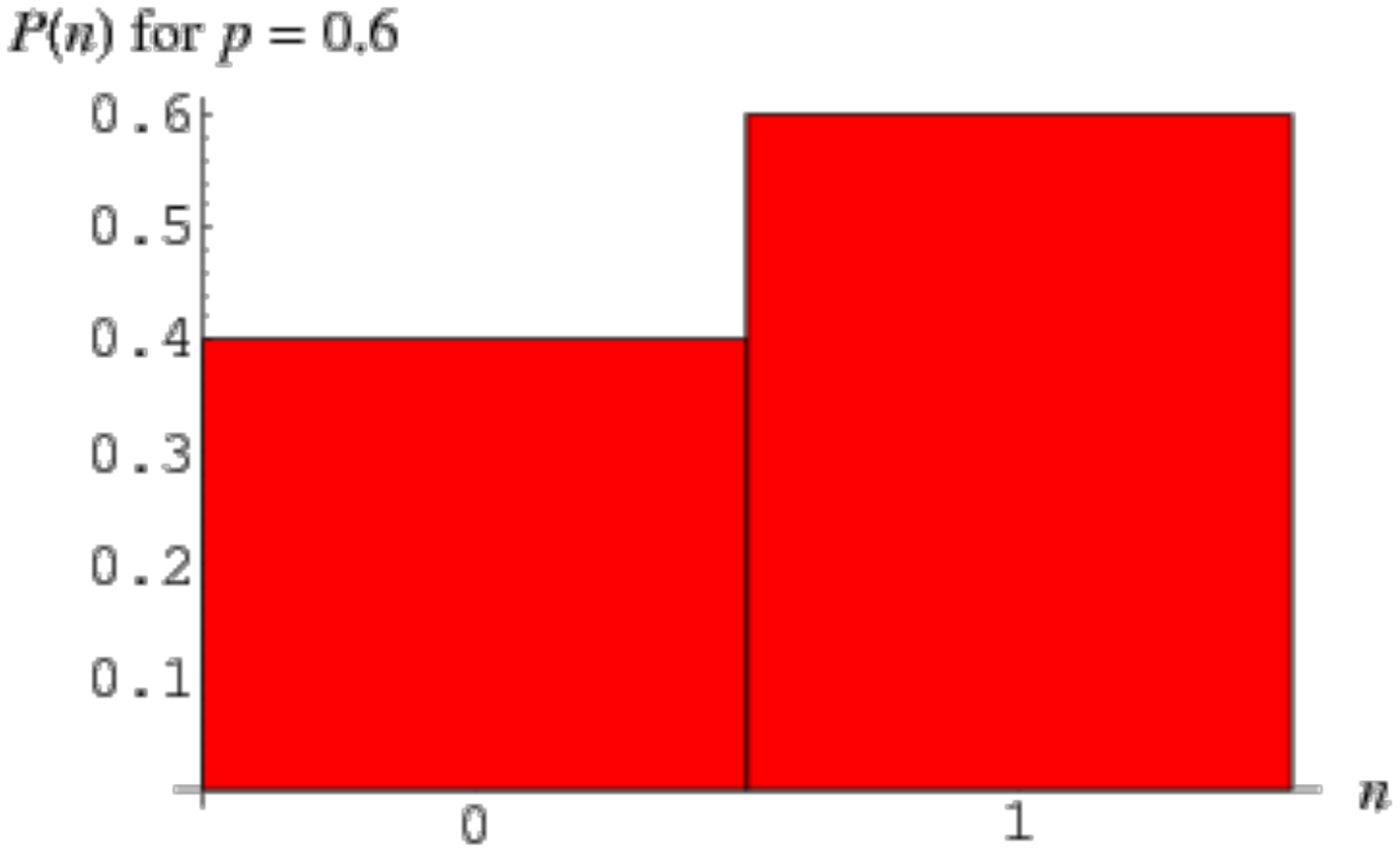
FIX 1: PROBABILITY

- 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

- Bernoulli Distribution



FIX 1: PROBABILITY

- 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 the data about it.

- Logistic regression estimates an unknown p for any given linear combination of predictors.

FIX 2: LINK FUNCTIONS

- Another advantage to OLS is that it allows for *generalized* 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

- 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* (/ˈloʊdʒɪt/ LOH-jit) function
  - Inverse of the *sigmoid* function.

FIX 2: LINK FUNCTIONS AND THE SIGMOID FUNCTION

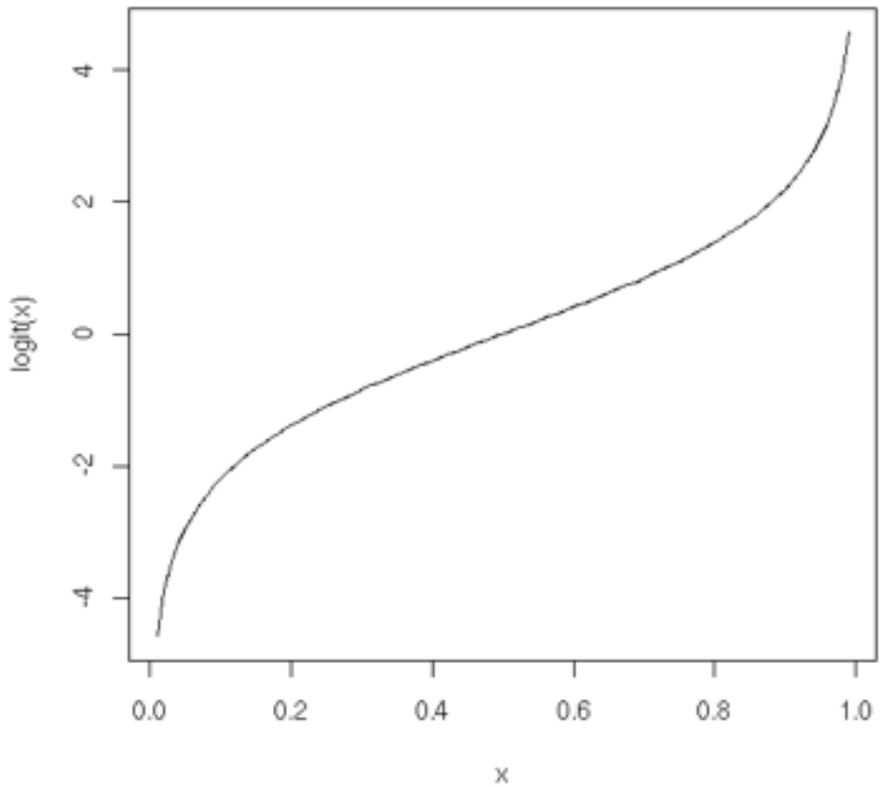
- Since x decides how to much to increase or decrease the value away from 0.5, x can be interpreted as something like a coefficient.

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

FIX 3: ODDS AND LOG-ODDS

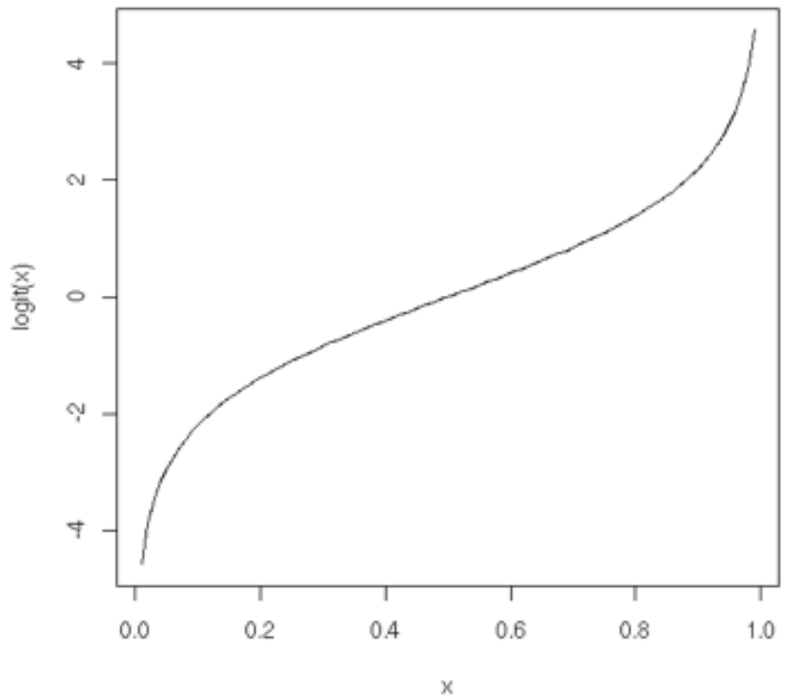
- This will act as our *link* function for logistic regression.
- Mathematically, the logit function is defined as

$$Ln\left(\frac{P}{1-P}\right)$$



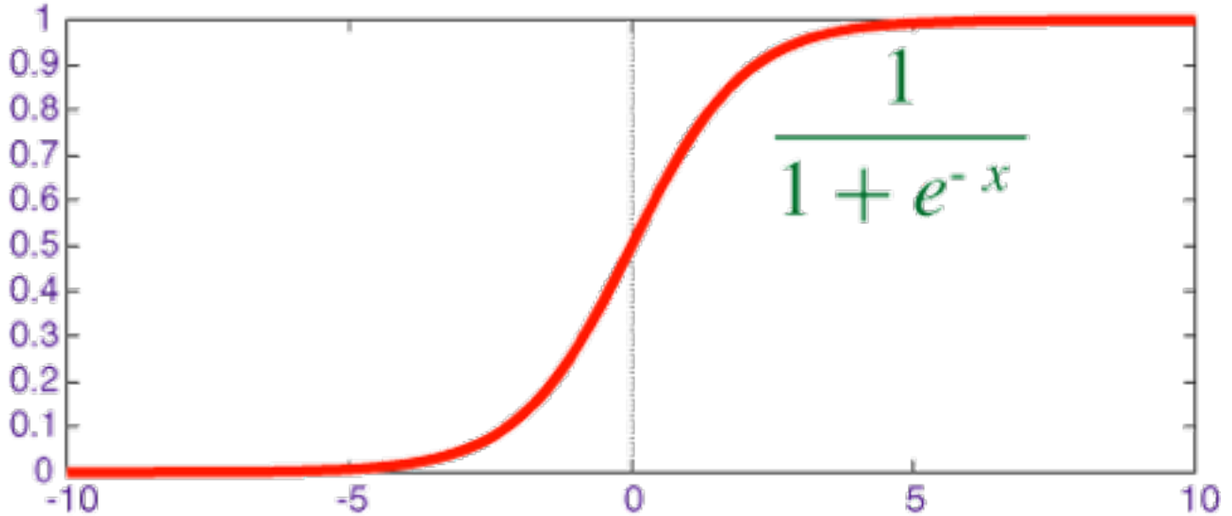
FIX 3: ODDS AND LOG-ODDS

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



THE SIGMOID FUNCTION

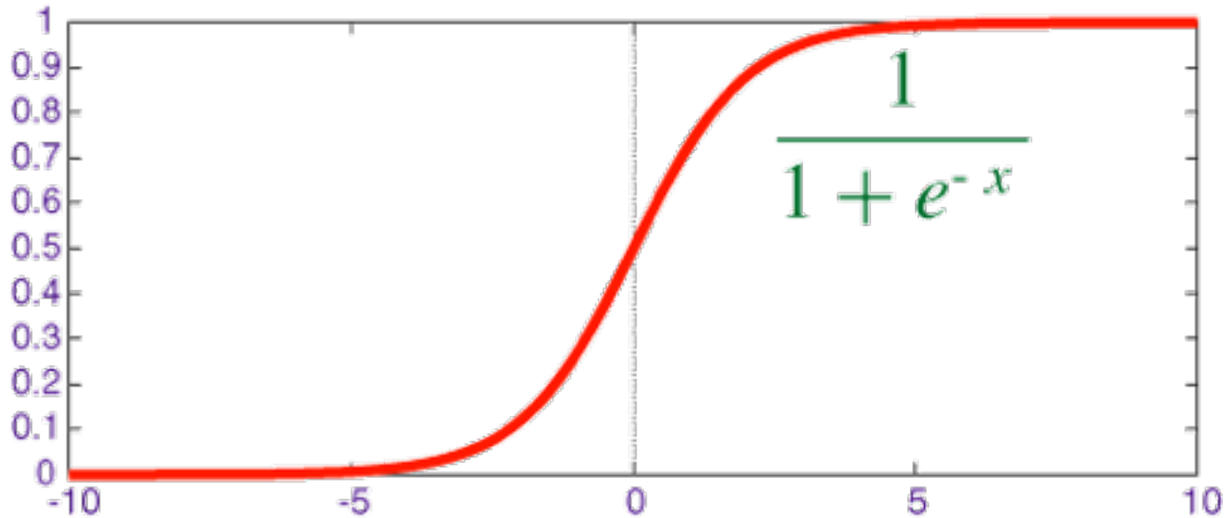
- A *sigmoid function* is a function that visually looks like an s.



- Mathematically, it is defined as $f(x) = \frac{1}{1+e^{-x}}$

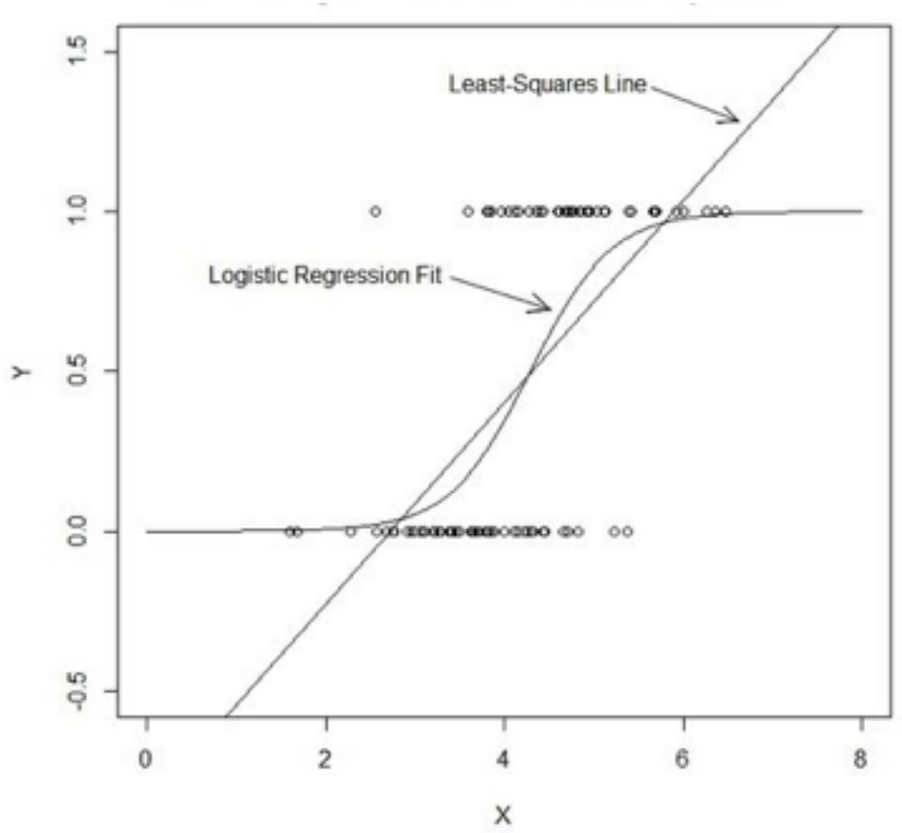
THE SIGMOID FUNCTION

- Recall that 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.



FIX 3: ODDS AND LOG-ODDS

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



FIX 3: ODDS AND LOG-ODDS

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

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

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

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

- 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.

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

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

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

GUIDED PRACTICE

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

ACTIVITY: LOGISTIC REGRESSION IMPLEMENTATION

DIRECTIONS (15 minutes)

Use the data `collegeadmissions.csv` and the `LogisticRegression` estimator in sklearn to predict the target variable `admit`.

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 admission rate? 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 = 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.

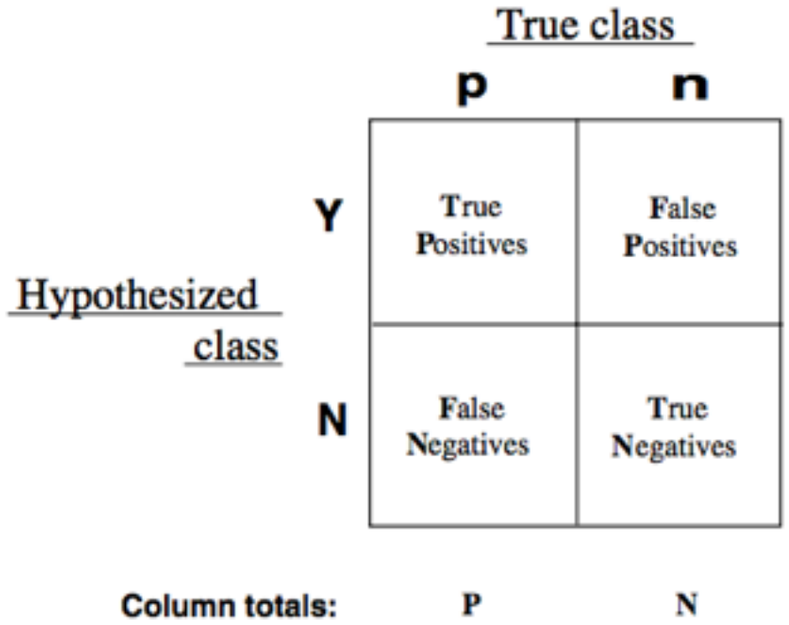
ADVANCED CLASSIFICATION METRICS

- 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*.

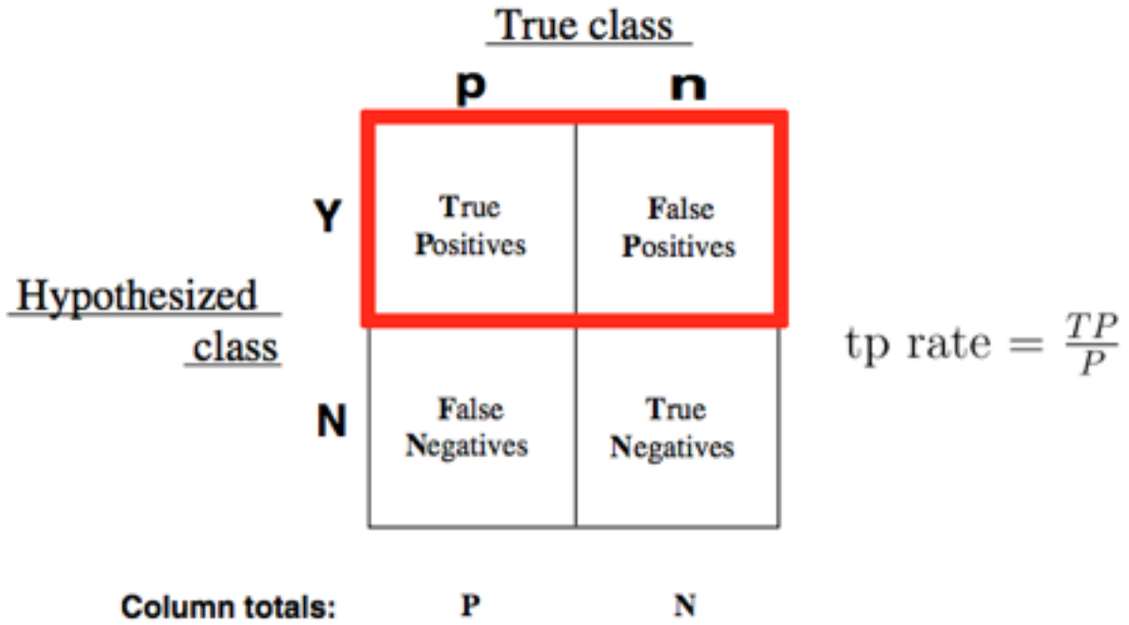
- For each label, we can put it into the category of a true positive, false positive, true negative, or false negative.



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?

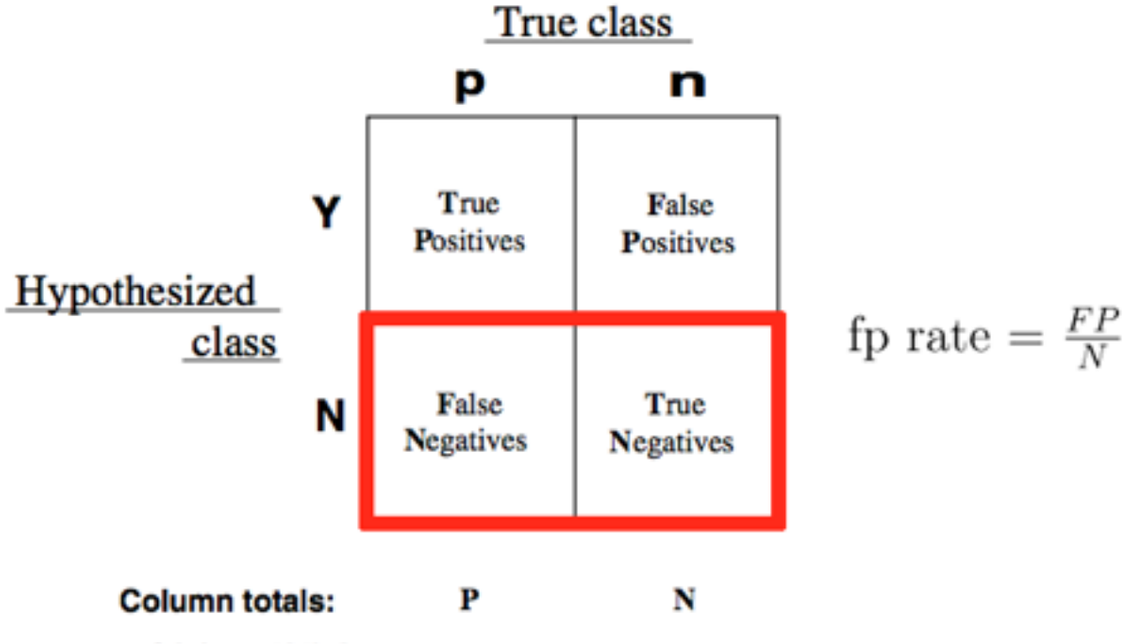


$$\text{tp rate} = \frac{TP}{P}$$

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?”

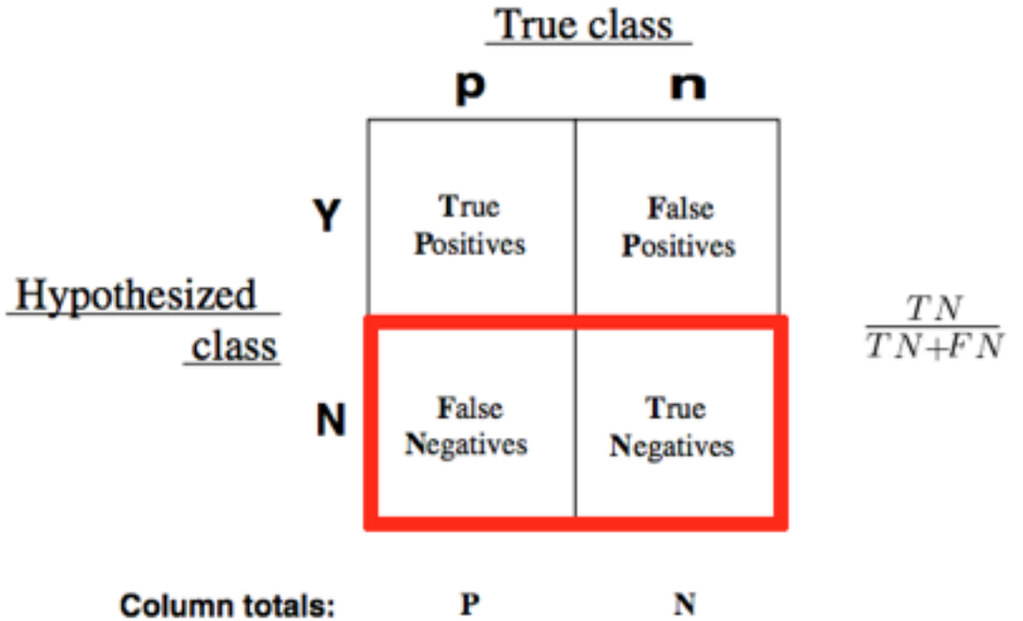
- For example, given a medical exam that tests for cancer, how often does it trigger a “false alarm” by incorrectly saying a patient has cancer?



$$\text{fp rate} = \frac{FP}{N}$$

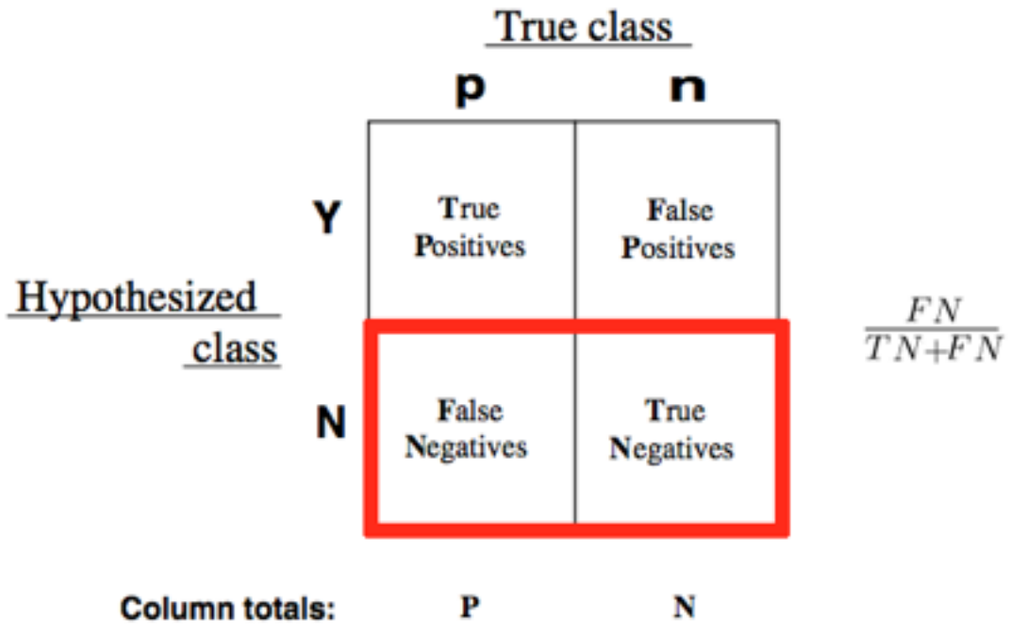
ADVANCED CLASSIFICATION METRICS

- 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 patient as cancer-free?



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

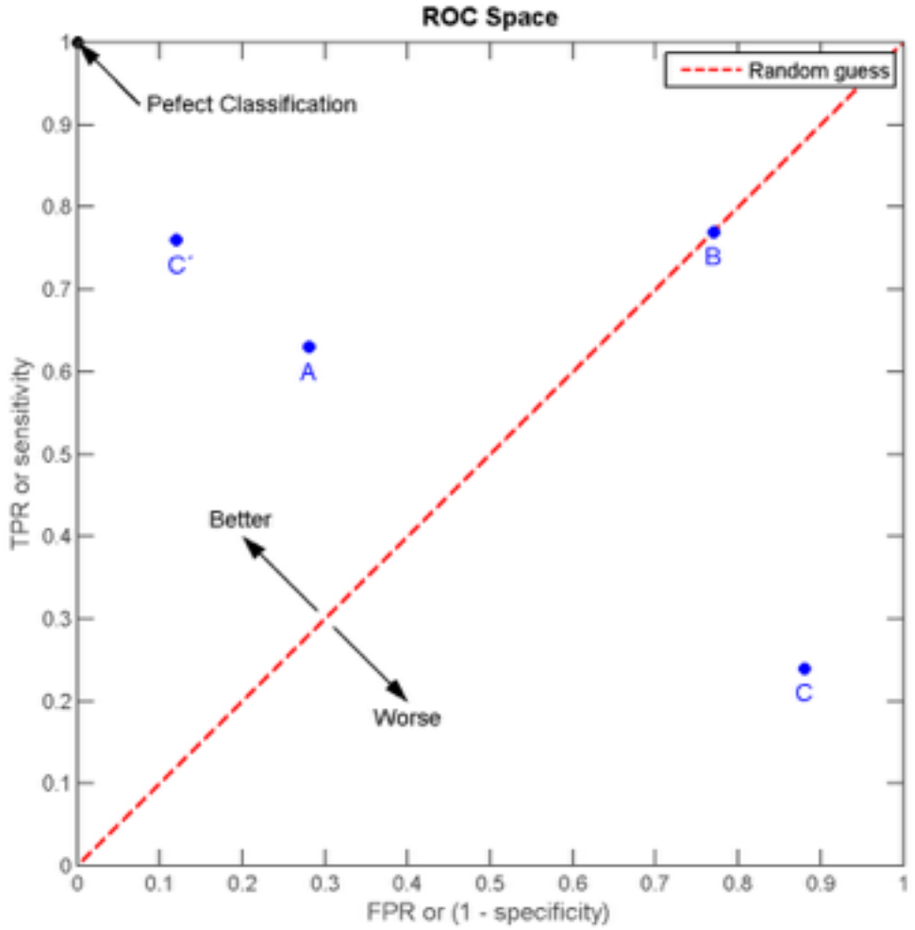
- We can vary the classification threshold for our model to get different predictions. But how do we know if a model is better overall than other model?
- We can compare the FPR and TPR of the models, but it can often be difficult to optimize two numbers at once.
- Logically, we would like a single number 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.

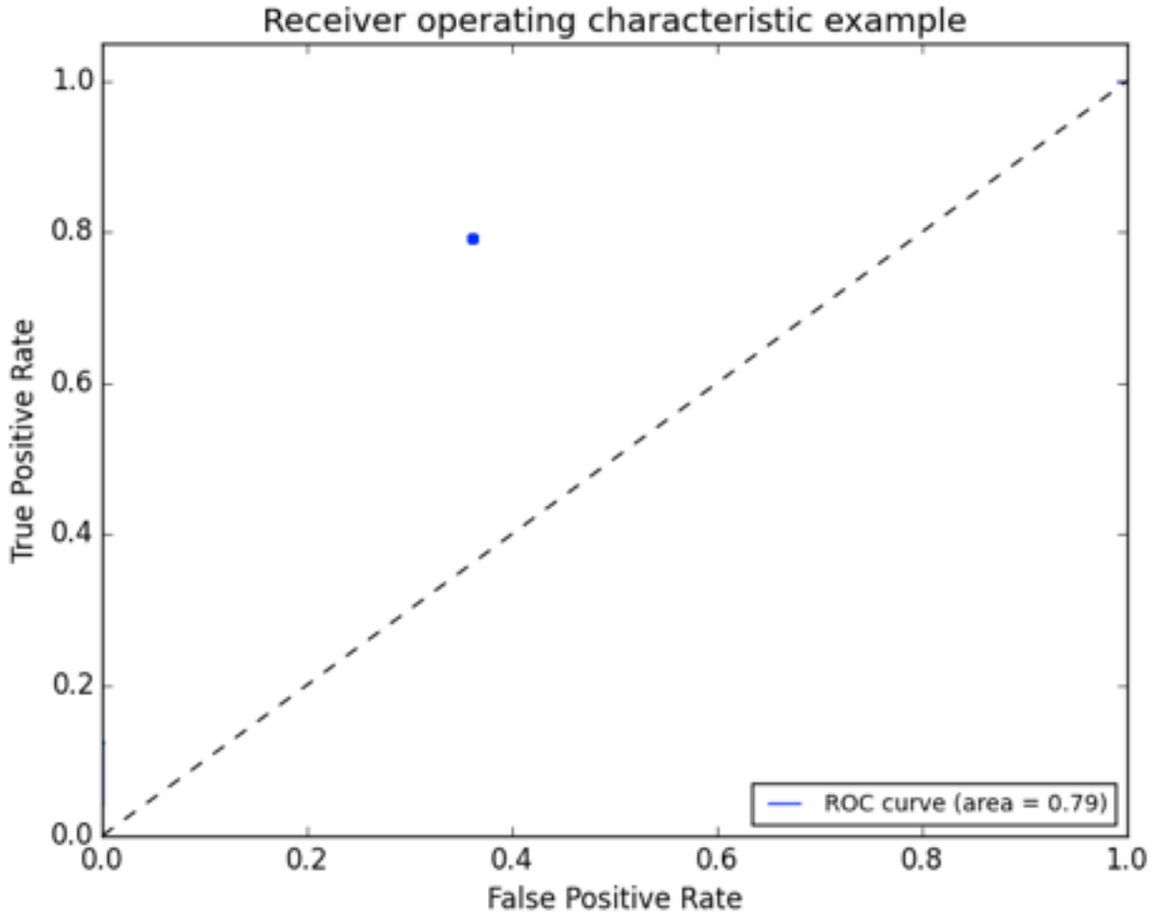
ADVANCED CLASSIFICATION METRICS

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



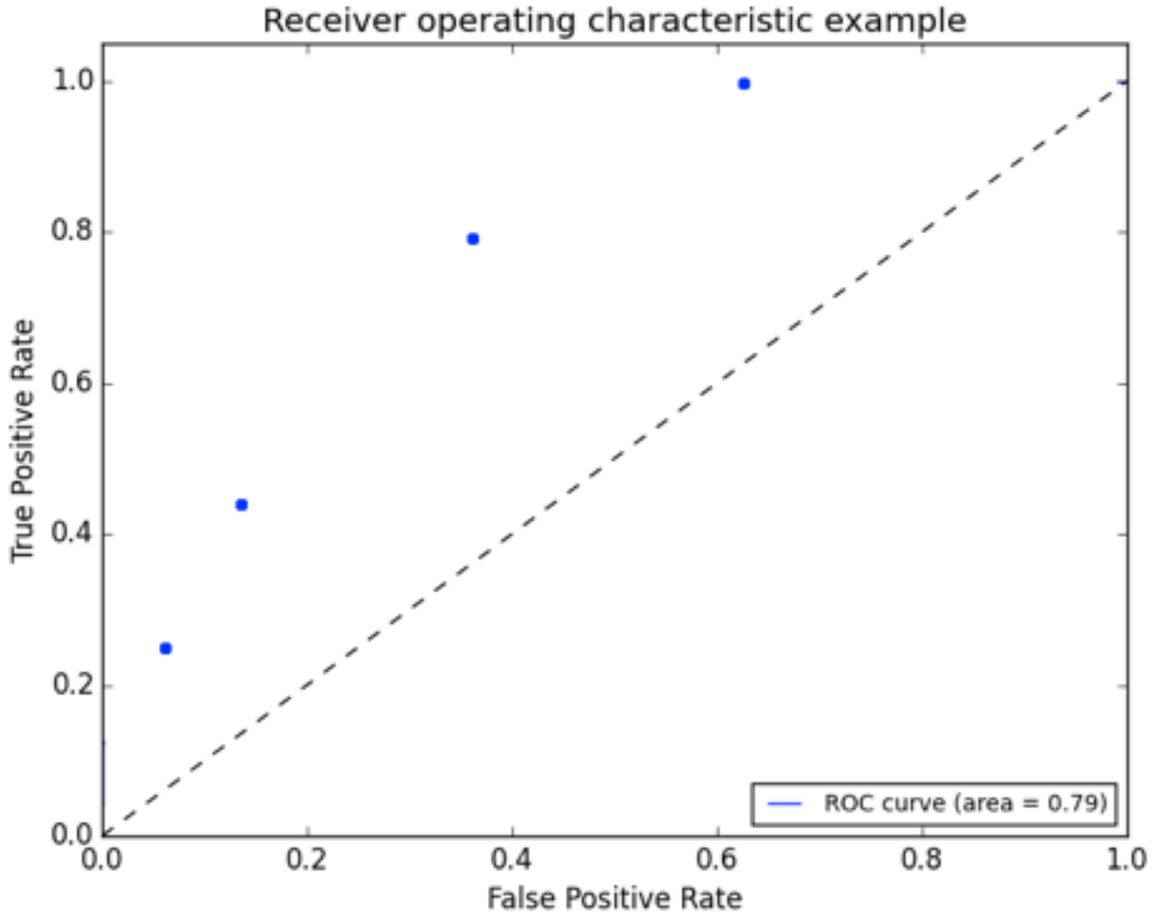
ADVANCED CLASSIFICATION METRICS

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



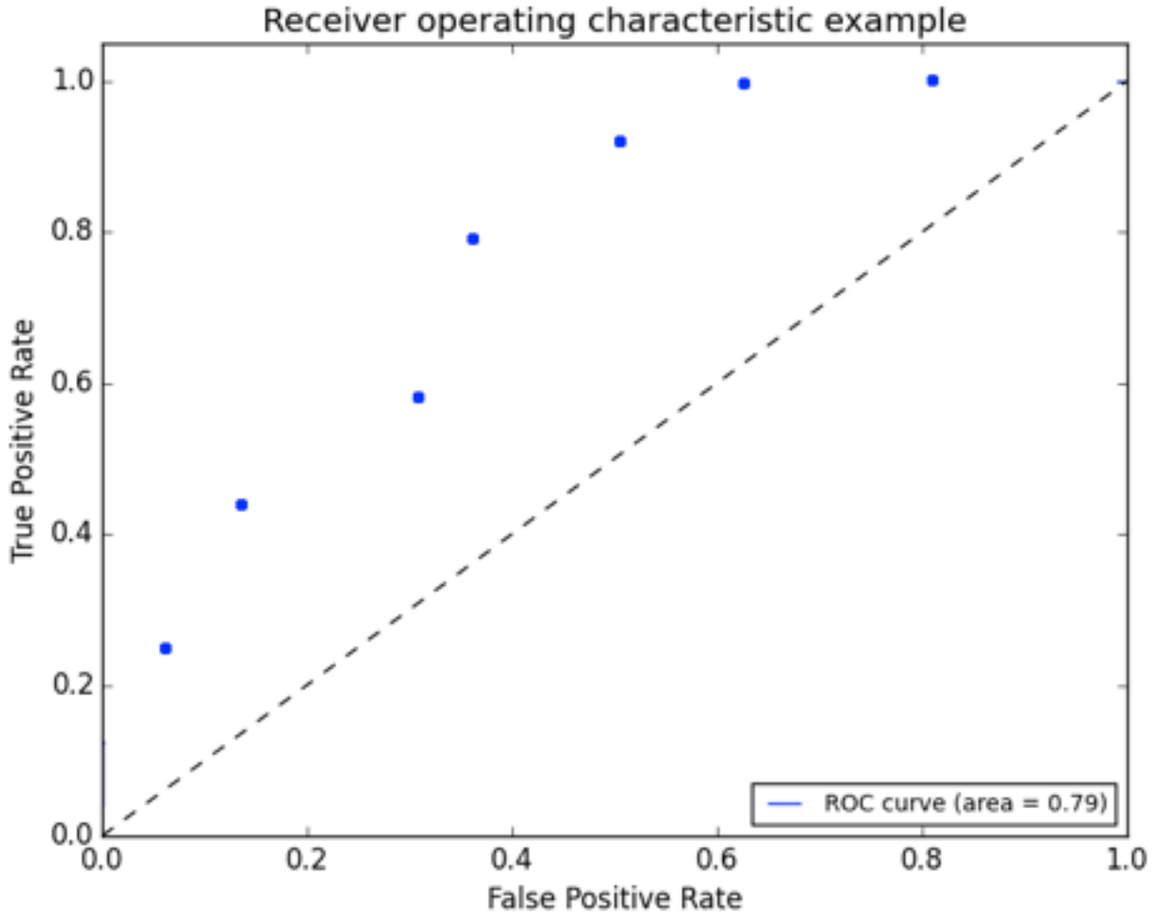
ADVANCED CLASSIFICATION METRICS

- We can continue adding pairs for different thresholds



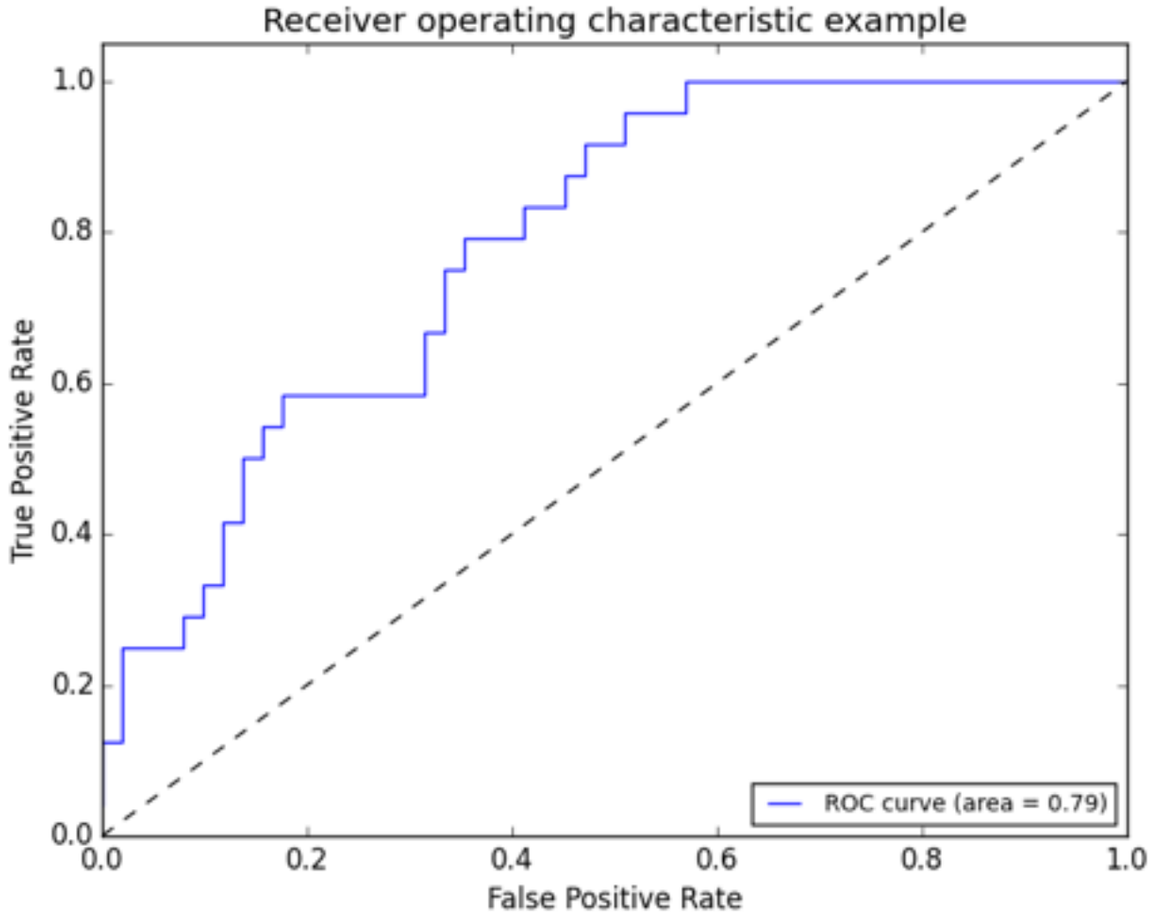
ADVANCED CLASSIFICATION METRICS

- We can continue adding pairs for different thresholds



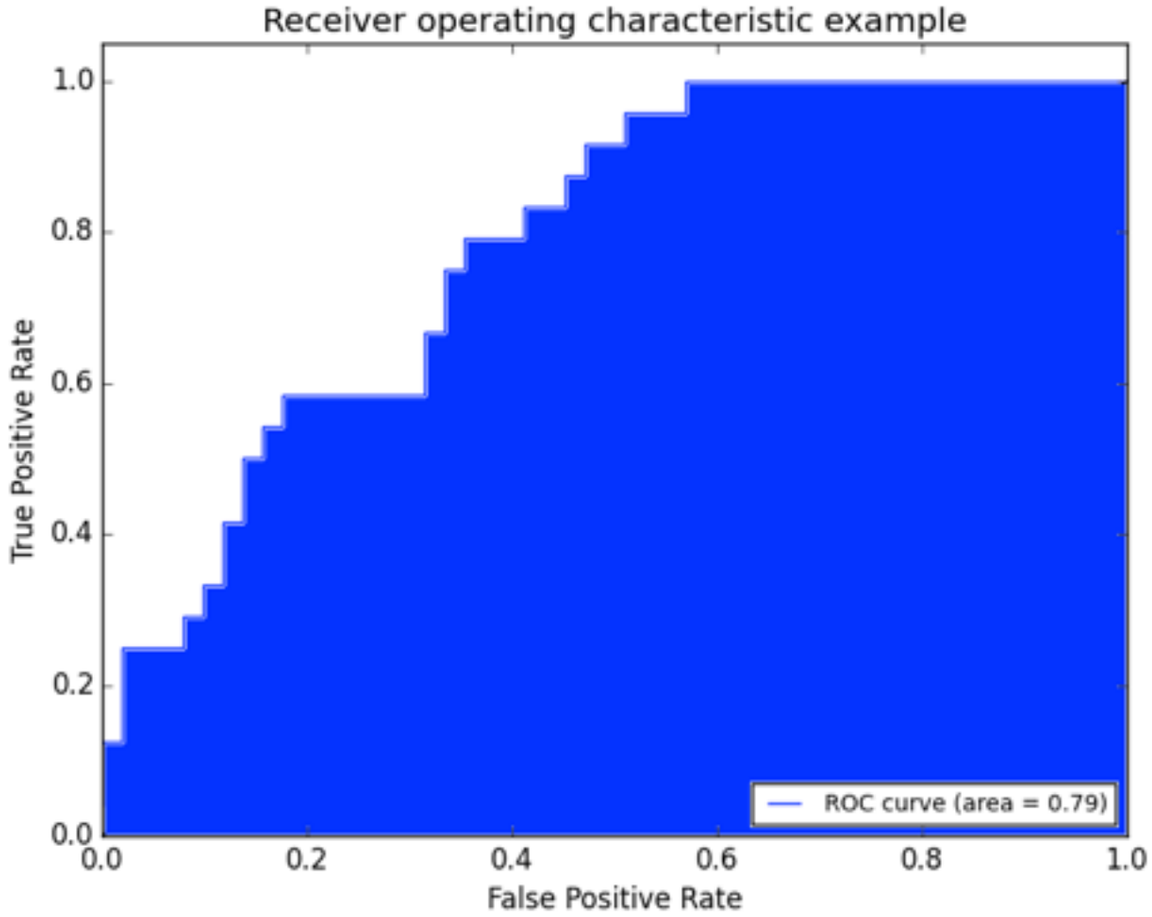
ADVANCED CLASSIFICATION METRICS

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



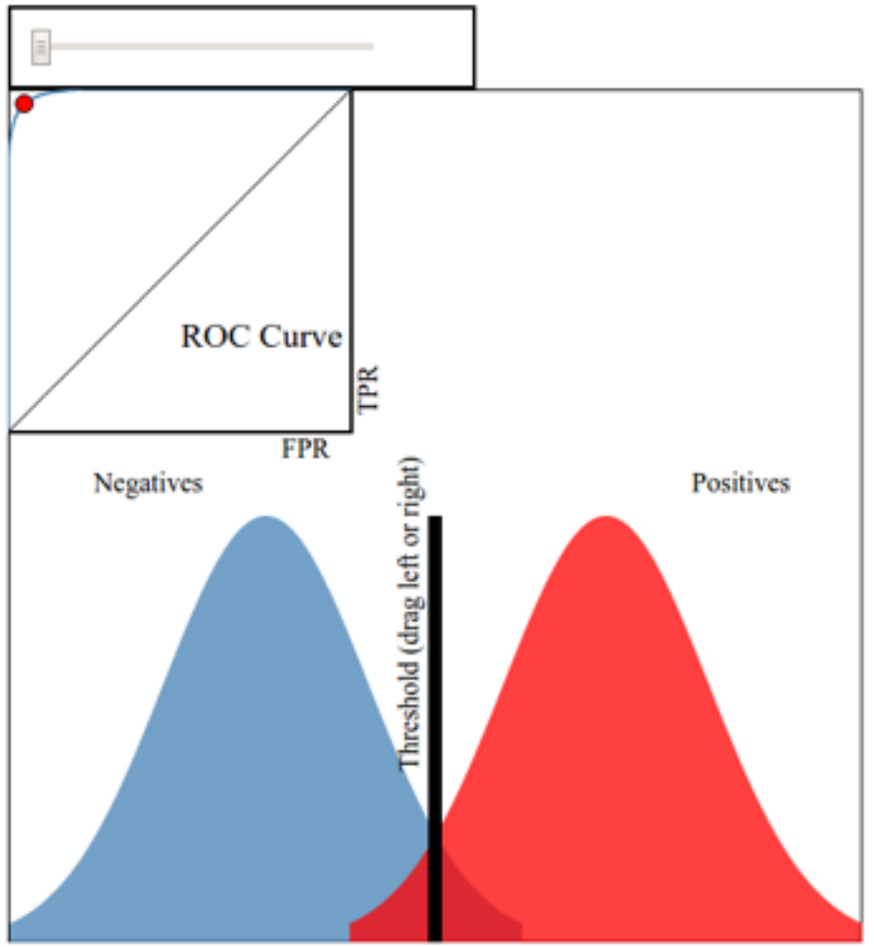
ADVANCED CLASSIFICATION METRICS

- 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

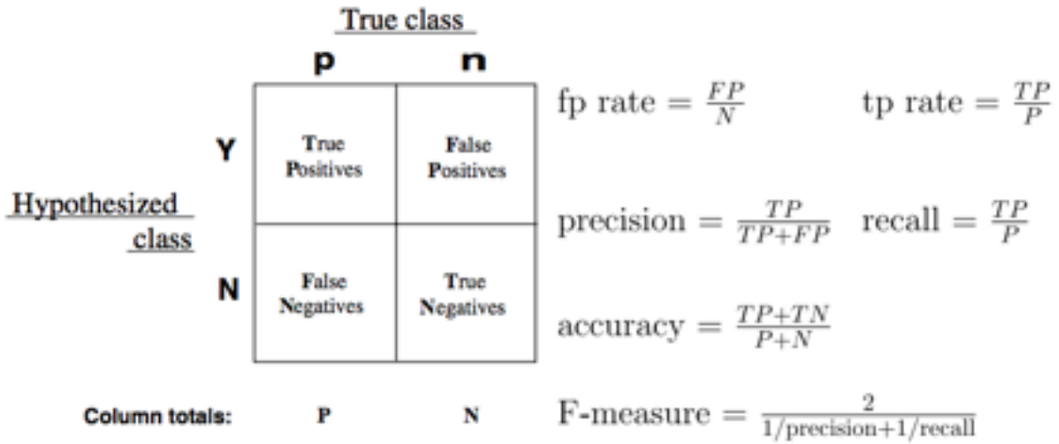
- If we have a TPR of 1 (all positives are marked positive) and FPR of 0 (all negatives are not marked positive), we'd have an AUC of 1. This means everything was accurately predicted.

- 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 all of the metrics located on one convenient page.

GUIDED PRACTICE

WHICH METRIC SHOULD I USE?

ACTIVITY: WHICH METRIC SHOULD I USE?

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.

ACTIVITY: WHICH METRIC SHOULD I USE?

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.

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

- What are True Positive and False Positive Rates?
- 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?
  - Examples?

COURSE

BEFORE L11

Admissions: Part 4

LESSON

Q & A

LESSON

EXIT TICKET

DON’T FORGET TO FILL OUT YOUR EXIT TICKET