GENERAL ASSEMBLY

INTRODUCTION TO LOGISTIC REGRESSION

Tan Kwan Chong

Chief Data Scientist, Booz Allen Hamilton

LEARNING OBJECTIVES

- Build a Logistic regression classification model using the statsmodels and sklearn libraries
- 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

OPENING

INTRODUCTION TO LOGISTIC REGRESSION

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?

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)?

DELIVERABLE

Answers to the above questions

INTRODUCTION

LOGISTIC REGRESSION

LOGISTIC REGRESSION

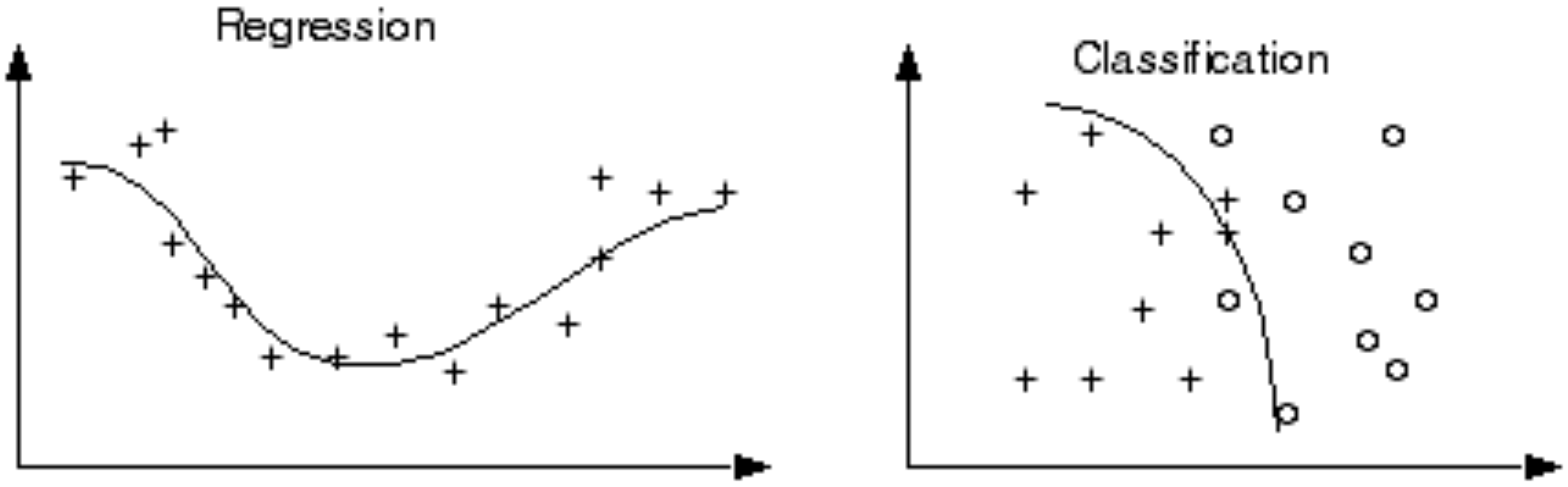
- Logistic regression is a *linear* approach to solving a *classification* problem.

- 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.

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.



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.

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 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.

ACTIVITY: KNOWLEDGE CHECK

EXERCISE

ANSWER THE FOLLOWING QUESTIONS

1. Recall the ordinary least squares formula.
2. The prior probability is most similar to which value in the ordinary least squares formula?

DELIVERABLE

Answers to the above questions

FIX 2: LINK FUNCTIONS AND THE SIGMOID FUNCTION

- 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.

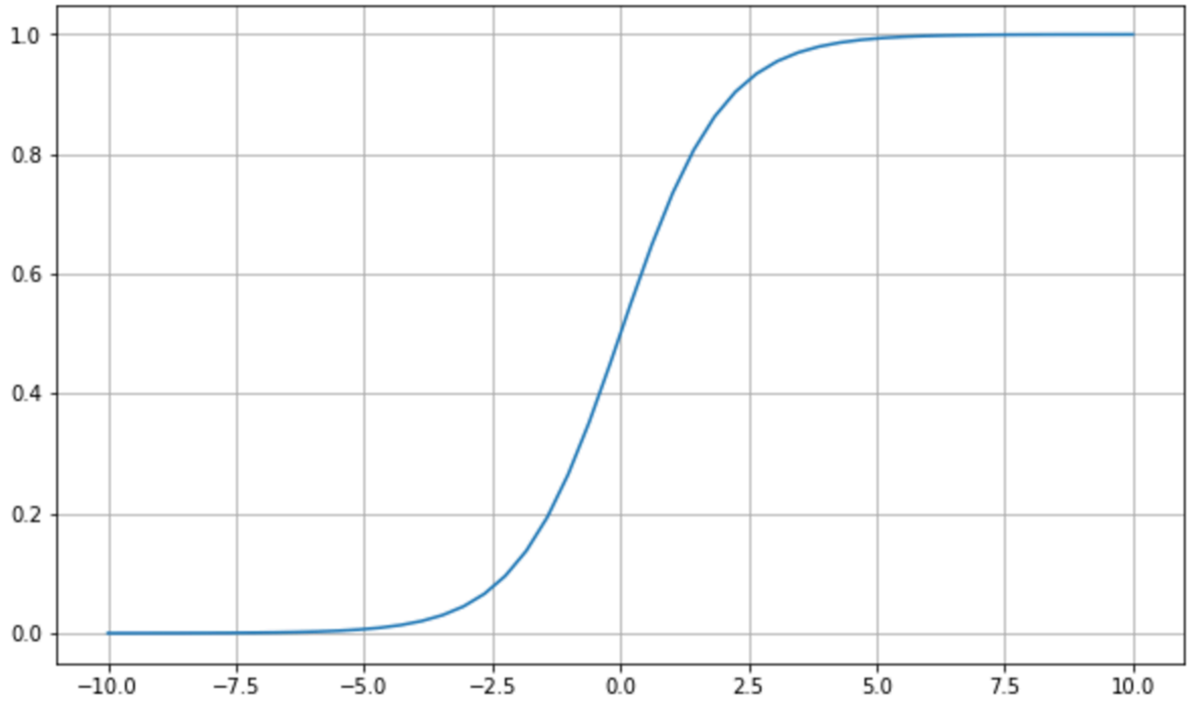
FIX 2: LINK FUNCTIONS AND THE SIGMOID FUNCTION

- 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

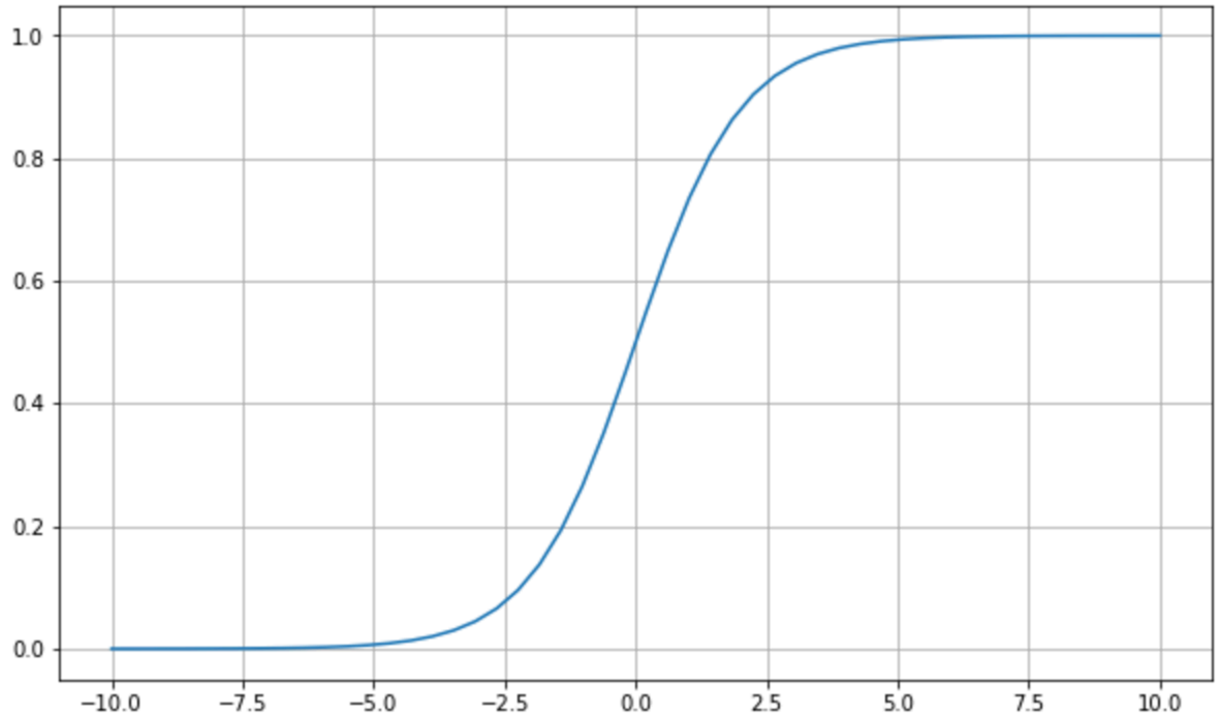
- A *sigmoid function* is a function that visually looks like an s.



- 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.



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.

DEMO

PLOTTING A SIGMOID FUNCTION

PLOTTING A SIGMOID FUNCTION

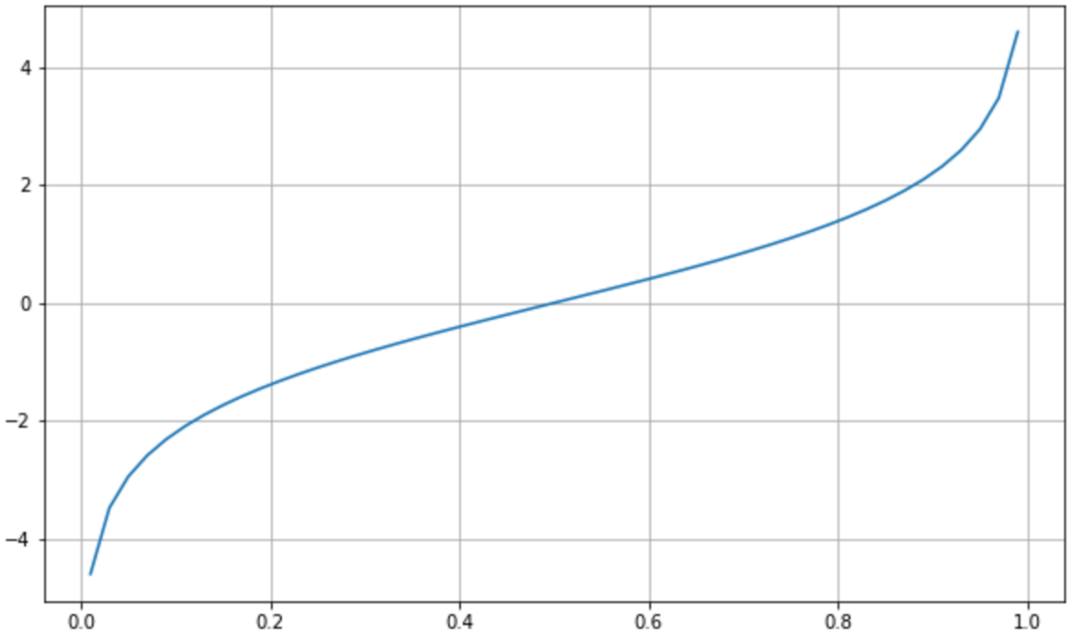
- Use the sigmoid function definition with values of x between -6 and 6 to plot it on a graph.
- Do this by hand or write Python code to evaluate it.
- Recall that e = 2.71.
- Do we get an the “S” shape we expect?

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 $\ln(\frac{p}{1-p})$



FIX 3: ODDS AND LOG-ODDS

- The value within the natural log, $\frac{p}{1-p}$ represents the *odds*. Taking the natural log of odds generates *log odds*.
- Odds for an event – 5:2 reflects the event happening 5 times and not happening 2 times i.e. probability = 5/7

FIX 3: ODDS AND LOG-ODDS

- The logit function allows for values between $-\infty$ 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?
2. What does this remind us of?

DELIVERABLE

Answers to the above questions

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.

$$\text{Log Odds: } \ln\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

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

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

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

FIX 3: ODDS AND LOG-ODDS

- Suppose we are looking at predicting the association between smoking status and gender
- Outcome:
  - Smoking = 1 for current smokers, 0 for current non-smokers
- Predictor:
  - Gender = 1 for men, 0 for women

FIX 3: ODDS AND LOG-ODDS

- Recall in linear regression, if we only had one binary X predictor i.e. gender, we would be predicting two means:

$$E(Y) = \beta_0 + \beta_1(Gender)$$

- β_0 would be the mean outcome when Gender = 0
- $\beta_0 + \beta_1$ would be the mean outcome when Gender = 1
- Therefore β_1 is the difference in the mean outcome when Gender = 1 vs Gender = 0

FIX 3: ODDS AND LOG-ODDS

- For logistic regression:

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

- Given sample model coefficients of -3.1 for β_0 and 1.0 for β_1

$$\ln\left(\frac{p(X)}{1-p(X)}\right) = -3.1 + 1.0 * (Gender)$$

- For women: $\ln\left(\frac{p(X)}{1-p(X)}\right) = -3.1 + 1.0 * 0 = -3.1$
- For men: $\ln\left(\frac{p(X)}{1-p(X)}\right) = -3.1 + 1.0 * 1 = -2.1$

FIX 3: ODDS AND LOG-ODDS

- β_1 represents the change in log odds comparing men to women
- β_0 would be the log odds when Gender = 0
- $\beta_0 + \beta_1$ would be the log odds when Gender = 1
- Therefore β_1 is the difference in log odds when Gender = 1 vs Gender = 0
- Odds of smoking for women = $e^{\beta_0} = e^{-3.1} = 0.045$
- Odds of smoking for men = $e^{\beta_0} = e^{-2.1} = 0.1224$
- Odds Ratio = $\frac{Odds\ for\ men}{Odds\ for\ women} = \frac{e^{\beta_0+\beta_1 X}}{e^{\beta_0}} = \frac{0.1224}{0.045} = e^1 = 2.72$

FIX 3: ODDS AND LOG-ODDS

- $Relative\ Risk = \frac{p_{men}}{p_{women}} = \frac{\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}}{\frac{e^{\beta_0}}{1+e^{\beta_0}}}$
- $p_{men} = \frac{0.1224}{1+0.1224} = 0.109$
- $p_{women} = \frac{0.045}{1+0.045} = 0.043$
- $Relative\ Risk = \frac{0.109}{0.043} = 2.53$

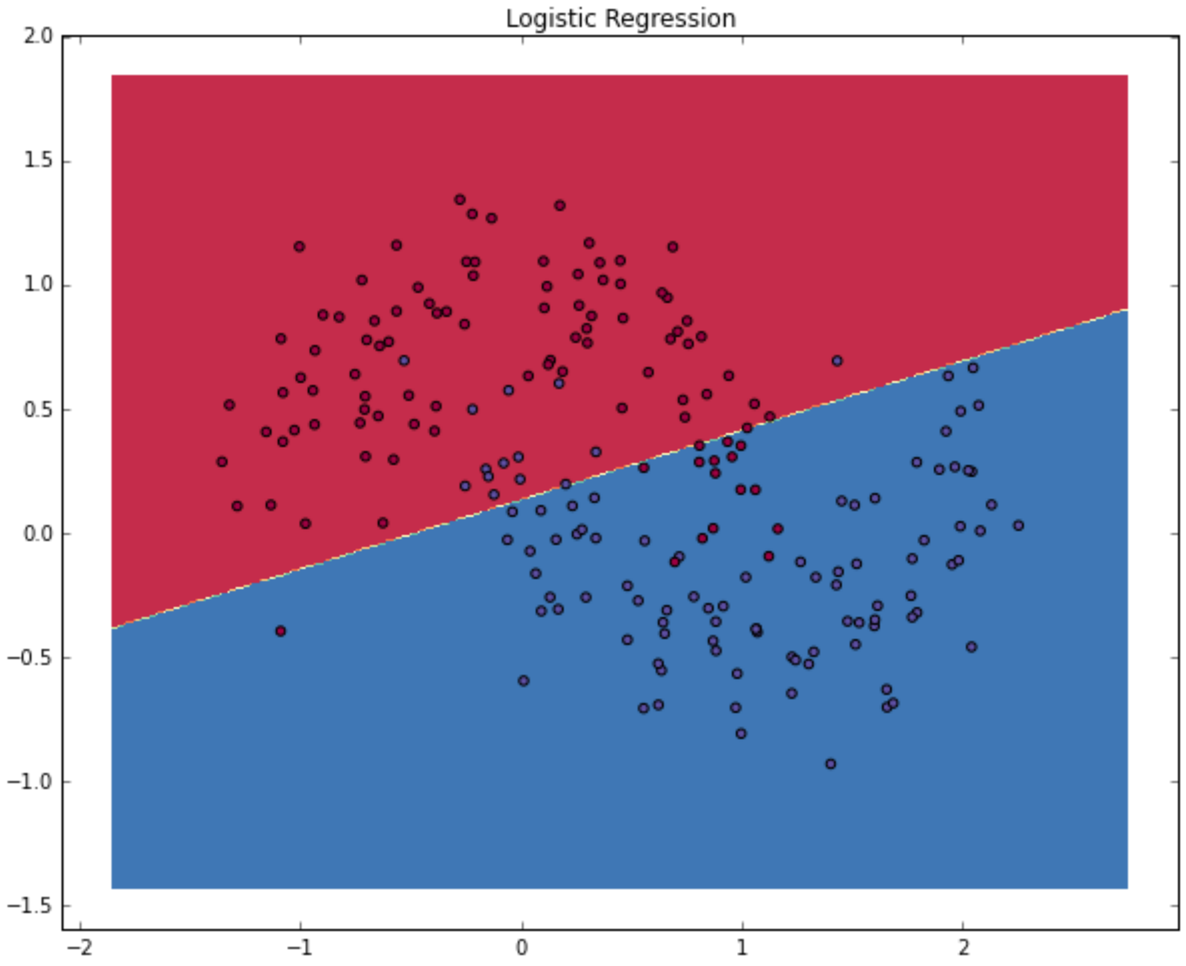
FIX 3: ODDS AND LOG-ODDS

- The coefficients are unknown and must be estimated based on the available training data
- Although we could use (non-linear) least squares to fit the model, the more general method of maximum likelihood is preferred since it has better statistical properties
- The basic intuition is that we seek estimates for the coefficients such that the predicted probability corresponds as closely to the actual observed value:

$$l(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} (1 - p(x_i))$$

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.



INDEPENDENT PRACTICE

LOGISTIC REGRESSION IMPLEMENTATION

ACTIVITY: LOGISTIC REGRESSION IMPLEMENTATION

EXERCISE

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?

DELIVERABLE

Answers to the above questions

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 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.

	Prediction Positive	Prediction Negative
Condition Positive	True Positive (TP)	False Negative (FN)
Condition Negative	False Positive (FP)	True Negative (TN)

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?

	Prediction Positive	Prediction Negative
Condition Positive	True Positive (TP)	False Negative (FN)
Condition Negative	False Positive (FP)	True Negative (TN)

True Positive Rate aka Sensitivity, Recall

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

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?

	Prediction Positive	Prediction Negative
Condition Positive	True Positive (TP)	False Negative (FN)
Condition Negative	False Positive (FP)	True Negative (TN)

False Positive Rate aka Specificity

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

ADVANCED CLASSIFICATION METRICS

- Precision reflects how many of the positive predictions are indeed positive

	Prediction Positive	Prediction Negative
Condition Positive	True Positive (TP)	False Negative (FN)
Condition Negative	False Positive (FP)	True Negative (TN)

Positive Prediction Value aka Precision

$$PPV = \frac{TP}{(TP + FP)}$$

	Prediction Positive	Prediction Negative
Condition Positive	True Positive (TP)	False Negative (FN)
Condition Negative	False Positive (FP)	True Negative (TN)

Negative Prediction Value

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

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

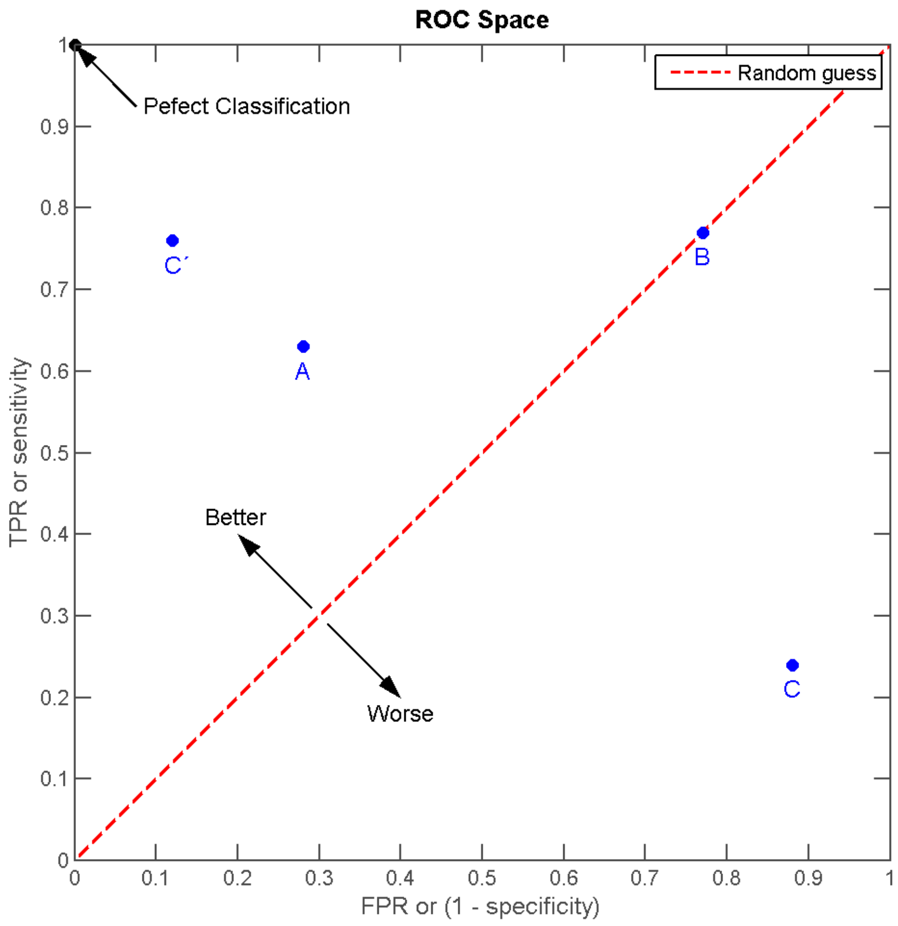
- This is where the Receiver Operation Characteristic (ROC) curve comes in handy.

- 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.

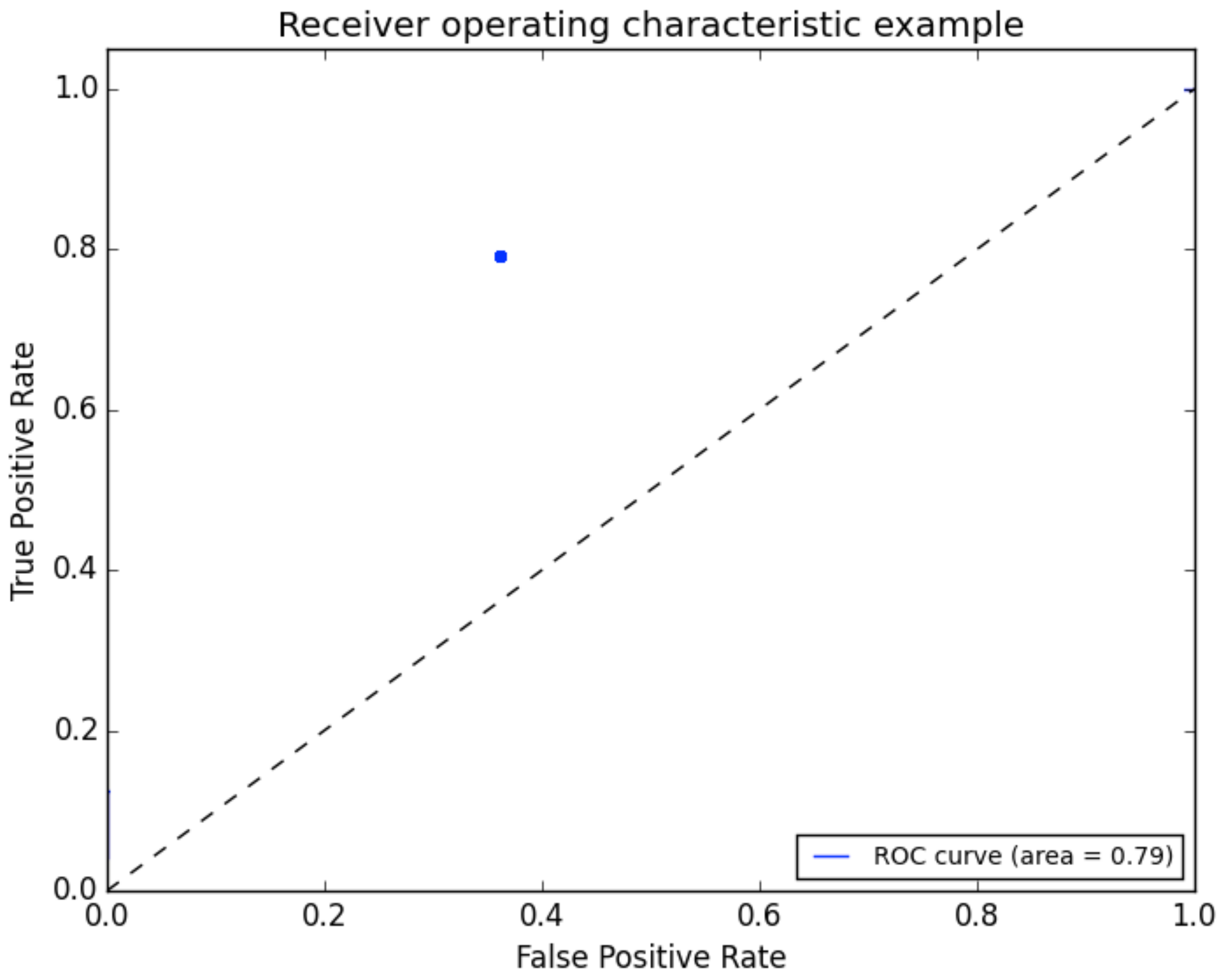
ADVANCED CLASSIFICATION METRICS

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



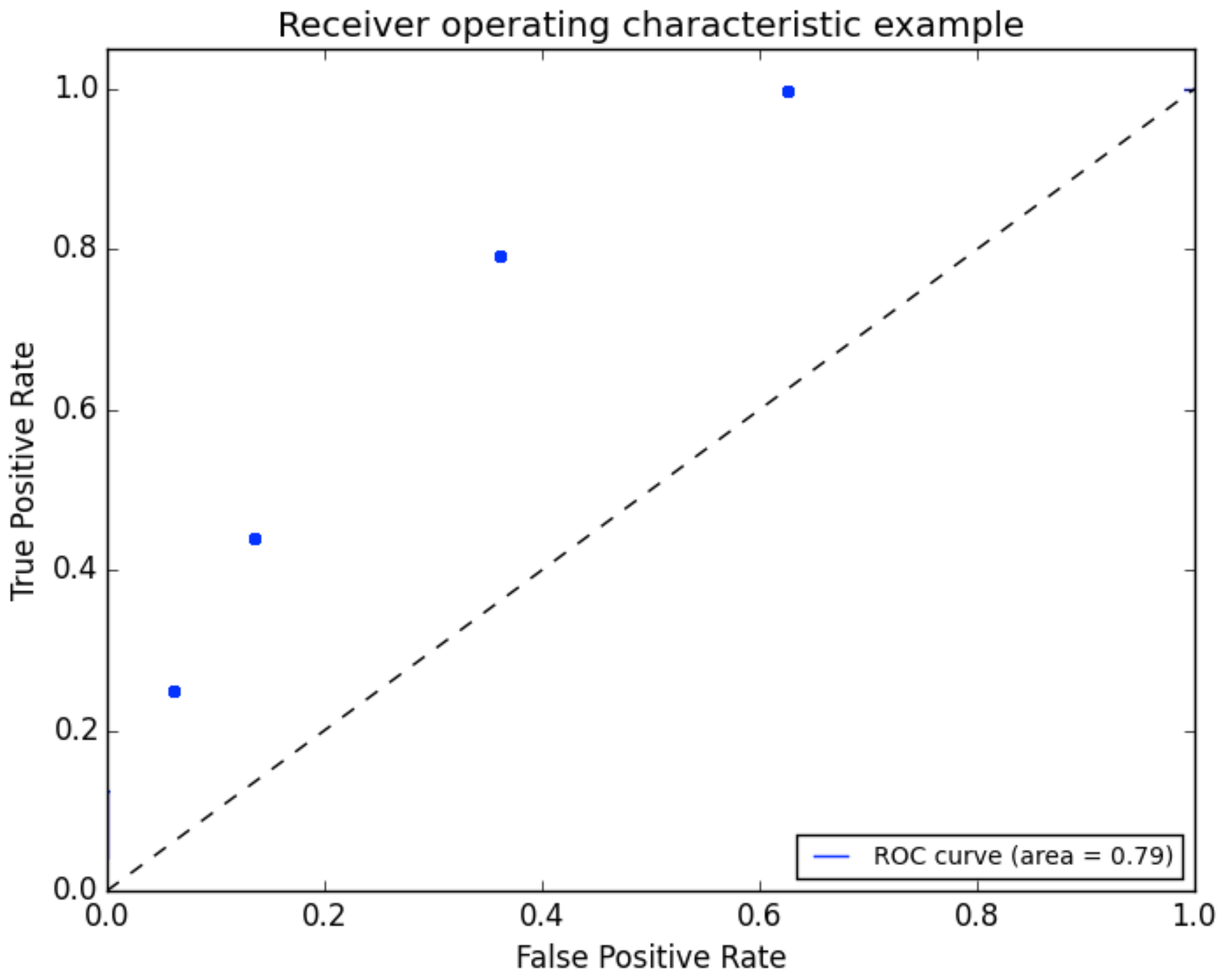
ADVANCED CLASSIFICATION METRICS

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



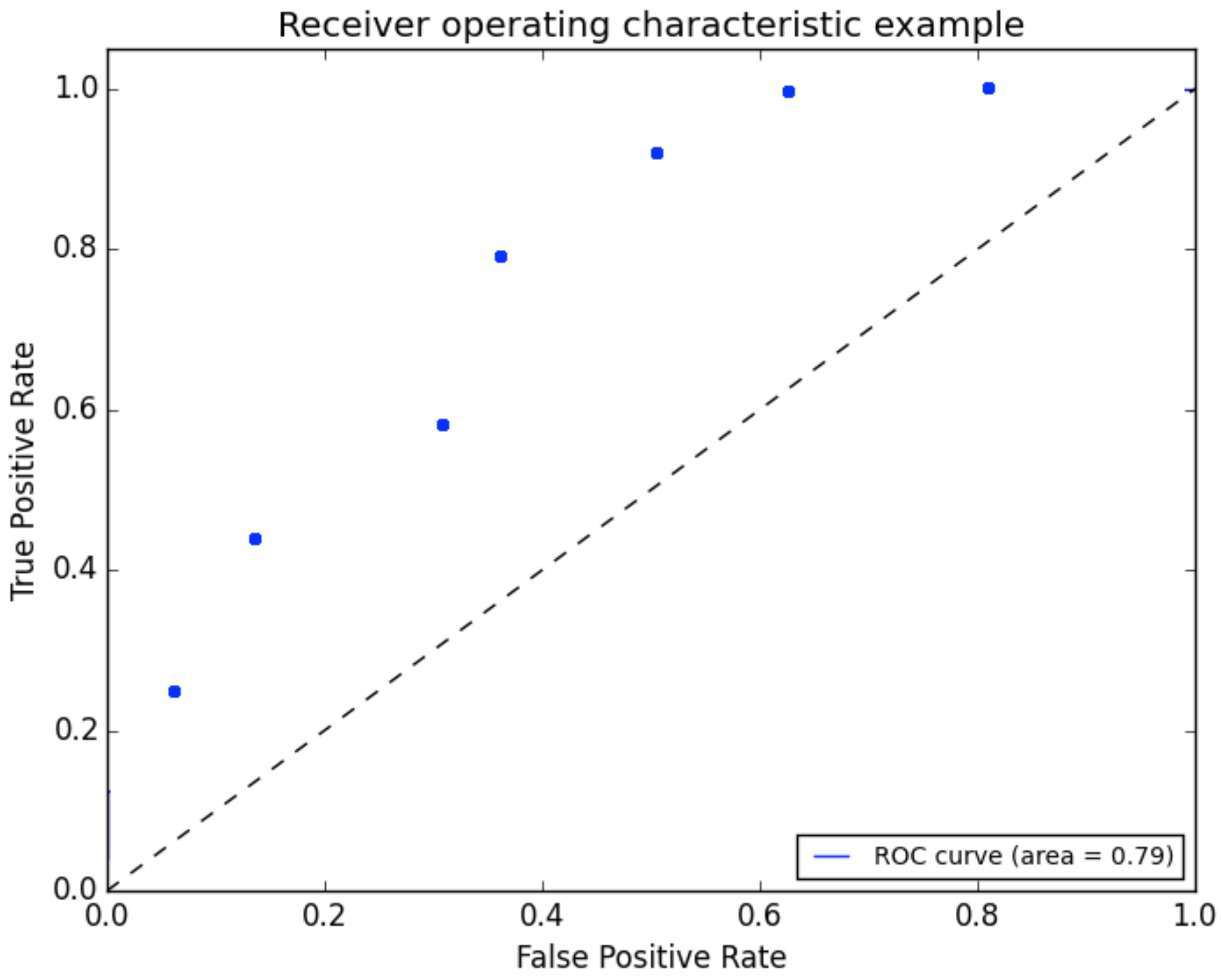
ADVANCED CLASSIFICATION METRICS

- We can continue adding pairs for different thresholds



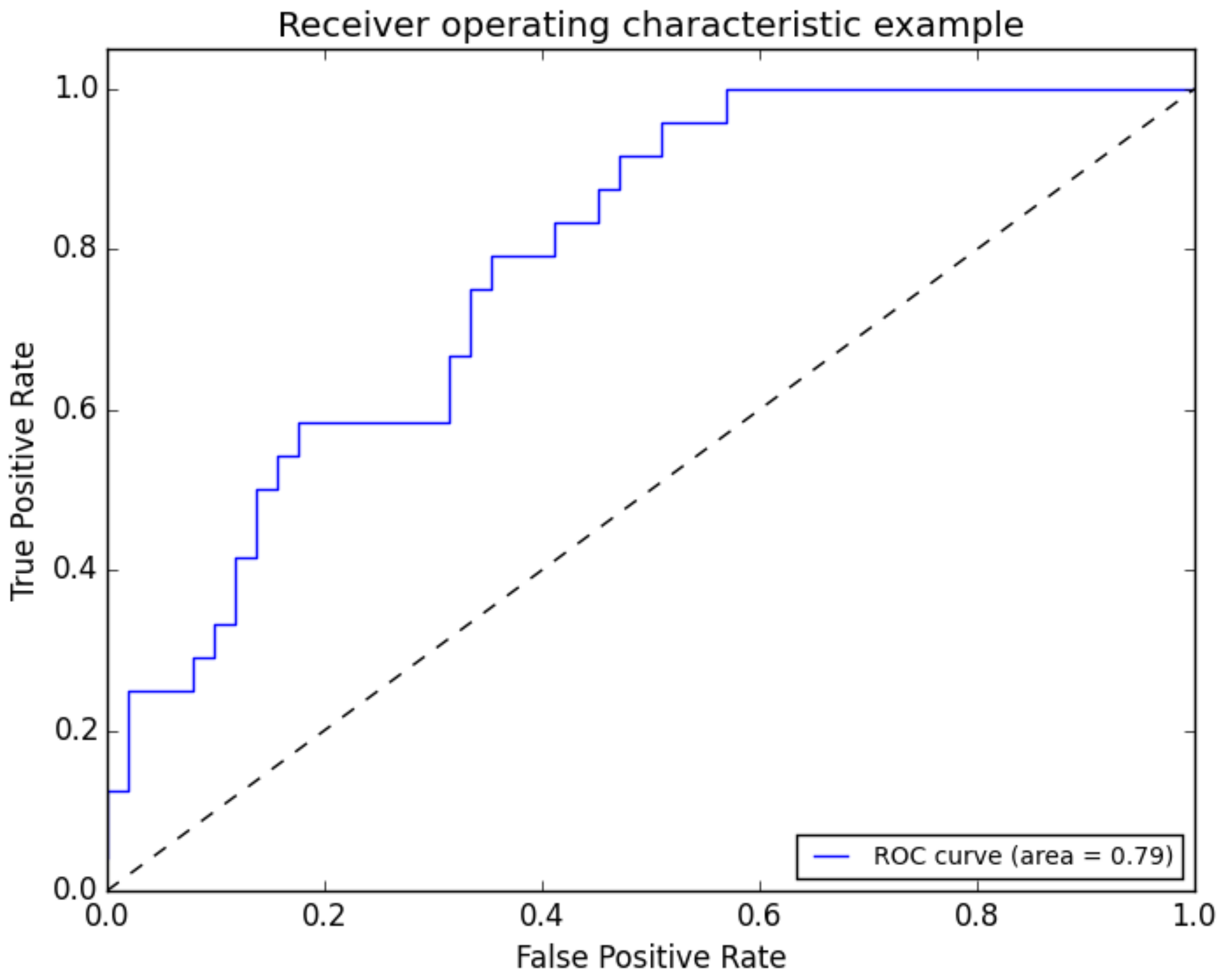
ADVANCED CLASSIFICATION METRICS

- We can continue adding pairs for different thresholds



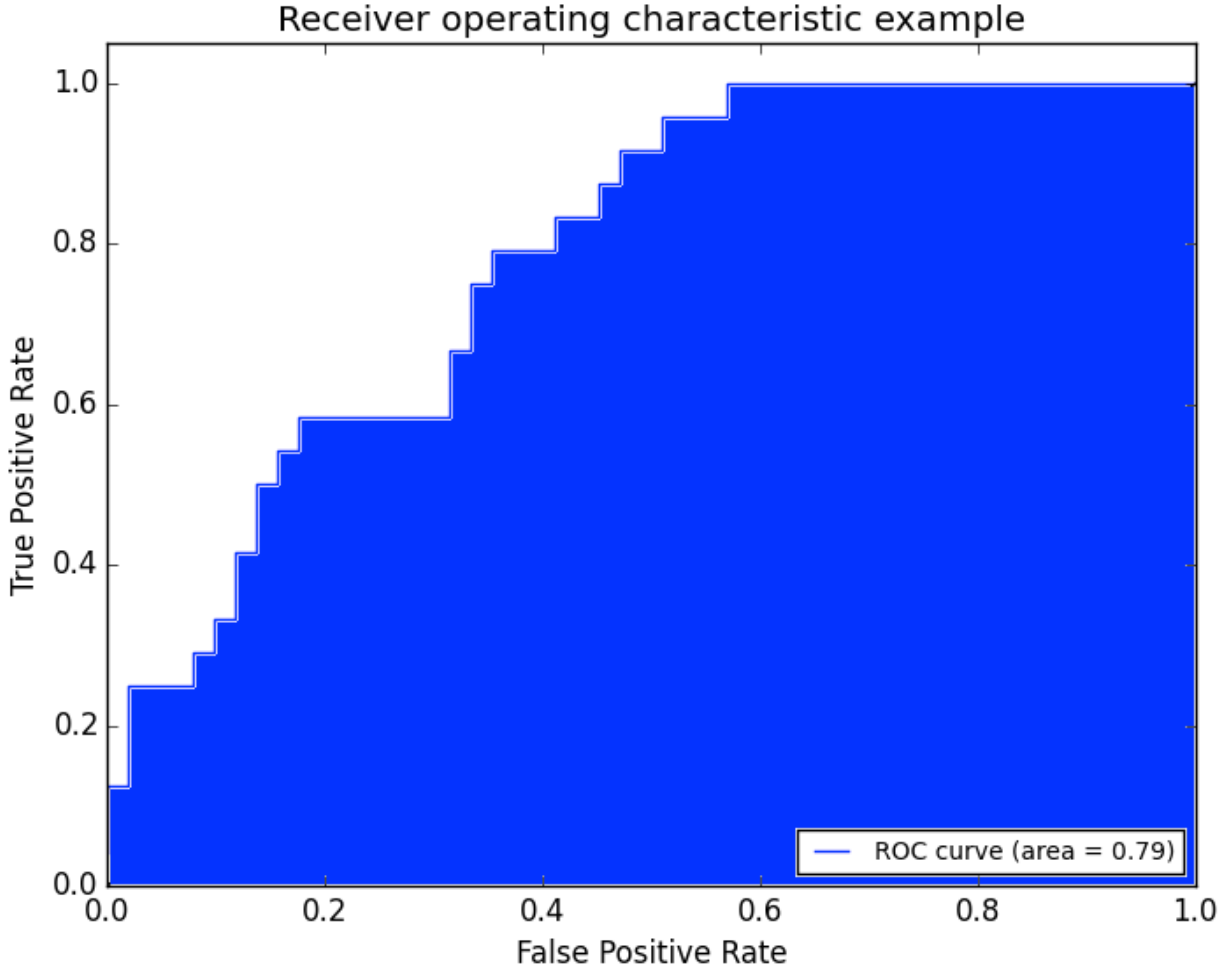
ADVANCED CLASSIFICATION METRICS

- 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

- 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.

True class

Hypothesized class	p	n
Y	True Positives	False Positives
N	False Negatives	True Negatives
Column totals:	P	N

$$\text{fp rate} = \frac{FP}{N} \qquad \text{tp rate} = \frac{TP}{P}$$

$$\text{precision} = \frac{TP}{TP+FP} \qquad \text{recall} = \frac{TP}{P}$$

$$\text{accuracy} = \frac{TP+TN}{P+N}$$

$$\text{F-measure} = \frac{2}{1/\text{precision}+1/\text{recall}}$$

- 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 (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?

2. 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

LESSON

Q & A

LESSON

EXIT TICKET

DON’T FORGET TO FILL OUT YOUR EXIT TICKET