GENERAL ASSEMBLY

INTRODUCTION TO LOGISTIC REGRESSION

Abbas Chokor, Ph.D.

Staff Data Scientist, Seagate Technology

OUR PROGRESS SO FAR

UNIT 1: RESEARCH DESIGN AND EXPLORATORY DATA ANALYSIS

- ~~What is Data Science~~ ~~Lesson 1~~
- ~~Research Design and Pandas~~ ~~Lesson 2~~
- ~~Statistics Fundamentals I~~ ~~Lesson 3~~
- ~~Statistics Fundamentals II~~ ~~Lesson 4~~
- ~~Flexible Class Session~~ ~~Lesson 5~~

UNIT 2: FOUNDATIONS OF DATA MODELING

- ~~Introduction to Regression~~ ~~Lesson 6~~
- ~~Evaluating Model Fit~~ ~~Lesson 7~~
- ~~Introduction to Classification~~ ~~Lesson 8~~
- Introduction to Logistic Regression Lesson 9
- Communicating Logistic Regression Results Lesson 10
- Flexible Class Session Lesson 11

Today's Class

UNIT 3: DATA SCIENCE IN THE REAL WORLD

- Decision Trees and Random Forests Lesson 12
- Natural Language Processing Lesson 13
- Dimensionality Reduction Lesson 14
- Time Series Data I Lesson 15
- Time Series Data II Lesson 16
- Database Technologies Lesson 17
- Where to Go Next Lesson 18
- Flexible Class Session Lesson 19
- Final Project Presentations Lesson 20

LAST CLASS

WHAT DID WE LEARN?

- ✓ Define class label and classification

You got all objectives?

- ✓ Build a K-Nearest Neighbors using the scikit-learn library

- ✓ Evaluate and tune model by using metrics such as classification accuracy/error

Not all of them…

Let’s form groups of 1’s and 2’s …

LAST CLASS

ANNOUNCEMENTS

- Mid-class survey
- Moving to Rino Station starting next Thursday December 14th
- Parking is free on the streets behind the Rino Station building, and all of your key fobs should work now at Rino Station as well
- You will need to return your parking garage passes by next week.

is it really useful to have ore than one distance formula, minkowski, etc?

I am still not sure what I am doing wrong on the formula for plotting the k vs s[1] to do optimization

how are sklearn datasets structured??

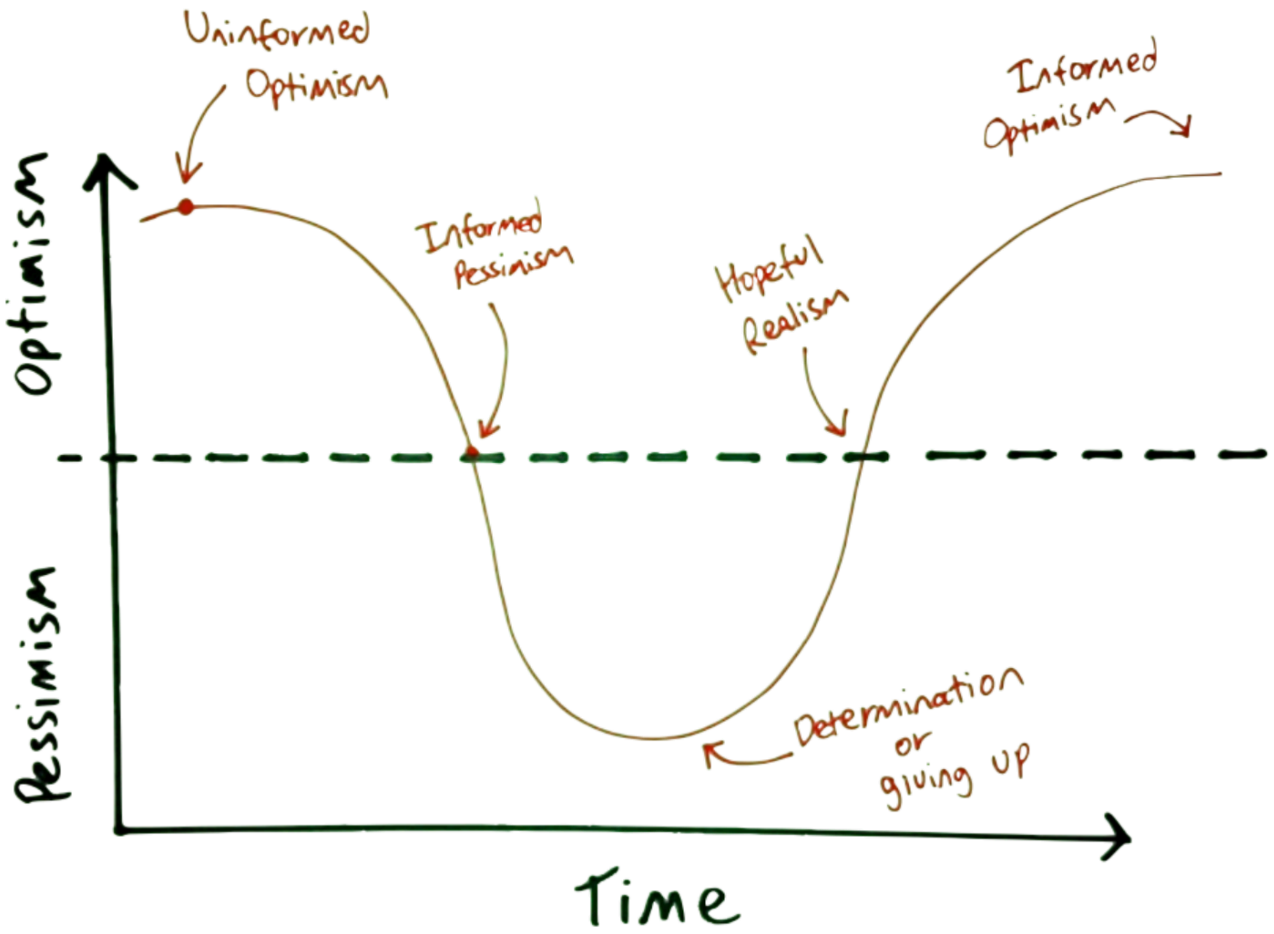
Others?

REFRESH YOUR KNOWLEDGE

BIG
IMAGE

CLASS EXPECTATIONS

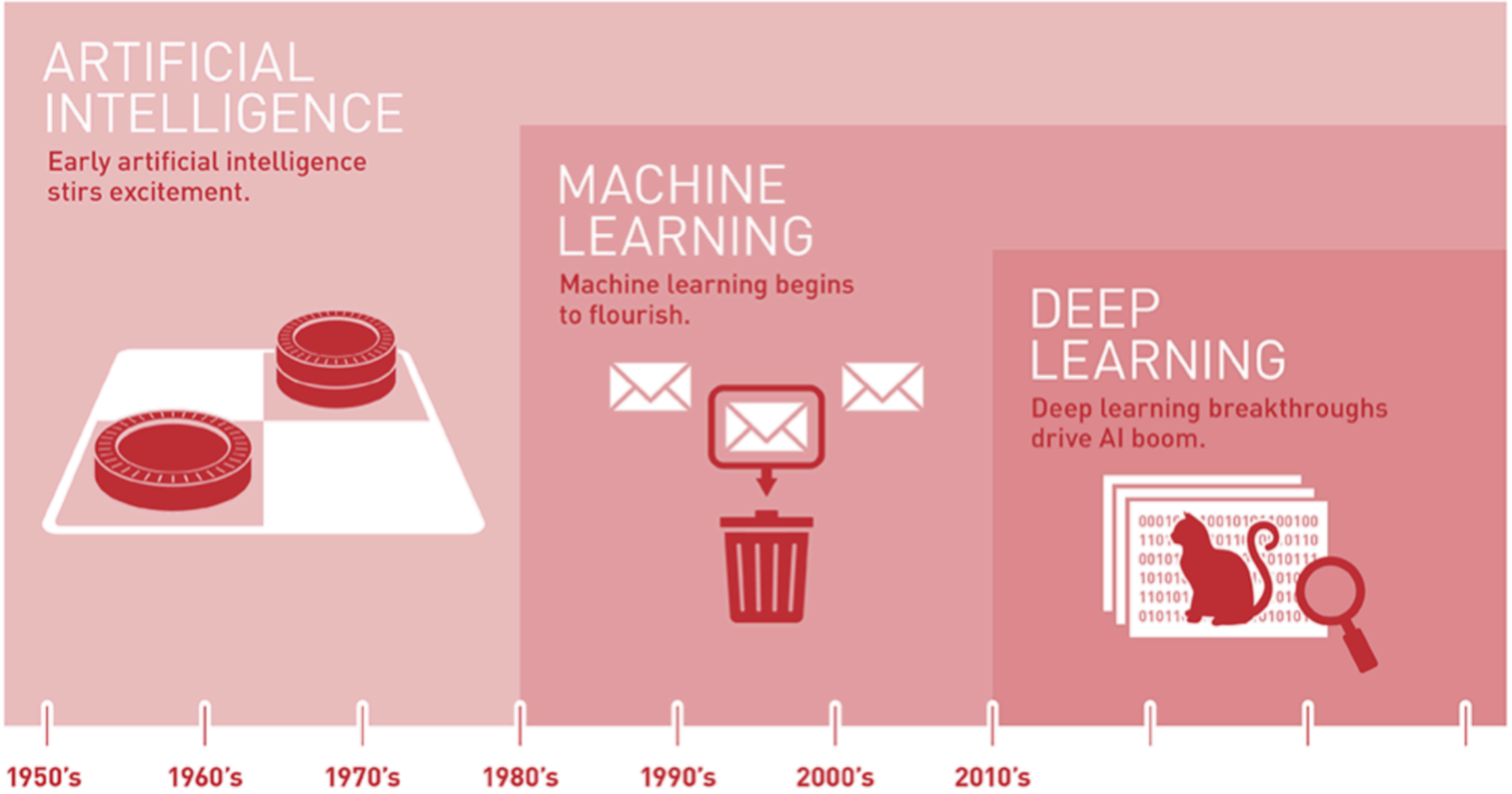
Where Are You Now?



WHAT IS MACHINE LEARNING

[Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed.

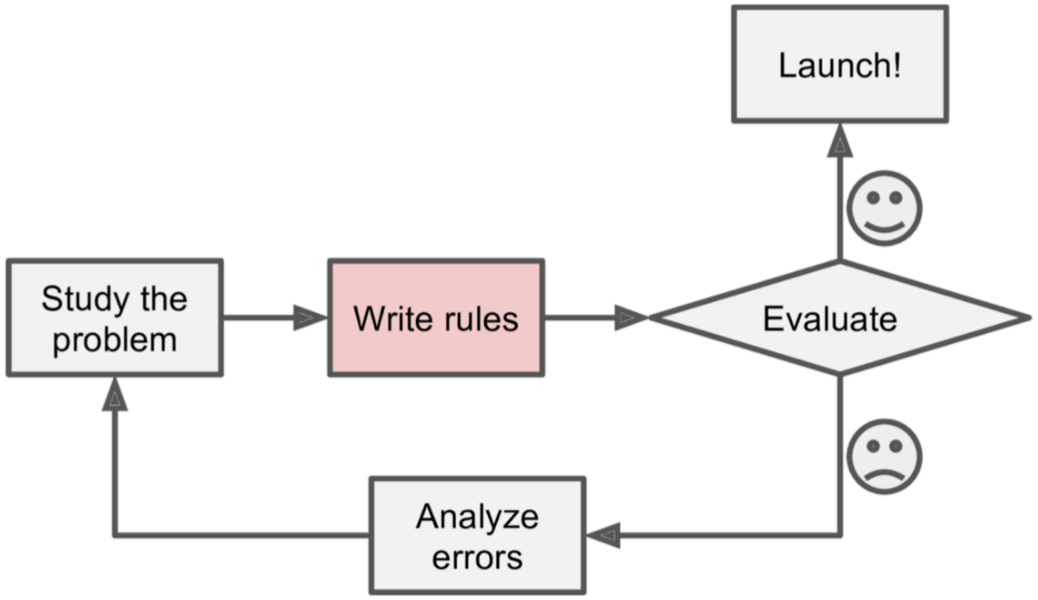
Arthur Samuel, *1959*

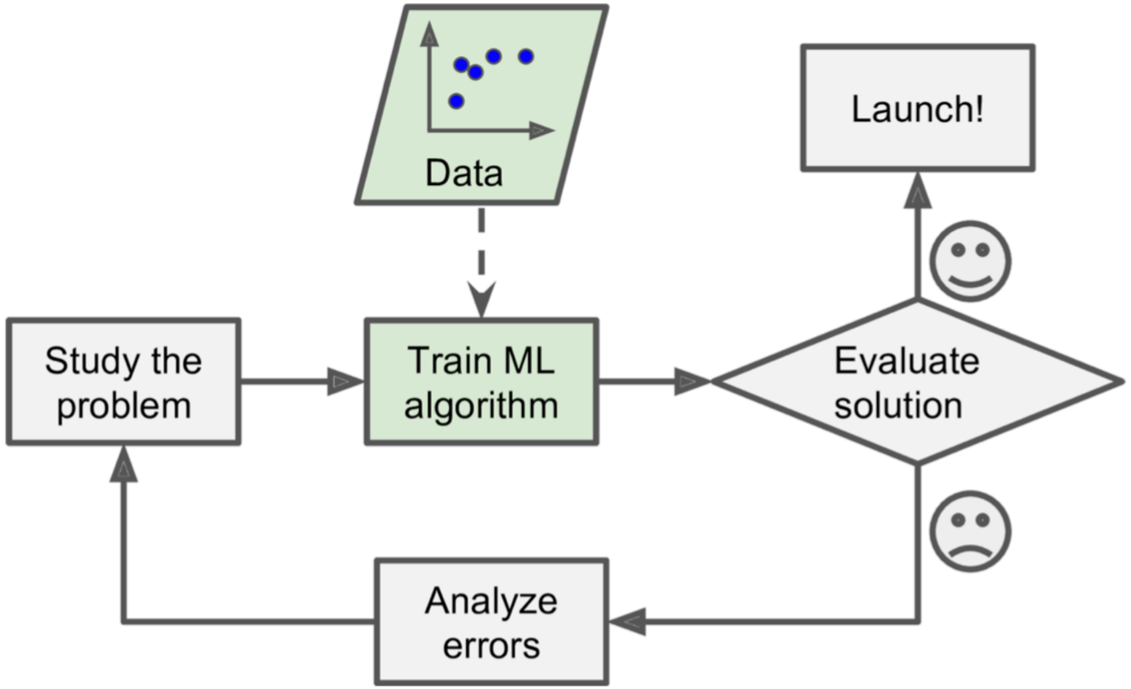


Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

<u>Machine Learning</u>: is the science (and art) of programming computers so they can *learn from data*.

WHY TO USE MACHINE LEARNING?



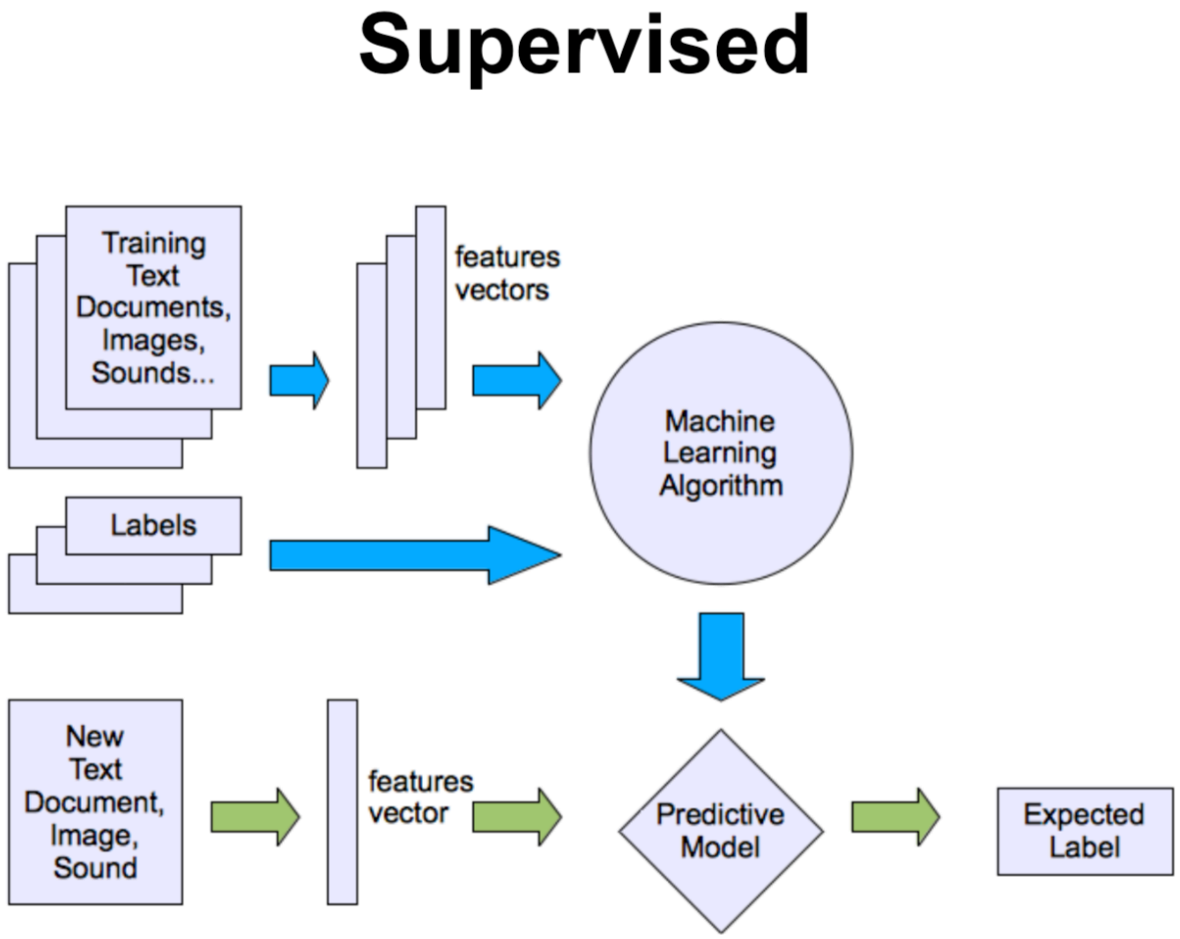


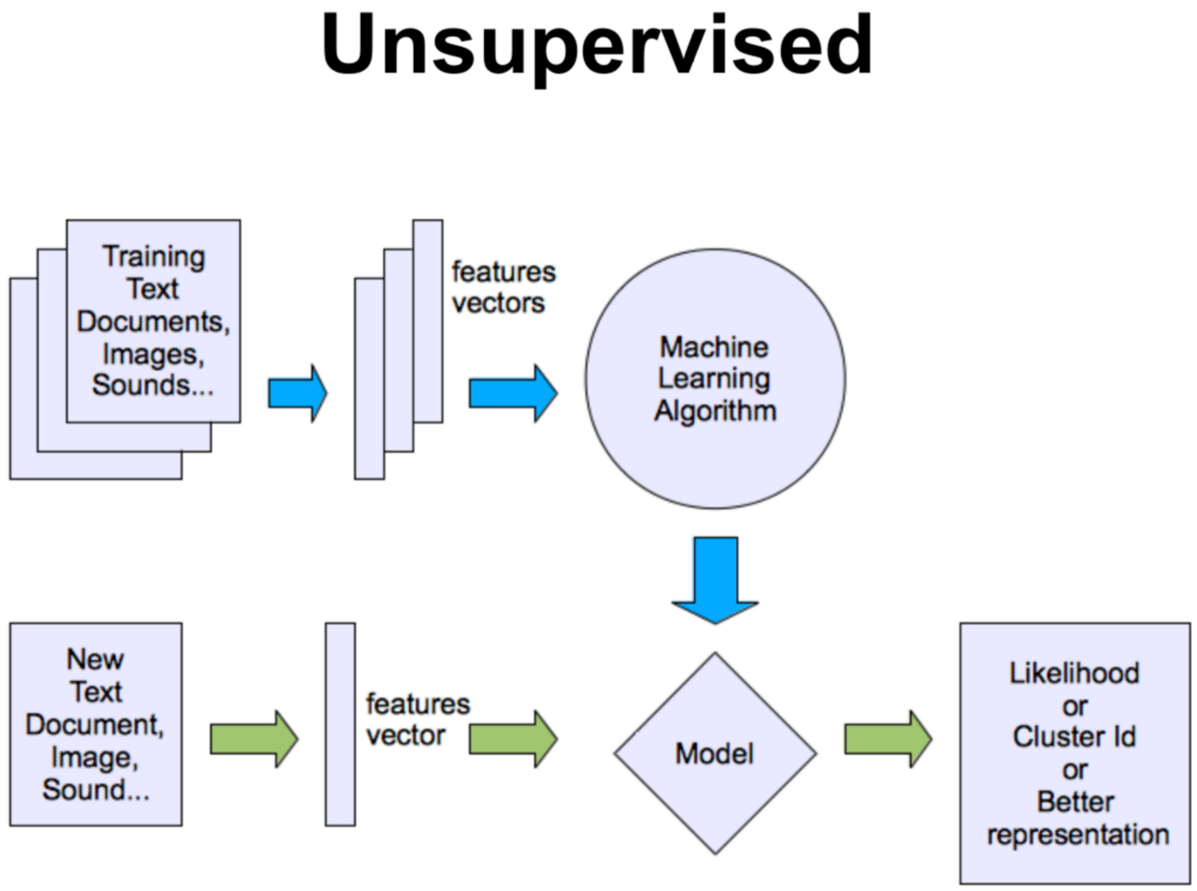
Traditional approach **Vs.** ***Machine Learning approach***

Did you know?

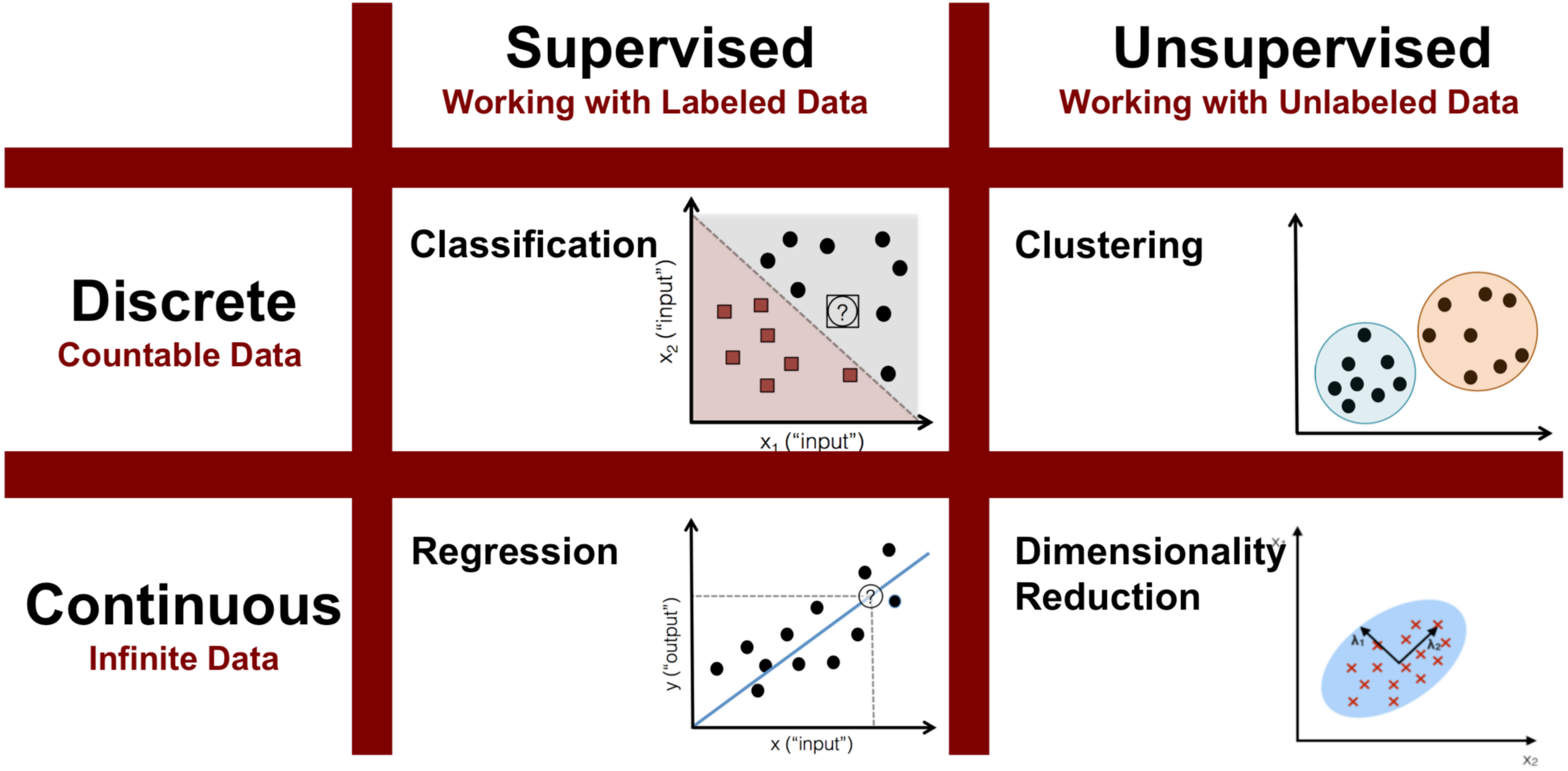
Machine learning algorithms are expected to replace 25% of the jobs across the world, in the next 10 years.

TYPES OF MACHINE LEARNING

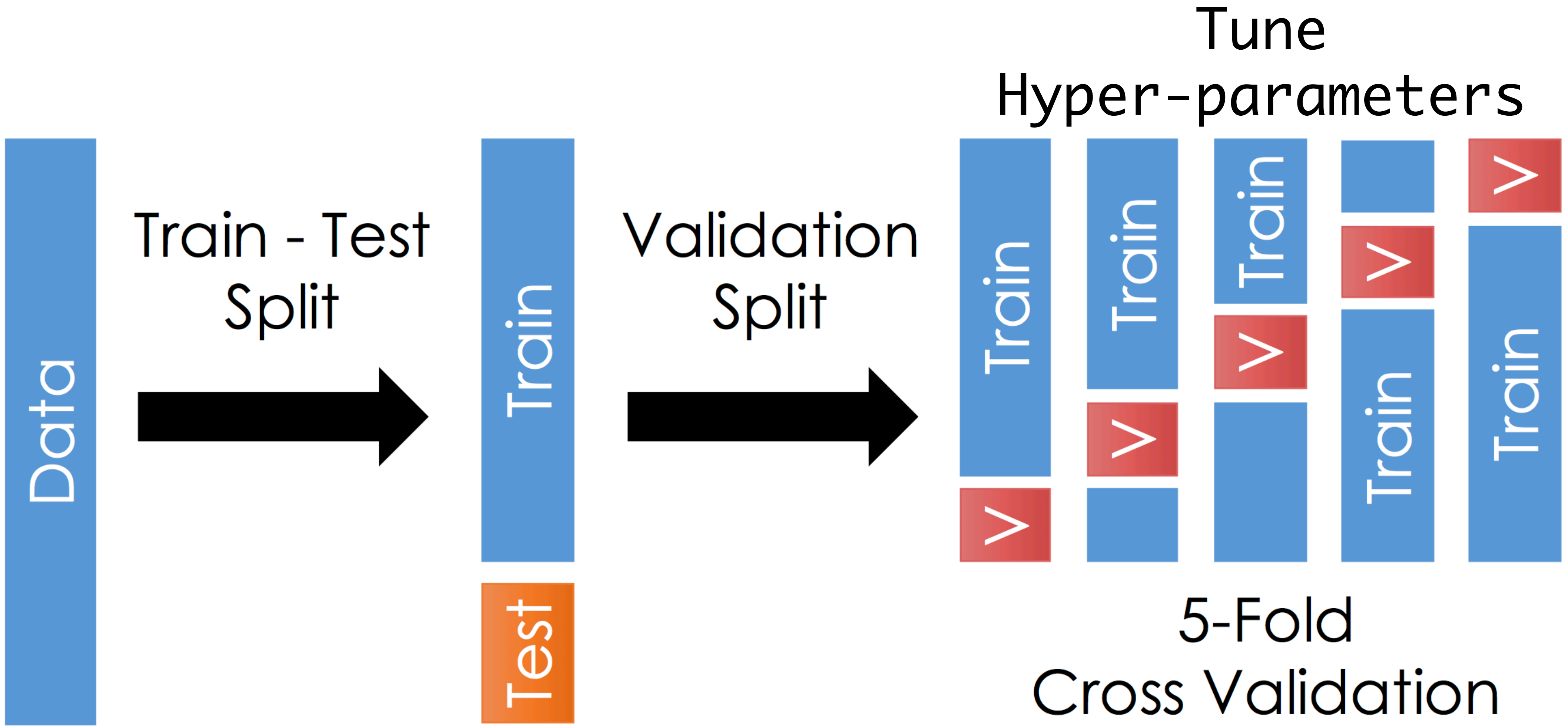




TYPES OF MACHINE LEARNING



TRAINING – VALIDATING - TESTING



COURSE

PRE-WORK

PRE-WORK REVIEW

- Did you do complete the Regression Quiz?
- What R^2 and MSE you got?

LEARNING OBJECTIVES

- Build a Logistic regression classification model using the statsmodels 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

INTRODUCTION

LOGISTIC REGRESSION

LOGISTIC REGRESSION

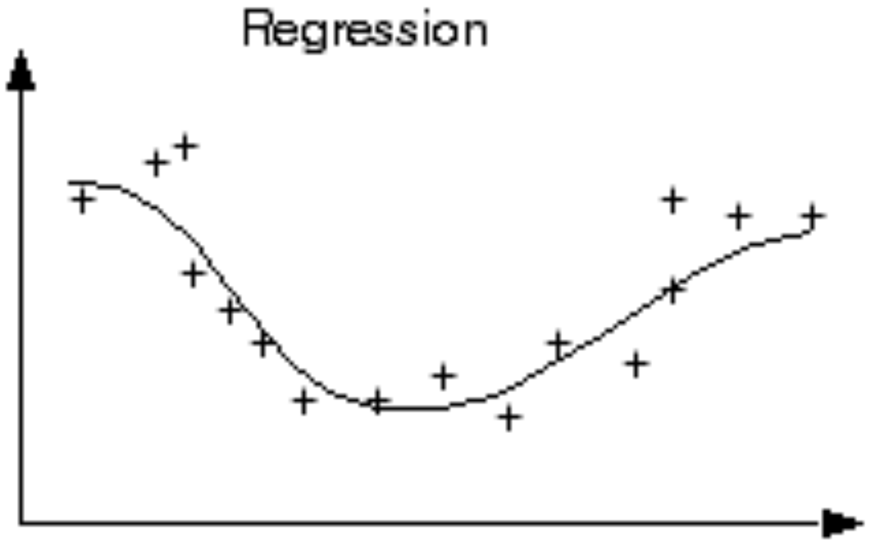
- Logistic regression is a *linear* approach to solving a *classification* problem.

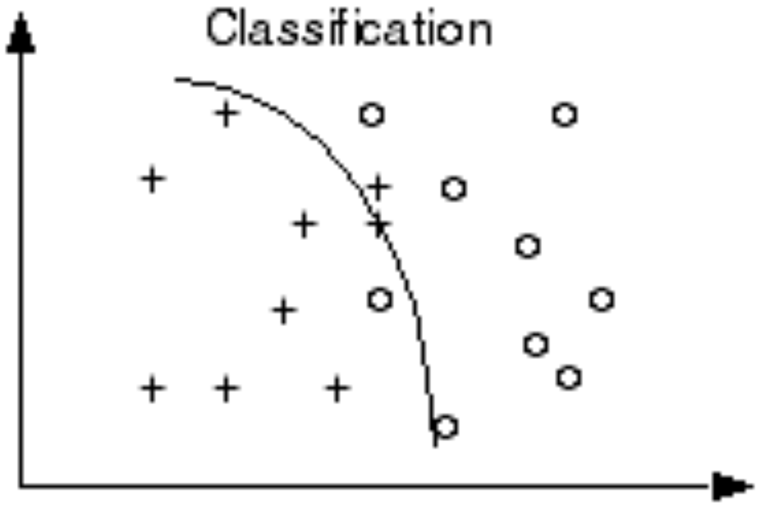
- That is, we can use a linear model, similar to Linear regression, in order to solve if an item *belongs* or *does not belong* to a class label.

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.



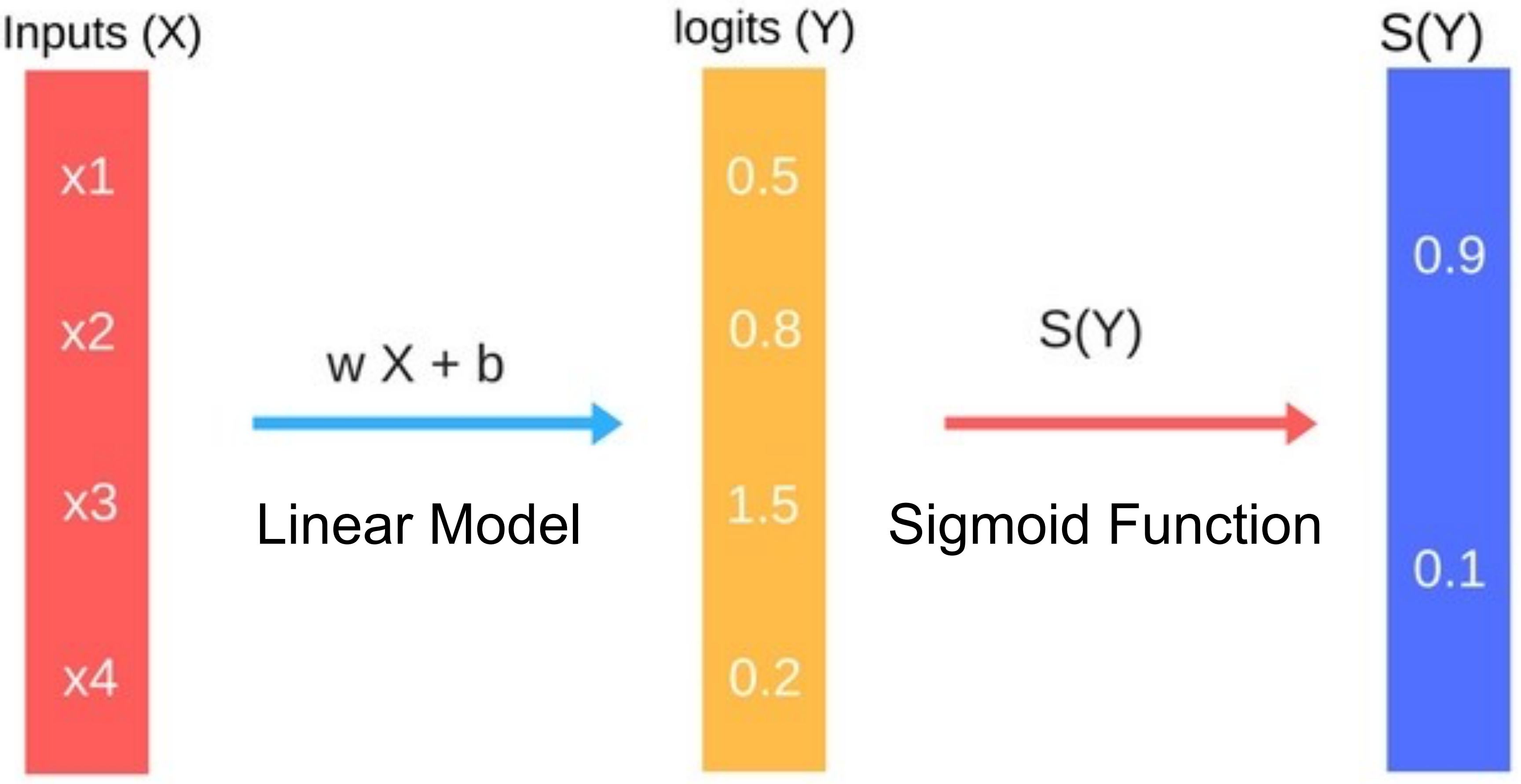


CHALLENGE! LINEAR REGRESSION RESULTS FOR CLASSIFICATION

- 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?

TO SUMMARIZE LOGISTIC REGRESSION



$$y = \ln\left(\frac{p}{1-p}\right) = b_0 + b_1 \bullet x$$

$$p = \frac{1}{1 + e^{-y}}$$

p / (1-p) represents the *odds*
ln (p / (1-p)) represents the *log of odds*

FIX 1: PROBABILITY

- One approach is predicting the probability that an observation belongs to a certain class.

- We could assume the *prior probability* 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.

FIX 2: LINK FUNCTIONS AND THE SIGMOID FUNCTION

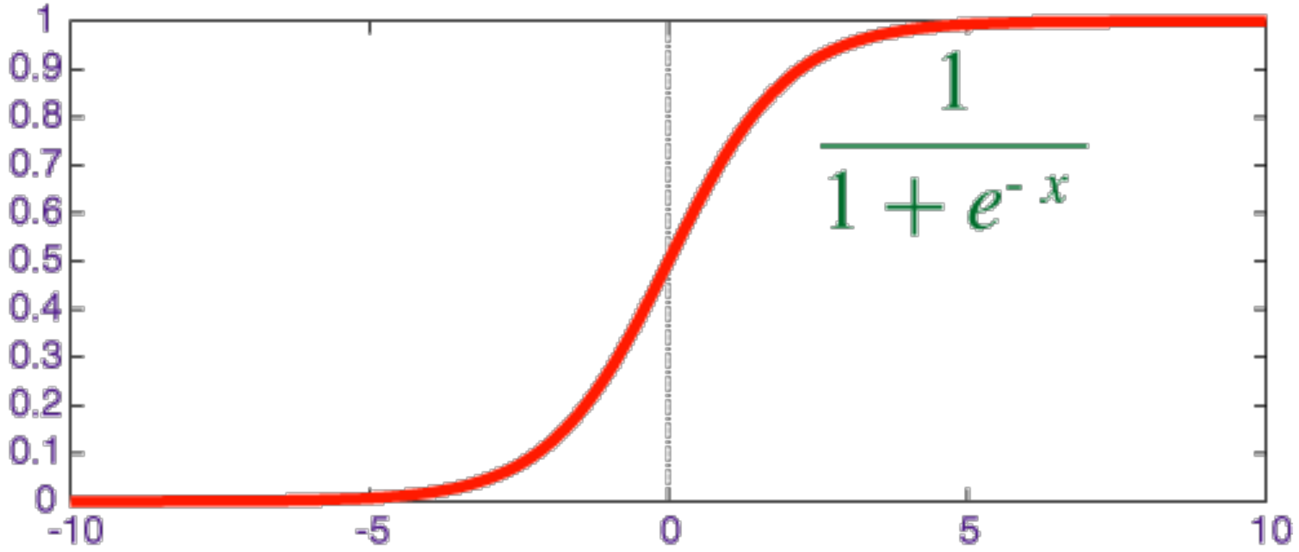
- Another advantage to OLS is that it allows for *generalized* models using a *link function*.

- 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

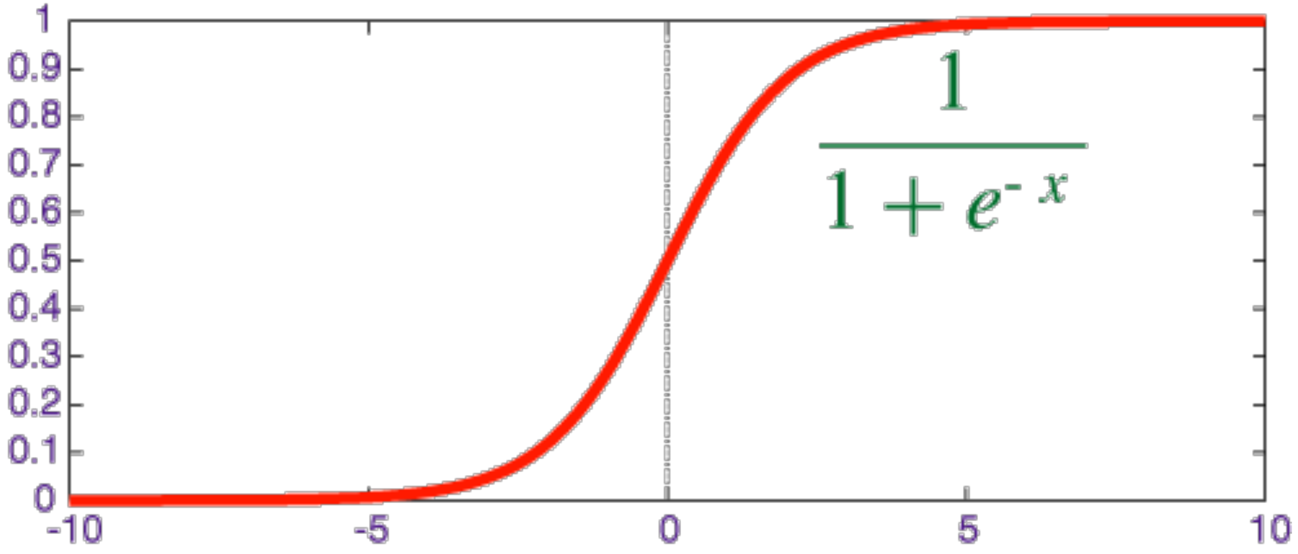
- A *sigmoid function* is a function that visually looks like an s.



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

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 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 ODDS, LOG ODDS, AND SIGMOID FUNCTIONS

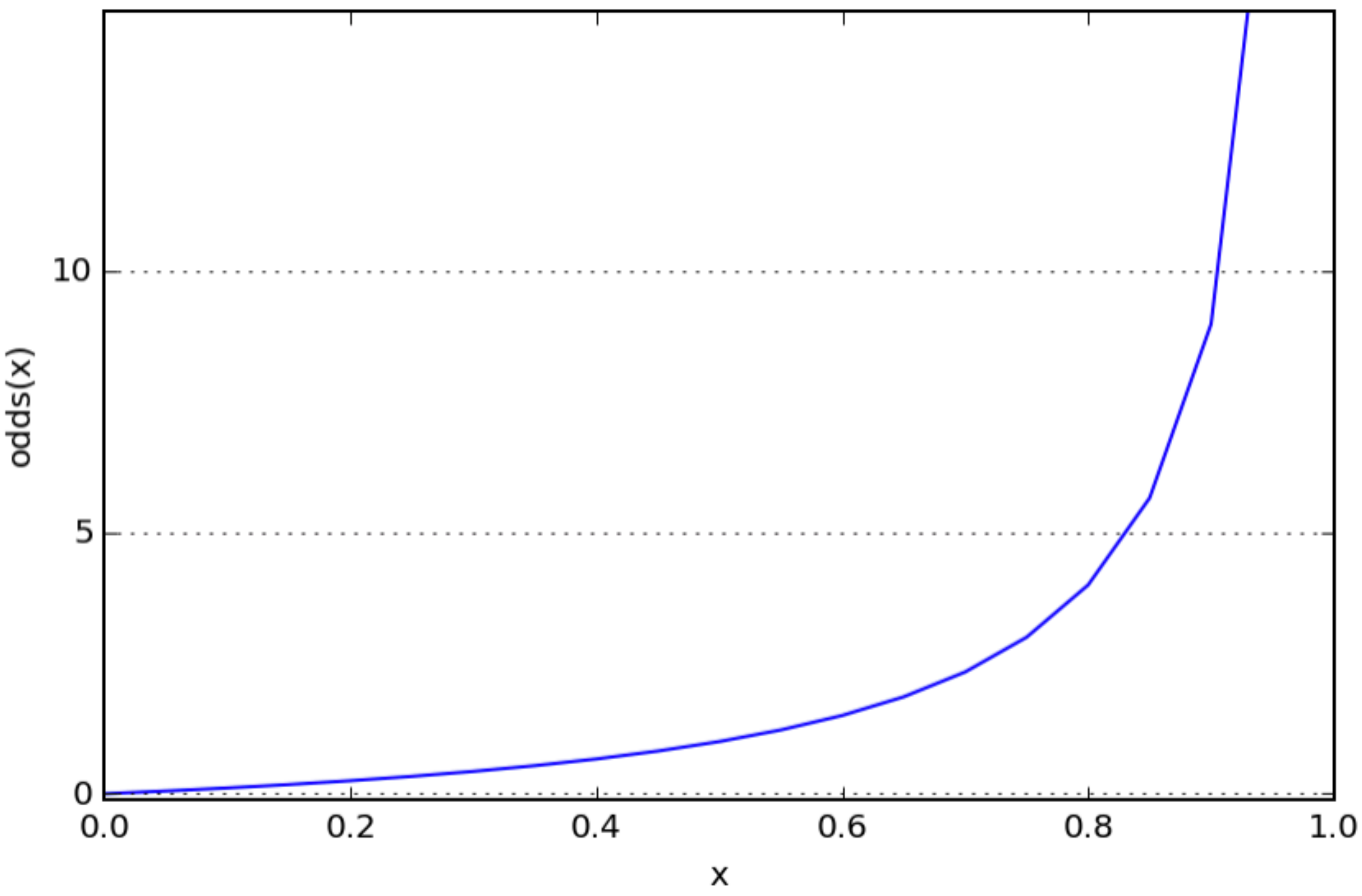
PLOTTING A SIGMOID FUNCTION

- Open the “basics” in lesson 9 - code
- Recall that e = 2.71.
- Do we get an the “S” shape we expect?

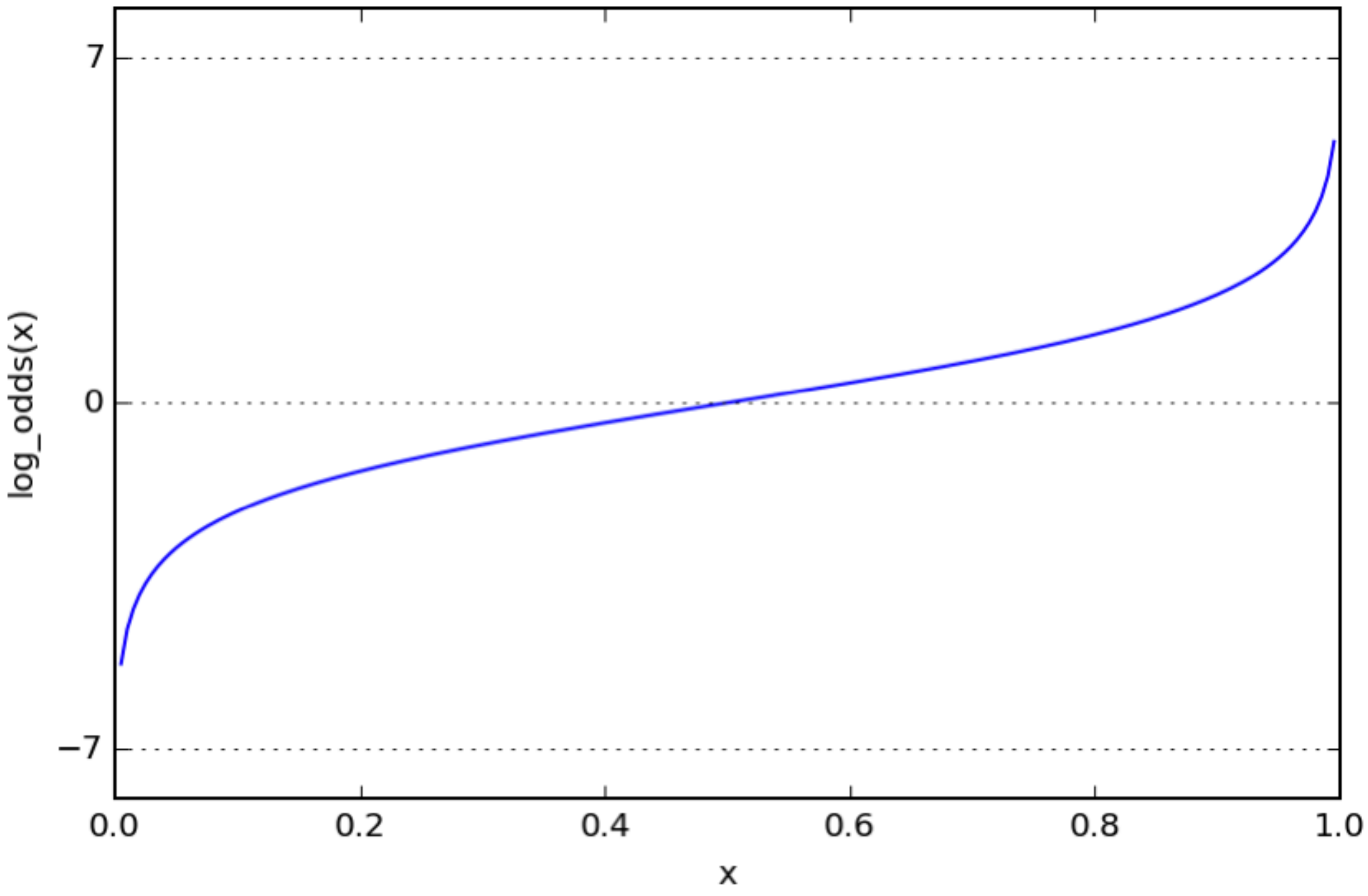
INTRODUCTION

LOGISTIC REGRESSION

FIX 3: ODDS AND LOG-ODDS

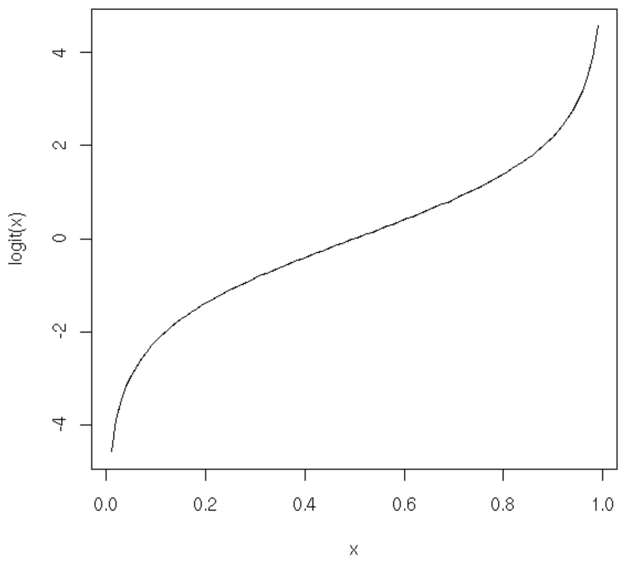


FIX 3: ODDS AND LOG-ODDS

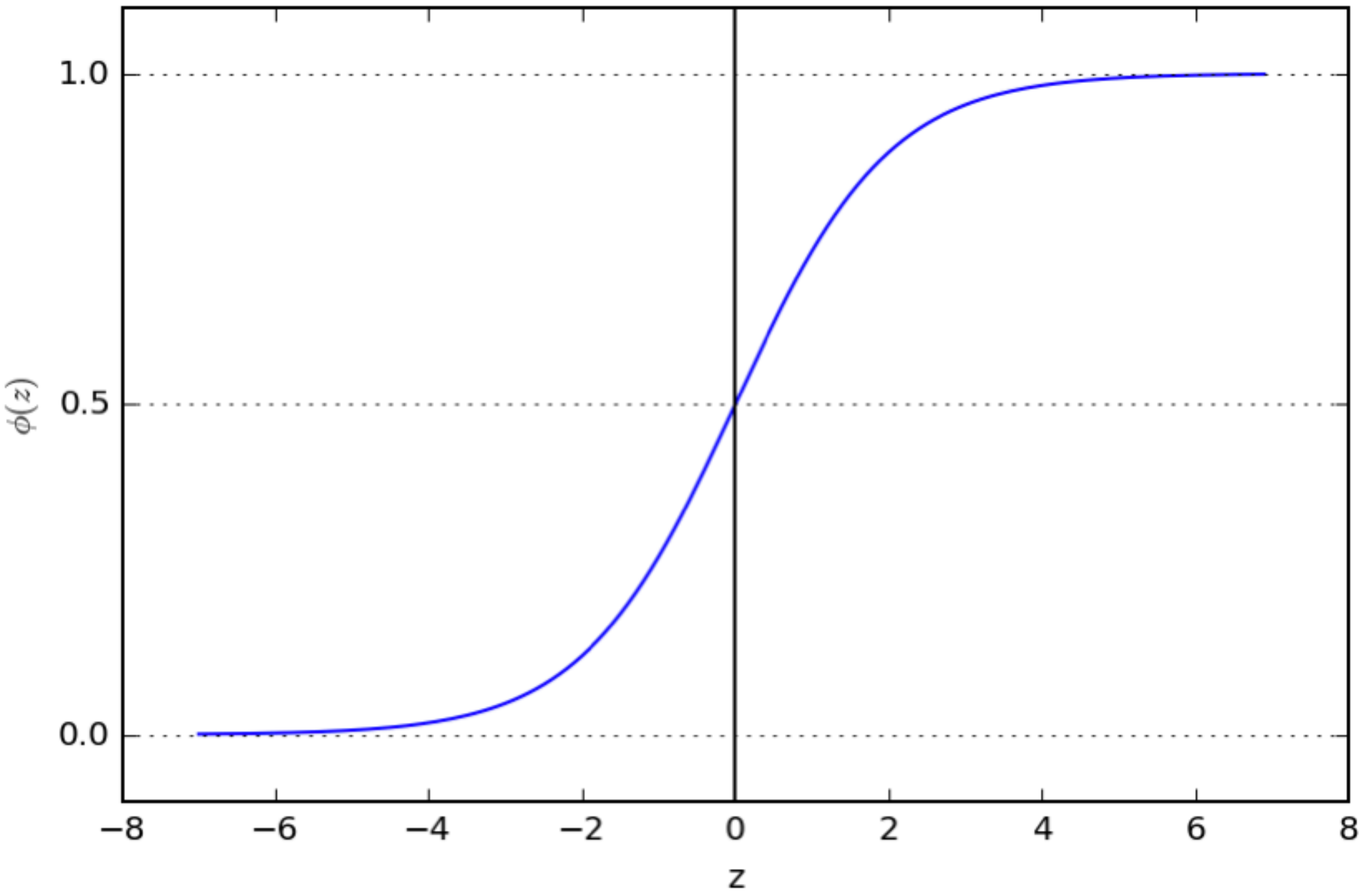


FIX 3: ODDS AND LOG-ODDS

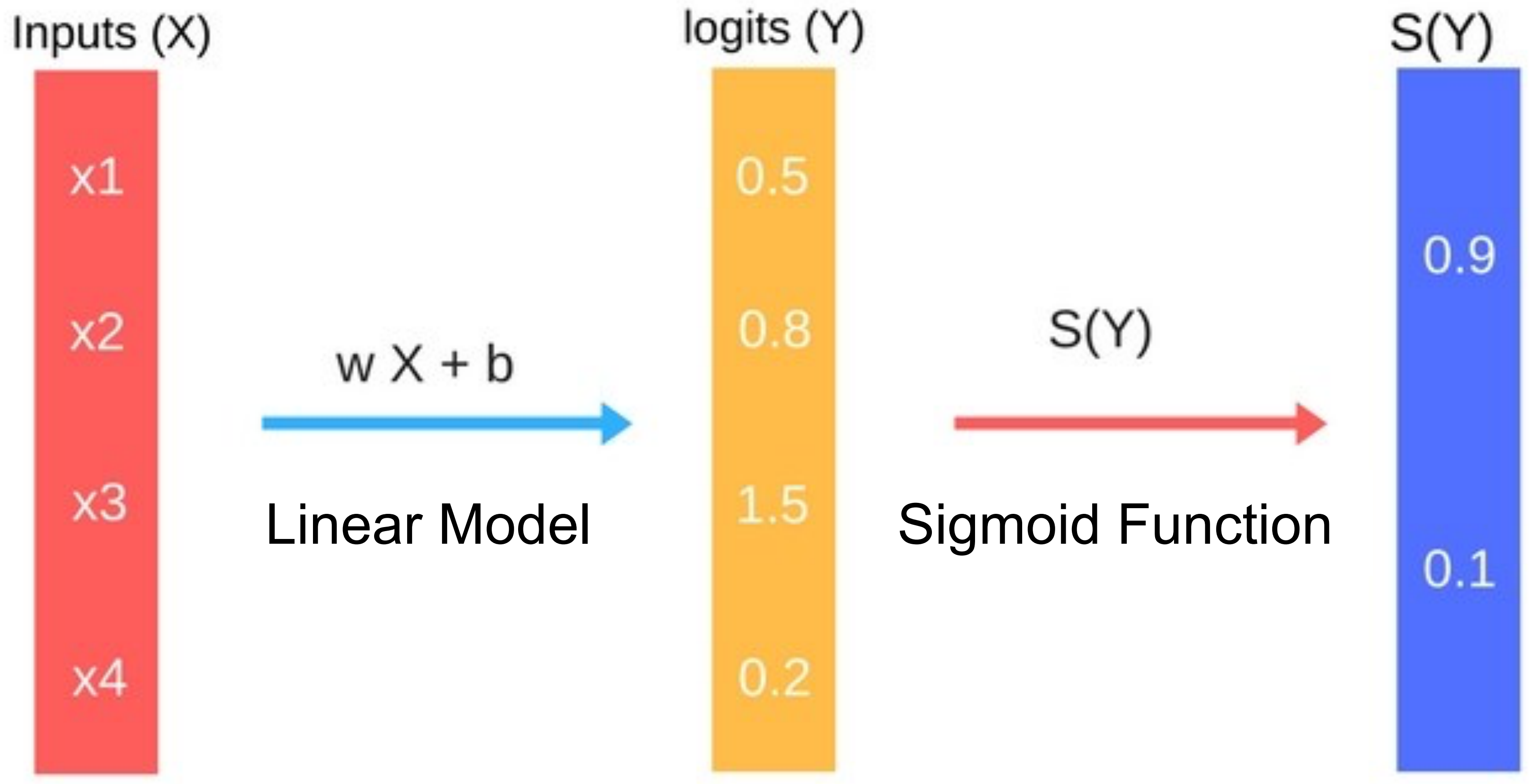
- The *logit* function is the inverse of the *sigmoid* function.
- This will act as our *link* function for logistic regression.
- Mathematically, the logit function is defined as $Ln\left(\dfrac{P}{1-P}\right)$



FIX 3: ODDS AND LOG-ODDS



TO SUMMARIZE LOGISTIC REGRESSION



$$\ln\left(\frac{p}{1-p}\right)=b_0+b_1\bullet x \qquad p=\frac{1}{1+e^{-y}}$$

INDEPENDENT PRACTICE

LOGISTIC REGRESSION IMPLEMENTATION

ACTIVITY: LOGISTIC REGRESSION IMPLEMENTATION

EXERCISE

DIRECTIONS (Part A of Starter Code)

Use the data `collegeadmissions.csv` and to predict the target variable `admit`.

1. Build a simple model with one feature and explore the `coef_` value. Does this represent the odds or logit (log odds)?

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*.

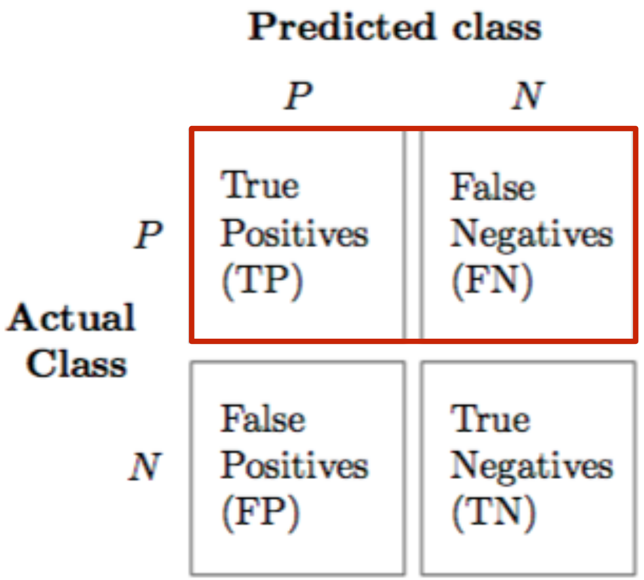
- For each label, we can put it into the category of a true positive, false positive, true negative, or false negative.

		Predicted class	
		P	N
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (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?



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

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?

		Predicted class	
		P	*N*
Actual Class	*P*	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

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

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.

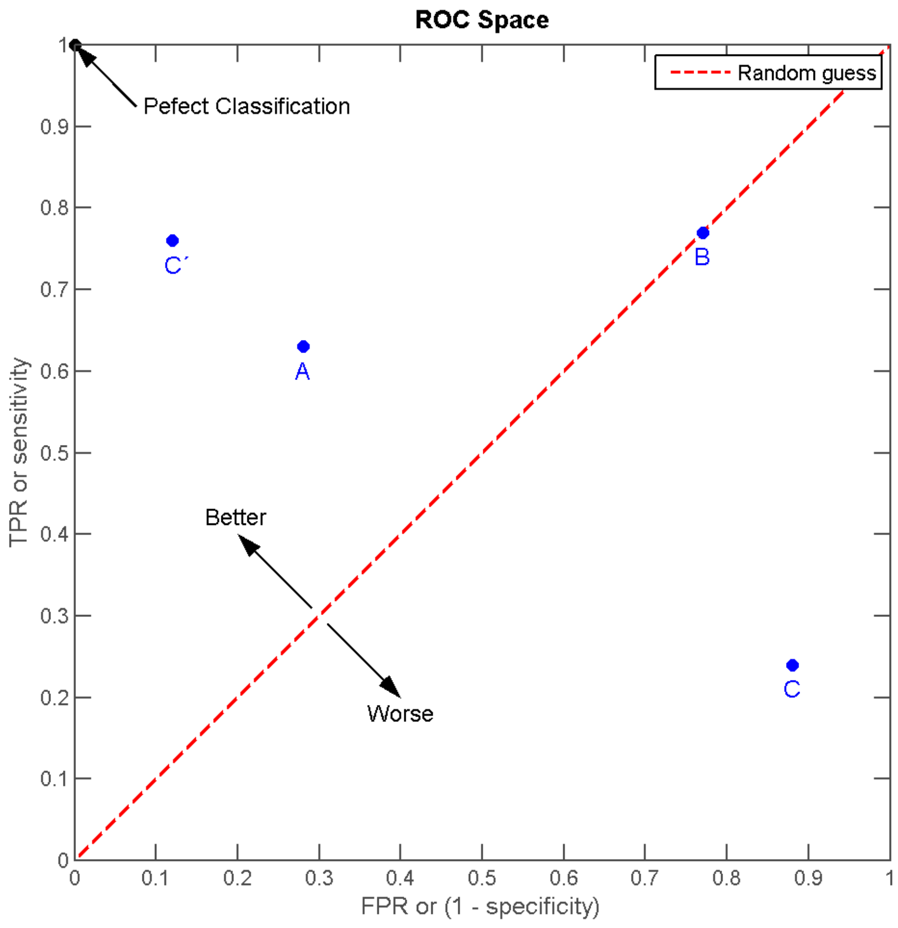
- Can you think of any ways to combine our two metrics?

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.

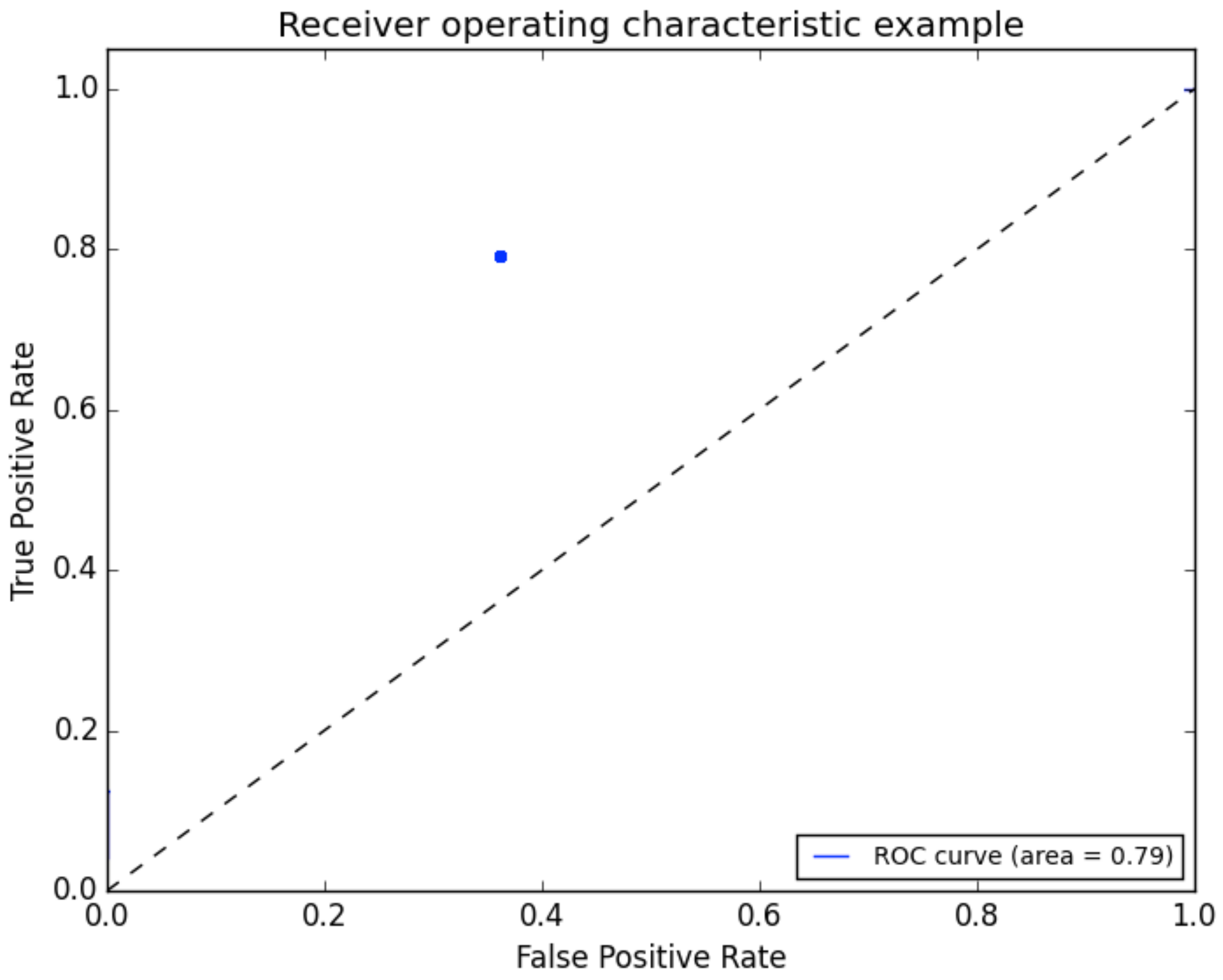
ADVANCED CLASSIFICATION METRICS

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



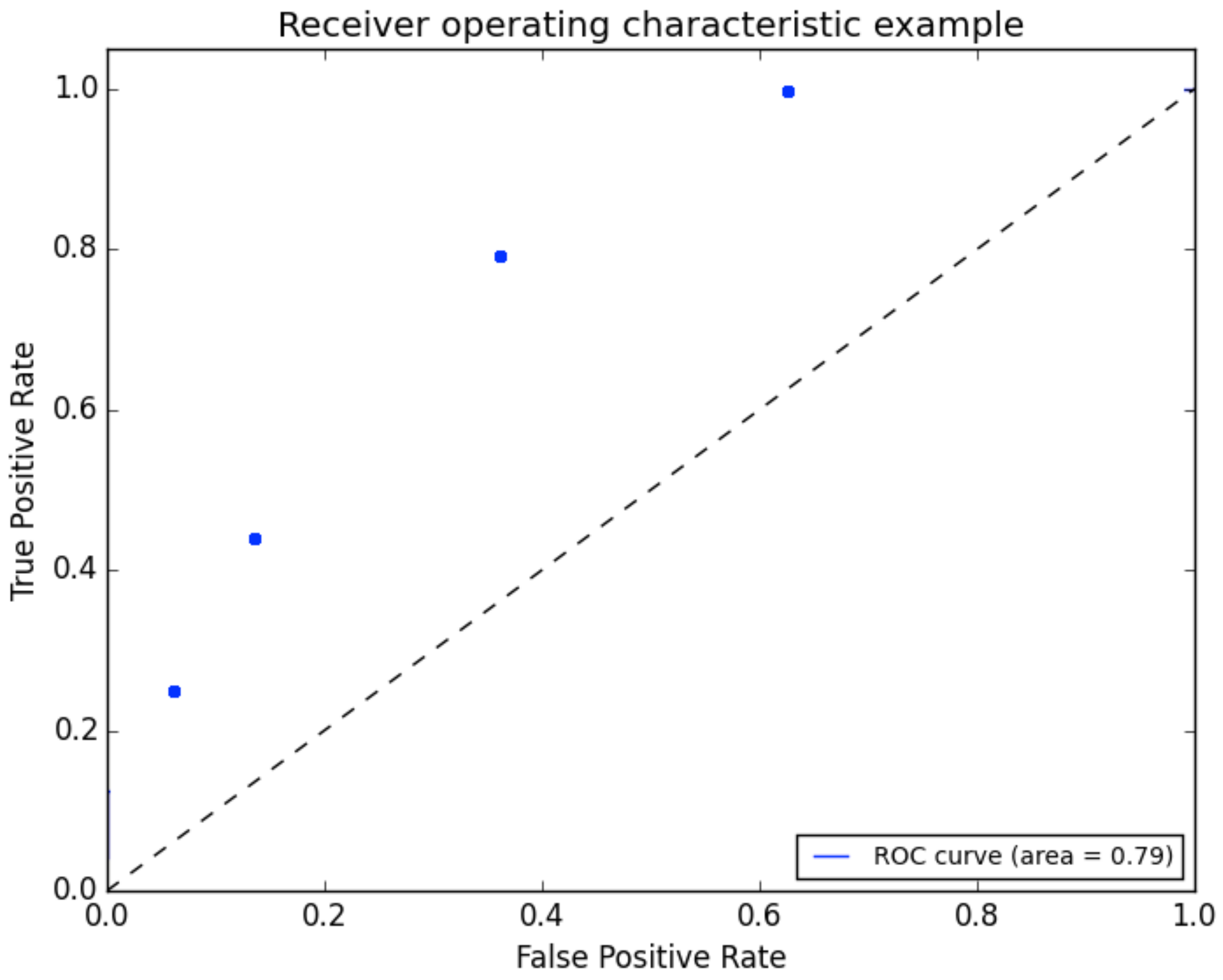
ADVANCED CLASSIFICATION METRICS

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



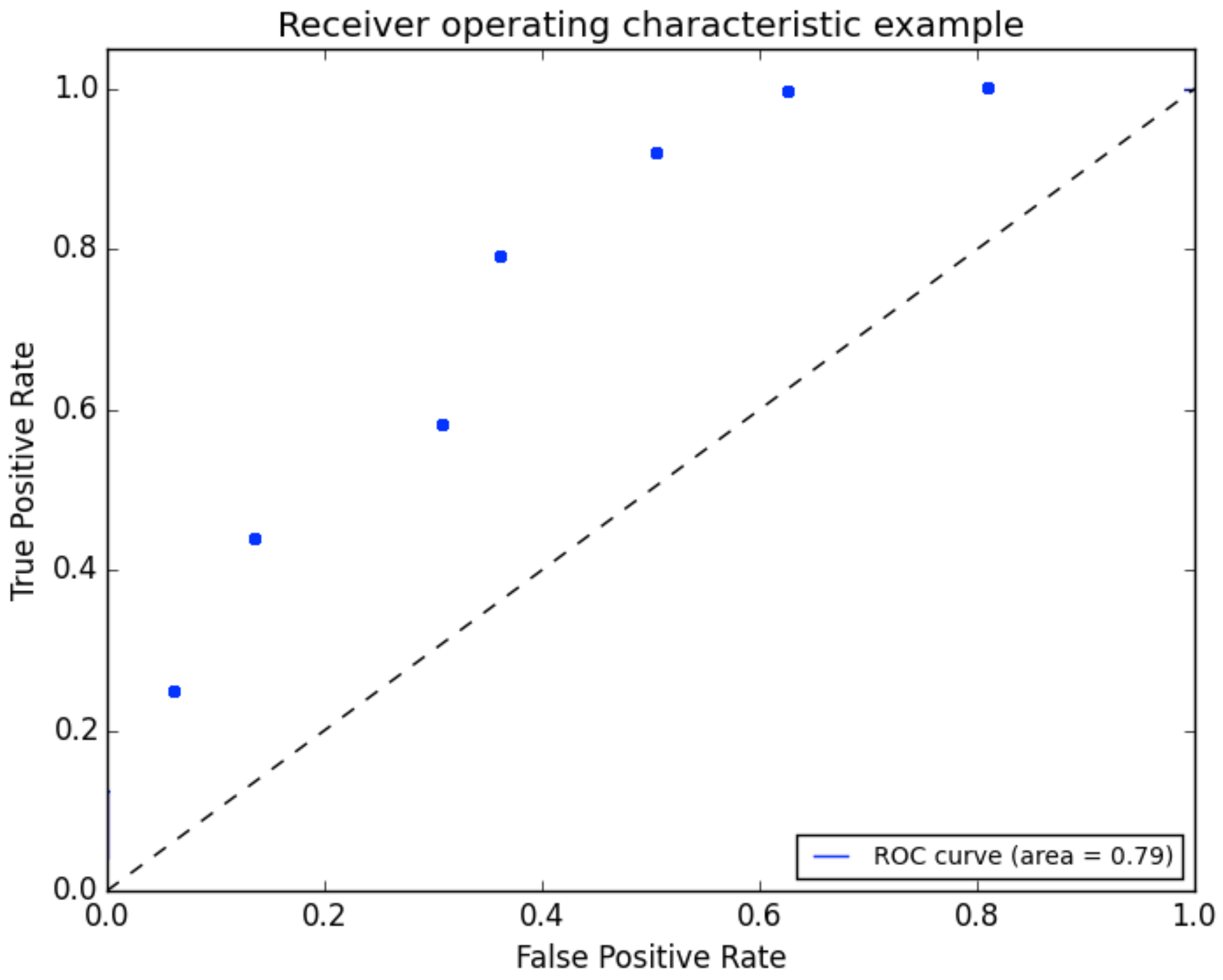
ADVANCED CLASSIFICATION METRICS

- We can continue adding pairs for different thresholds



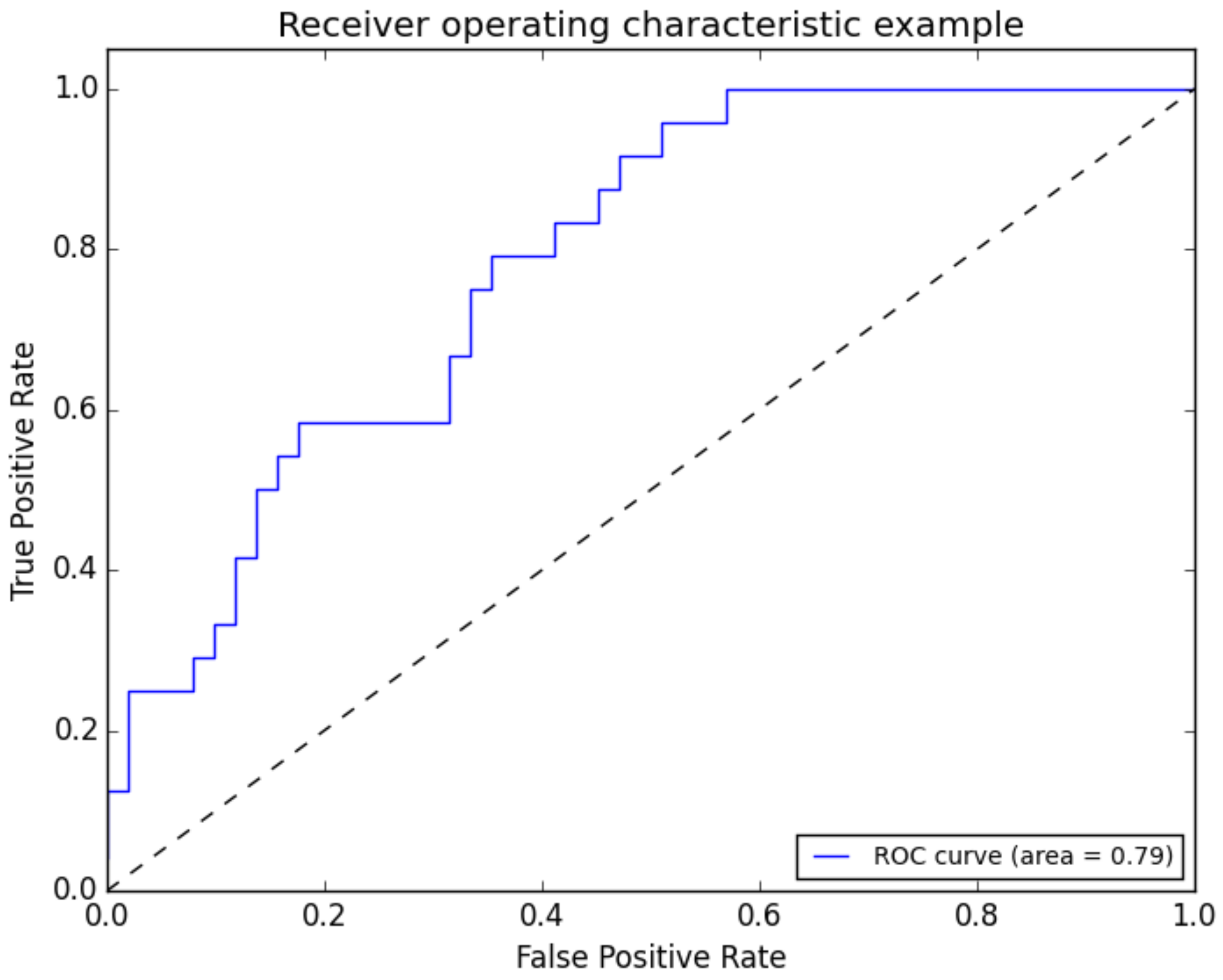
ADVANCED CLASSIFICATION METRICS

- We can continue adding pairs for different thresholds



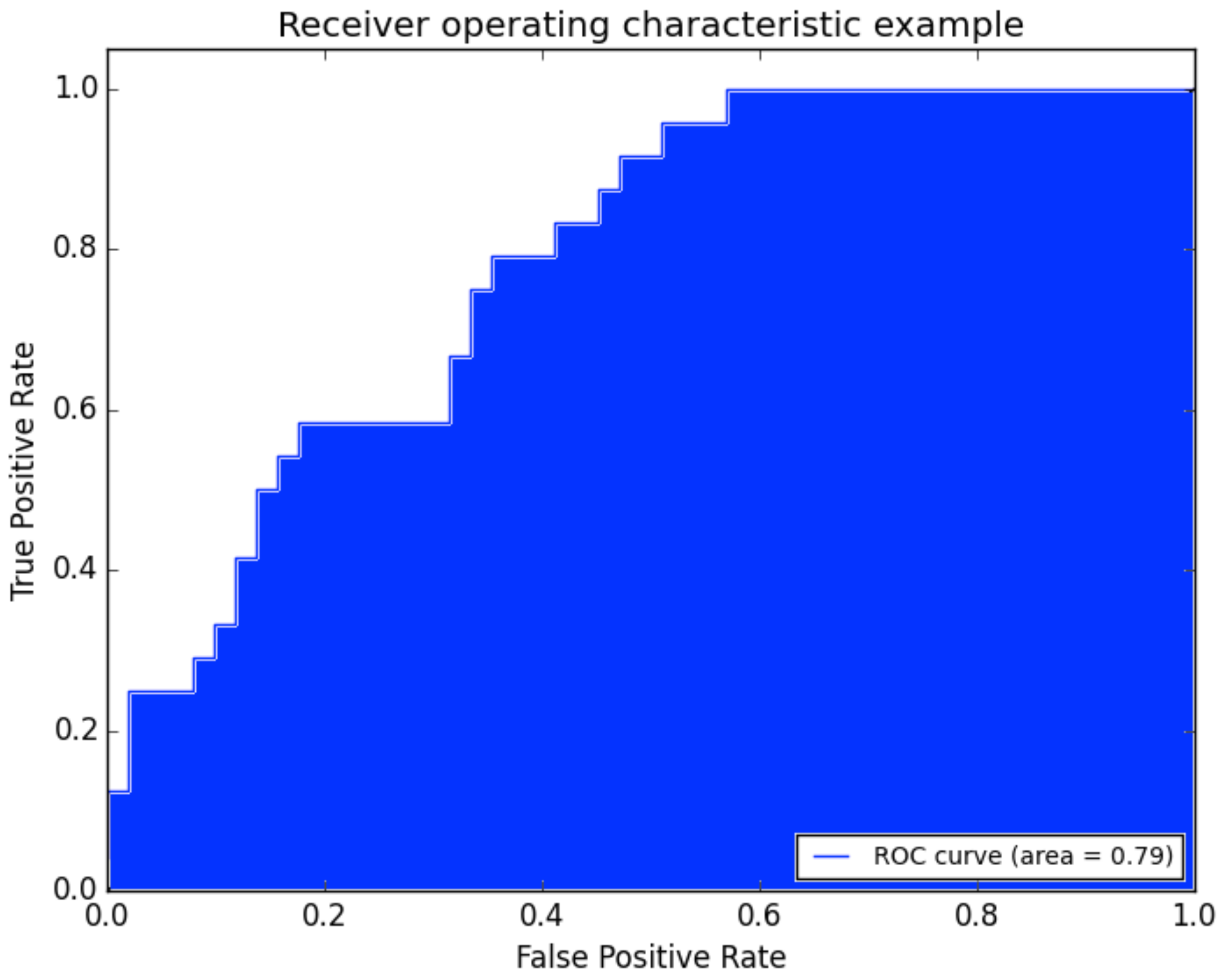
ADVANCED CLASSIFICATION METRICS

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



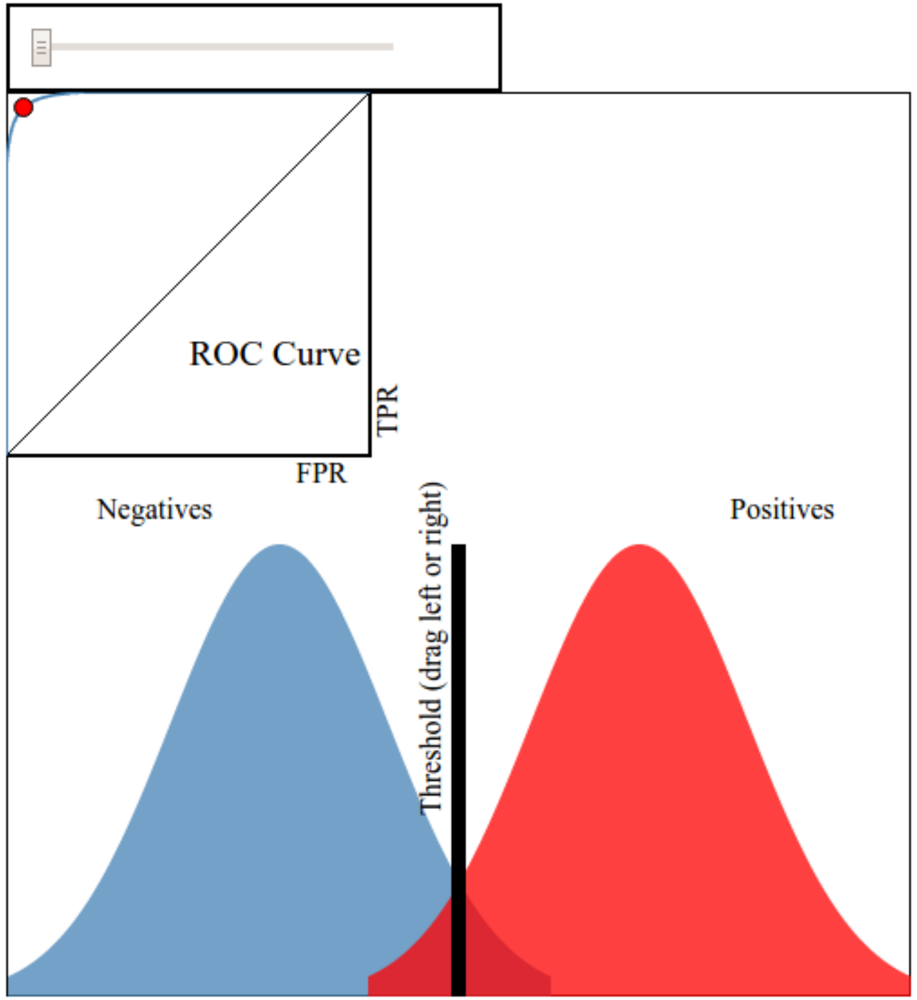
ADVANCED CLASSIFICATION METRICS

- 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.

	Predicted class		
		P	N
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

condition positive (P)

the number of real positive cases in the data

condition negatives (N)

the number of real negative cases in the data

---

true positive (TP)

eqv. with hit

true negative (TN)

eqv. with correct rejection

false positive (FP)

eqv. with false alarm, Type I error

false negative (FN)

eqv. with miss, Type II error

sensitivity, recall, hit rate, or true positive rate (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

specificity or true negative rate (TNR)

$$\mathrm{TNR} = \frac{\mathrm{TN}}{N} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

precision or positive predictive value (PPV)

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

negative predictive value (NPV)

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}}$$

accuracy (ACC)

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{P + N} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

F1 score

is the harmonic mean of precision and sensitivity

$$F_1 = 2 \cdot \frac{\mathrm{PPV} \cdot \mathrm{TPR}}{\mathrm{PPV} + \mathrm{TPR}} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

GUIDED PRACTICE

WHICH METRIC SHOULD I USE?

ACTIVITY: WHICH METRIC SHOULD I USE?

EXERCISE

DIRECTIONS

- Complete Part B of the starter code

- 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

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

In part C of the code, let's explore survival data from the Titanic.

1. Spend a few minutes determining which data would be most important to use in the prediction problem.
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.

TIPS FOR LOGISITC REGRESSION

Pros:

- Easy to interpret – the idea of regression is familiar and intuitive
- Low variance
- Provides probabilities for outcomes

Cons:

- Dependent variables to be categorical in nature
- Doesn’t handle large number of categorical features well
- Relies on transformations for non-linear features

COURSE

BEFORE NEXT CLASS

OUR PROGRESS SO FAR

UNIT 1: RESEARCH DESIGN AND EXPLORATORY DATA ANALYSIS

- ~~What is Data Science~~ ~~Lesson 1~~
- ~~Research Design and Pandas~~ ~~Lesson 2~~
- ~~Statistics Fundamentals I~~ ~~Lesson 3~~
- ~~Statistics Fundamentals II~~ ~~Lesson 4~~
- ~~Flexible Class Session~~ ~~Lesson 5~~

UNIT 2: FOUNDATIONS OF DATA MODELING

- ~~Introduction to Regression~~ ~~Lesson 6~~
- ~~Evaluating Model Fit~~ ~~Lesson 7~~
- ~~Introduction to Classification~~ ~~Lesson 8~~
- ~~Introduction to Logistic Regression~~ ~~Lesson 9~~
- Communicating Logistic Regression Results Lesson 10
- Flexible Class Session Lesson 11

Next Class

UNIT 3: DATA SCIENCE IN THE REAL WORLD

- Decision Trees and Random Forests Lesson 12
- Natural Language Processing Lesson 13
- Dimensionality Reduction Lesson 14
- Time Series Data I Lesson 15
- Time Series Data II Lesson 16
- Database Technologies Lesson 17
- Where to Go Next Lesson 18
- Flexible Class Session Lesson 19
- Final Project Presentations Lesson 20

DUE DATE

- Project: Unit Project 3 due Thursday Dec 14th

LESSON

Q & A

LESSON

EXIT TICKET

DON’T FORGET TO FILL OUT YOUR EXIT TICKET