GENERAL ASSEMBLY

# INTRODUCTION TO LOGISTIC REGRESSION

# TODAY’S LEARNING OBJECTIVES

- Logistic Regression & decision boundaries
- Gradient Descent and how it fits into Logistic Regression
- Evaluate a model using metrics such as classification accuracy/error, confusion matrix, ROC/AUC curves, and loss functions
- Build a Logistic Regression using sklearn

OPENING

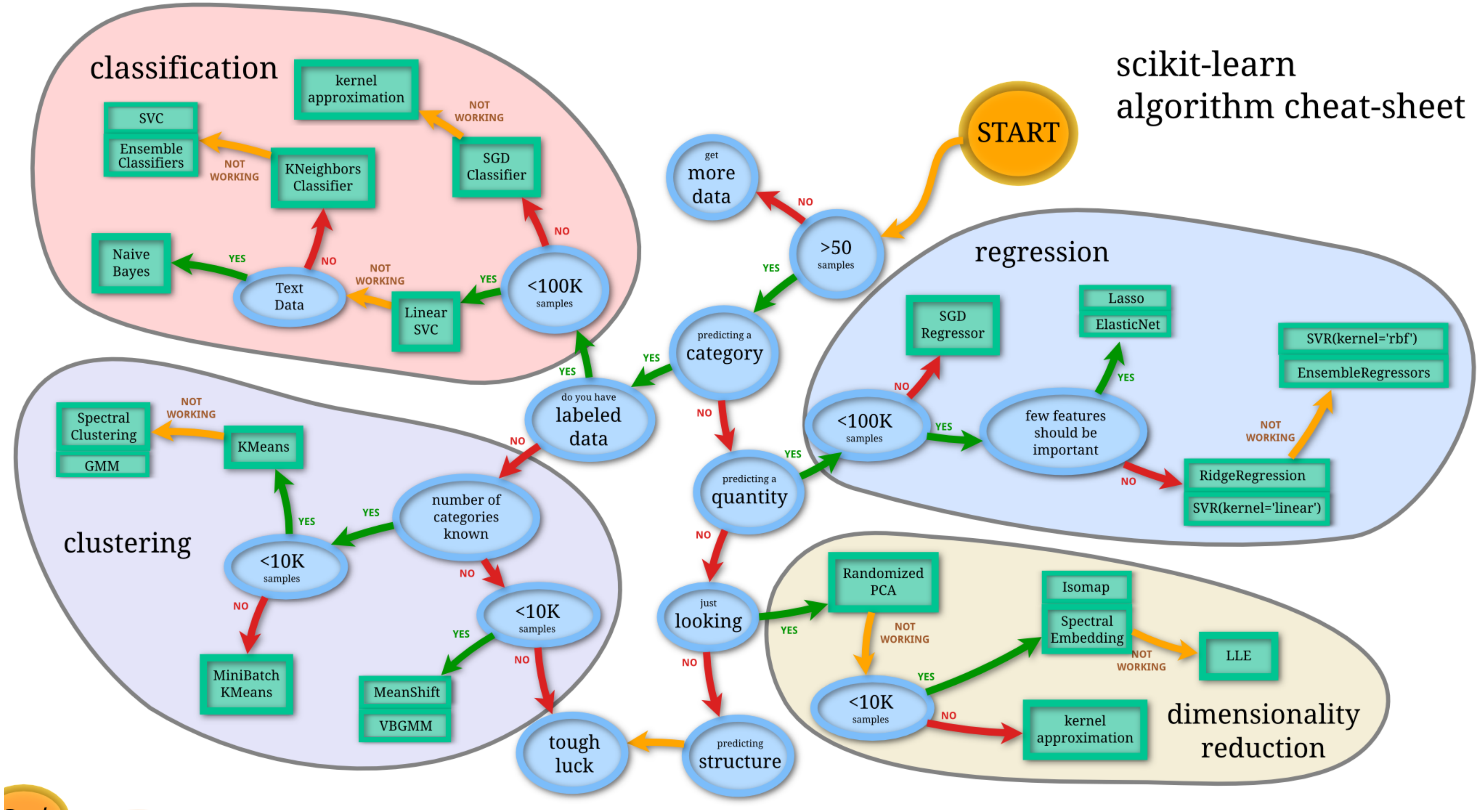
# MODEL DIFFERENCES

# WHERE ARE WE IN THE DATA SCIENCE WORKFLOW?

- Data has been **acquired** and **parsed.**

- Today we'll **refine** the data and **build** models (We'll also use plots to **represent** the results).

# WHERE ARE WE IN THE MACHINE LEARNING UNIVERSE?

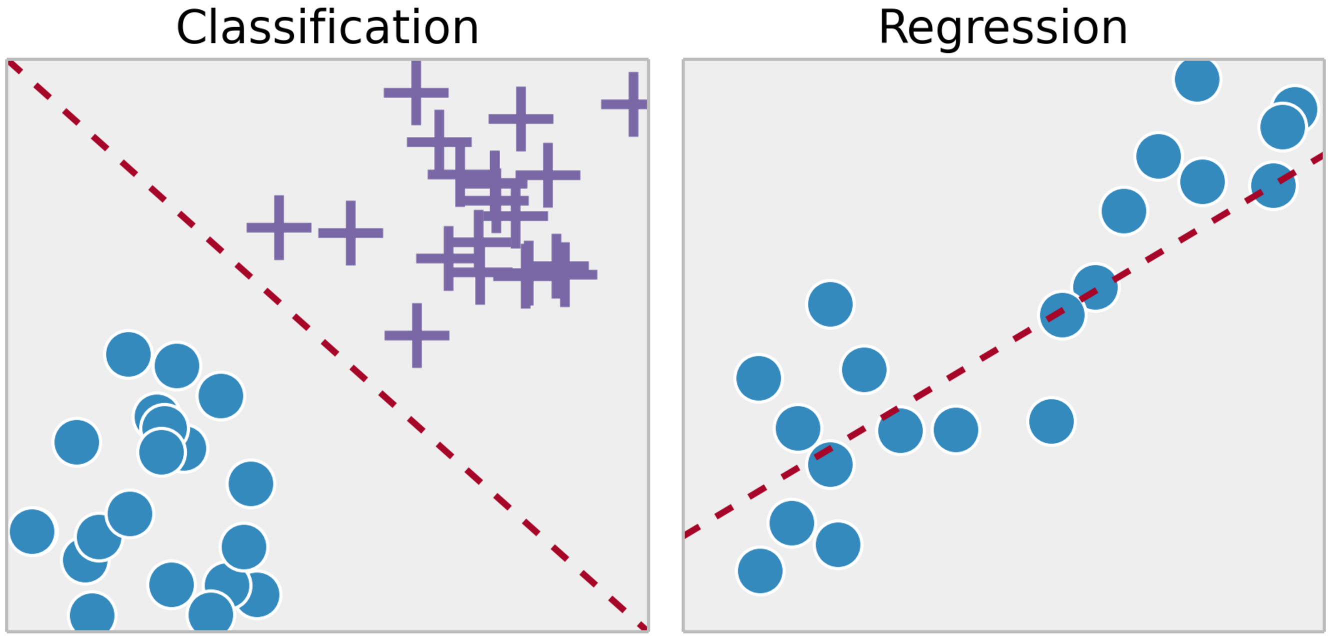


## INTRODUCTION

# LOGISTIC REGRESSION

# LOGISTIC REGRESSION

- Logistic regression is a linear approach to solving a classification problem. It will use a linear regression *style* approach to predict the class of an item, but retain the interpretability of linear regression model.

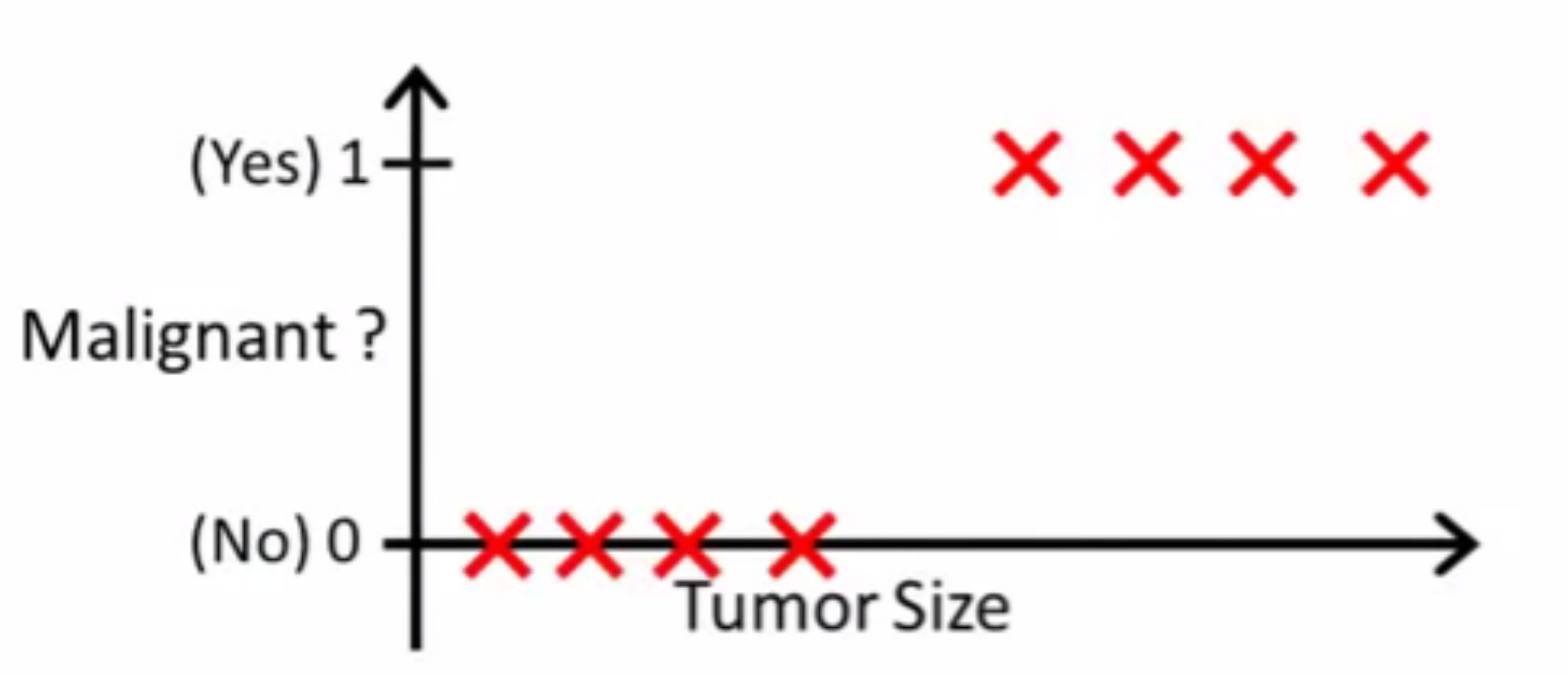


DEMO

# WHY NOT LINEAR REGRESSION?

# LOGISTIC REGRESSION

- **Motivation**: suppose we want to predict whether a tumor is malignant or benign based on its size.

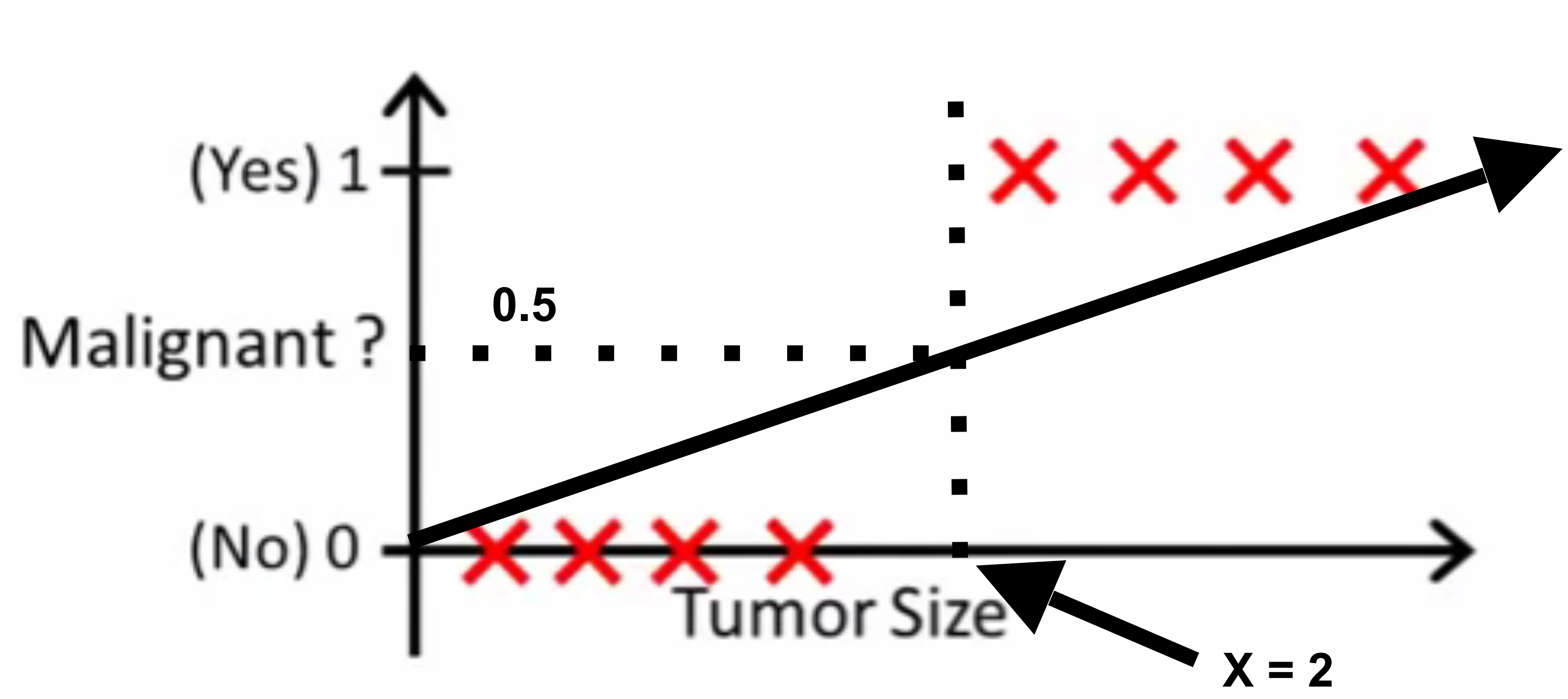


**Decision Boundary**

- Find where y = 0.5 => x = 2
- Anything larger than 2 we will classify as a malignant tumor
- Anything less than 2 we will classify as a benign tumor

# LOGISTIC REGRESSION

- Would Linear Regression work?



**Decision Boundary**

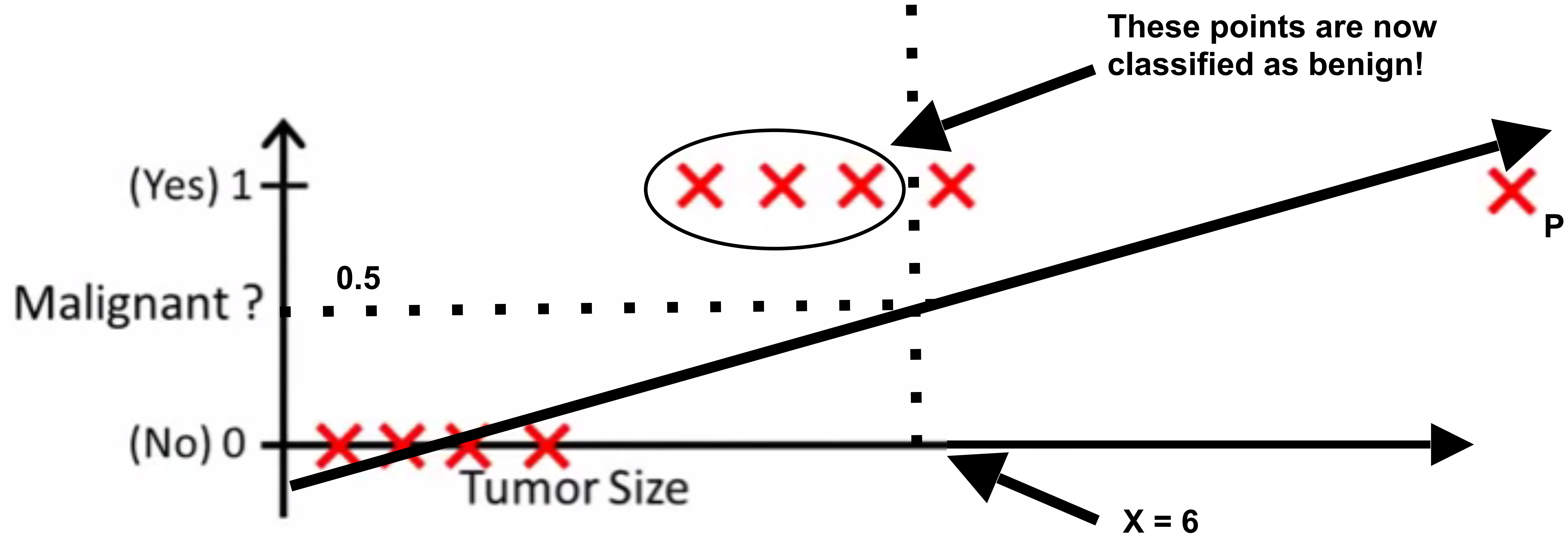
- Find where y = 0.5 => x = 2
- Anything larger than 2 we will classify as a malignant tumor
- Anything less than 2 we will classify as a benign tumor

- What types of problems can you foresee with this technique?

# LOGISTIC REGRESSION

**Problem 1**

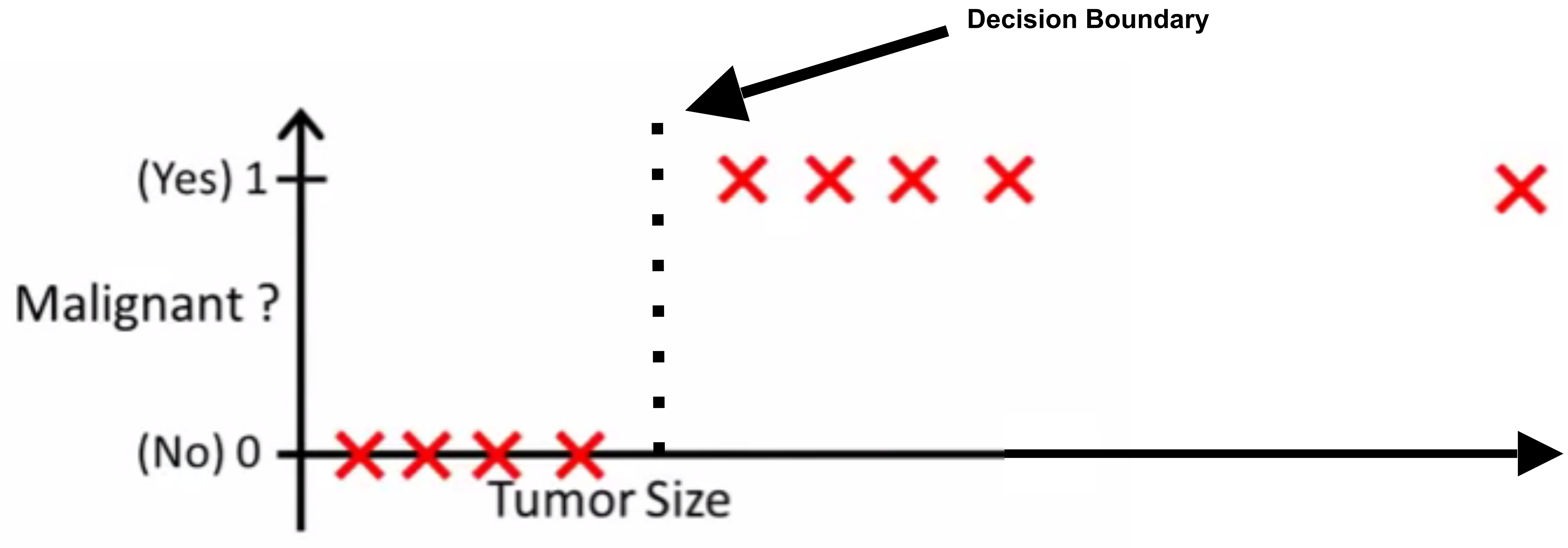
- Suppose we have another training example, p
- Because OLS is minimizing the residuals, it stretches out the line and moves our decision boundary!



# LOGISTIC REGRESSION

## Problem 2

- Linear Regression outputs predictions < 0 or > 1



- We want a model that predicts values between 0 and 1
- We want a decision boundary AKA “anything past this is class A and anything before it is class B”

INTRODUCTION

# LOGISTIC REGRESSION

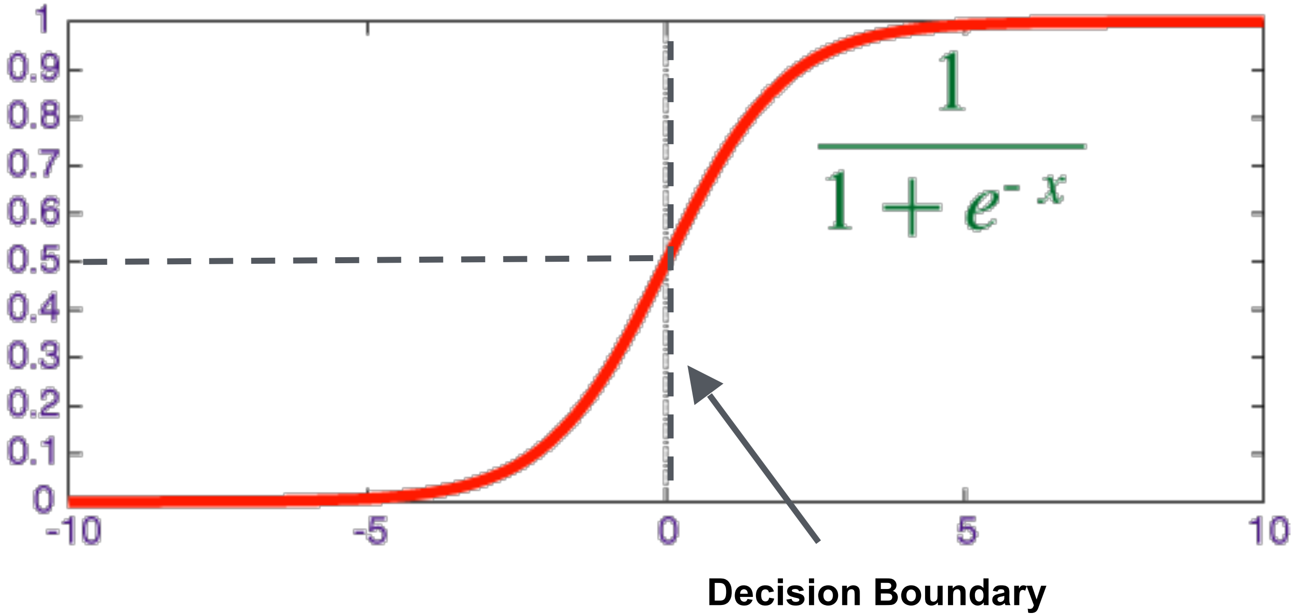
# LOGISTIC REGRESSION

- We will use **Logistic Regression** to predict a binary response (0 or 1).
- We use the **Sigmoid/Logistic** function to output probabilities for our predictions. It has the functional form of

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

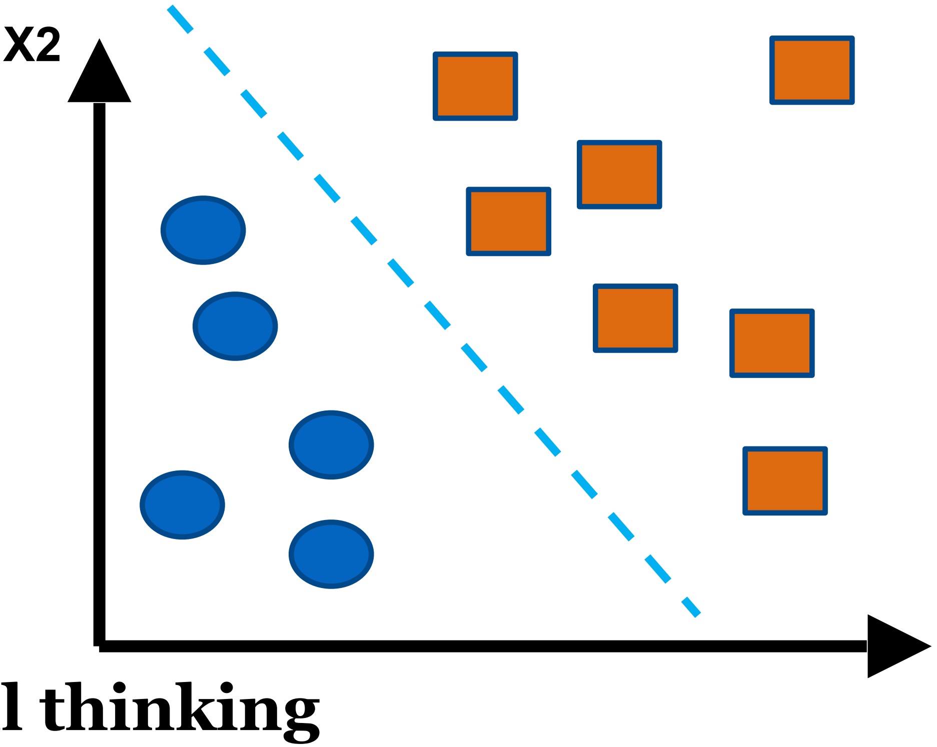
This is called our **Link** function.
We need it in order to find optimal parameters in our model

- A standard rule is to classify our “1” class as anything with a probability > 0.5 and our “0” class anything below that.
- The point at which the probability = 0.5 is called the **decision boundary**.



# TRANSFORMING LINEAR REGRESSION

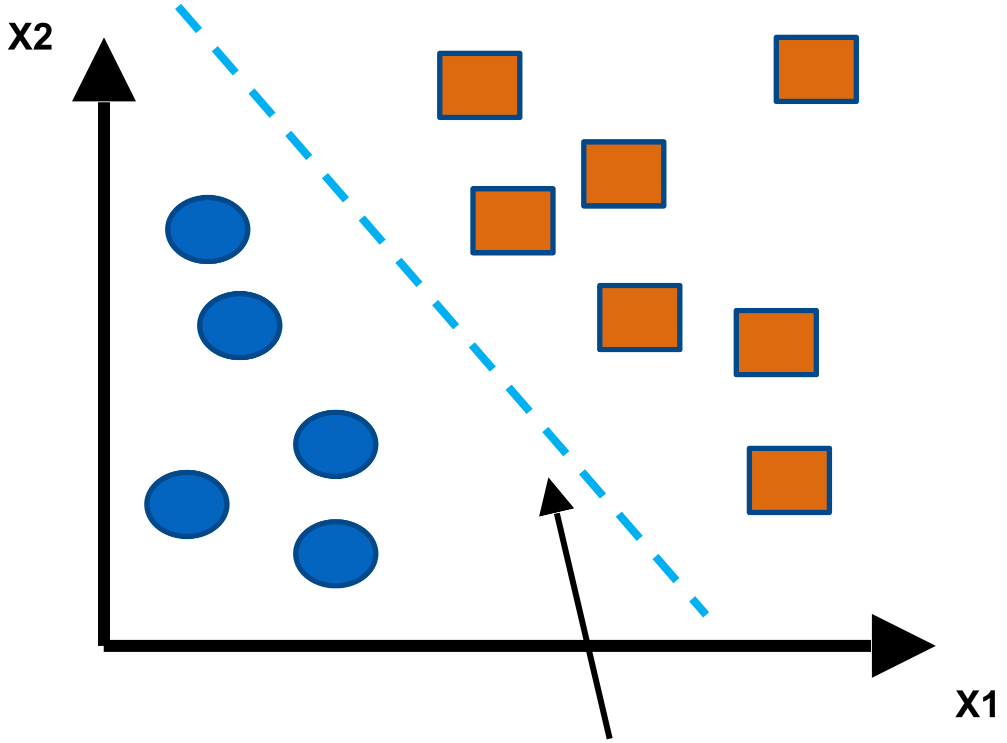
- **Intuition**: the “linearity” of logistic regression is in the decision boundary we develop. You can imagine it as a line (in 2-D) that separates the two classes.



**Questions you’re still thinking**
How do we develop this line and what’s the criteria for it? How does it relate to the Sigmoid function?

# TRANSFORMING LINEAR REGRESSION

**Question:** how do we develop the decision boundary?

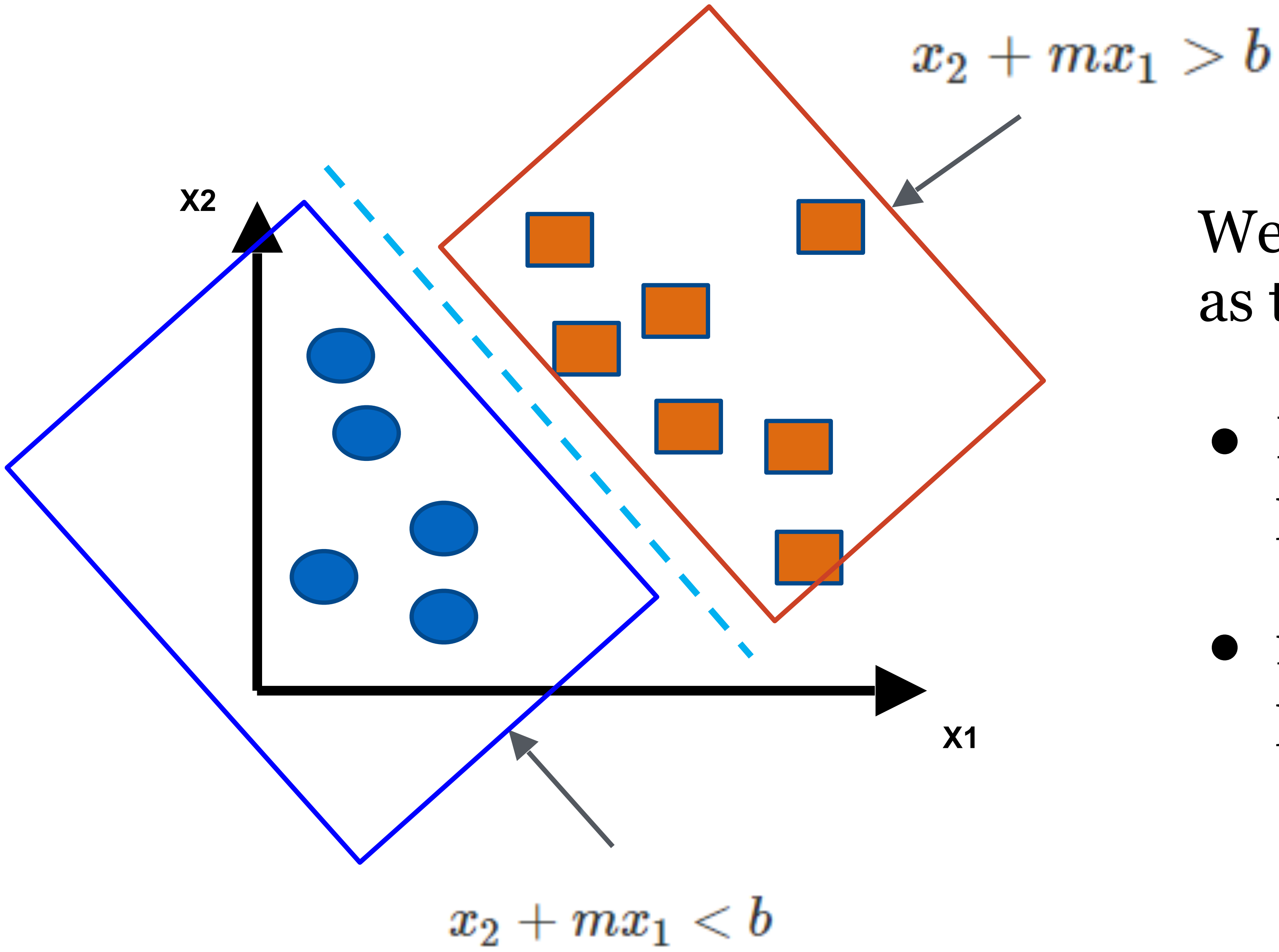


$$x_2 = -mx_1 + b$$

**Layman's definition**
We fit a line that best "divides" the two classes

**Mathematical definition**
We fit a line such that we minimize the sum of the residuals between the true label and the "strength" (aka probability) of our guess.

# TRANSFORMING LINEAR REGRESSION

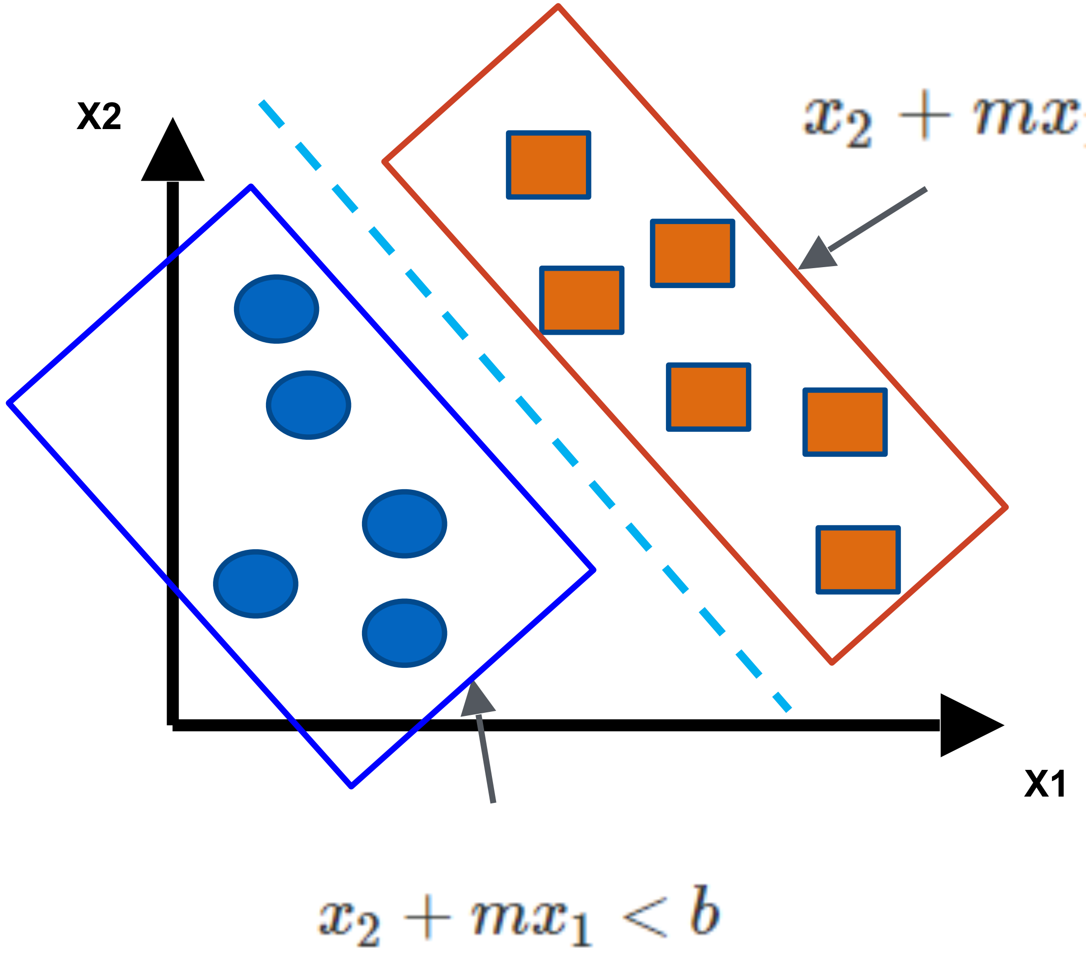


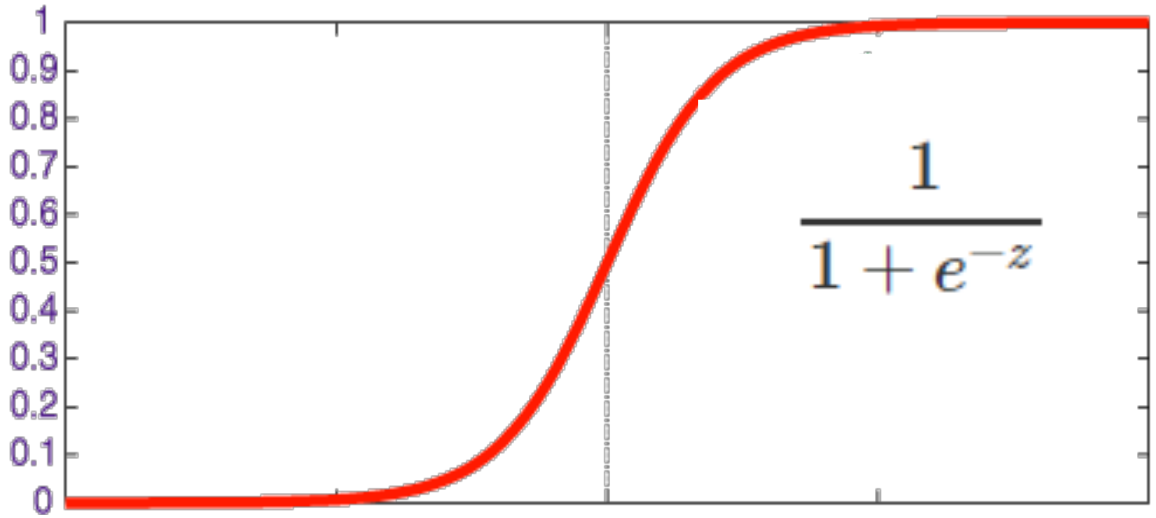
We can also rewrite the equation as this:

- If $x_2 + mx_1 > b$ then we predict 1.
- if $x_2 + mx_1 < b$ then we predict 0

# TRANSFORMING LINEAR REGRESSION

**Question:** how does our decision boundary relate to the sigmoid function?





**Where** $z = mx_1 + b - x_2$

The sigmoid function outputs values from [0,1] that represent the "strength" aka probability of our prediction

# KNOWLEDGE CHECK

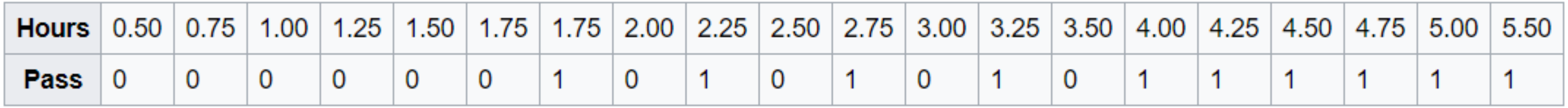
EXERCISE

1. Why is Linear Regression not an appropriate tool to use in classification?
2. How does Logistic Regression account for the failings of Linear Regression?
3. What is the main criteria for how Logistic Regression develops its decision boundary?
4. How does the Sigmoid function relate to the decision boundary?

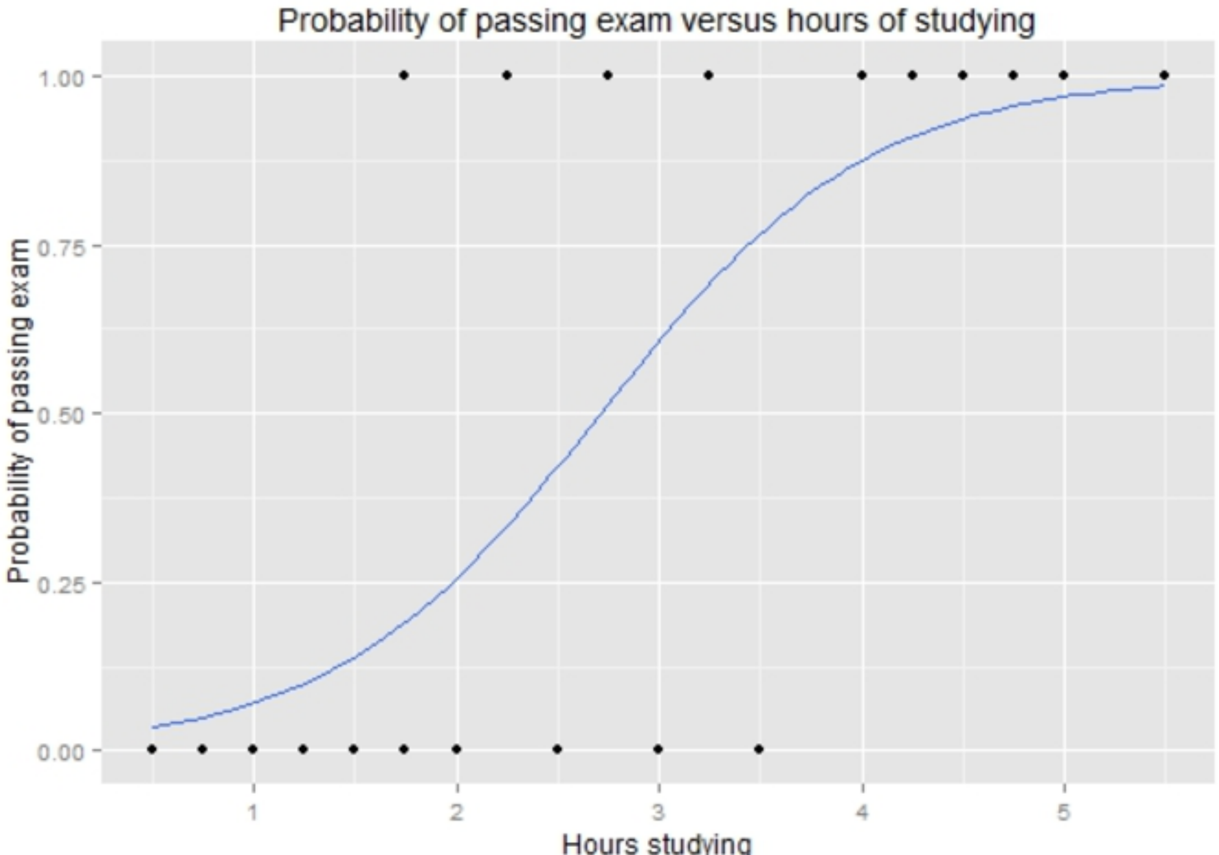
# INTERPRETING LOGISTIC REGRESSION

# TRANSFORMING LINEAR REGRESSION

**Example:** how does the number of hours spent studying affect the probability that the student will pass the exam?

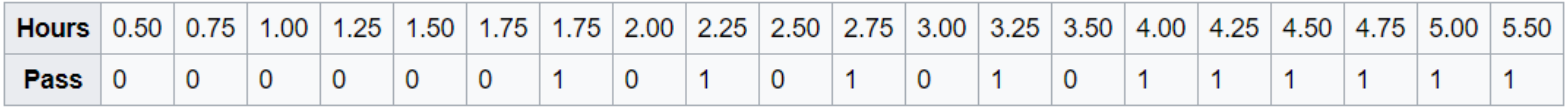


| Hours | 0.50 | 0.75 | 1.00 | 1.25 | 1.50 | 1.75 | 1.75 | 2.00 | 2.25 | 2.50 | 2.75 | 3.00 | 3.25 | 3.50 | 4.00 | 4.25 | 4.50 | 4.75 | 5.00 | 5.50 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Pass | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |

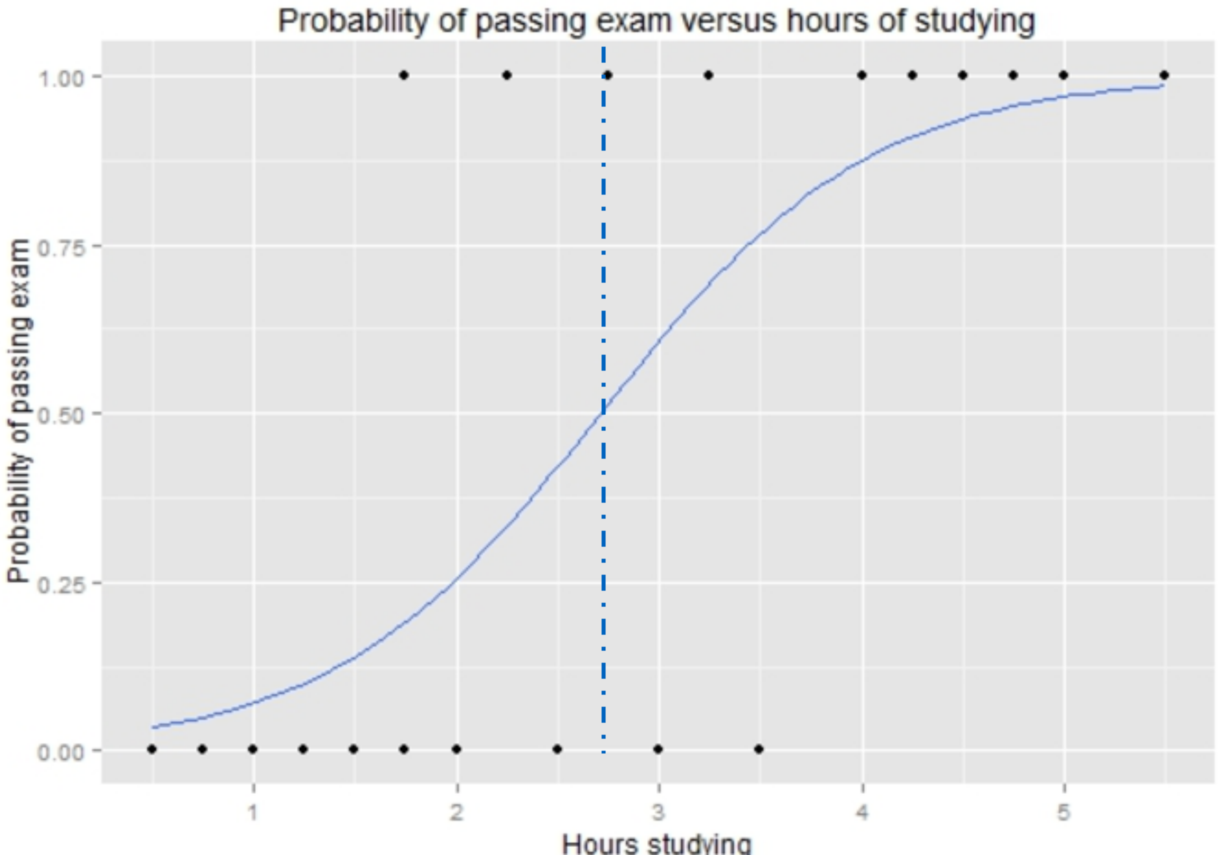


# TRANSFORMING LINEAR REGRESSION

**Example:** how does the number of hours spent studying affect the probability that the student will pass the exam?

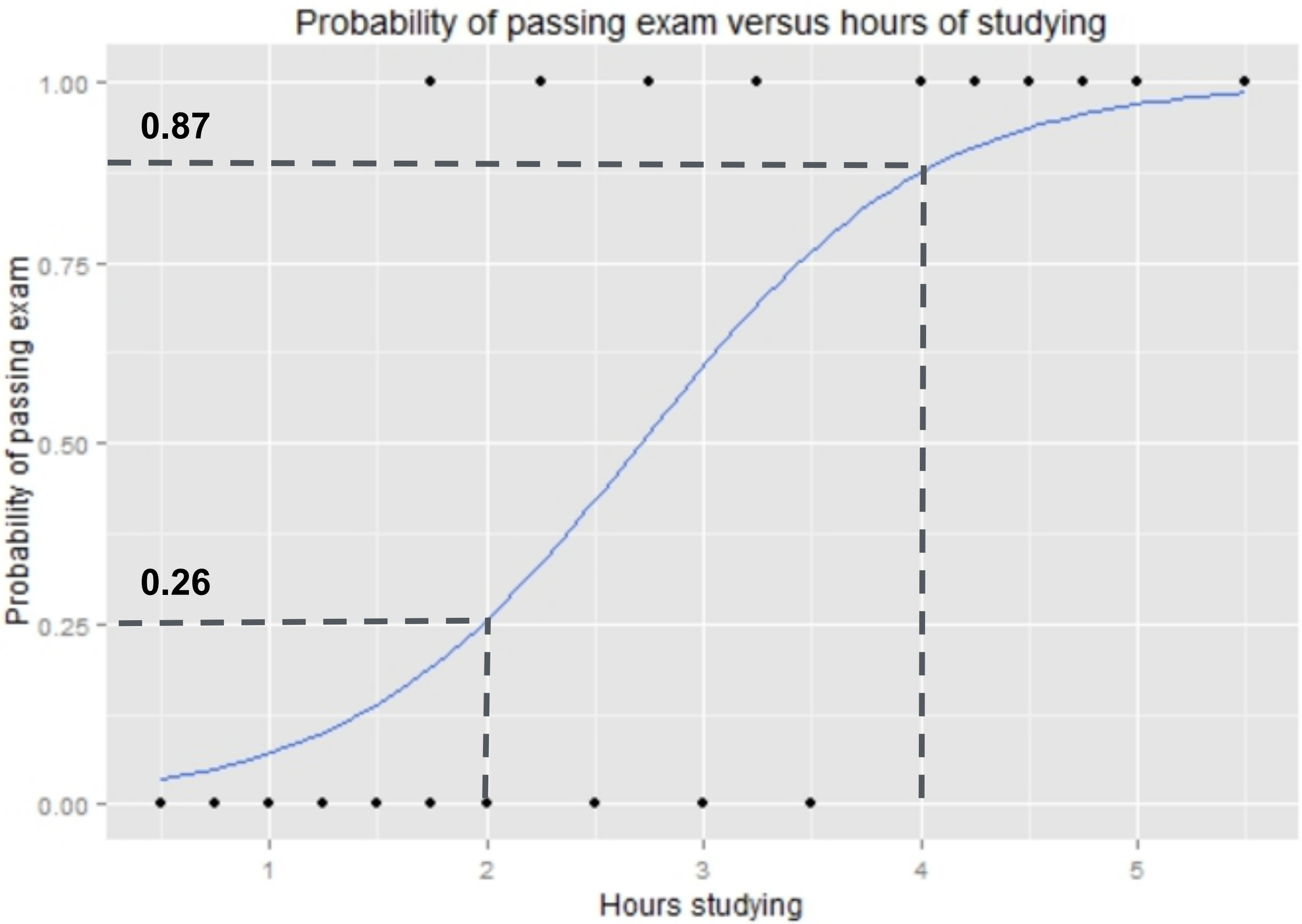


| Hours | 0.50 | 0.75 | 1.00 | 1.25 | 1.50 | 1.75 | 1.75 | 2.00 | 2.25 | 2.50 | 2.75 | 3.00 | 3.25 | 3.50 | 4.00 | 4.25 | 4.50 | 4.75 | 5.00 | 5.50 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Pass | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |



$$\frac{1}{1+e^{-(1.5046*hours-4.0777)}}$$

# TRANSFORMING LINEAR REGRESSION



For a student who studies **2 hours**, the equation gives an estimated probability of passing an exam at **0.26**:

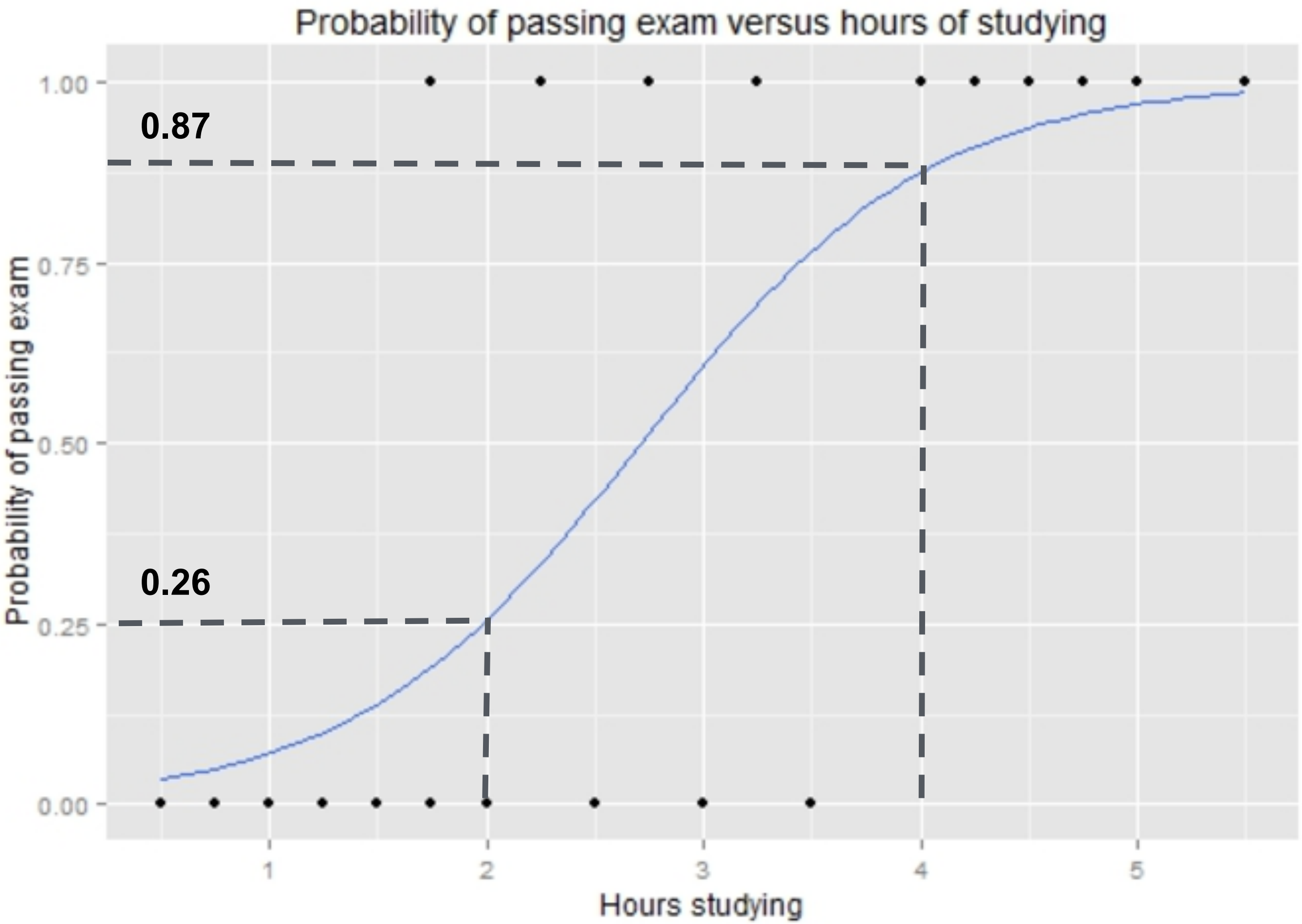
$$\frac{1}{1+e^{-(1.5046*2-4.0777)}} = 0.26$$

While someone that studies for **4 hours**, their probability of passing is **0.87**:

$$\frac{1}{1+e^{-(1.5046*4-4.0777)}} = 0.87$$

*Standard rule is to predict the "1" class if the **probability > 0.5**

# TRANSFORMING LINEAR REGRESSION

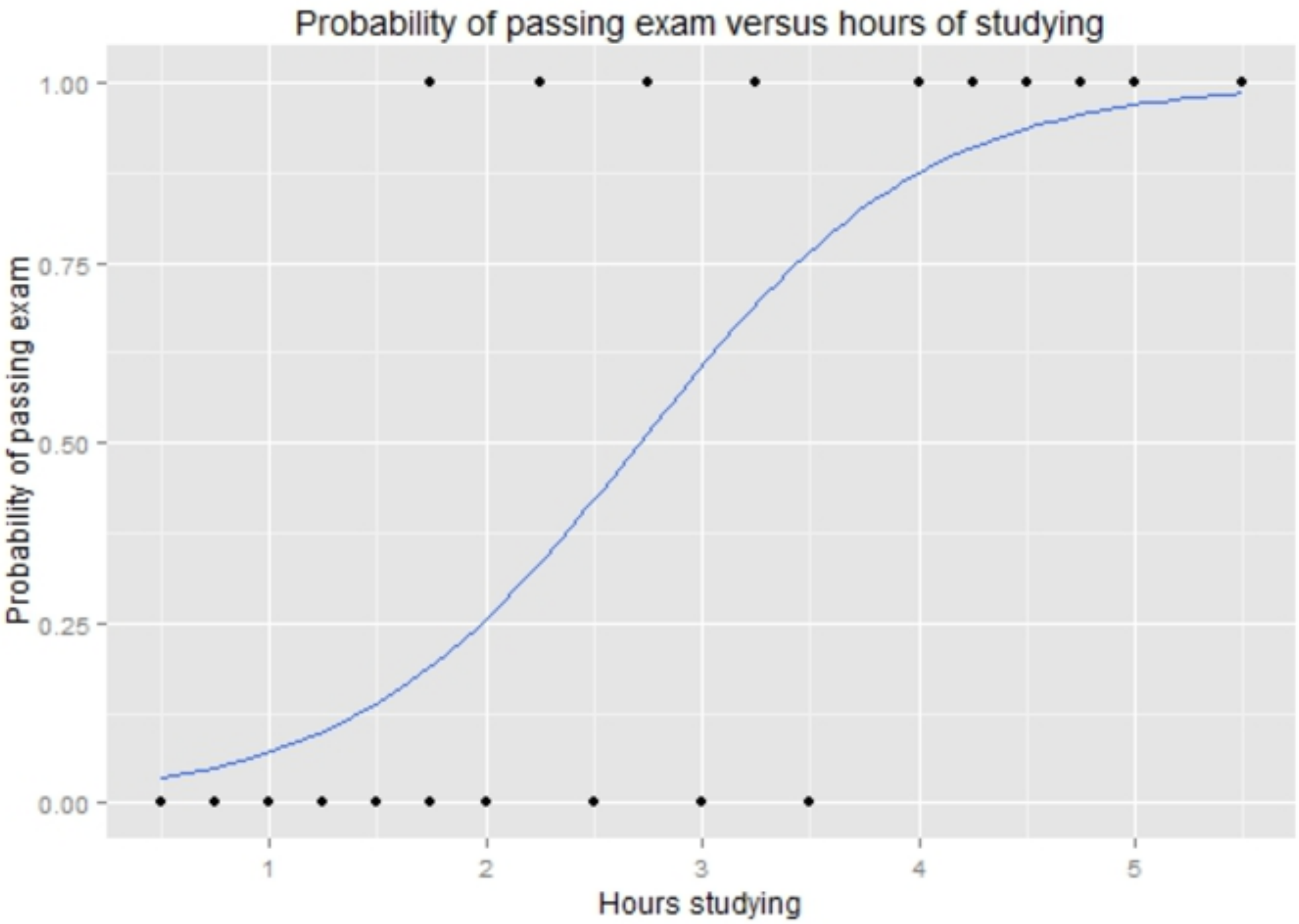


Standard rule is to predict the “1” class if the **probability > 0.5**

- What would you predict for a student that studies **2** hours?
- What would you predict for a student that studies **4** hours?

# TRANSFORMING LINEAR REGRESSION

## How do we interpret the coefficients?



$$\frac{1}{1 + e^{-(1.5046 * hours - 4.0777)}}$$

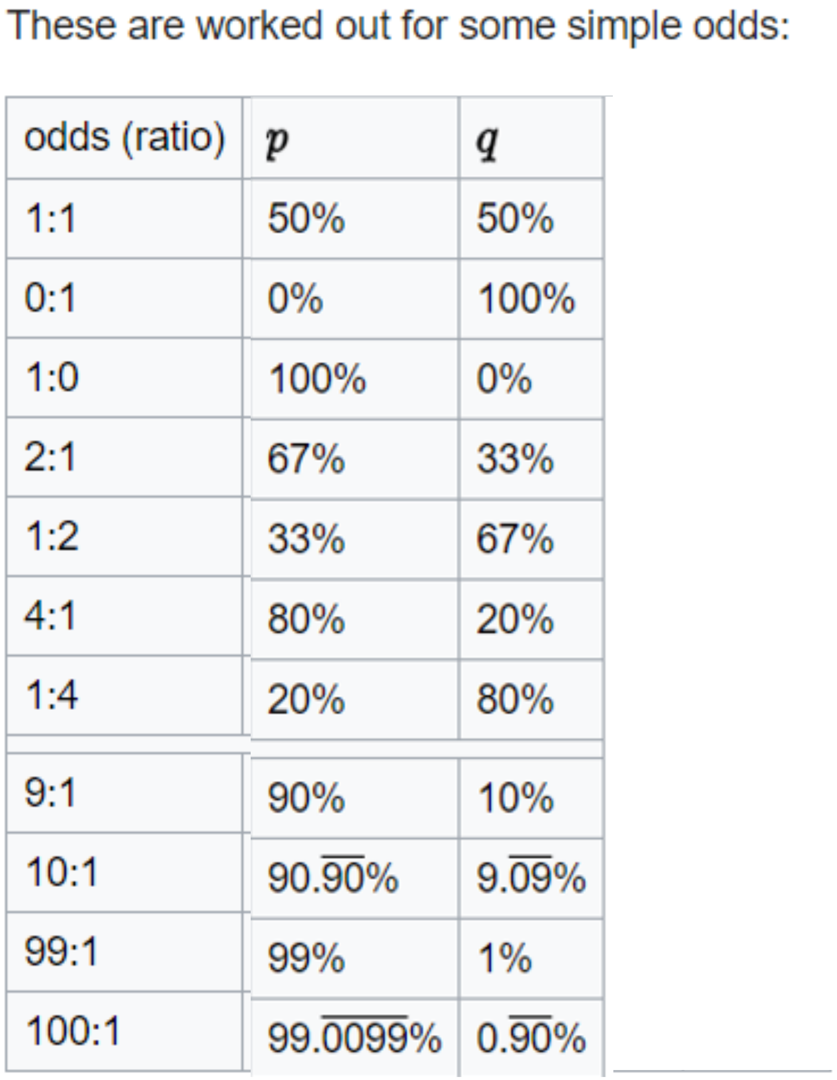
For every 1 hour increase in studying, a student multiplies their **odds** of passing an exam by:

$$e^{1.5046} \approx 4.5$$

**Bottom line:** Positive coefficients increase the log-odds of the response (and thus increase the probability), and negative coefficients decrease the log-odds of the response (and thus decrease the probability).

# TRANSFORMING LINEAR REGRESSION

**Odds** represent the likelihood that an event will take place.



These are worked out for some simple odds:

| odds (ratio) | $p$ | $q$ |
|---|---|---|
| 1:1 | 50% | 50% |
| 0:1 | 0% | 100% |
| 1:0 | 100% | 0% |
| 2:1 | 67% | 33% |
| 1:2 | 33% | 67% |
| 4:1 | 80% | 20% |
| 1:4 | 20% | 80% |
| 9:1 | 90% | 10% |
| 10:1 | $90.\overline{90}\%$ | $9.\overline{09}\%$ |
| 99:1 | 99% | 1% |
| 100:1 | $99.\overline{0099}\%$ | $0.\overline{90}\%$ |

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

$$probability = \frac{odds}{1 + odds}$$

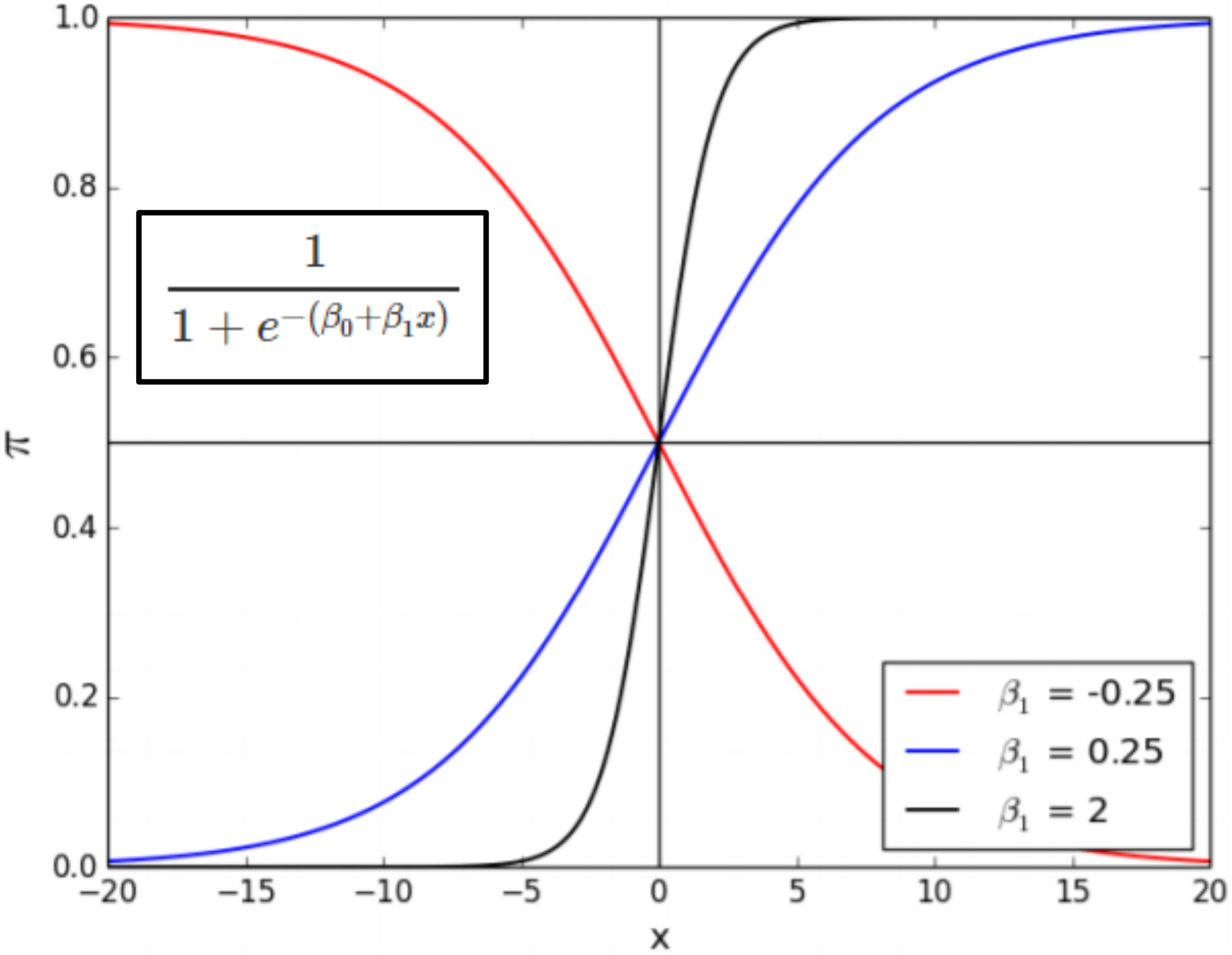
- **What would be the odds for an event with a *p* = 0.75?**

- **What about the odds for something with a *p* = 0.40?**

# TRANSFORMING LINEAR REGRESSION

*Changing the $\beta_1$ value changes the slope of the curve*

- **The bigger the $\beta_1$ the steeper the curve**
- **The smaller $\beta_1$ the more gradual the curve**

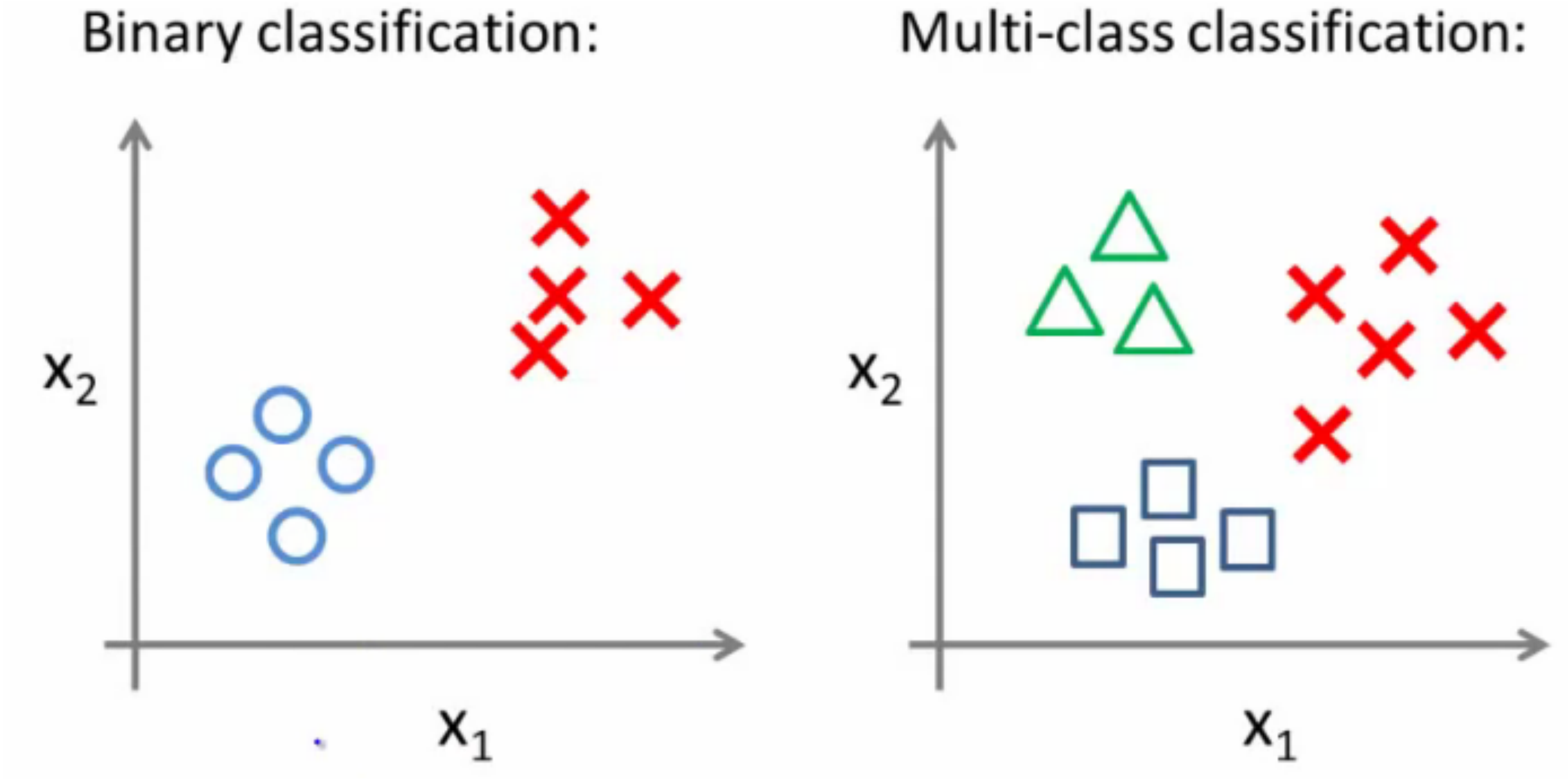


INDEPENDENT PRACTICE

# MULTICLASS LOGISTIC REGRESSION

# MULTICLASS LOGISTIC REGRESSION

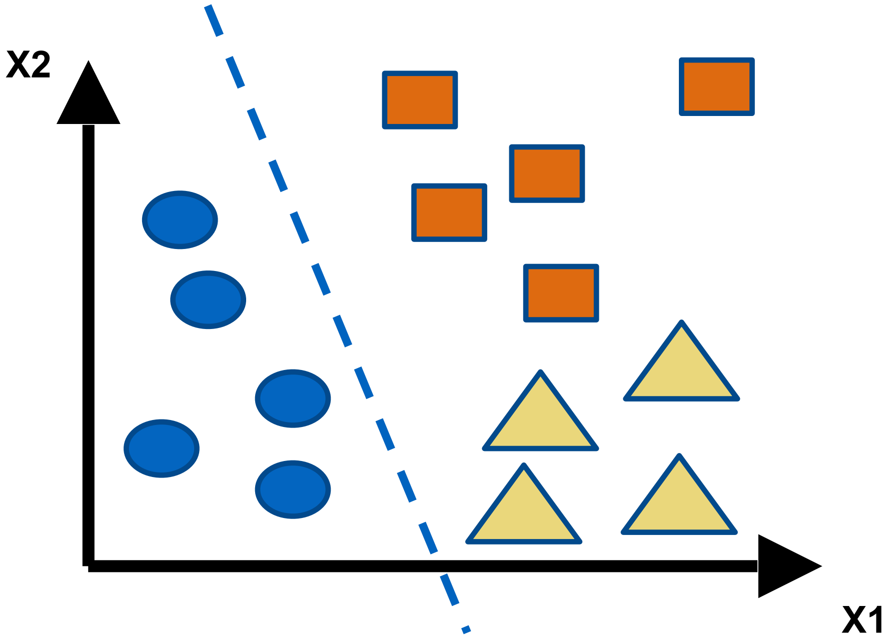
- Getting logistic regression for multiclass classification using **one vs. all**
- Multiclass - more than yes or no (1 or 0)
    - Classification with multiple classes for assignment



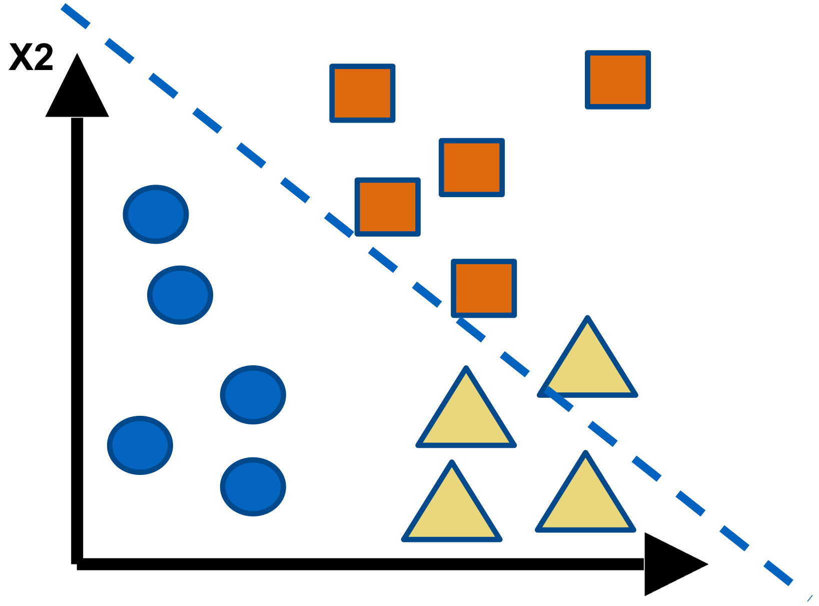
# MULTICLASS LOGISTIC REGRESSION

**One vs. all classification**

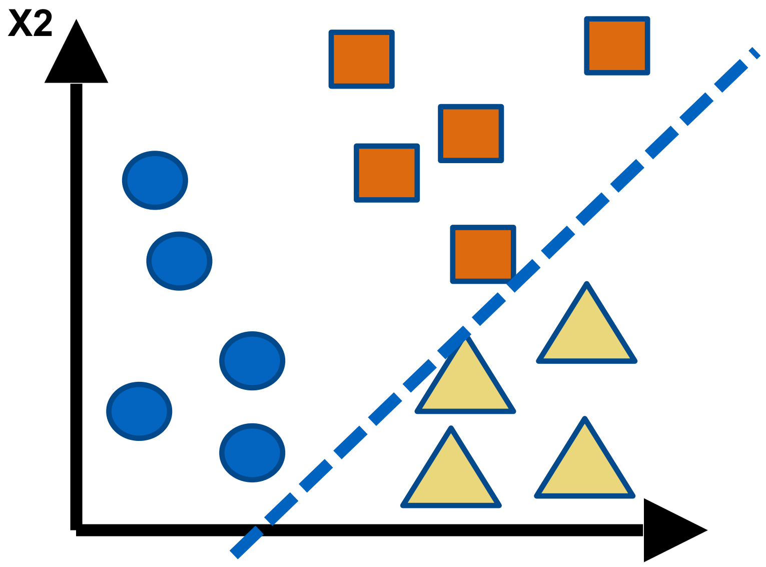
- Split the training set into three separate binary classification problems



Circle vs. everything else
(squares and triangles)



Squares vs. everything else
(circles and triangles)

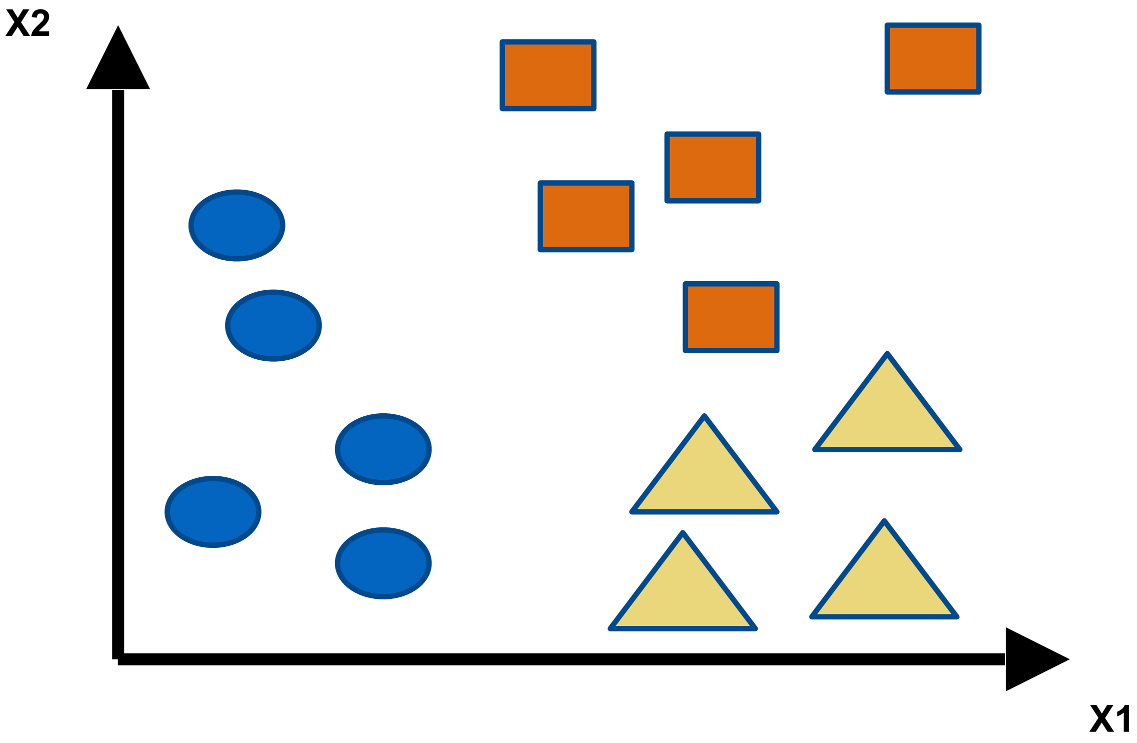


Triangle vs. everything else
(circles and squares)

# MULTICLASS LOGISTIC REGRESSION

**Overall**

- Train a logistic regression classifier for each class, i.
- Make predictions based on which ever model outputs the highest probability for y=1

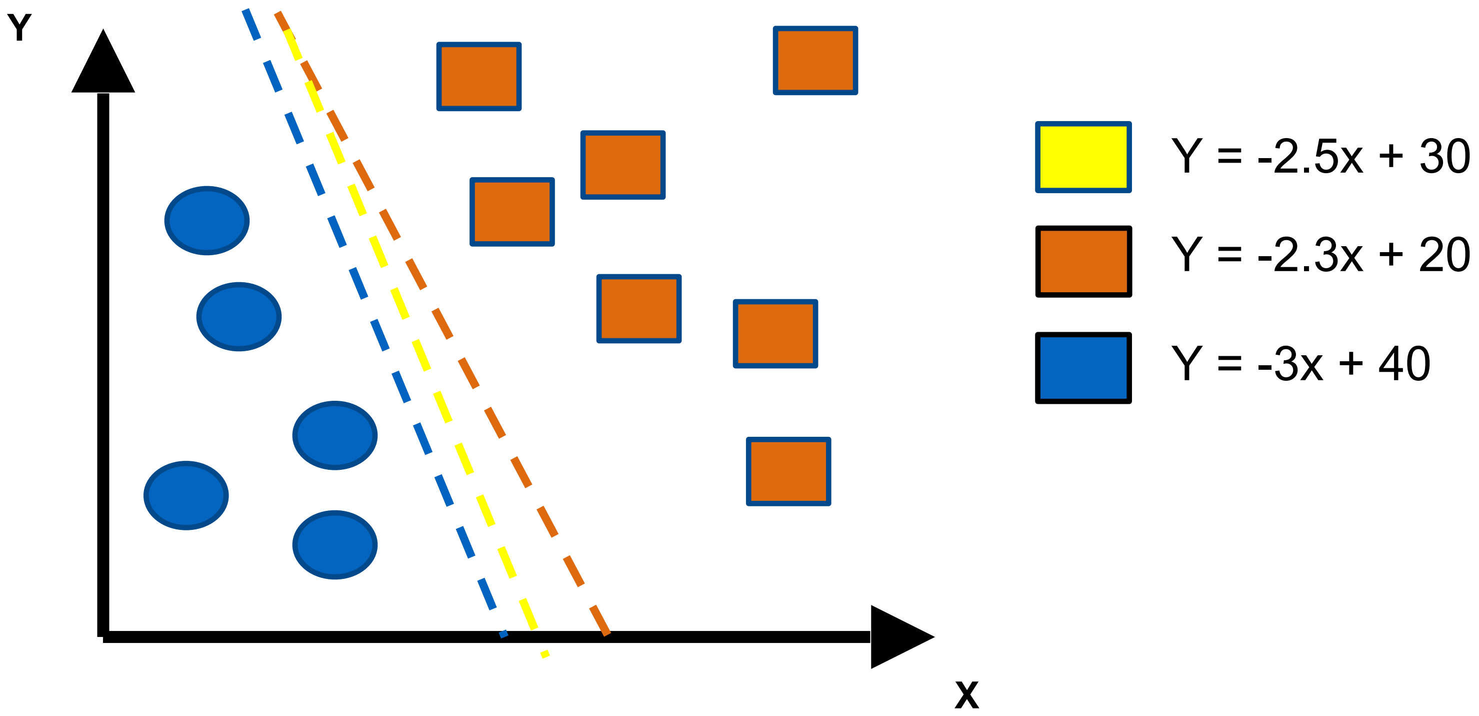


INTRODUCTION

# GRADIENT DESCENT

# GRADIENT DESCENT

**Intuition:** how do we find the decision boundary that best separates our two classes? AKA what are the coefficients for the linear boundary?



We want to build a decision boundary that best divides the two classes. There are two ways we can think of this:

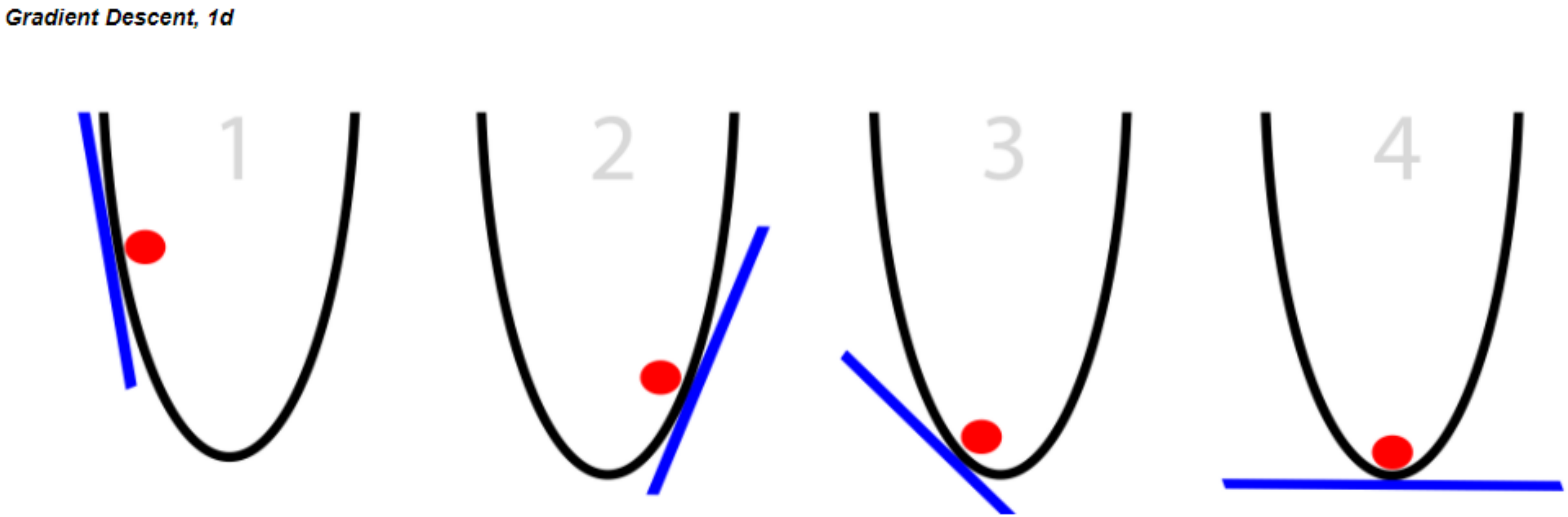
(1) We fit a line such that we **minimize** the sum of the residuals between the true label and the "strength" (aka probability) of our guess.

(2) We **optimize** the probabilities

- We make our probabilities as **large** as possible when y = 1.
- We make our probabilities as **small** as possible when y = 0.

# GRADIENT DESCENT

- Gradient Descent is an **iterative** process that we use to pick our decision boundary
- You can see it as almost a guess and check – do these sets of coefficients minimize the error between my guess and the true point? If no, then keep looking for better coefficients. If yes, then stop!
- The y-axis represents the residuals AKA the **cost function**, and the x-axis represents our parameters for the decision boundary
- Notice how there is only **one** point that minimizes the cost function



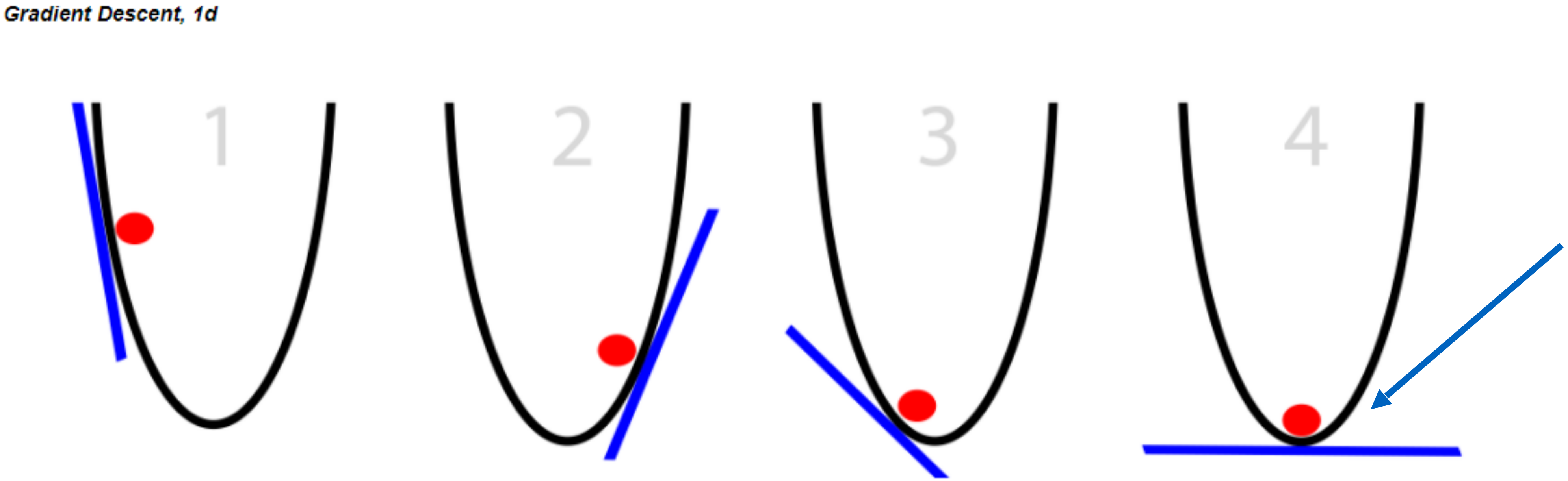
# GRADIENT DESCENT

1. A random linear solution is provided as a starting point (usually a "flat" line or solution)

2. The solver then attempts to find a next step: we take a step in any direction and measure each performance.

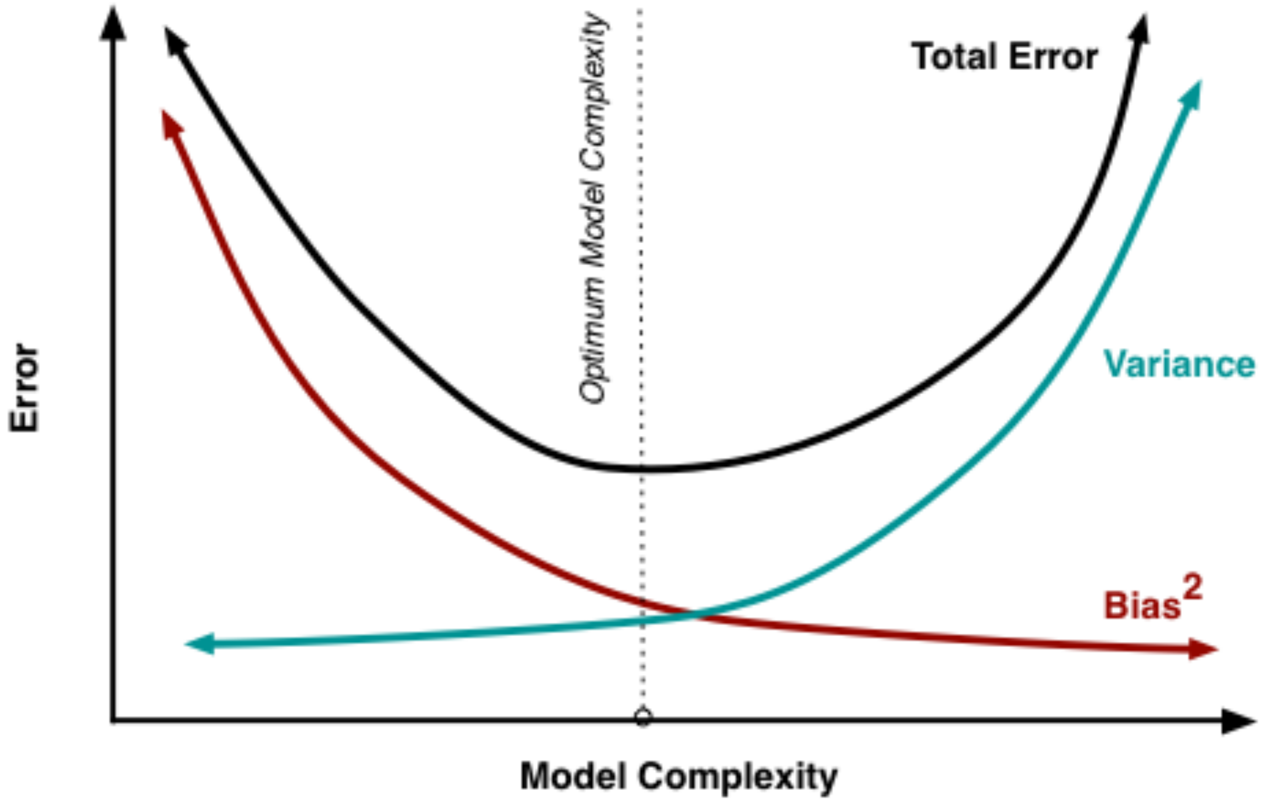
3. If the solver finds a better solution (optimizing toward a metric such as mean squared error), this is the new starting point.

4. Repeat these steps until the performance is optimized and no "next steps" perform better. The size of the steps will shrink over time.



Because we've minimized the residuals at this point, the set of parameters we use to define our decision boundary are here!

# DOES THE CURVE LOOK FAMILIAR?



# LOGISTIC REGRESSION

## Advantages of logistic regression

- Highly interpretable (if you remember how)
- Model training and prediction are fast
- No tuning is required (excluding regularization)
- Features don't need scaling
- Can perform well with a small number of observations
- Outputs well-calibrated predicted probabilities

## Disadvantages of logistic regression

- Presumes a linear relationship between the features and the log-odds of the response
- Performance is (generally) not competitive with the best supervised learning methods
- Can't automatically learn feature interactions

INTRODUCTION

# ADVANCED CLASSIFICATION METRICS

# ADVANCED CLASSIFICATION METRICS

- What if we wanted to know exactly how a classifier was performing (e.g. what is predicting correctly vs incorrectly)?

# THE CONFUSION MATRIX

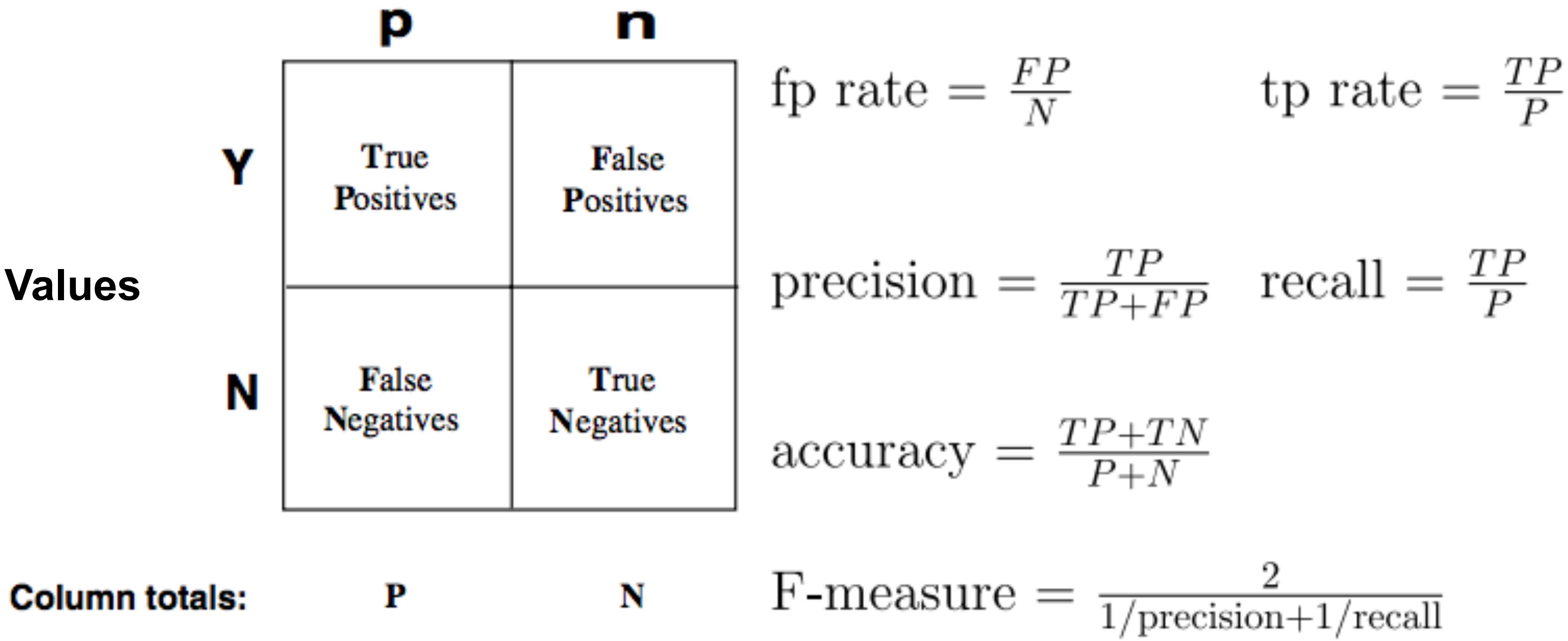
- Confusion matrices allow for the interpretation of correct and incorrect predictions for *each class label*.

- It is the first step for the majority of classification metrics and goes deeper than just accuracy.

# INTRO TO THE CONFUSION MATRIX

- Let’s review our confusion matrix.

**Prediction**



| | | p | n |
|---|---|---|---|
| **Actual Values** | Y | True Positives | False Positives |
| | N | False Negatives | True Negatives |
| Column totals: | | P | N |

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

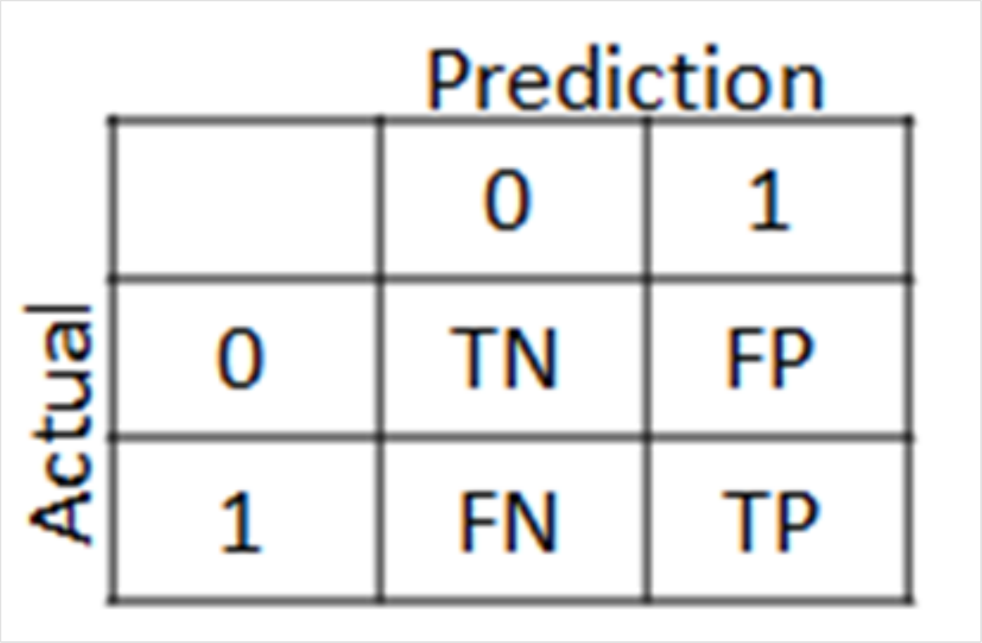
$$\text{precision} = \frac{TP}{TP+FP} \qquad \text{recall} = \frac{TP}{P}$$

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

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

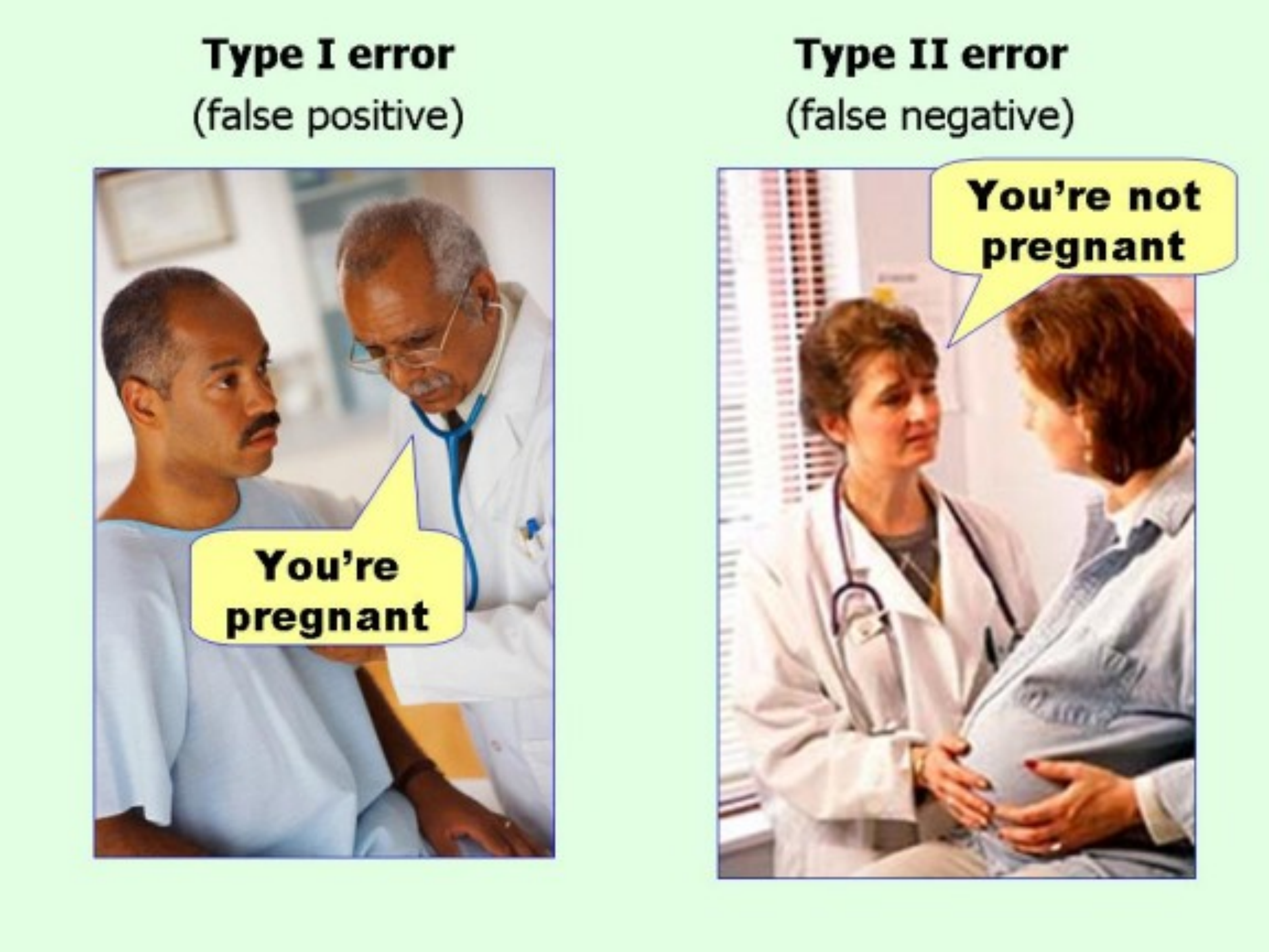
# ADVANCED CLASSIFICATION METRICS

- We can use a **confusion matrix** to obtain more granular accuracy ratings for of each class by using the *true positive rate* and the *false positive rate*.



| | | Prediction | |
|---|---|---|---|
| | | 0 | 1 |
| Actual | 0 | TN | FP |
| | 1 | FN | TP |

# CONFUSION MATRIX – ERROR TYPES



# ADVANCED CLASSIFICATION METRICS

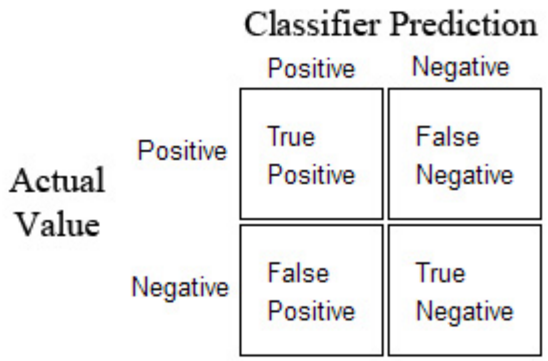
|  |  | **Predictions** |  |
|---|---|---|---|
|  |  | Not Pregnant | Pregnant |
| **True Class** | Not Pregnant | 50 | 24 |
|  | Pregnant | 10 | 62 |

- What are the number of true positives?
- What are the number of false negatives?
- What are the number of false positives?
- What are the number of true negatives?

# ADVANCED CLASSIFICATION METRICS

- The **true positive rate (TPR)** asks, “Out of all of the target classes, how many were accurately predicted to belong to that class?”

- Using our example, the TPR would be how often does our model correctly identify customer who will default on their credit card debt.



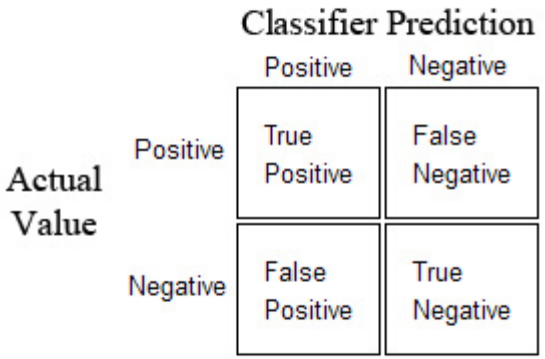
|  |  | Classifier Prediction |  |
|---|---|---|---|
|  |  | Positive | Negative |
| Actual Value | Positive | True Positive | False Negative |
|  | Negative | False Positive | True Negative |

$$TPRate = \frac{TP}{TP + FN}$$

# ADVANCED CLASSIFICATION METRICS

- The **false positive rate (FPR)** asks, “Out of all items not belonging to a class, how many were predicted as belonging to that target class label?”

- Using our example, the FPR would be how often the model predict that a customer will default when they end up not doing so.



| | | Classifier Prediction | |
|---|---|---|---|
| | | Positive | Negative |
| Actual Value | Positive | True Positive | False Negative |
| | Negative | False Positive | True Negative |

$$FPRate = \frac{FP}{FP + TN}$$

# PRECISION AND RECALL

- Our previous metrics were primarily designed for less biased data problems: we could be interested in both outcomes, so it was important to generalize our approach.

- For example, we may be interested if a person will vote for a Republican or Democrat. This is a binary problem, but we're interested in both outcomes.

# PRECISION AND RECALL

- Precision and recall, metrics built from the confusion matrix, focus on *information retrieval*, particularly when one class is more interesting than the other.

- For example, we may want to predict if a person will be a customer. We care much more about people who will be a customer of ours than people who won't.

# PRECISION AND RECALL

- *Precision* aims to product a high amount of relevancy instead of irrelevancy.

- Precision asks, “Out of all of our positive predictions (both true positive and false positive), how many were correct?”

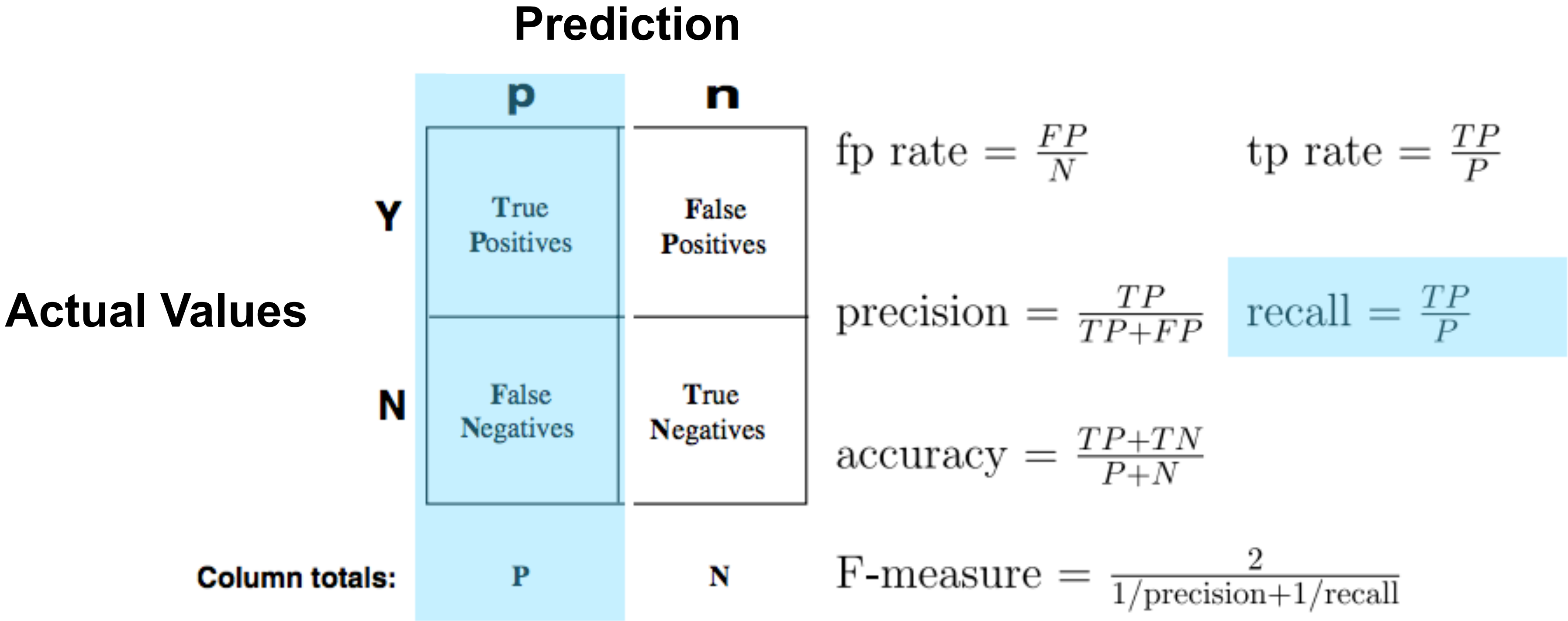
- *Recall* aims to see how well a model returns specific data (literally, checking whether the model can *recall* what a class label looked like).

- Recall asks, “Out of all of our positive class labels, how many were correct?”

# THE MATH FOR RECALL

- Recall is the count of predicted *true positives* over the total count of that class label.

- This is the same as True Positive Rate or *sensitivity*.



Prediction

| Actual Values | | p | n |
|---|---|---|---|
| | Y | True Positives | False Positives |
| | N | False Negatives | True Negatives |
| Column totals: | | P | N |

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

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

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

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

# THE MATH FOR RECALL

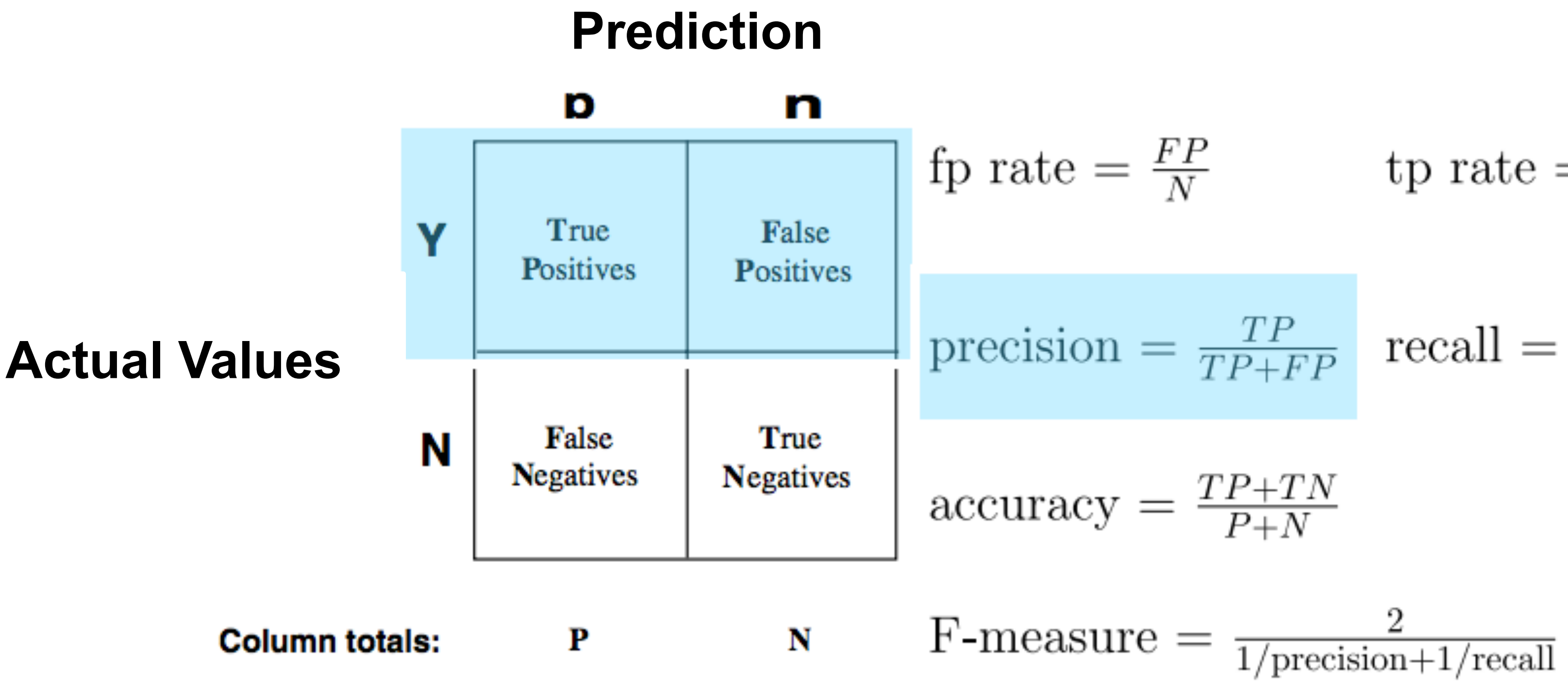
- Imagine predicting the color of a marble as either red or green.  There are 10 of each.

- If the model identifies 8 identifies 8 of the green marbles as green, the recall is 8 / 10 = 0.80.

- However, this says nothing of the number of *red* marbles that are also identified as green.

# THE MATH FOR PRECISION

- Precision, or positive predicted value, is calculated as the count of predicted true positives over the count of all values predicted to be positive.



Prediction

| Actual Values | | p | n |
|---|---|---|---|
| | Y | True Positives | False Positives |
| | N | False Negatives | True Negatives |
| Column totals: | | P | N |

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

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

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

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

# THE MATH FOR PRECISION

- Let’s use our marble example again.

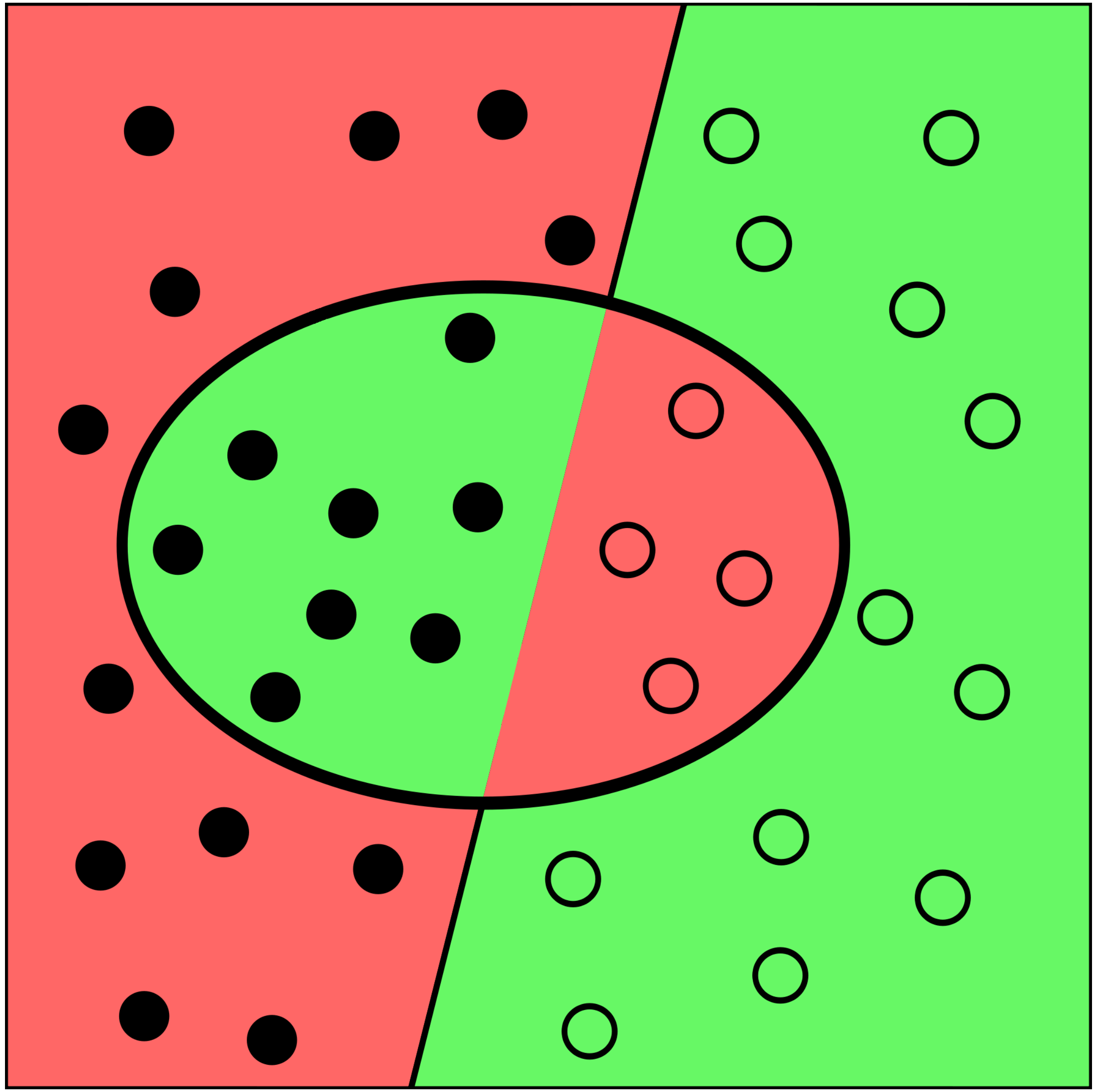
- If a model predicts 8 of the green marbles as green, then precision would be 1.00, because all marbles predicted as green were in fact green.

- Let’s assume all red marbles were predicted correctly, and 2 green were predicted as red.

- The precision of red marbles would be 10 / (10 + 2) = 0.833.

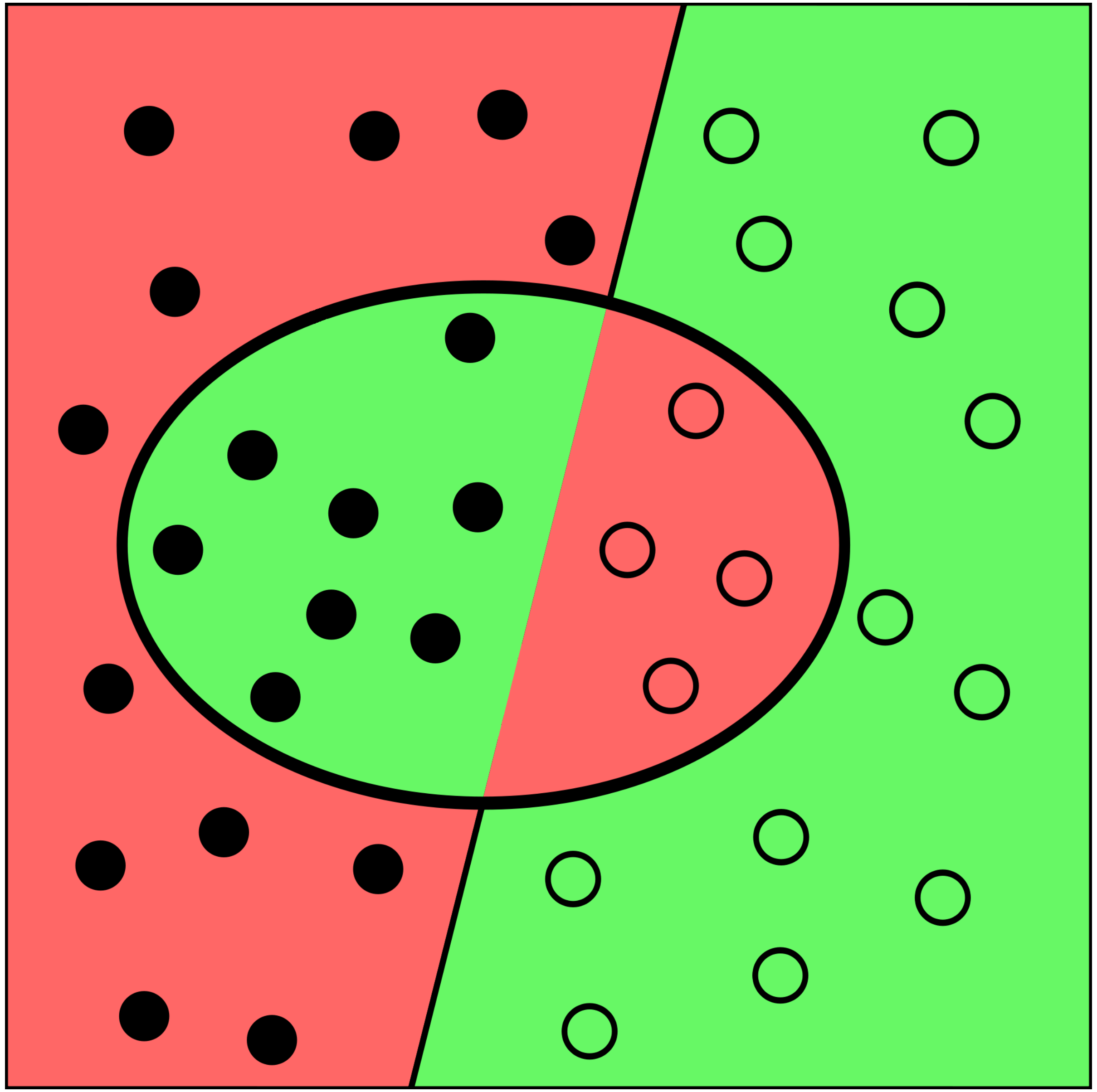
## ANOTHER EXAMPLE

- Imagine we have another marble problem where we consider green to be our positive class. The diagram to the right shows our results.
- Shaded circles represent correct predictions (e.g. green was predicted as green).
- Unshaded circles represent incorrect predictions (e.g. green was predicted red).



# ANOTHER EXAMPLE

- The background shows the color predicted.
- A shaded circle on a green background represents a green marble that was predicted as green.
- An unshaded circle on a red background represents a red marble that was predicted as green.



# ANOTHER EXAMPLE

- For this example, we would have the following confusion matrix.

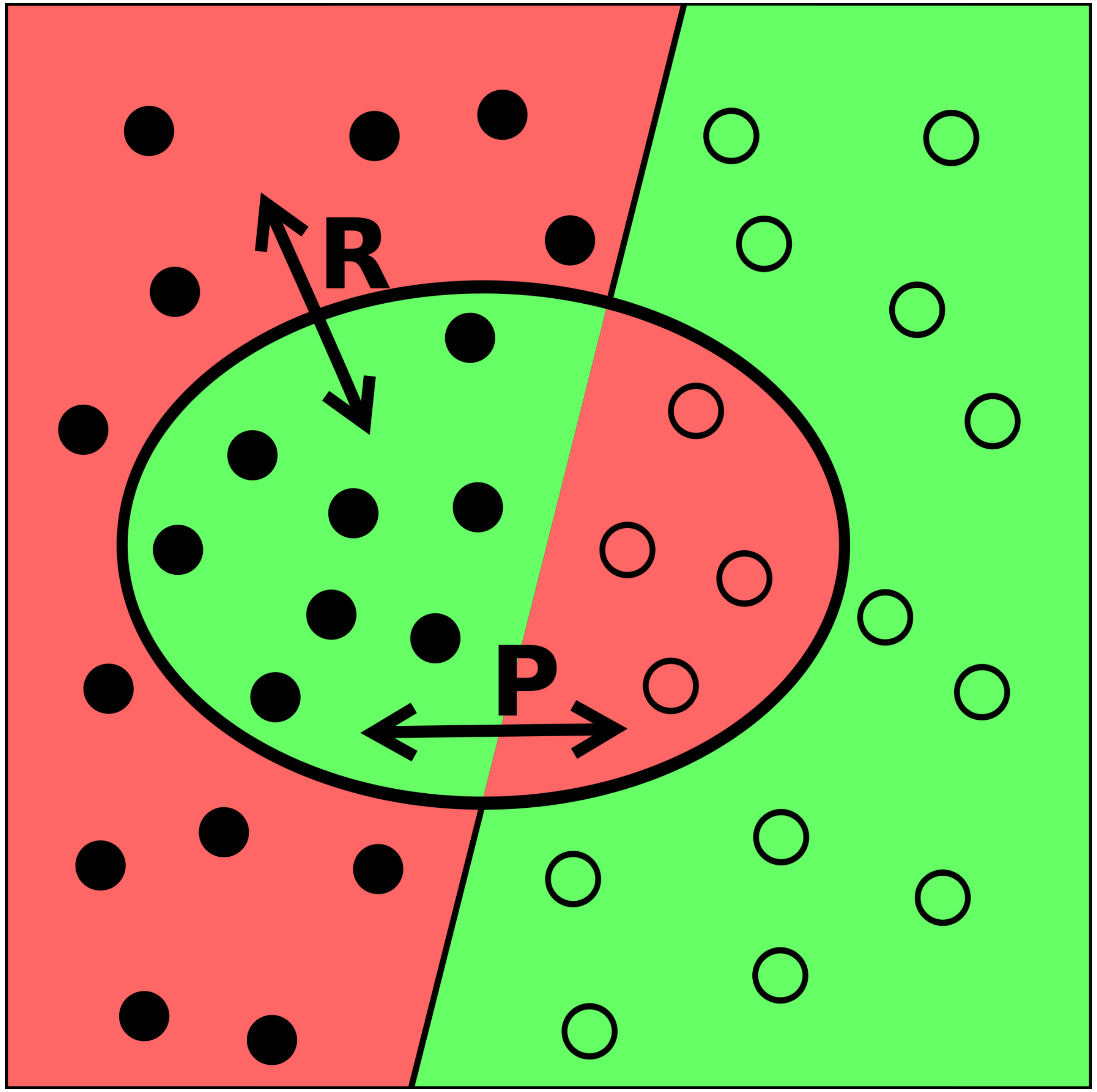
| | | True Class | |
|---|---|---|---|
| | | Green | Red |
| Predicted Class | Green | 8 | 4 |
| | Red | 12 | 12 |

- We could calculate precision for green marbles as 8 / (8 + 4) = 0.6666.

- We could calculate recall for green marbles as 8 / (8 + 12) = 0.4000.

# ANOTHER EXAMPLE

- We could update our diagram to reflect these calculations.
- Notice we don't talk about the red marbles predicted as green.
- We've chosen to focus on our model's accuracy as it relates to predicting green marbles.



INTRODUCTION

# ROC CURVE

# THE ROC CURVE

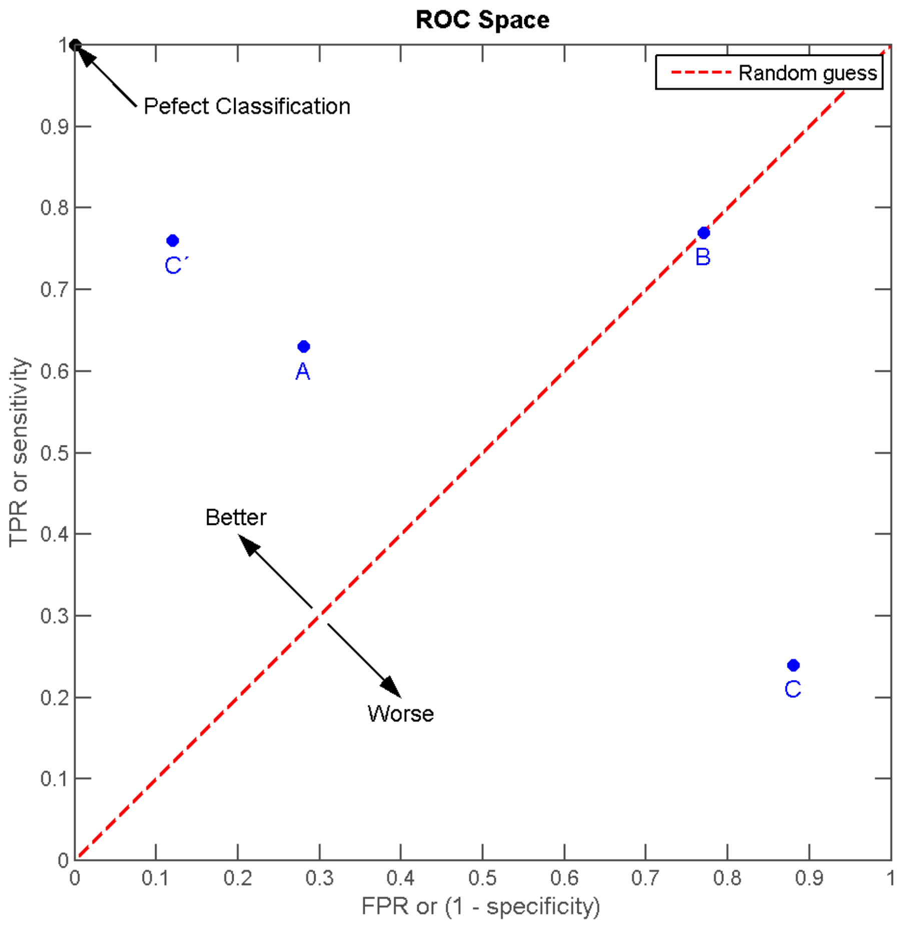
- This is where the **Receiver Operation Characteristic** (ROC) curve comes in handy.

- The curve is created by plotting the TPR against the FPR at various probability thresholds.
  - This is where we toggle our threshold of $p = 0.5$

- Area Under the Curve (AUC) summarizes the impact of TPR and FPR in a single value.

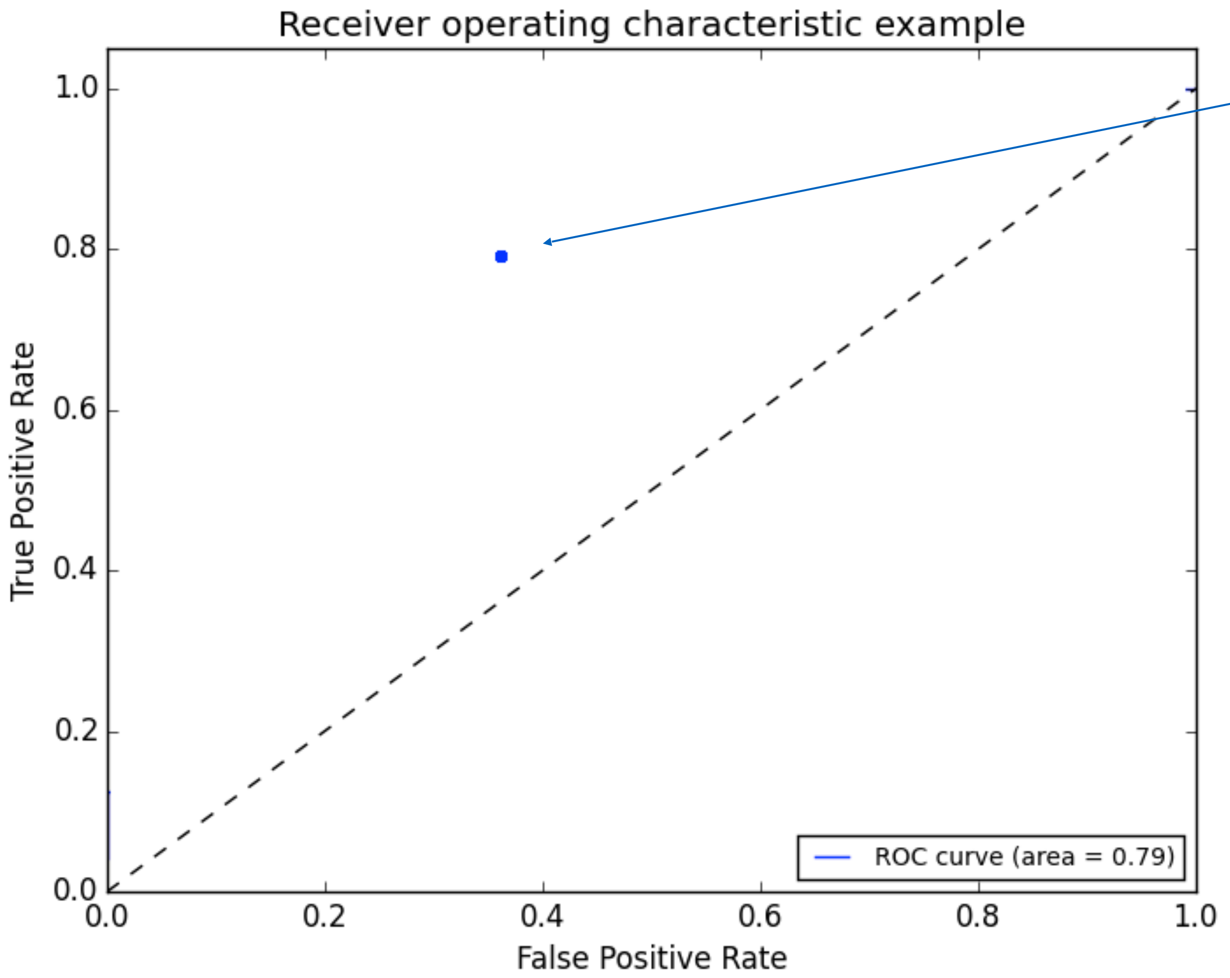
# THE ROC CURVE

- There can be a variety of points on an ROC curve.
- Each point is determined by a unique probability threshold



# THE ROC CURVE

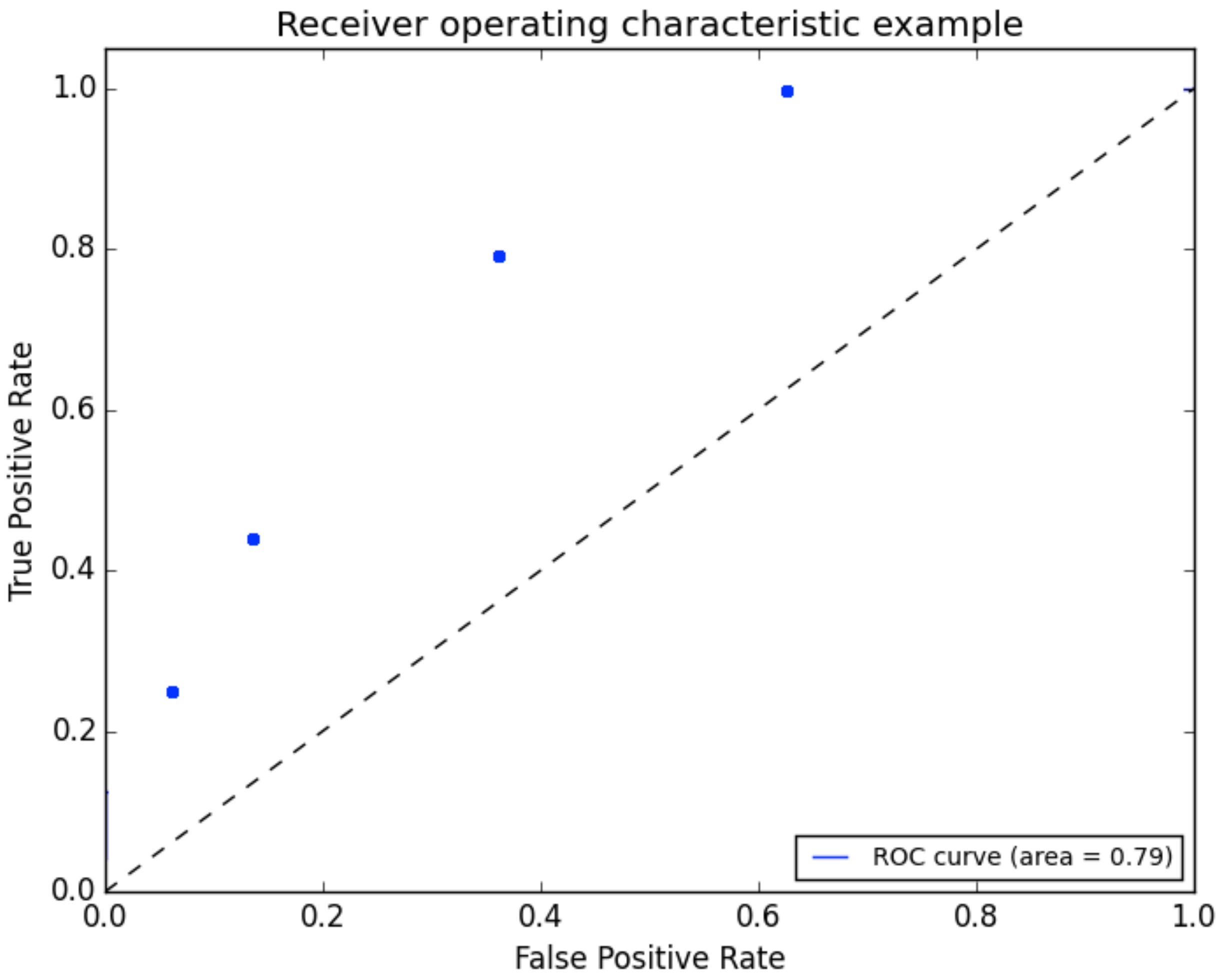
- We can begin by plotting an individual TPR/FPR pair for one threshold.



Let’s say this is for a default p = 0.5

# THE ROC CURVE

- We can continue adding pairs for different thresholds.

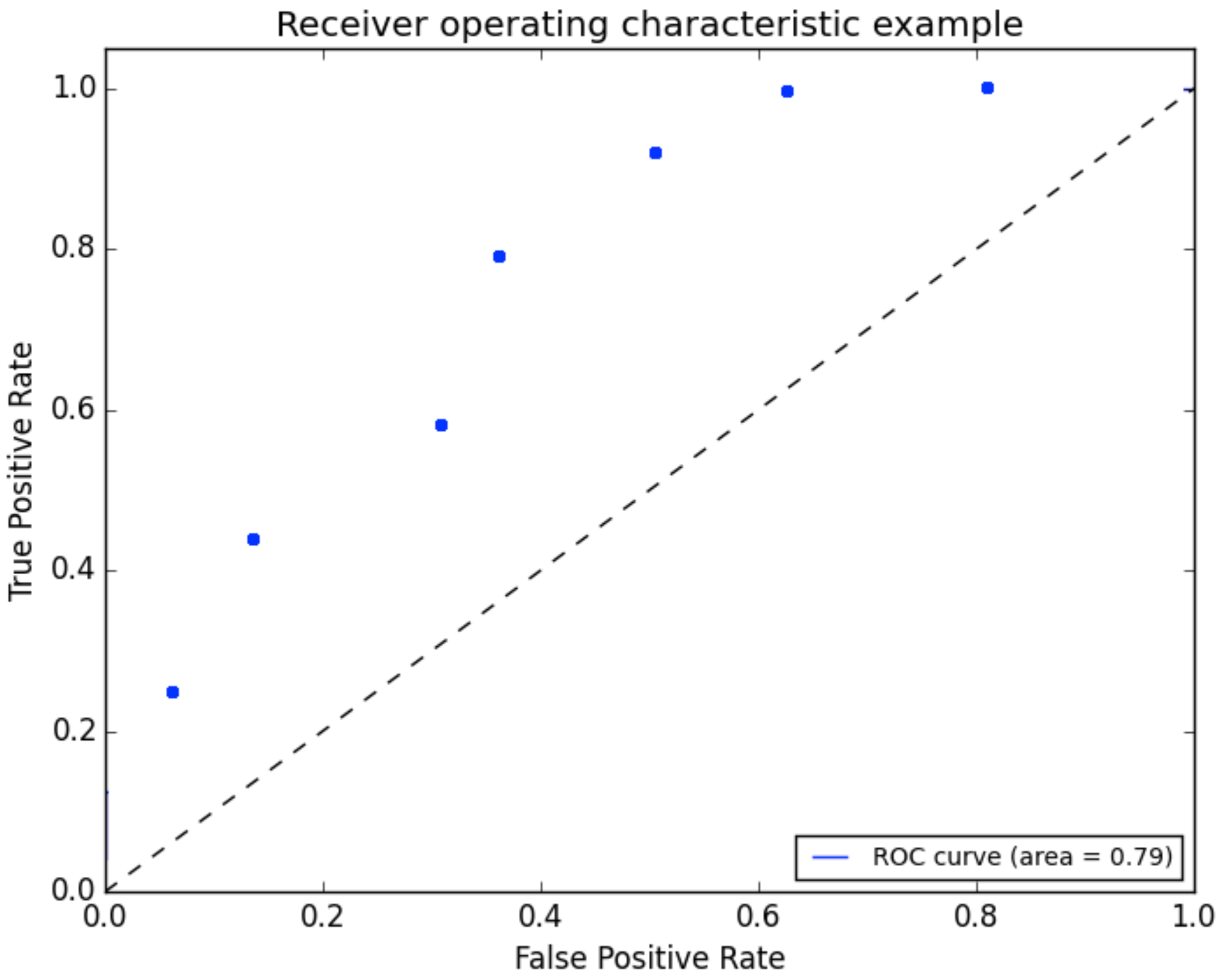


Let’s say we wanted to set a probability threshold of p > 0.9. AKA I’m only making a prediction of class ‘0’ when the model outputs a probability that’s > 0.9.

- Which point would best represent this case?

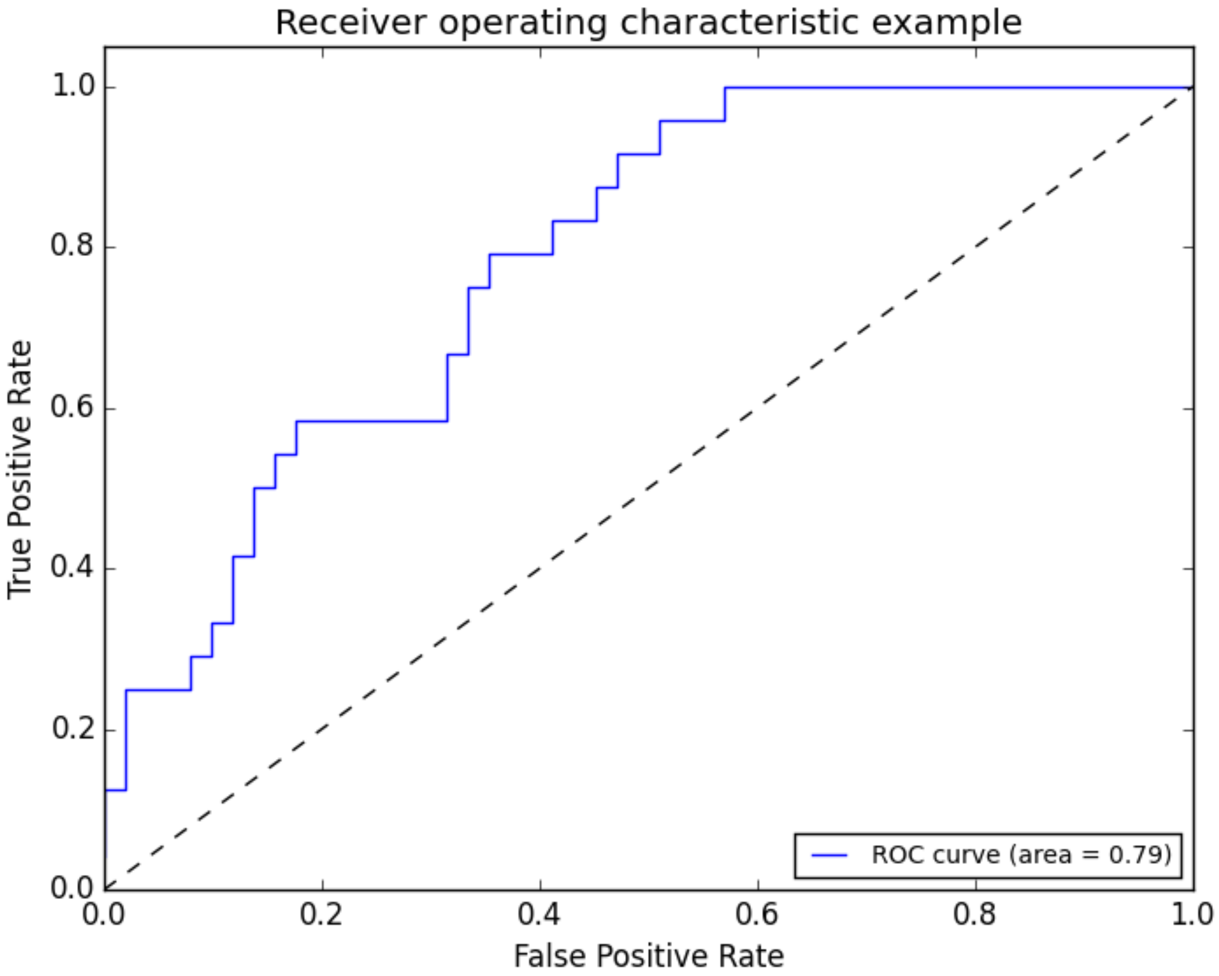
# THE ROC CURVE

- We can continue adding pairs for even more thresholds.



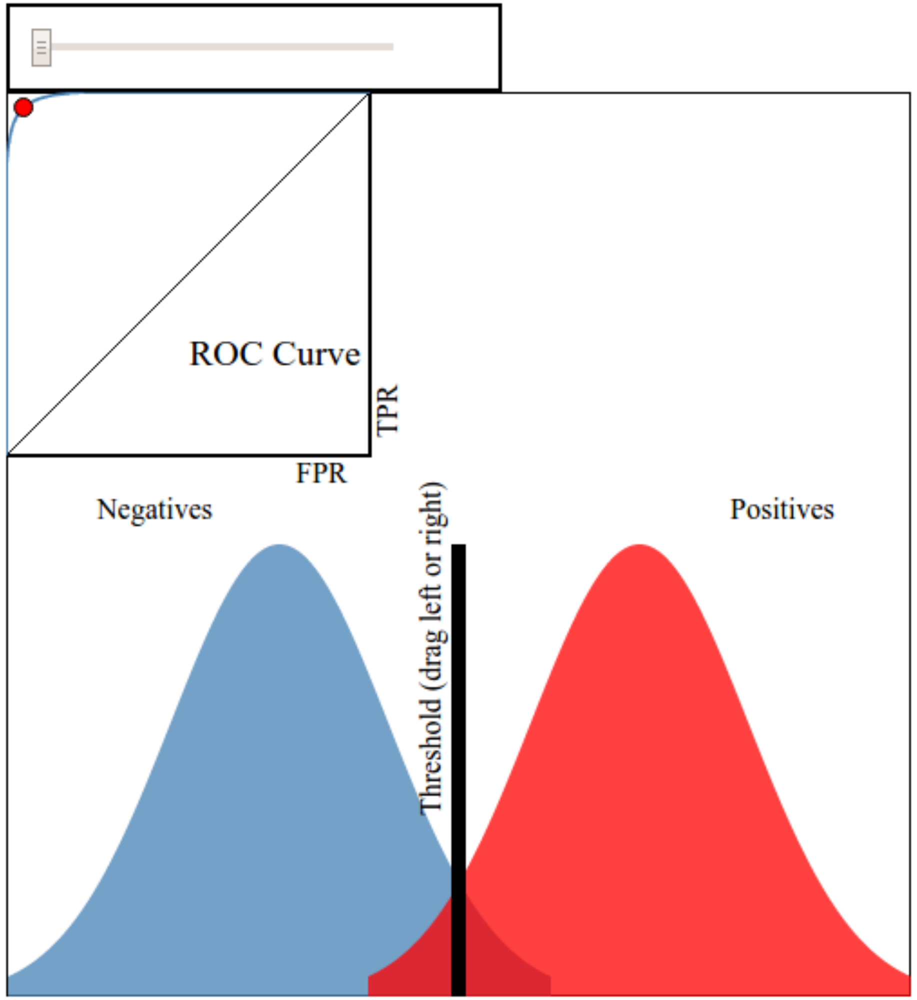
# THE ROC CURVE

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



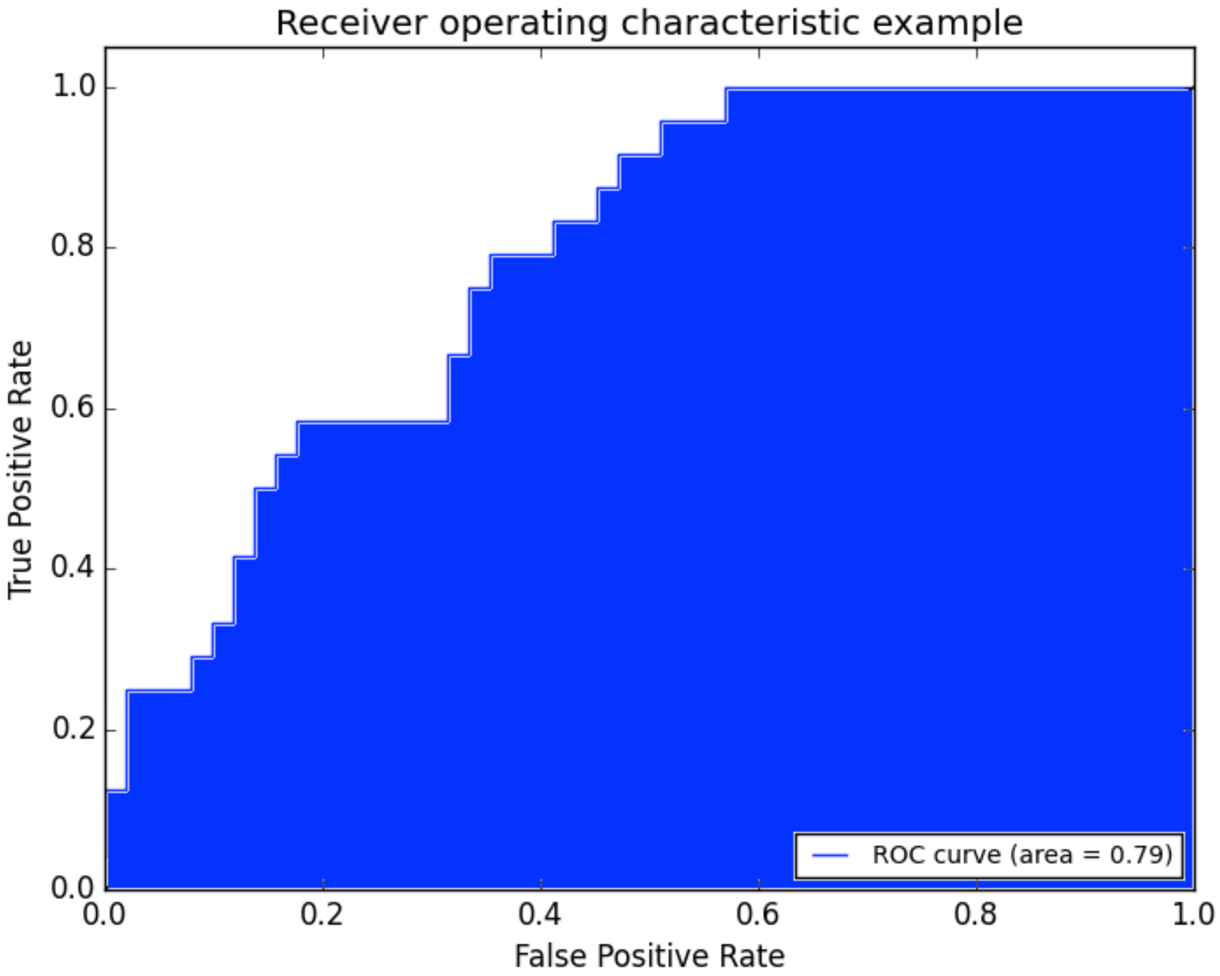
# THE ROC CURVE

- This interactive visualization can help practice visualizing ROC curves.



# AREA UNDER THE CURVE

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



➢ As you increase your sensitivity (true positives) and can identify more cases with a certain condition, you also **sacrifice** accuracy on identifying those without the condition (specificity).

CONCLUSION

# TOPIC REVIEW

# REVIEW QUESTIONS

- What's the link function used for in logistic regression?
- What criteria do we use to create our decision boundary?
- What do the *coefficients* in a logistic regression represent? How does the interpretation differ from ordinary least squares? How is it similar?
- How do we implement a multivariate logistic regression?
- How does True Positive Rate and False Positive Rate help explain accuracy?

Knowledge Review Section

# Can you answer these questions for mastery?

# CALCULATING ALL THREE

- **Probability** is the probability an event happens. For example, there might be an 80% chance of rain today.

- **Odds** (more technically the odds of success) is defined as probability of success/probability of failure. So the odds of a success (80% chance of rain) has an accompanying odds of failure (20% chance it doesn’t rain); as an equation (the “odds ratio“), that’s .8/.2 = 4.

- **Log odds** is the logarithm of the odds. Ln(4) = 1.38629436 ≅ 1.386.

# ACTIVITY: ~~NEVER~~ TELL ME THE ODDS

EXERCISE

## ANSWER THE FOLLOWING QUESTIONS

1. Given a standard deck of cards, calculate the probability and odds of obtaining the following cards

    - The 2 of clubs
    - Any diamond card
    - A face card (any J, Q, K)

2. What do each of these measures tell you about how likelihood of each scenario?

# KNOWLEDGE CHECK

EXERCISE

## DIRECTIONS (5 minutes)

Suppose we're trying to predict whether a widget is defective based on its manufacturing time in hours, $x_1$, and number of dongles, $x_2$. We run a Logistic Regression (where a "1" means the widget is defective and "0" means it is not defective) on a dataset and get the following sigmoid function:

$$\frac{1}{1+e^{-(2.5x_1-3.4x_2-4)}}$$

1. What's the probability of a defective dongle when the manufacturing time is 2 hours and there are 5 dongles?

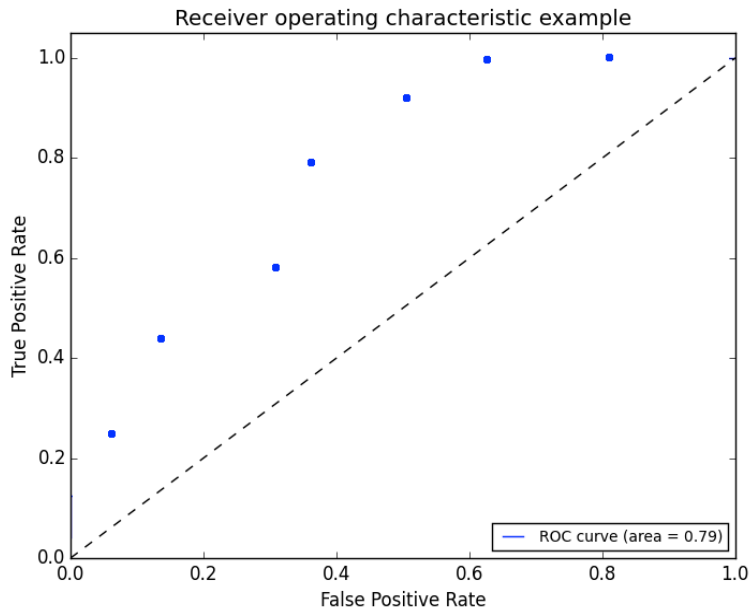
2. What happens to the odds of a widget being defective as we increase the number of dongles by 1 unit?

3. What happens to the odds of a widget being defective as we increase the manufacturing time by 1 unit?

# KNOWLEDGE CHECK

EXERCISE

1. What do all of the points on the ROC curve represent?



2. As we decrease our probability threshold, we (increase/decrease) our true positive rate and (increase/decrease) our false positive rate.

3. An AUC (Area under the curve) close to 0.5 means our model is (good/bad)?

4. An AUC (Area under the curve) close to 1 means our model is (good/bad)?

# ACTIVITY: KNOWLEDGE CHECK

EXERCISE

## ANSWER THE FOLLOWING QUESTIONS

1. What would the precision and recall be for the following confusion matrix (with "green" being "true")?

| | **predicted_green** | **predicted_not_green** |
|---|---|---|
| is_green | 13 | 7 |
| is_not_green | 8 | 12 |

## DELIVERABLE

Answers to the above question

# ACTIVITY

EXERCISE

For each of the example:

1. Write a confusion matrix: true positive, false positive, true negative, false negative. Then decide what each square represents for that specific example.

**Examples**:

1. A test is developed for determining if a patient has cancer or not.

1. A newspaper company is targeting a marketing campaign for "at risk" users that may stop paying for the product soon.

1. You build a spam classifier for your email system.