GENERAL ASSEMBLY

INTRODUCTION TO LOGISTIC REGRESSION

Stefan Jansen

DAT-NYC

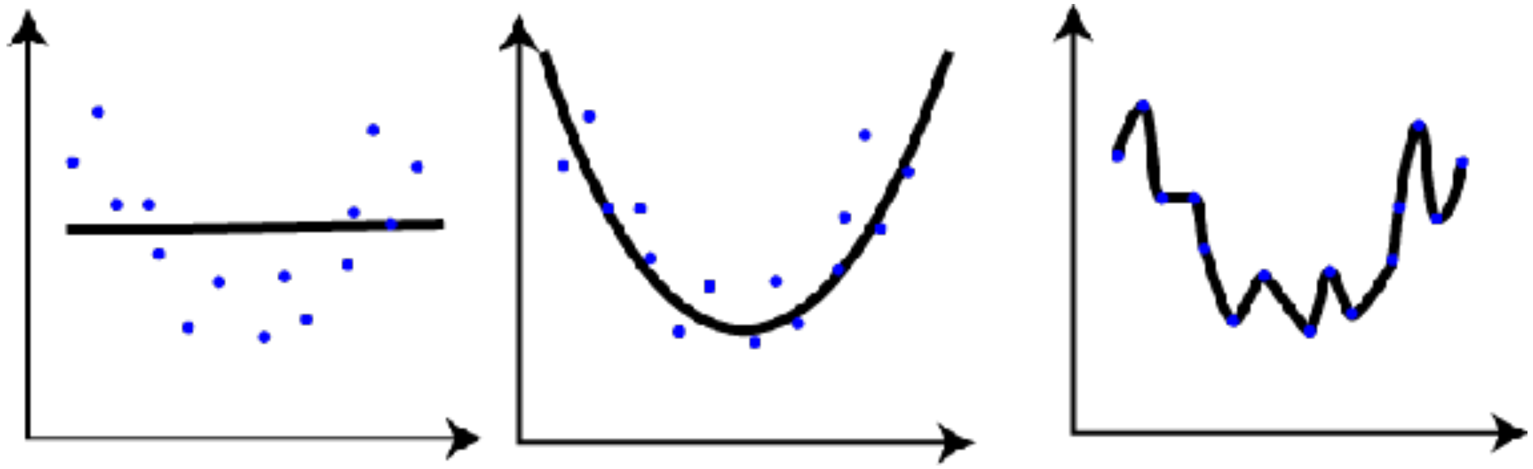
LEARNING OBJECTIVES

- 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, confusion matrix, ROC/AUC curves, and loss functions

COURSE

REGULARIZATION: REVIEW

WHAT IS OVERFITTING?



- The first model poorly explains the data.
- The second model explains the general curve of the data.
- The third model drastically overfits the model, bending to every point.
- Regularization helps prevent the third model, which is overly complex.

WHAT IS REGULARIZATION? AND WHY DO WE USE IT?

- Regularization protects against over-fitting by adding a penalty to the sum of squared residuals that depends on the size of the parameters

- This 'penalty for complexity' shrinks coefficients closer to zero.

    - Lasso => some coefficients to zero

    - Ridge => proportionally closer to zero

$$L1 = \alpha \sum_{p=1}^{P} \left| \beta_p \right|$$

$$L2 = \lambda \sum_{p=1}^{P} \beta_p^2$$

- Scale matters - standardize inputs!

- Use Lasso when # features > # observations, Ridge otherwise.

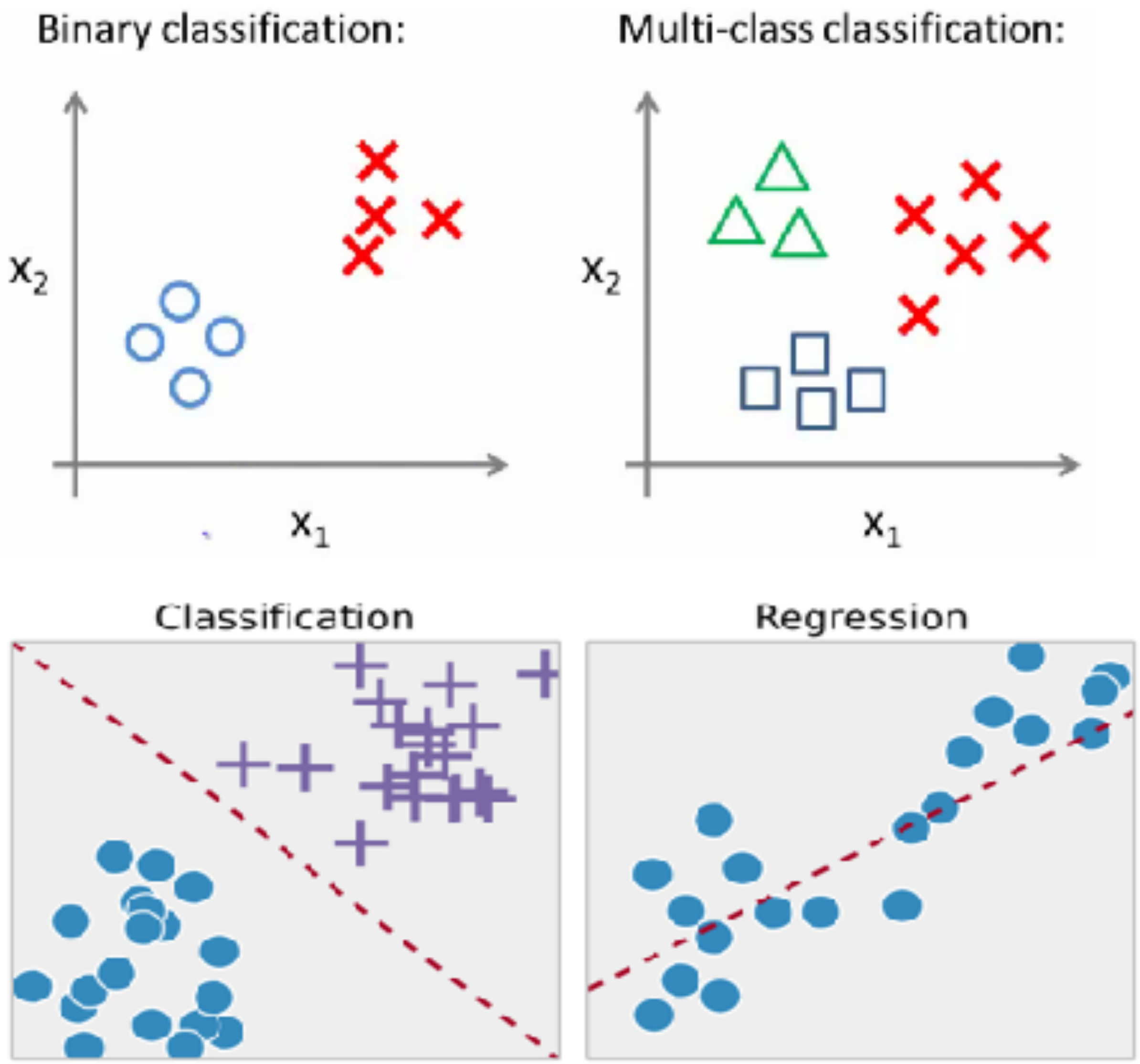
COURSE

CLASSIFICATION: REVIEW

WHAT IS CLASSIFICATION?

- **Classification** is a machine learning problem to assign one of 2 or more discrete values (categories) to observations using other information on these data points.

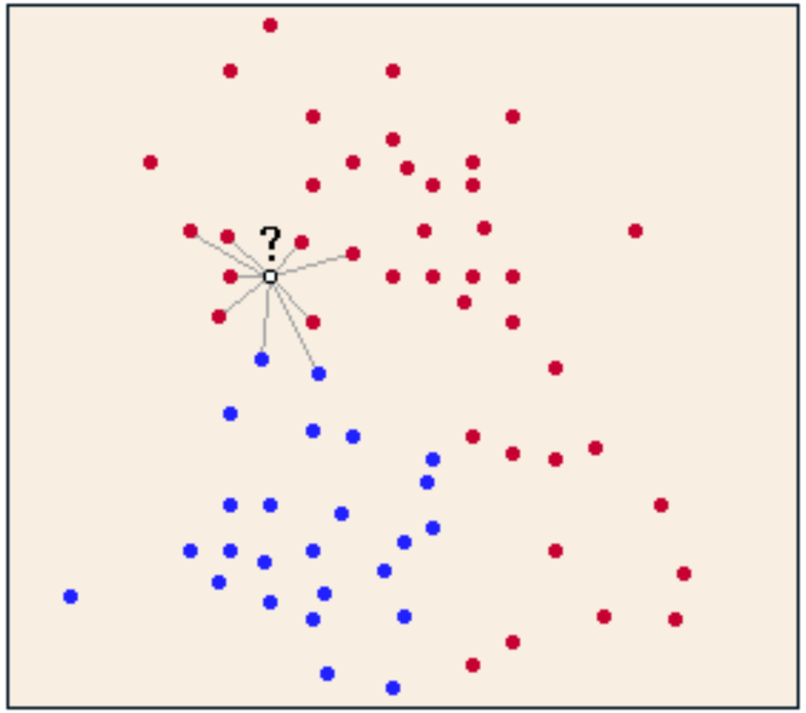
- Linear regression fits a line or hyper plane to the data; classification identifies one or more **decision boundaries**.



WHAT IS K NEAREST NEIGHBORS?

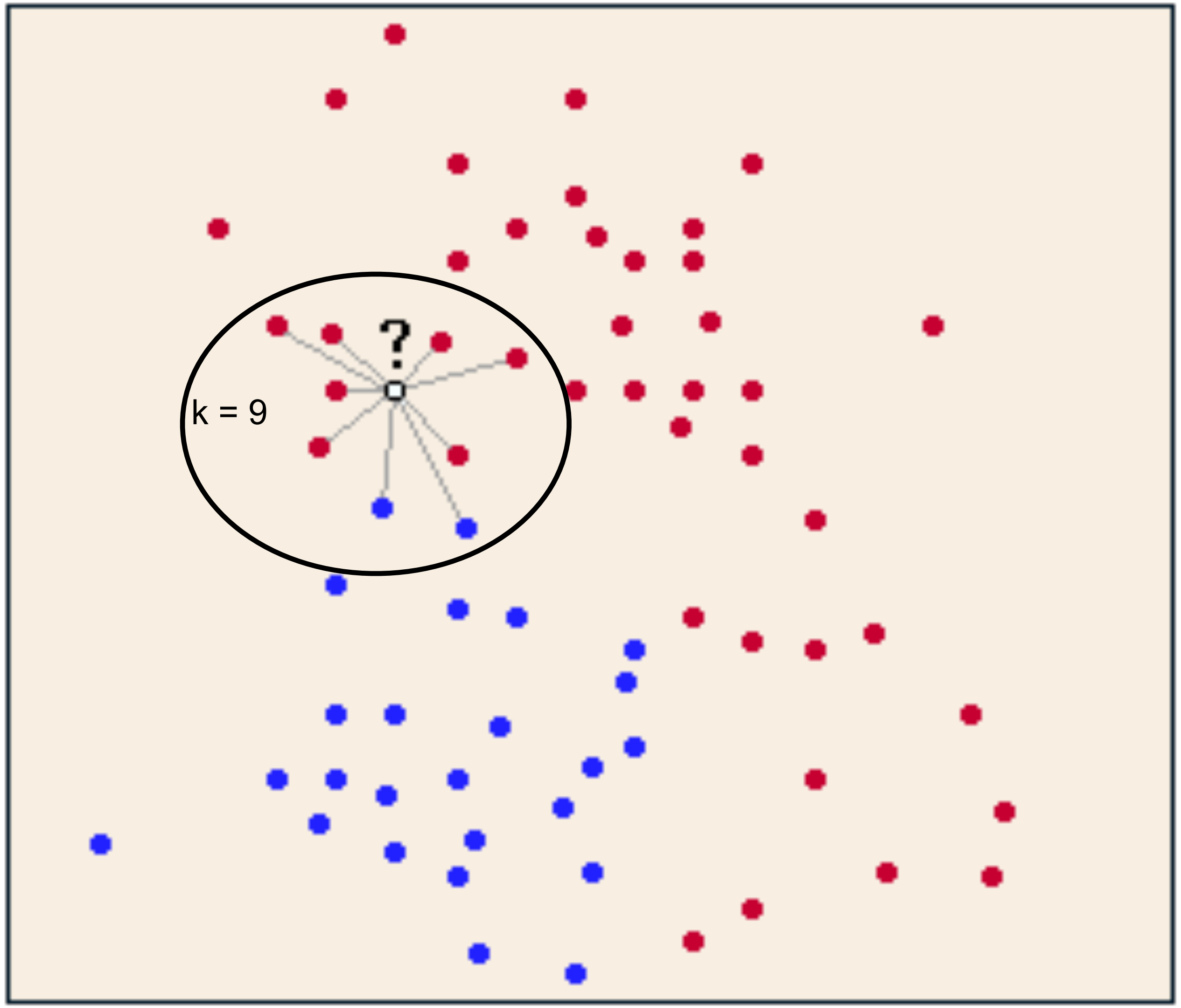
- **K Nearest Neighbors (KNN)** is a classification algorithm that makes a prediction based upon the closest data points as follows:

  - For a given point, calculate the distance to all other points.
  - Given these distances, pick the *k* closest points.
  - Calculate the probability of each class label given these points.
  - The original point is classified as the class label with the largest probability ("votes").



WHAT IS K NEAREST NEIGHBORS?

- KNN uses distance as a measure of similarity to predict a class label.
- Think of using shared traits to identify the most likely class label.
- Optimization concerns the choice of *k*, the size of the neighborhood.



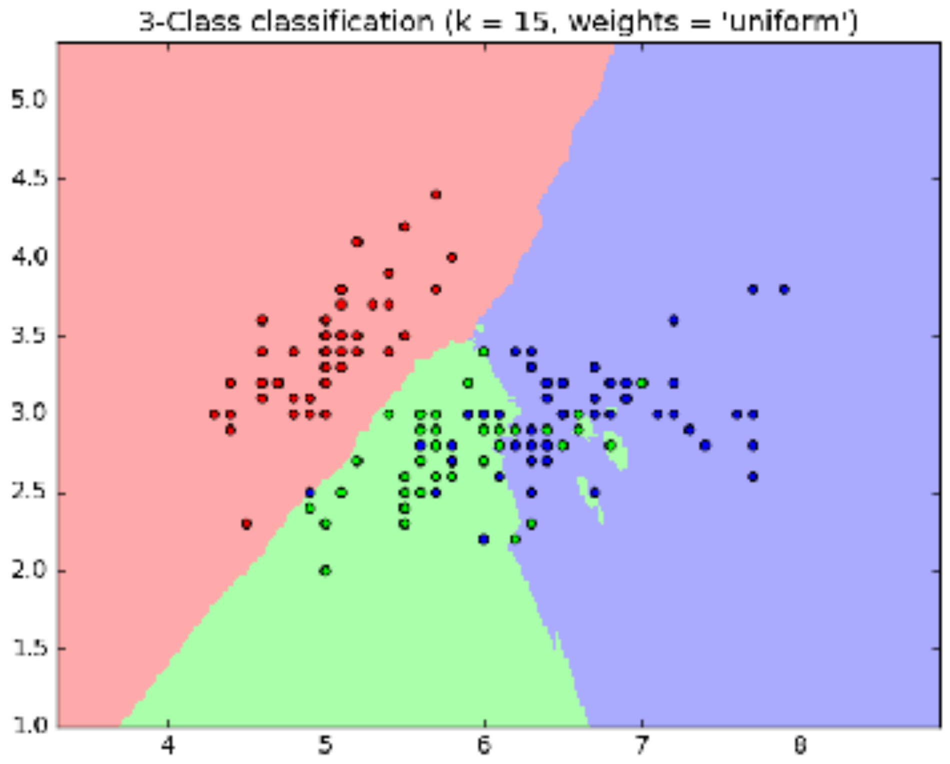
WHAT HAPPENS IN HIGH DIMENSIONALITY?

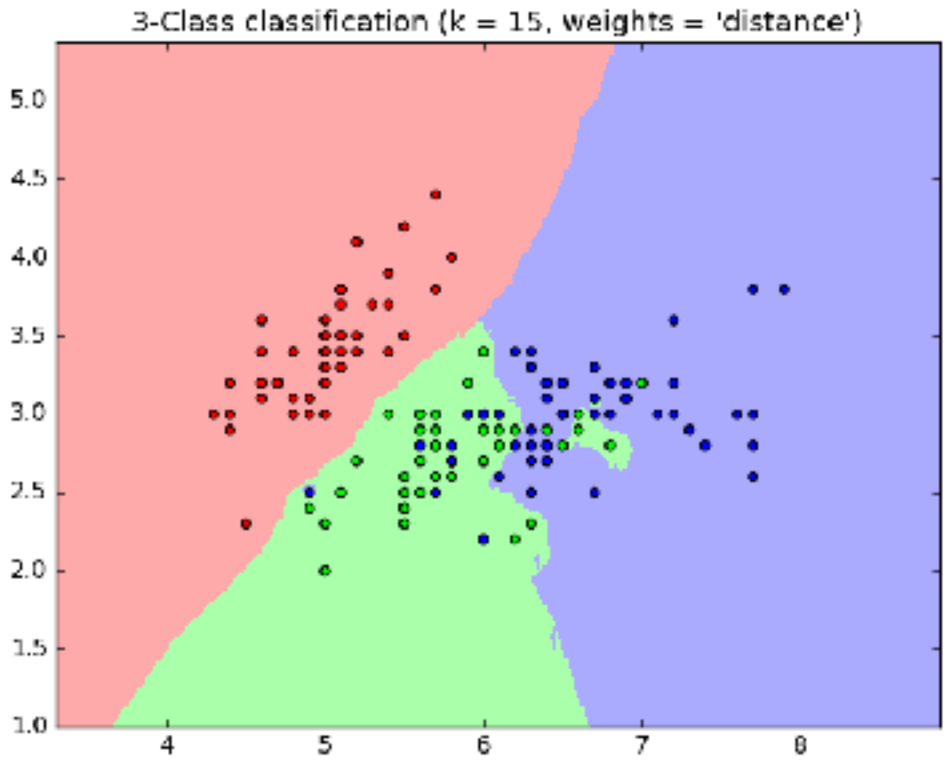
- Since KNN works with distance, higher dimensionality of data (i.e. more features) requires *significantly* more samples for the same predictive power.
- With a single feature measured on a scale 0-100, you need 100 data points to be on average within one unit of a new data point. With two such features, you need 100 * 100, or 10,000 data points. And so on.
- Hence: keep the feature space limited and KNN will do well. Exclude extraneous features when using KNN.

KNN IN SCIKIT-LEARN

- `scikit-learn` provides `KNeighborsClassifier` and `RadiusNeighborsClassifier` (if data not uniformly sampled).

- The `weights` parameter impacts how votes are cast by the k neighbors.





INTRODUCTION TO CLASSIFICATION METRICS

- We'll use two primary metrics: *accuracy* and *misclassification rate*.
- **Accuracy** is the number of *correct* predictions out of all predictions in the sample. This is a value we want to *maximize*.
- **Misclassification rate** is the number of *incorrect* predictions out of all predictions in the sample. This is a value we want to *minimize*.
- These two metrics are directly opposite of each other.
- 1 - misclassification rate = accuracy

OPENING

INTRODUCTION TO LOGISTIC REGRESSION

INTRODUCTION TO LOGISTIC REGRESSION

EXERCISE

ANSWER THE FOLLOWING QUESTIONS (5 min)

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

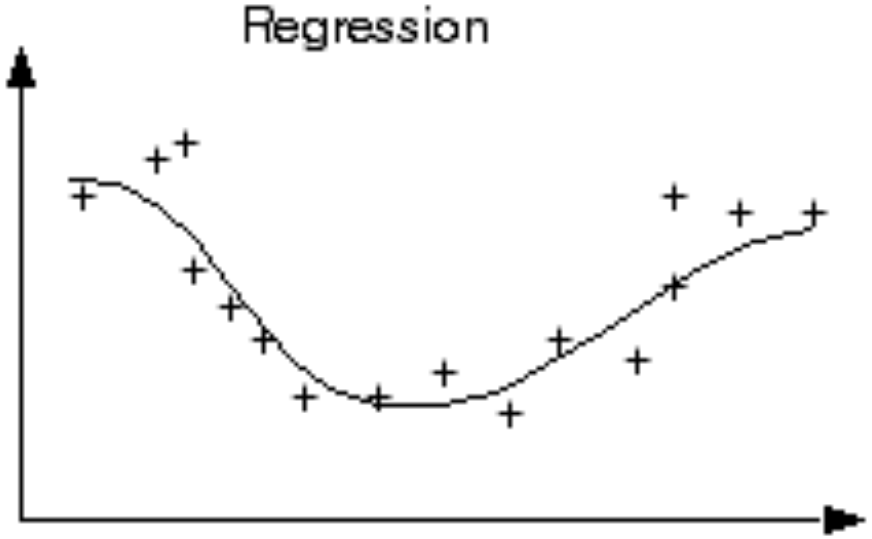
FROM LINEAR TO LOGISTIC REGRESSION

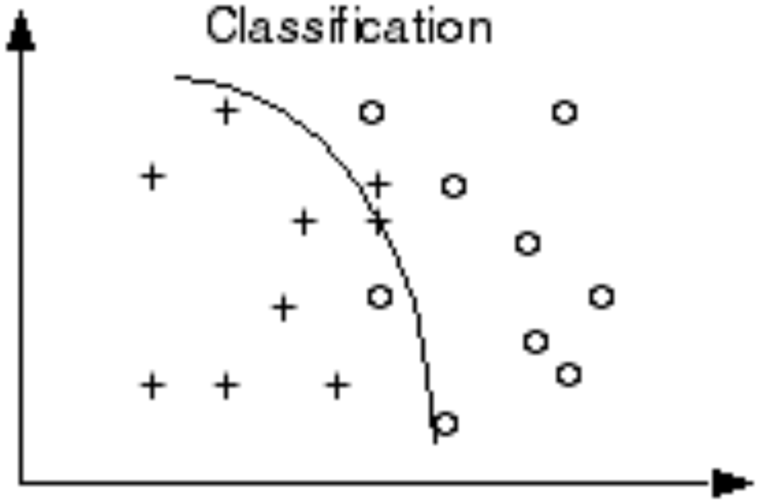
LOGISTIC REGRESSION

- 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 decide if an item *belongs* or *does not belong* to a class label.
- *Linear Model* means that the prediction is a linear function of parameters that we need to estimate.
- It also implies that the decision boundaries that we use to separate classes are linear.

LINEAR REGRESSION RESULTS FOR CLASSIFICATION

- Regression results can range from -∞ to ∞.
- Classification is used when predicted values (i.e. class labels) assume a limited number of values, and cannot be ordered in a meaningful way.





LINEAR REGRESSION RESULTS FOR CLASSIFICATION

- But, since most classification problems are binary (0 or 1) and 1 does not mean 'greater than 0', does it make sense to apply the concept of regression to solve classification?

- How might we contain our predictions with [0, 1] bounds and align our modeling approach more closely with the outcome of categories as opposed to quantities?

- Let's review some elements of our approach to make classification 'compatible' with a linear input function.

ELEMENT 1: PREDICTING CLASS 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 frequency.

- 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%).

ACTIVITY: KNOWLEDGE CHECK

EXERCISE

ANSWER THE FOLLOWING QUESTIONS

1. Recall the ordinary least squares formula.

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

DELIVERABLE

Answers to the above questions

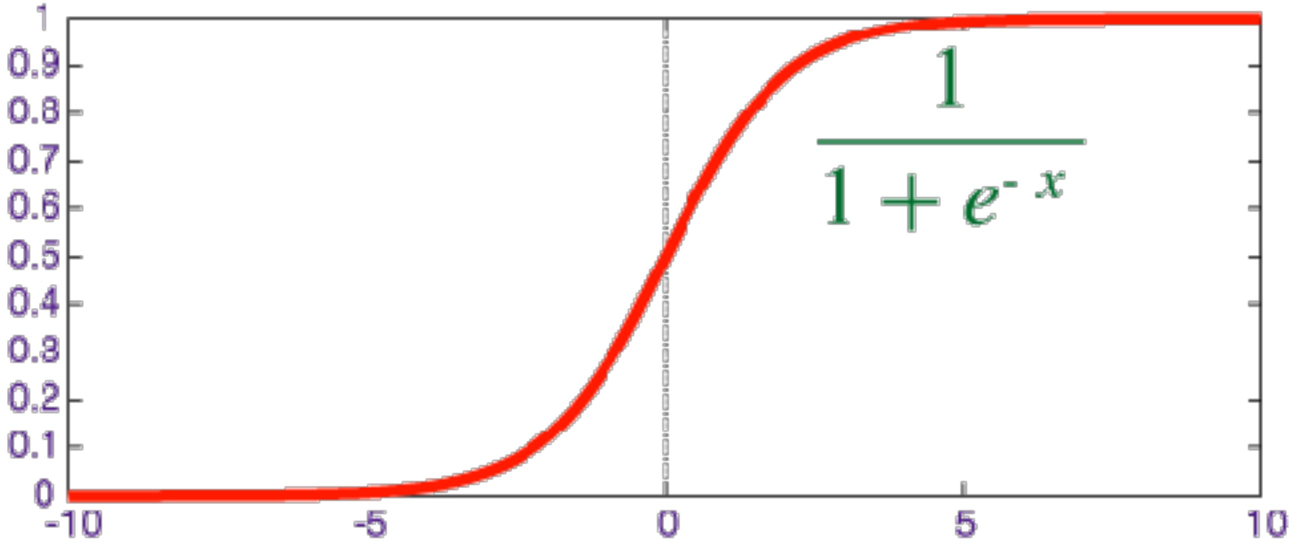
ELEMENT 2: PROBABILITY AS A FUNCTION OF INPUTS

- However, we would still like to use a (linear) function to increase or decrease the probability of an observation given the data about it.

- So we need a function that produces values between 0 and 1 from arbitrary inputs to relate our linear predictors to the response variable.

- One such function is the logistic function that produces an S-shaped sigmoid curve.

- It varies between [0, 1] as x varies between [-∞, ∞].



DEMO

PLOTTING A SIGMOID FUNCTION

ACTIVITY: PLOTTING A SIGMOID FUNCTION

EXERCISE

INSTRUCTIONS (5 min):

- Write Python code to evaluate the sigmoid function definition with values of x between -6 and 6 and plot it on a graph.
- Do we get the "S" shape we expect?

DELIVERABLE

Answers to the above questions

INTRODUCTION

MORE ON LOGISTIC REGRESSION

ACTIVITY: KNOWLEDGE CHECK

EXERCISE

ANSWER THE FOLLOWING QUESTIONS

1. What was the distribution most aligned with Linear Regression?
2. Where did the ‘random’ element appear in our Linear Regression model?
3. Which distribution could we use to model a binary outcome?

DELIVERABLE

Answers to the above questions

BUILDING A MODEL FOR CLASS PROBABILITIES

- We have two classes, labeled as 0 and 1. We are now modeling the class probabilities according to the Bernoulli distribution for binary variables:

$$P(y=1 \mid x) = h(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1)}}$$

$$P(y=0 \mid x) = 1 - h(x)$$

- We can't use least squares to find the parameters. Instead, we use a more general method called 'Maximum Likelihood' that finds parameters that make the sample 'most likely' given our probability model.

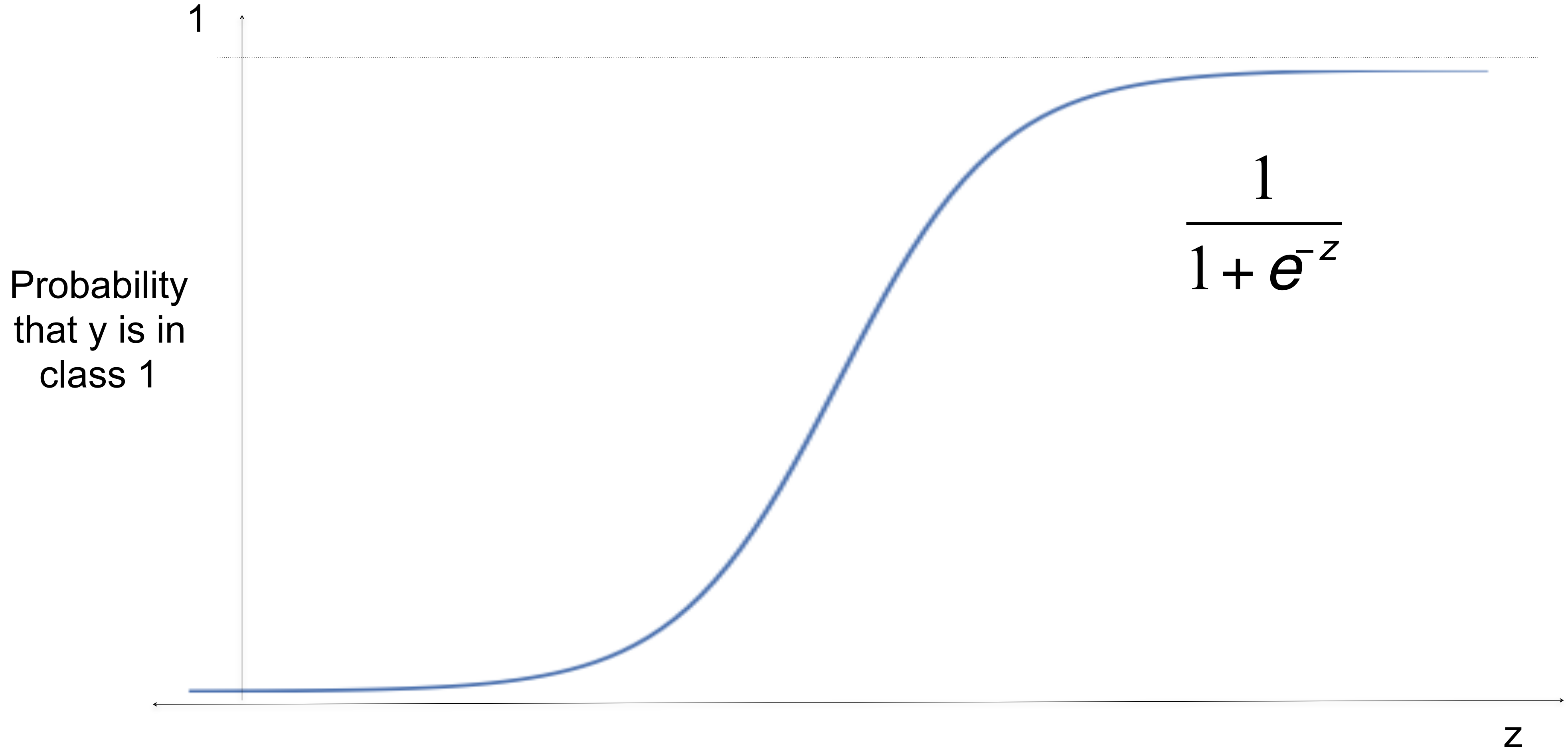
BUILDING A LINEAR MODEL FOR CLASSIFICATION

- Problem: Classify observations into 2 (or more) categories given the information provided by various features

- Solution: Model the probability of observations belonging to either class by mapping a linear combination of the features to values [0, 1] that also sum to 1 for each observation. The logistic function does this for us:

$$\underbrace{P(y=1 \mid x)}_{\substack{\text{Probability that} \\ \text{y is in class 1} \\ \text{given } x_1, x_2, \ldots, x_p}} = \frac{1}{1+e^{\underbrace{-(\beta_0+\beta_1 x_1+\ldots+\beta_p x_p)}_{\substack{\text{Linear function of } x_1, x_2, \ldots, x_p \\ \text{Need to estimate } \beta_1, \beta_2, \ldots, \beta_p}}}}$$

LOGISTIC CURVE AS LINEAR FUNCTION OF FEATURES



$$\frac{1}{1+e^{-z}}$$

$$z = f(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

HOW CAN WE INTERPRET THE PARAMETERS?

- Linear Regression: increase of x_1 by 1 unit changes y by β_1.

- Logistic Regression: increase of x_1 by 1 unit changes input z to logistic function. How can we interpret the input?

Input: $$z = f(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

Class Probability: $$\underbrace{p = P(y=1 \mid x)}_{\text{Probability that y is in class 1}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\ldots+\beta_p x_p)}} = \underbrace{\frac{1}{1+e^{-z}}}_{\text{Linear function of } x_1, x_2, \ldots, x_p}$$

HOW CAN WE INTERPRET THE PARAMETERS?

Let's invert the logistic function to get a mathematical expression for z:

$$\boxed{p = \frac{1}{1+e^{-z}}} \xrightarrow{(\ )^{-1}} \frac{1}{p} = 1 + e^{-z} \xrightarrow{-1} \frac{1}{p} - 1 = e^{-z}$$

$$\frac{1}{p} - 1 = e^{-z} \xrightarrow{\frac{p}{p}=1} \frac{1-p}{p} = e^{-z} \xrightarrow{(\ )^{-1}} \frac{p}{1-p} = e^{z}$$

$$\frac{p}{1-p} = e^{z} \xrightarrow{\ln(\ )} \ln\left(\frac{p}{1-p}\right) = \ln\left(e^{z}\right) \longrightarrow \boxed{z = \ln\left(\frac{p}{1-p}\right)}$$

HOW CAN WE INTERPRET THE PARAMETERS?

The linear function models the log odds:

$$z = f(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p = \underbrace{\ln \overbrace{\left(\frac{p}{1-p}\right)}^{\text{odds: y in class 1}}}_{\text{Log odds}}$$

Converting probability to odds and back

$$p = 0.8 \implies odds = \frac{0.8}{1-0.8} = \frac{0.8}{0.2} = \frac{4}{1}$$

$$odds = \frac{4}{1} \implies p = \frac{4}{4+1} = \frac{4}{5} = 0.8$$

Try for yourself converting p=0.6 to odds and back!

Odds vs Probability

$$P(x) = \frac{\text{Probability for x}}{\text{Total Probability}} \quad \text{=> [0\%, 100\%]}$$

$$Odds(x) = \frac{\text{Probability for x}}{\text{Probability against x}} \quad \text{=> } [0, +\infty]$$

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

HOW CAN WE INTERPRET THE PARAMETERS?

The linear function models the log odds:

$$z = f(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p = \underbrace{\ln\left(\overbrace{\frac{p}{1-p}}^{\text{odds: y in class 1}}\right)}_{\text{Log odds}}$$

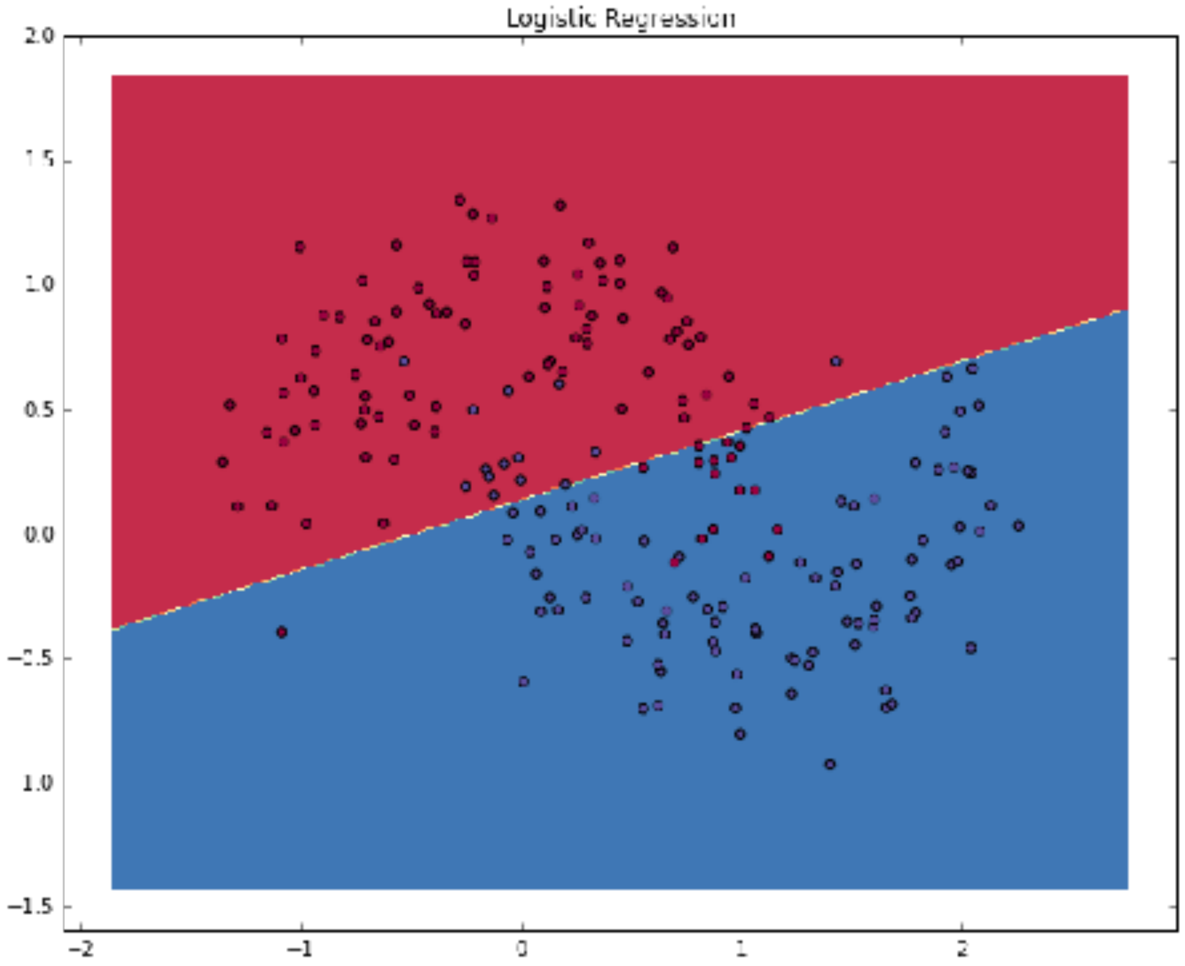
Hence - holding all other variables constant - a unit change in x_i changes:

- **the log odds by** β_i
- **the odds by** e^{β_i}

The impact of x_i on the actual class probability is non-linear and depends on the current value of the (log) odds; hence, it depends on the values of all x!

ODDS AND LOG-ODDS

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



ACTIVITY: KNOWLEDGE CHECK

EXERCISE

ANSWER THE FOLLOWING QUESTIONS

1. How did we interpret the coefficients in a linear regression model?
2. How would you interpret a coefficient in a logistic regression model?

DELIVERABLE

Answers to the above questions

GUIDED PRACTICE

WAGER THOSE ODDS!

CODING ACTIVITY 01: 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

CODING ACTIVITY 01: 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

INDEPENDENT PRACTICE

LOGISTIC REGRESSION IMPLEMENTATION

CODING ACTIVITY 02: LOGISTIC REGRESSION

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

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.
- 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.

EVALUATING OUTCOMES FOR BINARY CLASSIFICATION

		Outcome (Truth): Class 1	Outcome (Truth): Class 0
Prediction	**Class 1**	**True Positive (TP)**	**False Positive (FP)**
	Class 0	**False Negative (FN)**	**True Negative (TN)**

For all cases:

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{All Cases}} = \frac{TP + TN}{TP + FP + TN + FN}$$

For true Class 1 cases:

$$\text{True Positive Rate (Sensitivity, Recall)} = \frac{\text{Correct Class 1 Predictions}}{\text{All Class 1 Cases}} = \frac{TP}{TP + FN}$$

$$\text{False Negative Rate (Miss Rate)} = 1 - \text{True Positive Rate}$$

For true Class 0 cases:

$$\text{True Negative Rate (Specificity)} = \frac{\text{Correct Class 0 Predictions}}{\text{All Class 0 Cases}} = \frac{TN}{TN + FP}$$

$$\text{False Positive Rate (Fall-Out)} = 1 - \text{True Negative Rate}$$

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 another 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.

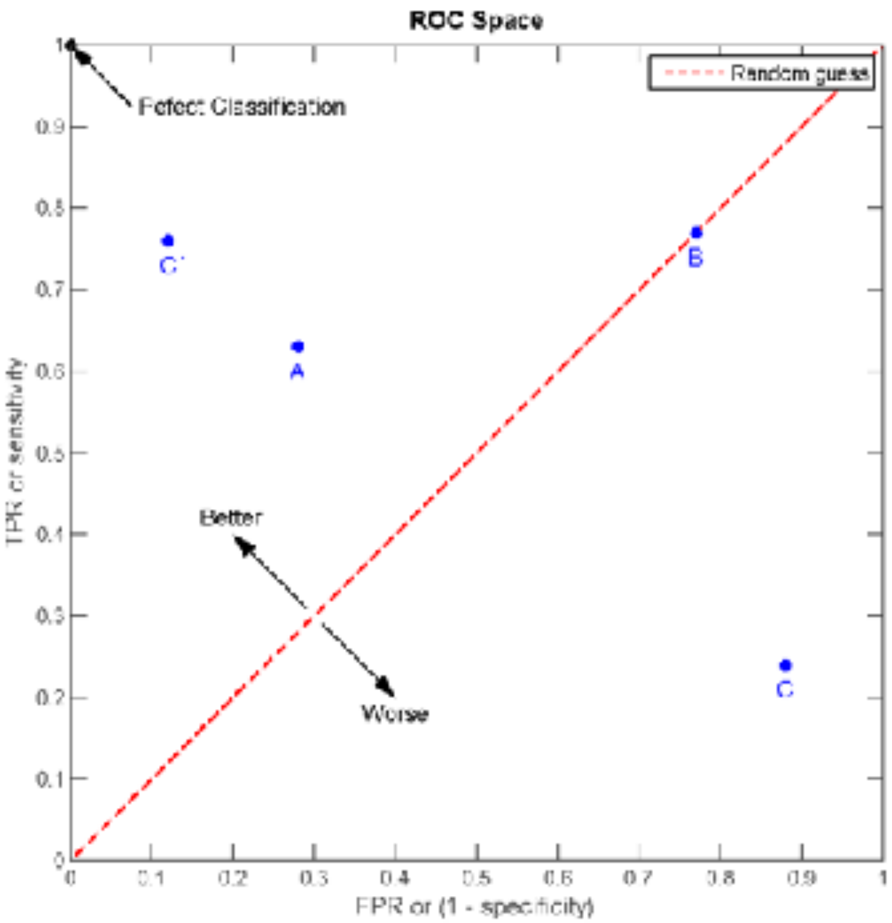
- Can you think of any ways to combine our two metrics?

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.

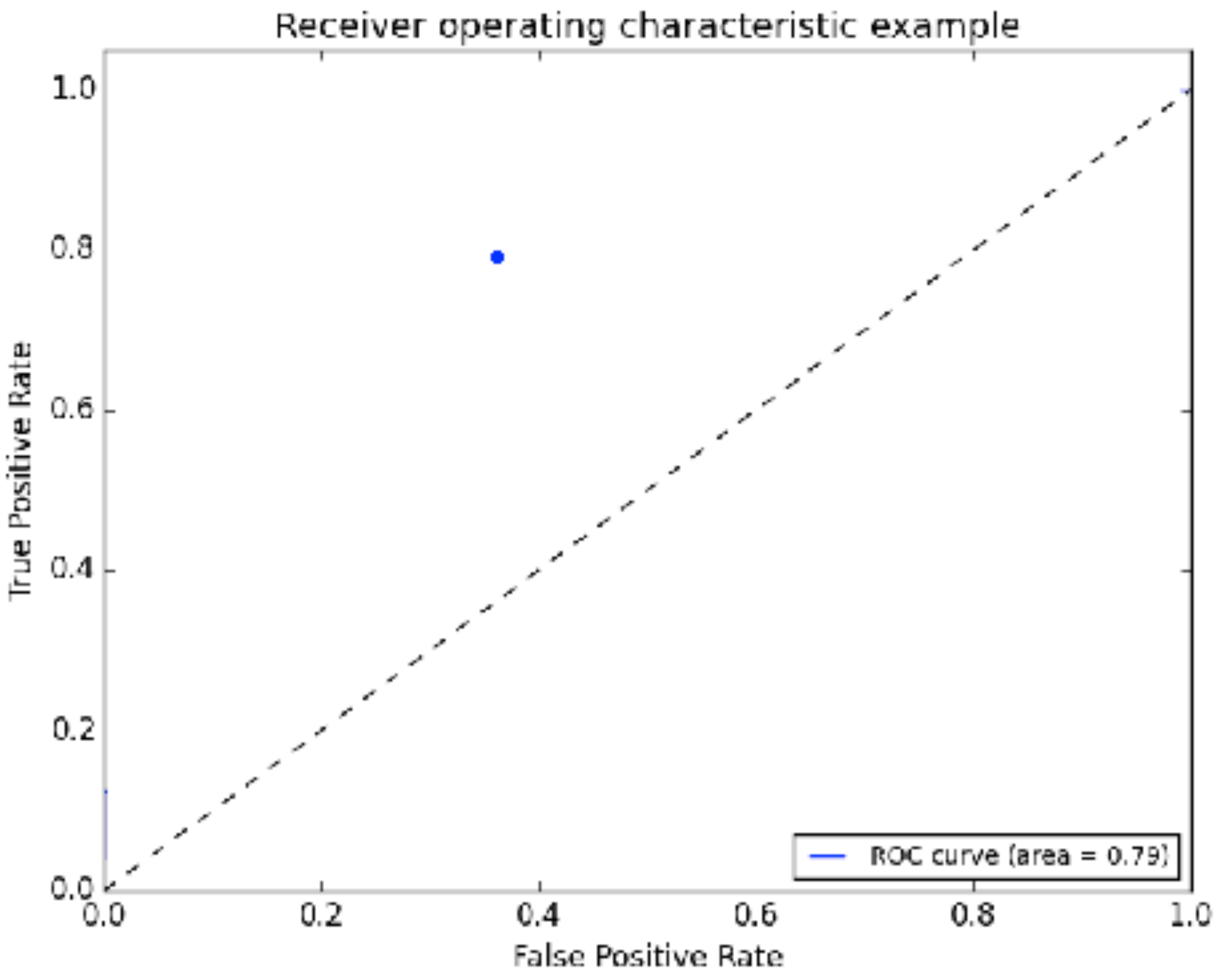
ADVANCED CLASSIFICATION METRICS

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



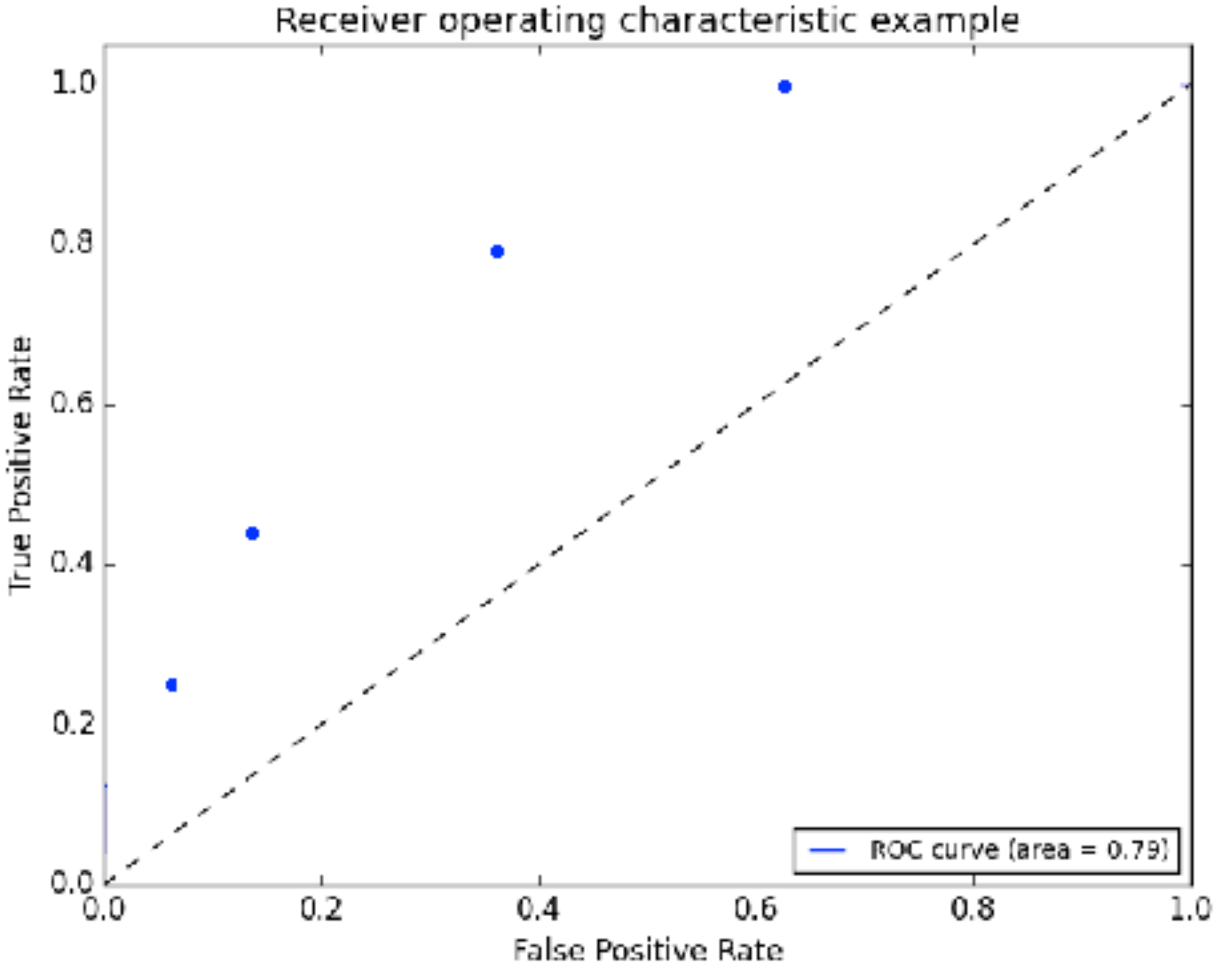
ADVANCED CLASSIFICATION METRICS

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



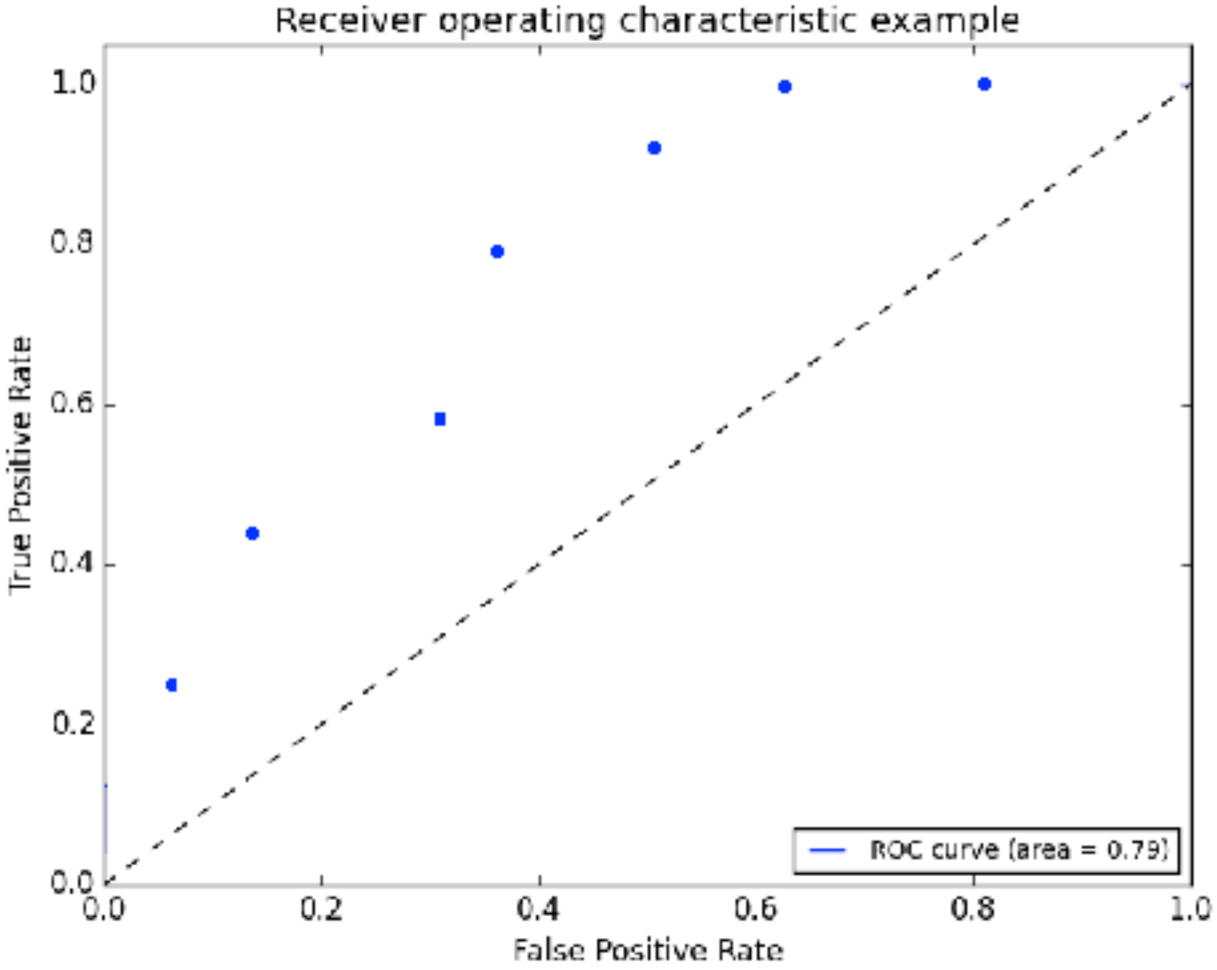
ADVANCED CLASSIFICATION METRICS

- We can continue adding pairs for different thresholds



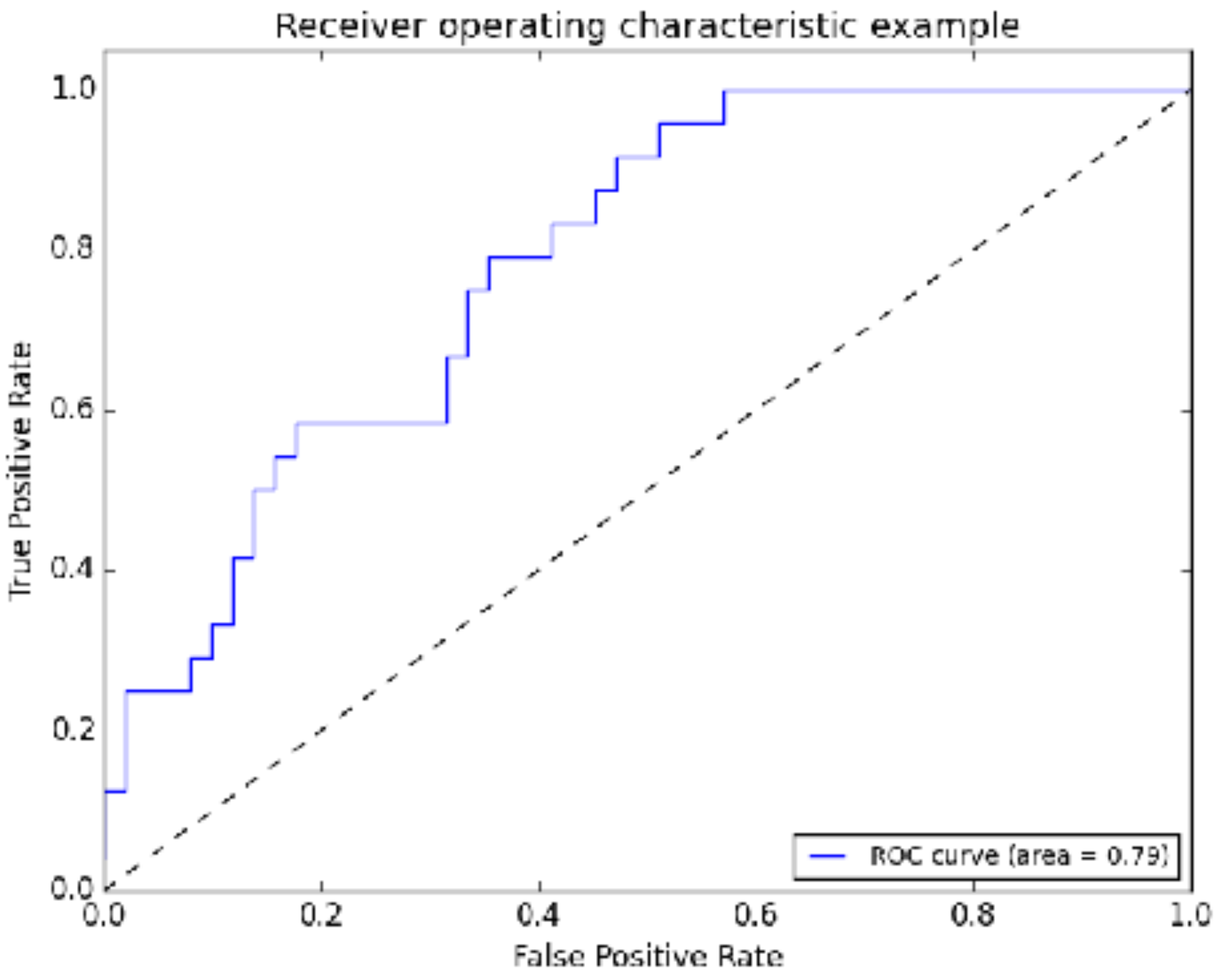
ADVANCED CLASSIFICATION METRICS

- We can continue adding pairs for different thresholds



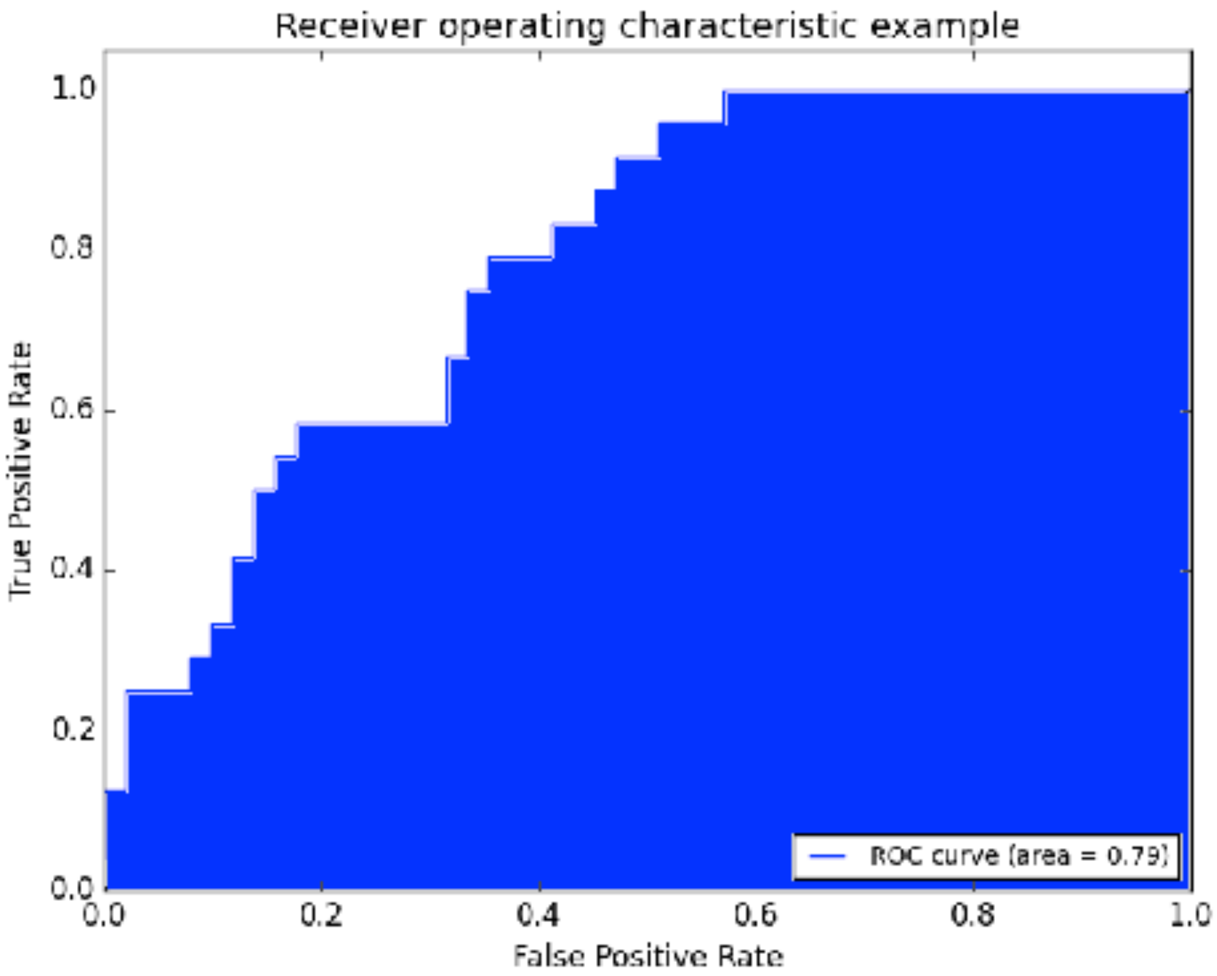
ADVANCED CLASSIFICATION METRICS

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



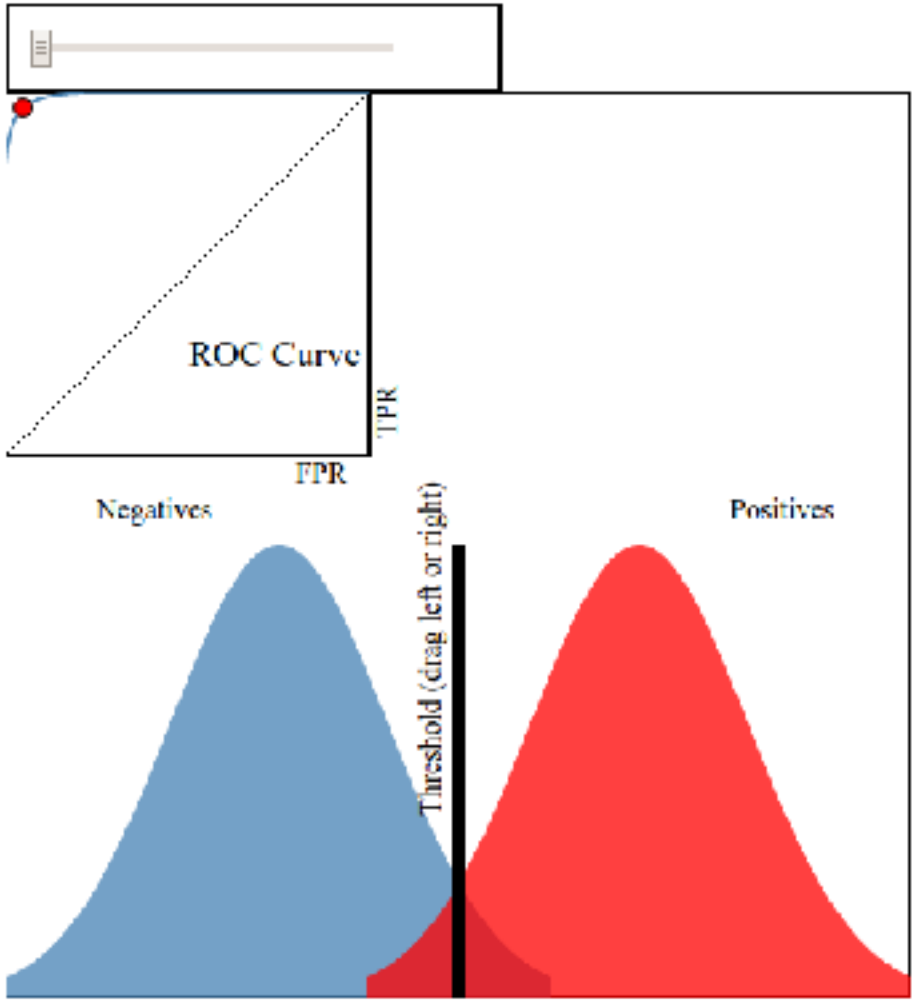
ADVANCED CLASSIFICATION METRICS

- 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?).

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.

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: 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.

BEFORE NEXT CLASS

BEFORE NEXT CLASS

DUE DATE:

- Project: Unit Project 03 – Logistic Regression

LESSON

Q & A

EXIT TICKET

DON’T FORGET TO FILL OUT YOUR EXIT TICKET