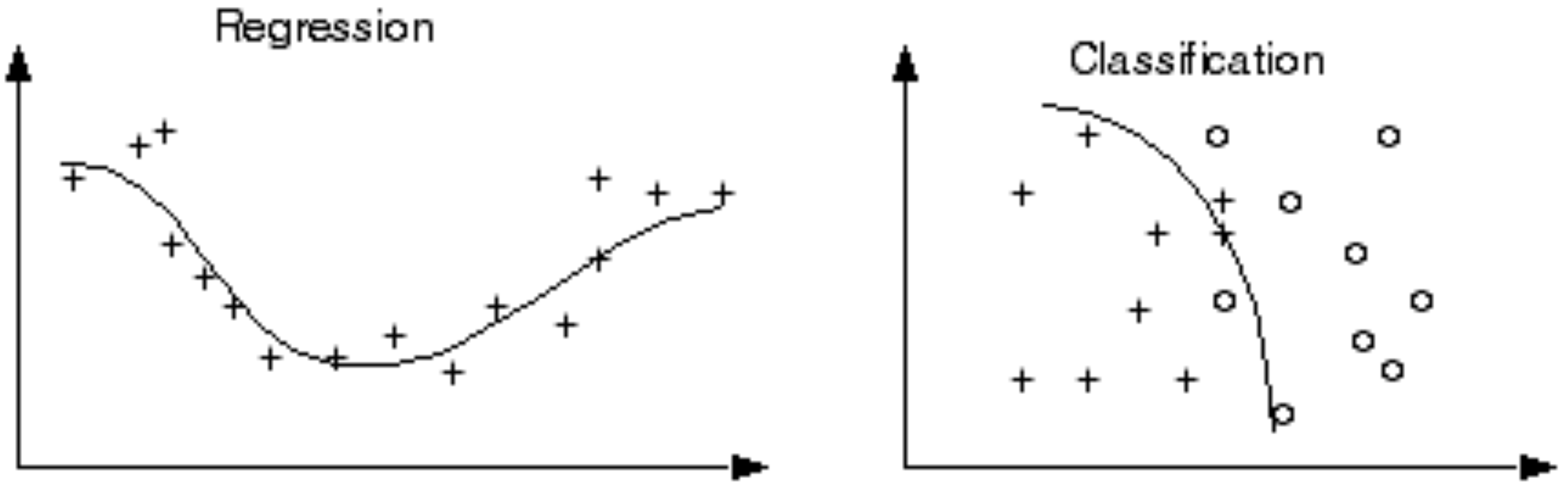
DSI

LOGISTIC REGRESSION

LINEAR REGRESSION RESULTS FOR CLASSIFICATION

- Regression results can have a value range from -∞ to ∞.
- Classification is used when predicted values (i.e. class labels) are not greater than or less than each other.



LINEAR REGRESSION RESULTS FOR CLASSIFICATION

- But, since many classification problems are binary (0 or 1) and 1 is greater than 0, does it make sense to apply the concept of regression to solve classification?
  - Output isn't normal
  - Predictions can be outside of [0,1], which violates laws of probability
  - Probability can also be U-shaped: Flu by Age
- How might we contain those bounds?
- Let's review some "fixes" to make classification with regression feasible.

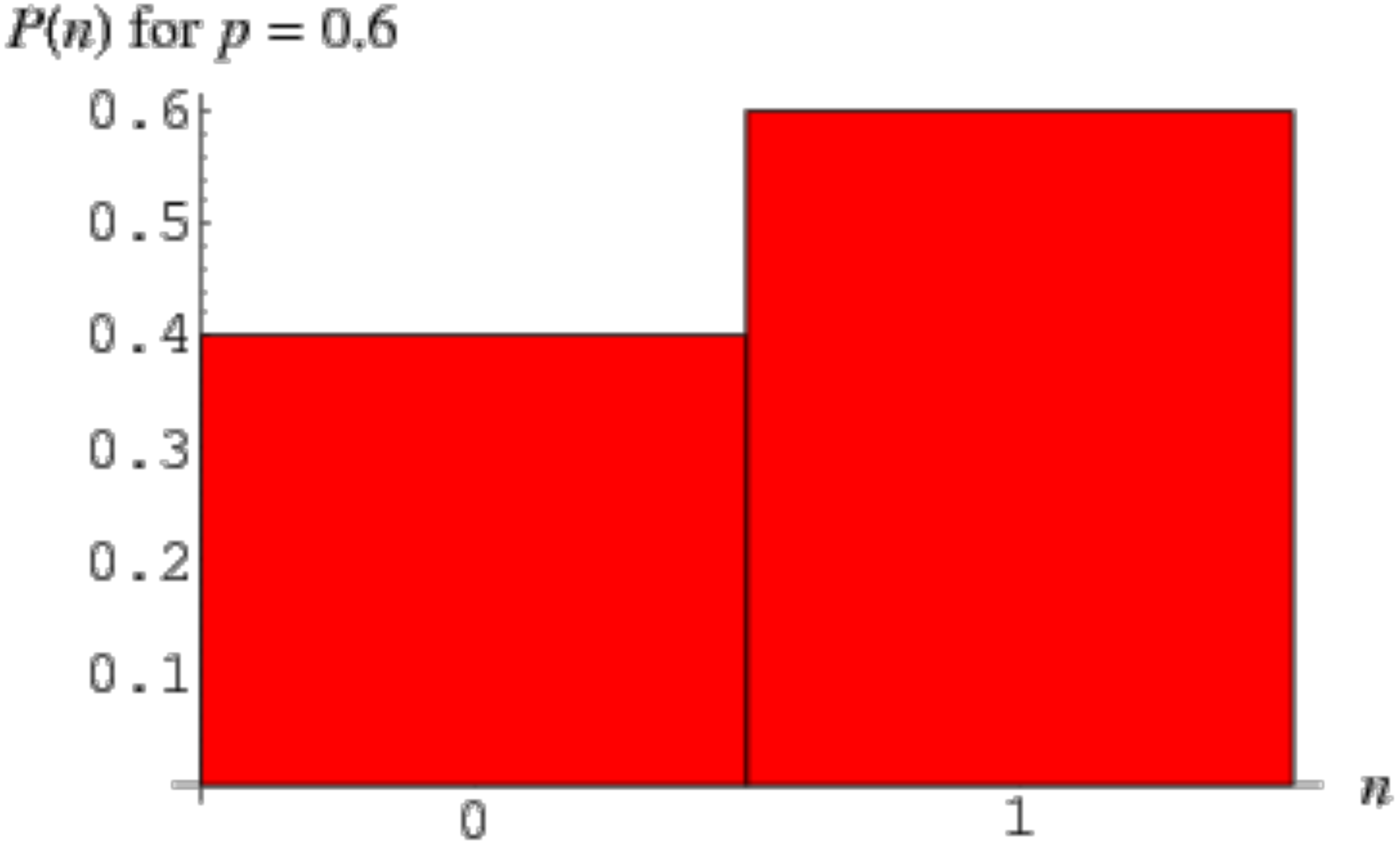
FIX 1: PROBABILITY

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

- We could assume the *prior probability* (the *bias*) of a class is the class distribution.

FIX 1: PROBABILITY

- Bernoulli Distribution



FIX 1: PROBABILITY

- For example, suppose we know that roughly 700 of 2200 people from the Titanic survived. Without knowing anything about the passengers or crew, the probability of survival would be ~0.32 (32%).

- However, we still need a way to use a linear function to either increase or decrease the probability of an observation given the data about it.

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

FIX 2: LINK FUNCTIONS

- Another advantage to OLS is that it allows for *generalized* models using a *link function*.
- Link functions allows us to build a relationship between a linear function and the mean of a distribution.
- We can now form a specific relationship between our linear predictors and the response variable.

FIX 2: LINK FUNCTIONS

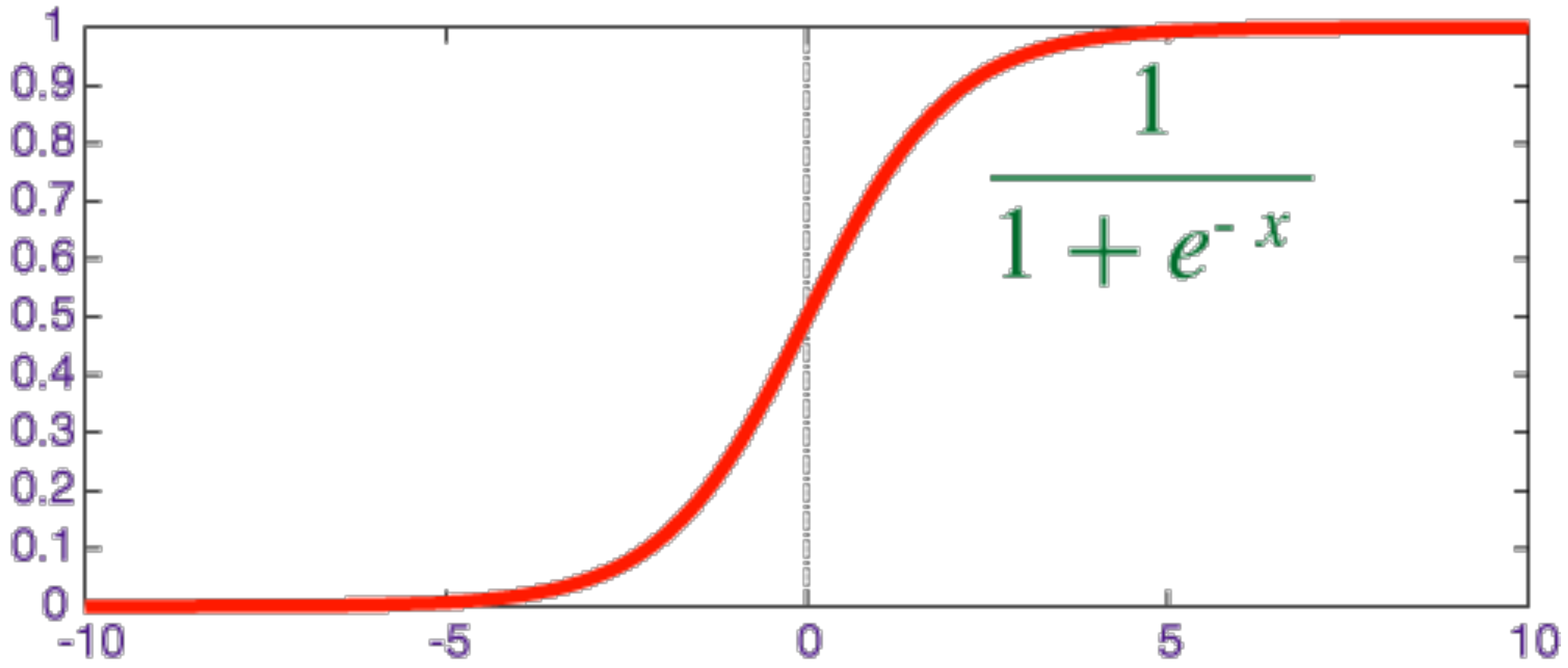
- For classification, we need a distribution associated with categories: given all events, what is the probability of a given event?

- The link function that best allows for this is the *logit* (/ˈloʊdʒɪt/ LOH-jit) function

  - Inverse of the *sigmoid* function; this is fortuitous.

THE SIGMOID FUNCTION

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

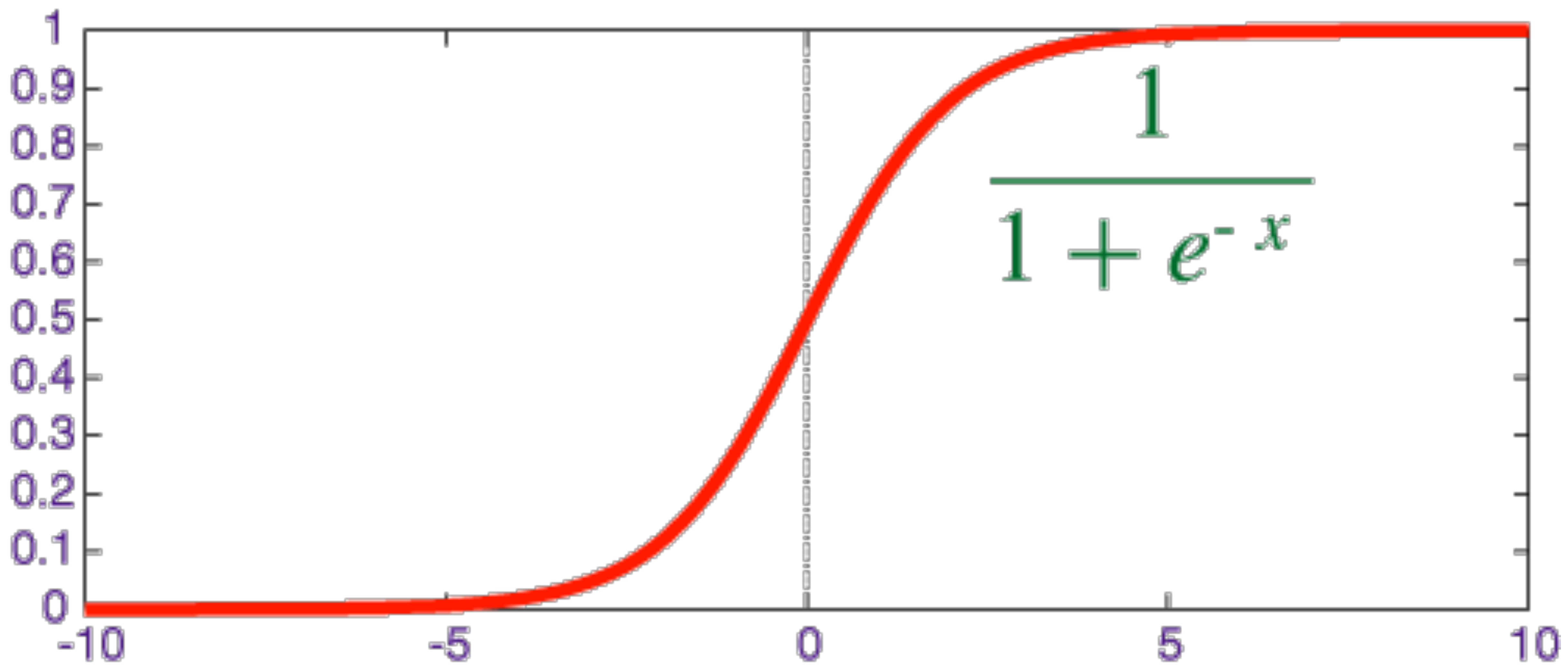


- Mathematically, it is defined as

$$f(x) = \frac{1}{1 + \mathrm{e}^{-x}}$$

THE SIGMOID FUNCTION

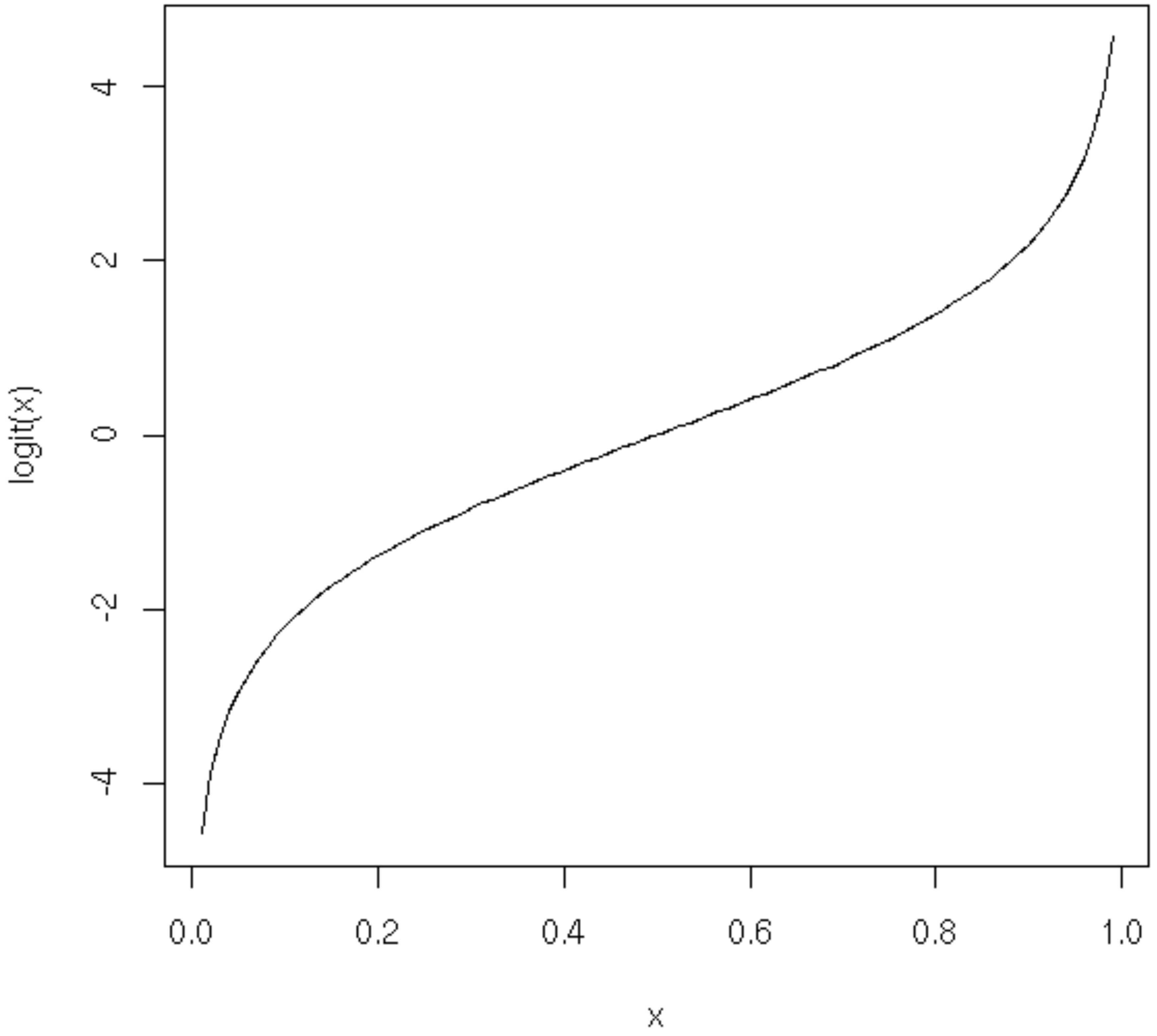
- Recall that e is the *inverse* of the natural log.
- As x increases, the results is closer to 1. As x decreases, the result is closer to 0.
- When x = 0, the result is 0.5.



FIX 3: ODDS AND LOG-ODDS

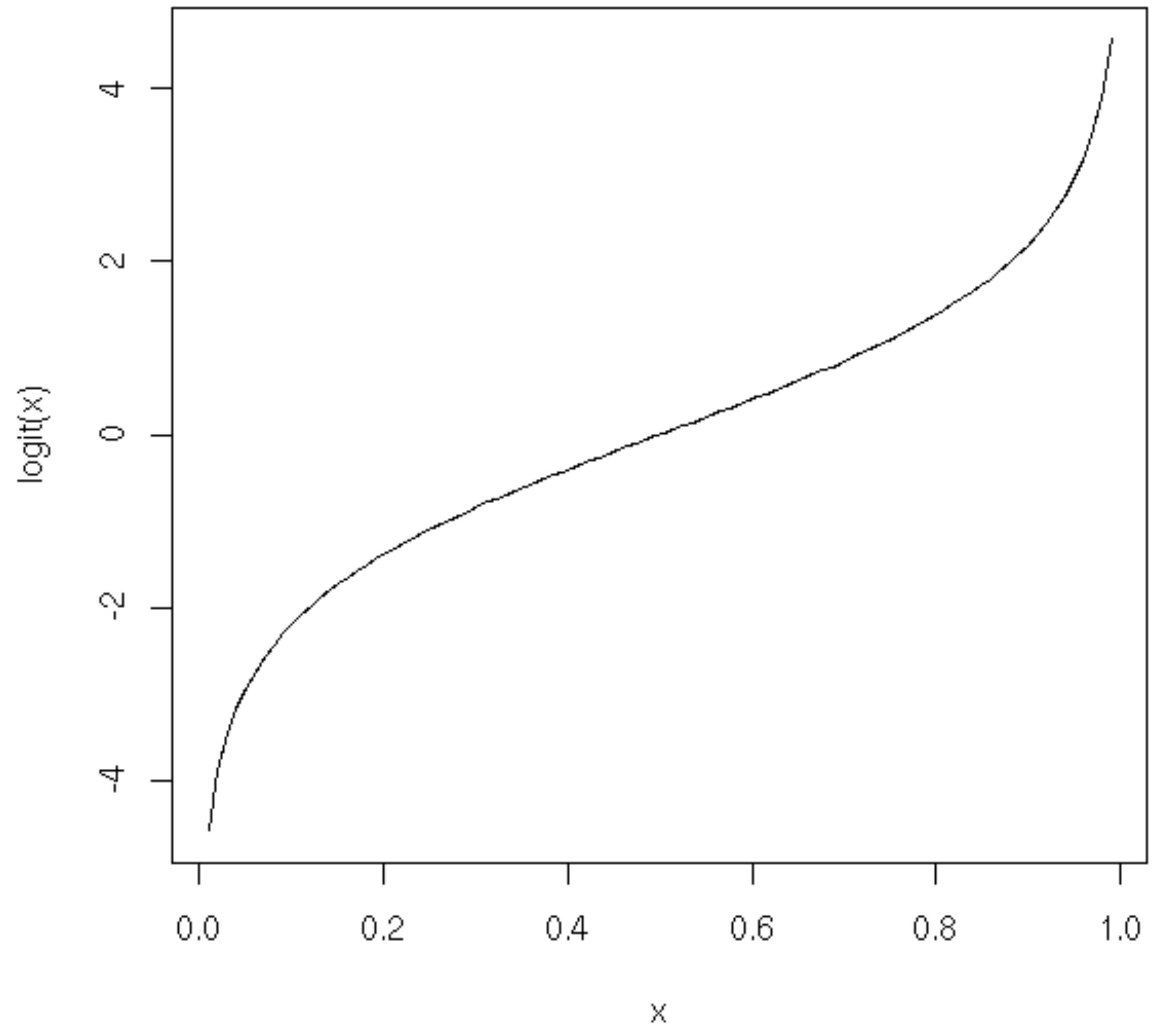
- This will act as our *link* function for logistic regression.
- Mathematically, the logit function is defined as

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



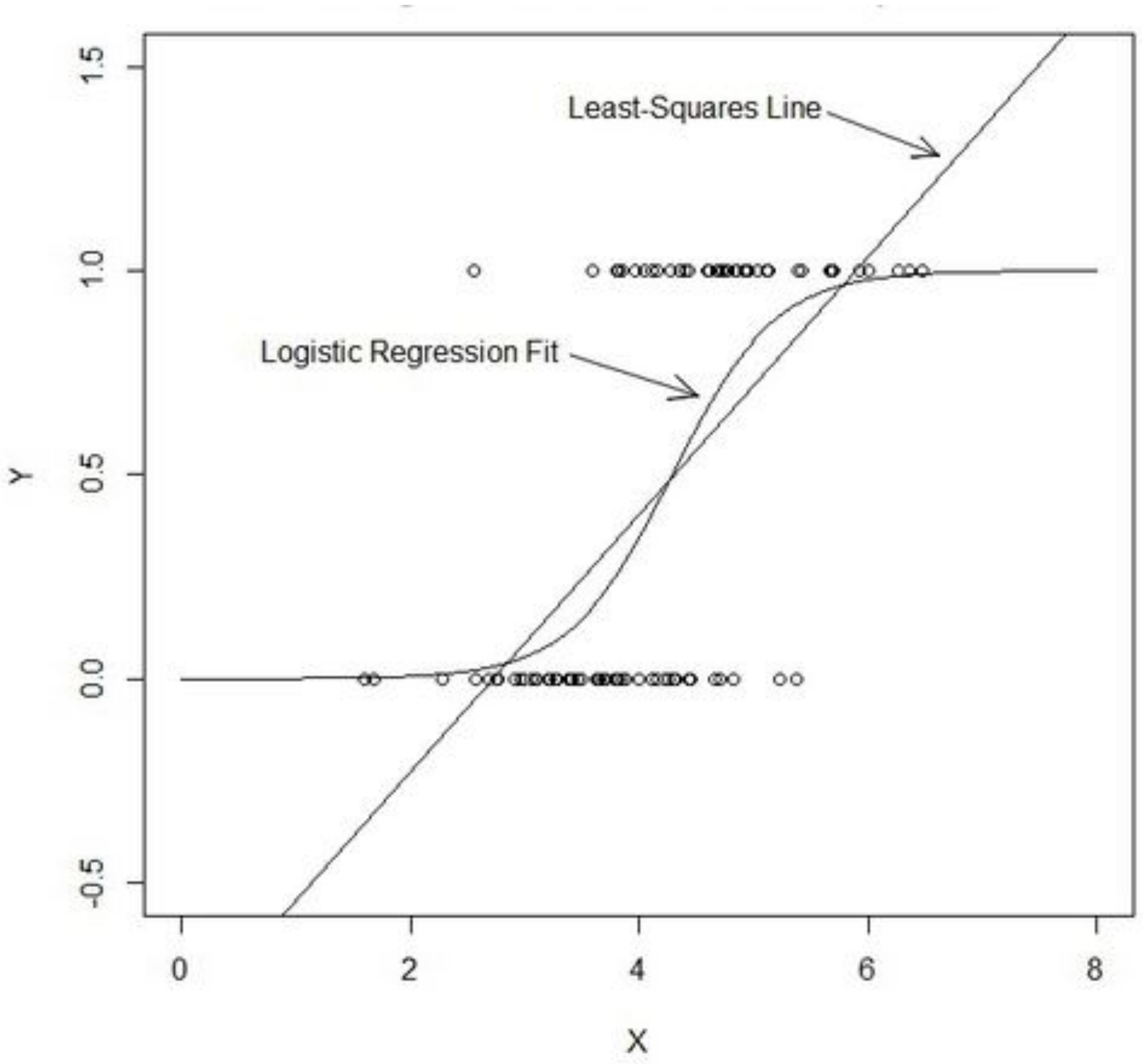
FIX 3: ODDS AND LOG-ODDS

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



FIX 3: ODDS AND LOG-ODDS

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



FIX 3: ODDS AND LOG-ODDS

- While the logit value represents the *coefficients* in the logistic function, we can convert them into odds ratios that make them more easily interpretable.

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

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

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

FIX 3: ODDS AND LOG-ODDS

The natural logarithm of the odds ratio is equivalent to a *linear* function of the independent variables. The antilog of the logit function allows us to find the estimated regression equation.

$$\text{logit}\,(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1$$

antilog

$$\frac{p}{1-p} = e^{\beta_0+\beta_1 x_1}$$

$$p = e^{\beta_0+\beta_1 x_1}(1-p)$$

$$p = e^{\beta_0+\beta_1 x_1} - e^{\beta_0+\beta_1 x_1} * p$$

$$p + e^{\beta_0+\beta_1 x_1} * p = e^{\beta_0+\beta_1 x_1}$$

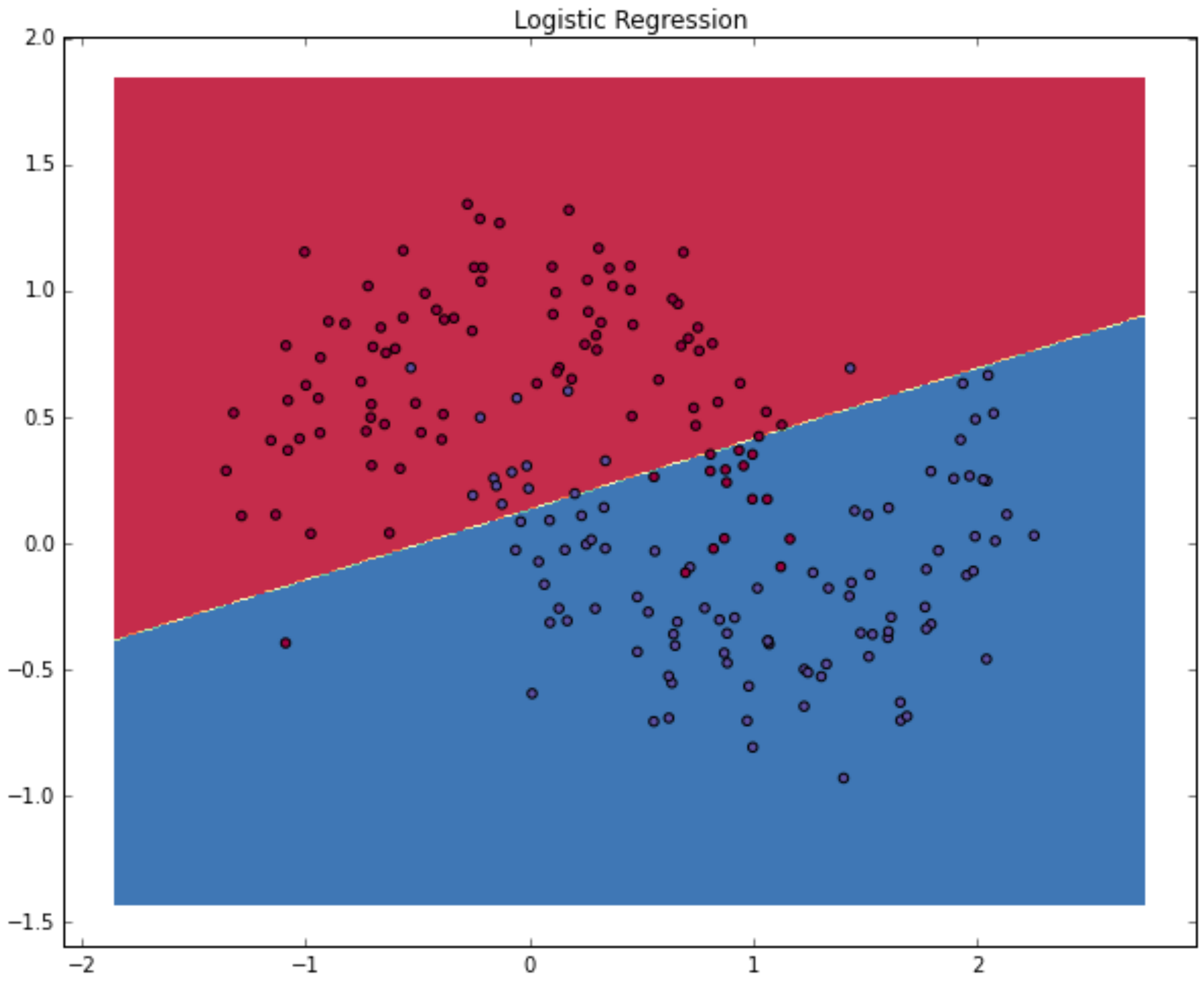
$$p(1 + e^{\beta_0+\beta_1 x_1}) = e^{\beta_0+\beta_1 x_1}$$

$$\hat{p} = \frac{e^{\beta_0+\beta_1 x_1}}{1 + e^{\beta_0+\beta_1 x_1}}$$

Estimated Regression Equation

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.



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.

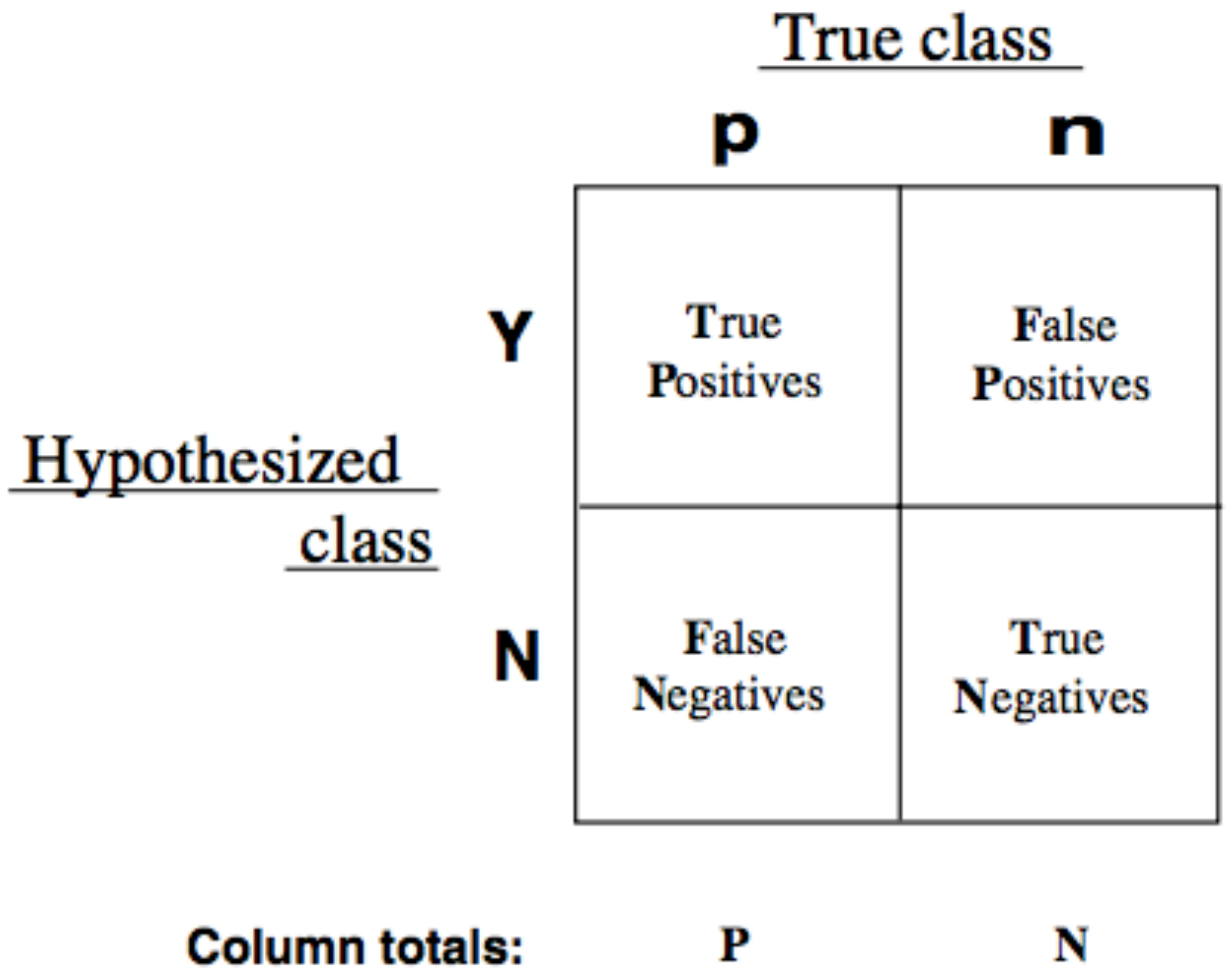
ADVANCED CLASSIFICATION METRICS

- 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*.

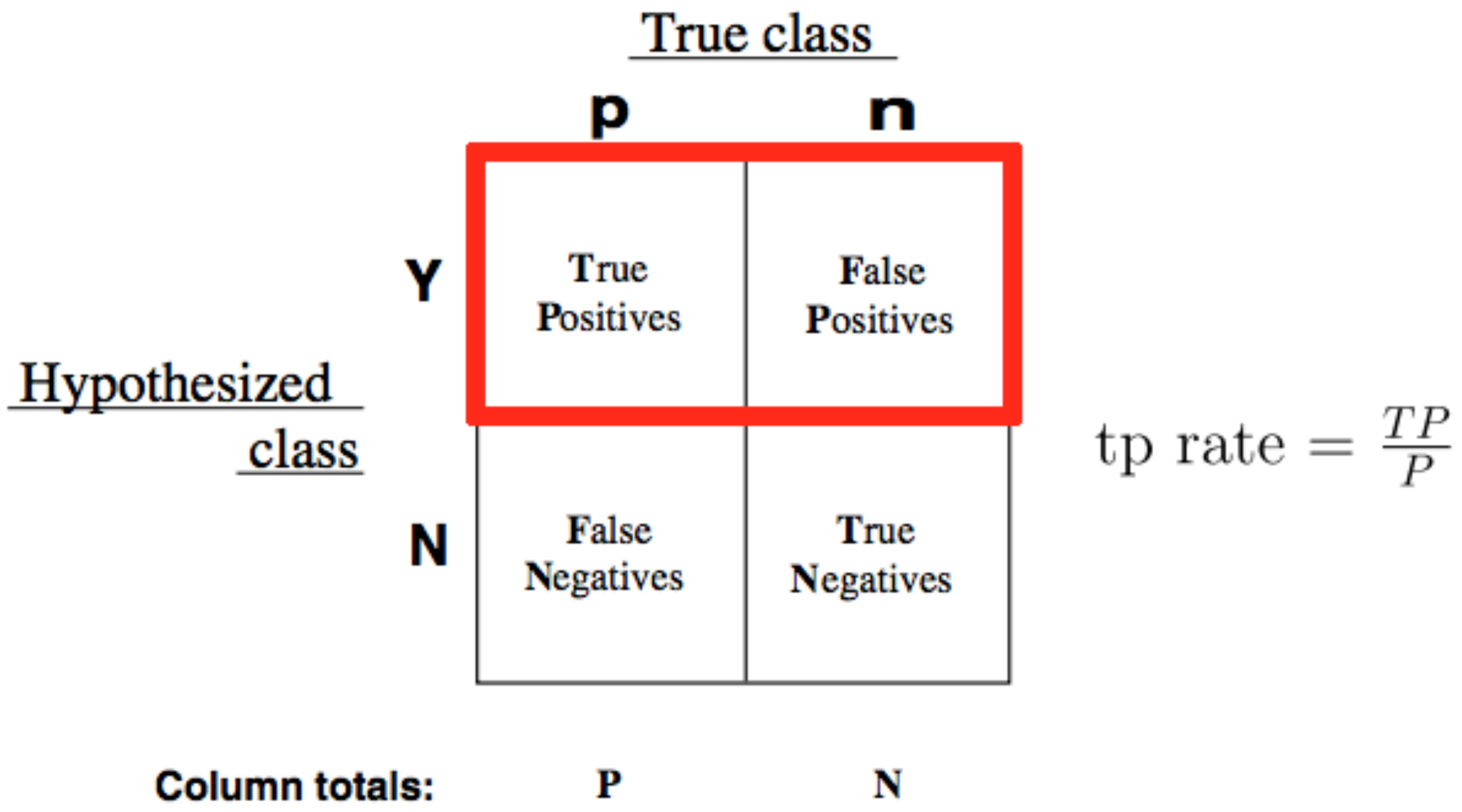
- For each label, we can put it into the category of a true positive, false positive, true negative, or false negative.



ADVANCED CLASSIFICATION METRICS

- True Positive Rate (TPR) asks, “Out of all cases in the target class label, how many were accurately predicted to belong to that class?”

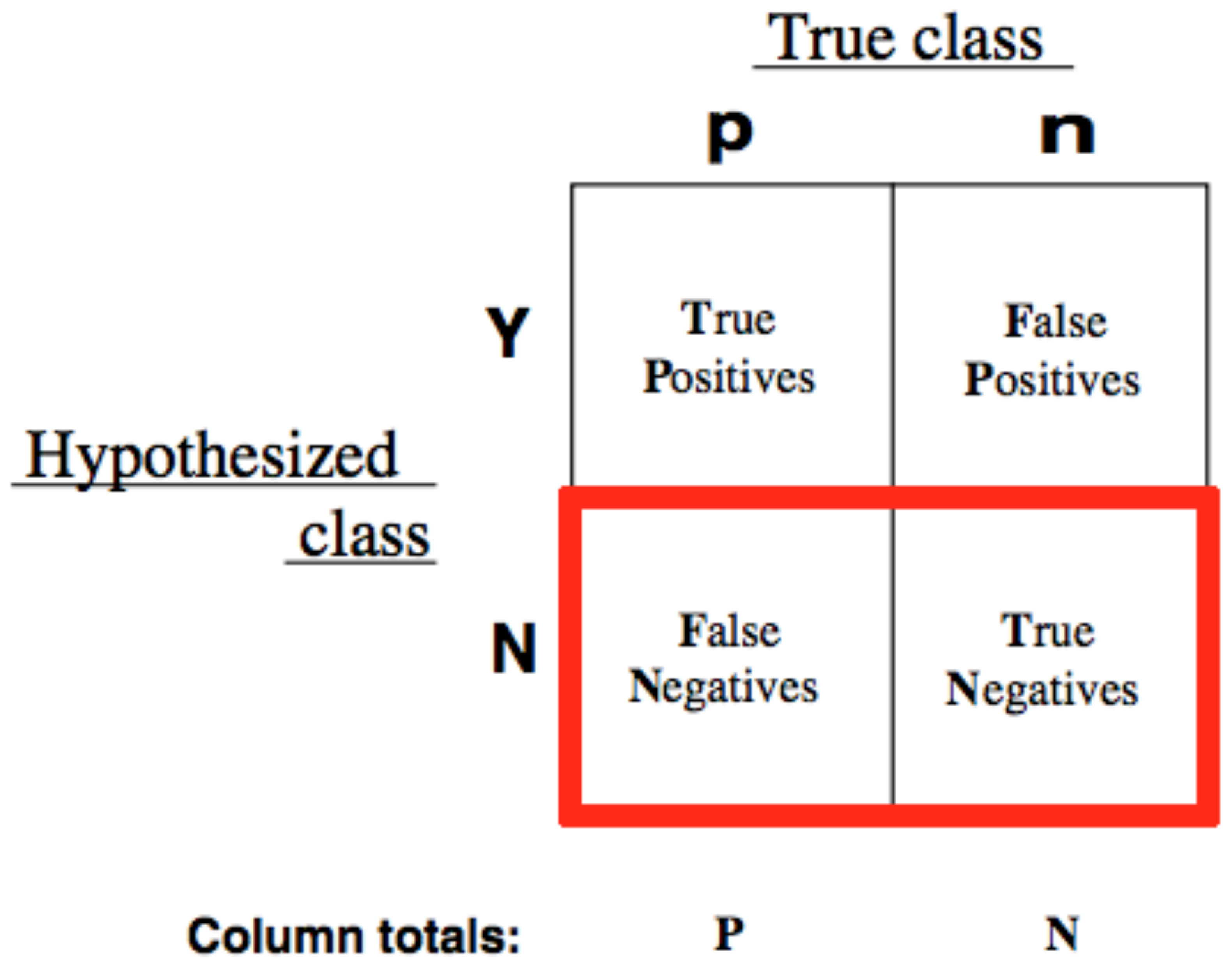
- For example, given a medical exam that tests for cancer, how often does it correctly identify patients with cancer?



ADVANCED CLASSIFICATION METRICS

- False Positive Rate (FPR) asks, “Out of all cases not belonging to a class label, how many were predicted as belonging to that target class label?”

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



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

<table>
<tr><td></td><td></td><td colspan="2">Predicted condition</td><td></td><td></td></tr>
<tr><td></td><td>Total population</td><td>Predicted Condition positive</td><td>Predicted Condition negative</td><td>Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$</td><td></td></tr>
<tr><td rowspan="2">True condition</td><td>condition positive</td><td>True positive</td><td>False Negative (Type II error)</td><td>True positive rate (TPR), Sensitivity, Recall $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$</td><td>False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$</td></tr>
<tr><td>condition negative</td><td>False Positive (Type I error)</td><td>True negative</td><td>False positive rate (FPR), Fall-out $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$</td><td>True negative rate (TNR), Specificity (SPC) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$</td></tr>
<tr><td></td><td rowspan="2">Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$</td><td>Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Test outcome positive}}$</td><td>False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Test outcome negative}}$</td><td>Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$</td><td rowspan="2">Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR}-}$</td></tr>
<tr><td></td><td>False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Test outcome positive}}$</td><td>Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Test outcome negative}}$</td><td>Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$</td></tr>
</table>

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 a smoking problem, this model would accurately predict *all* smokers as smokers and not accidentally predict any of the nonsmokers as smokers

ADVANCED CLASSIFICATION METRICS

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

ADVANCED CLASSIFICATION METRICS

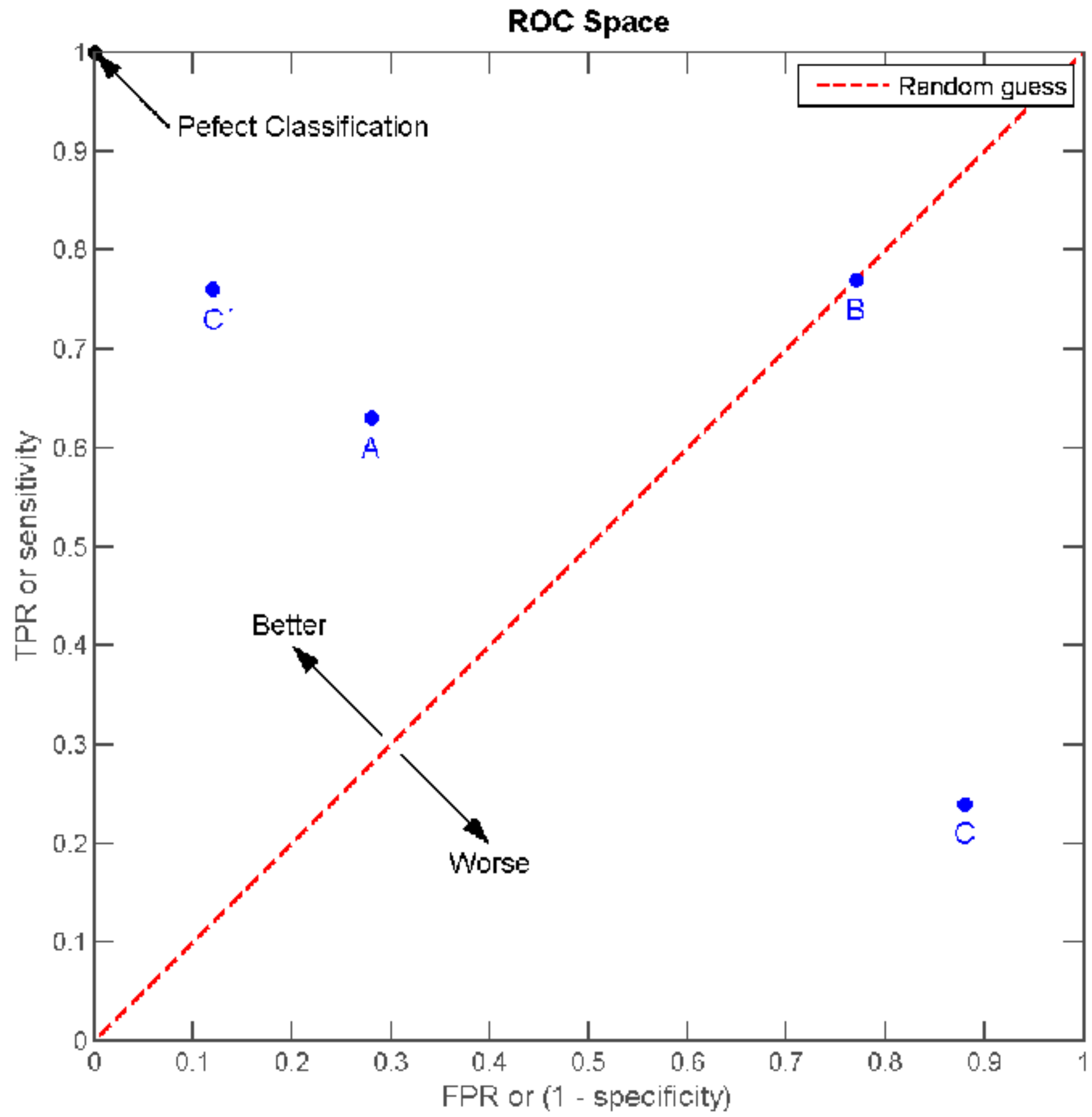
- This is where the Receiver Operation Characteristic (ROC) curve comes in handy.

- The curve is created by plotting the true positive rate against the false positive rate at various model threshold settings.

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

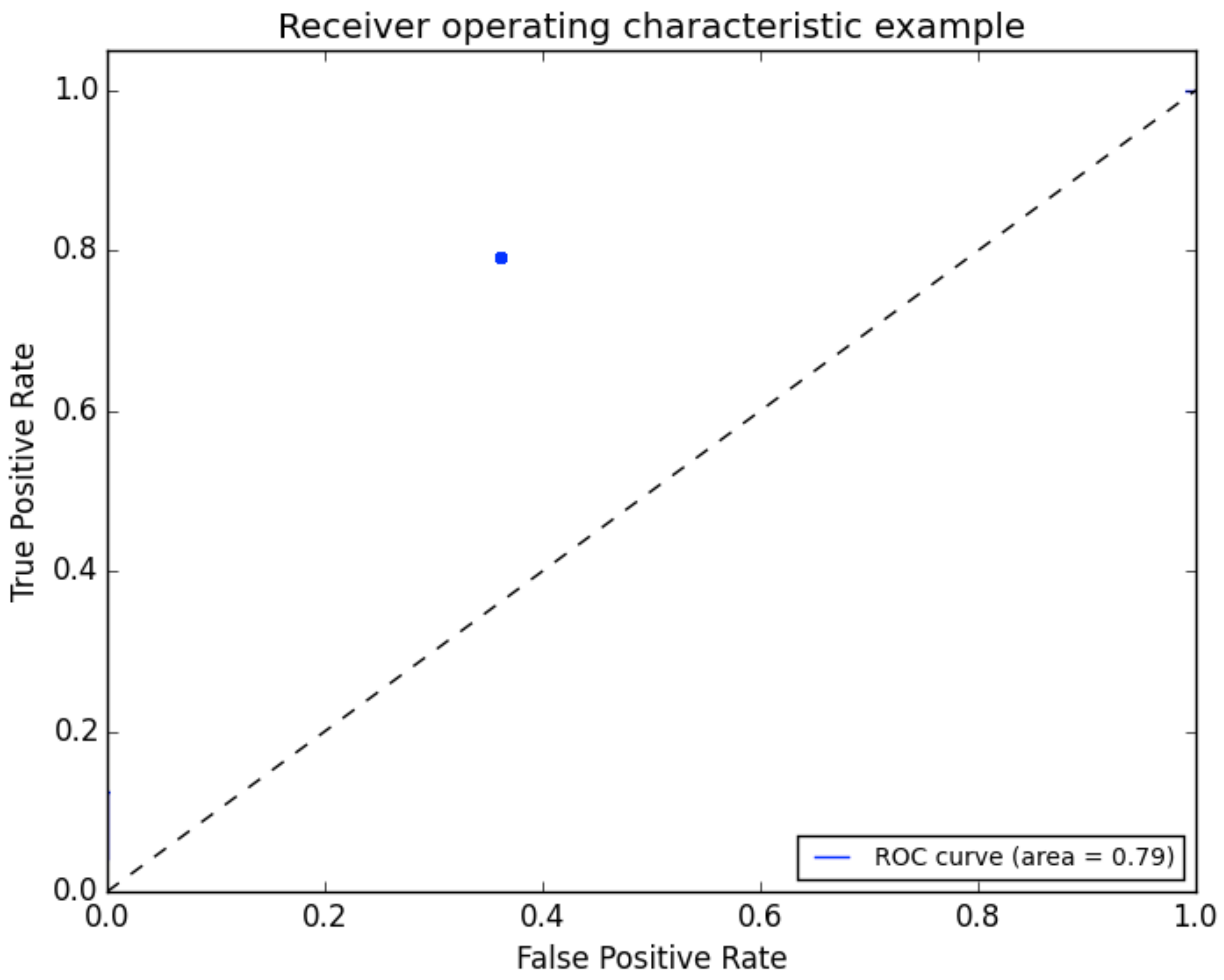
ADVANCED CLASSIFICATION METRICS

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



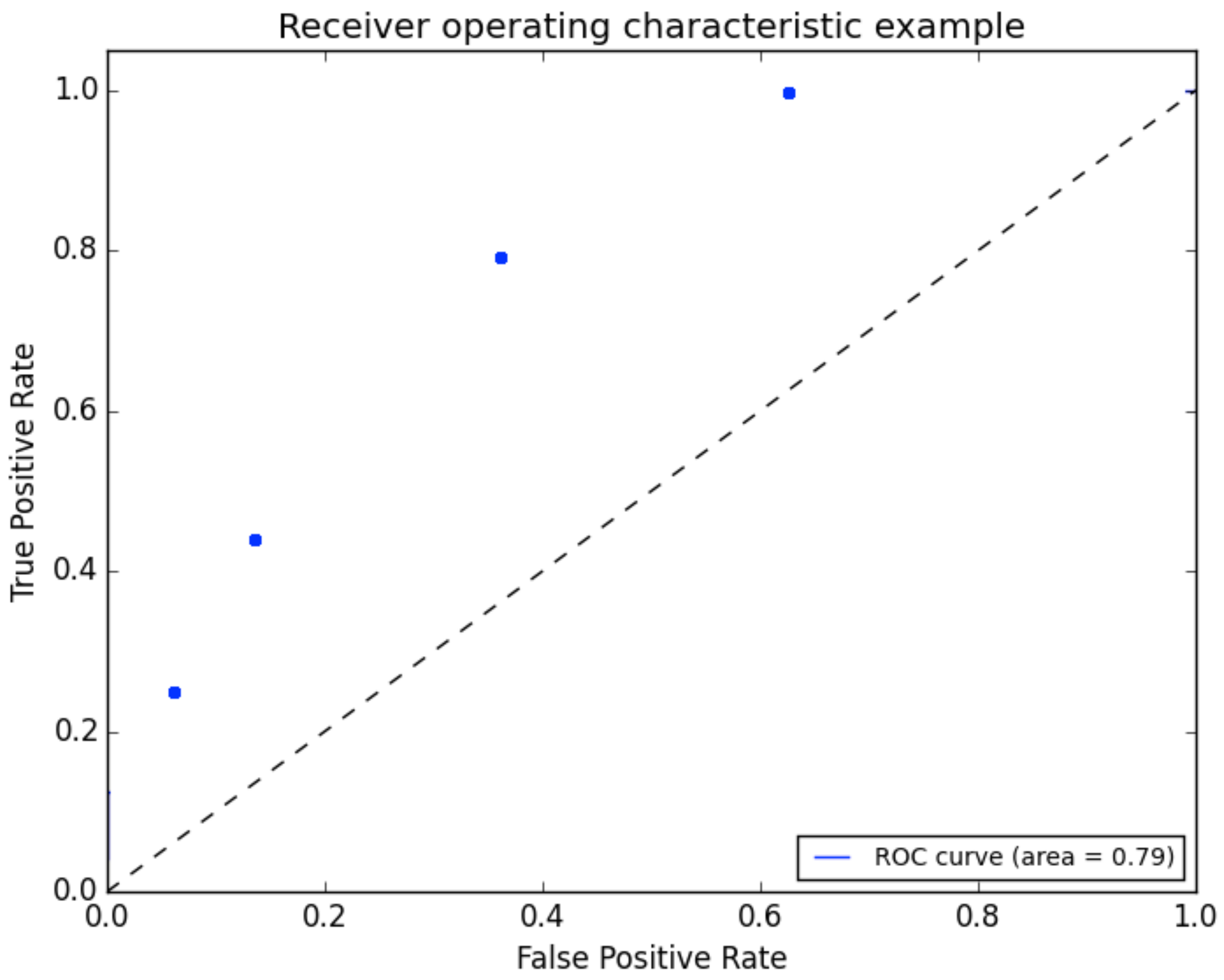
ADVANCED CLASSIFICATION METRICS

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



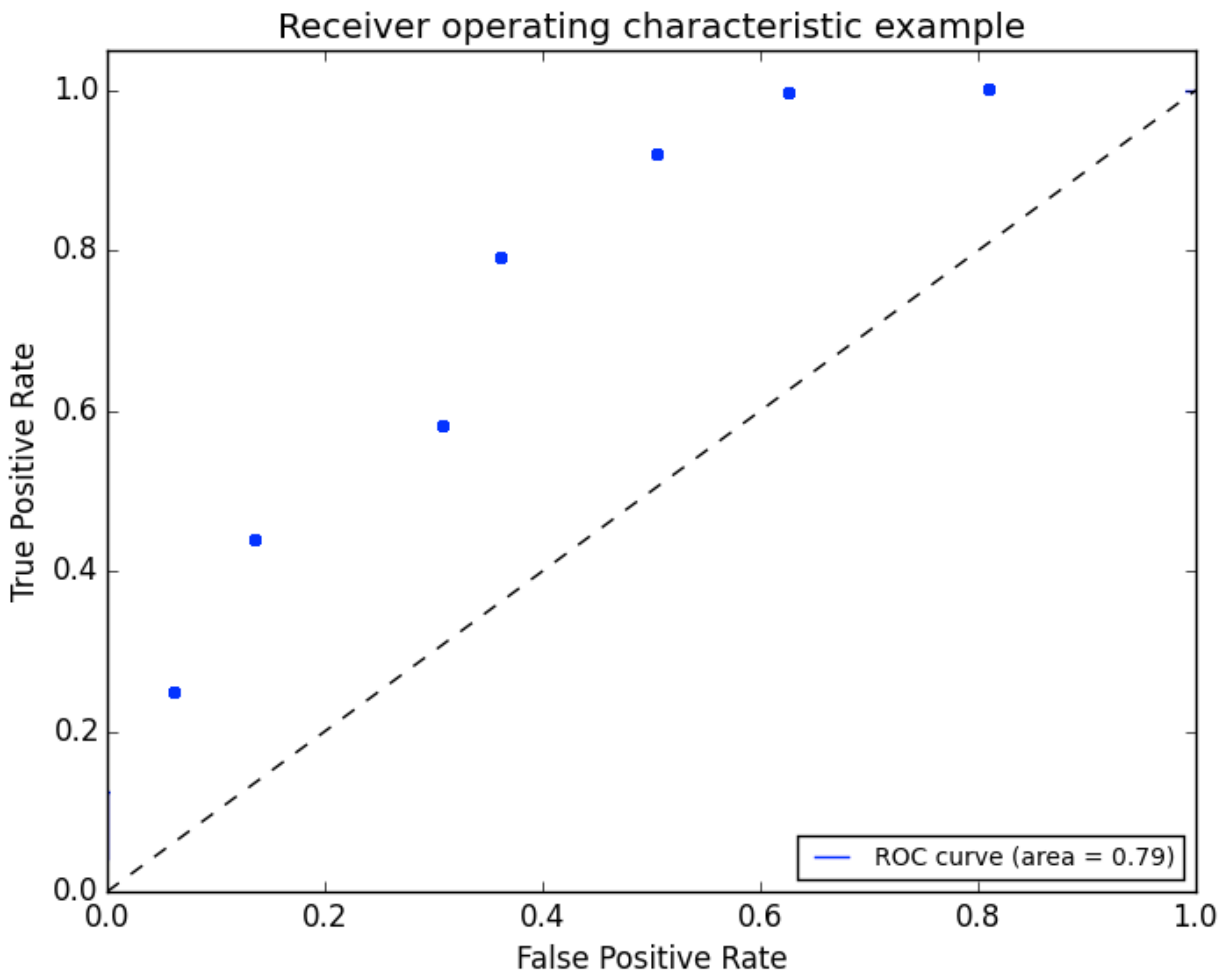
ADVANCED CLASSIFICATION METRICS

- We can continue adding pairs for different thresholds



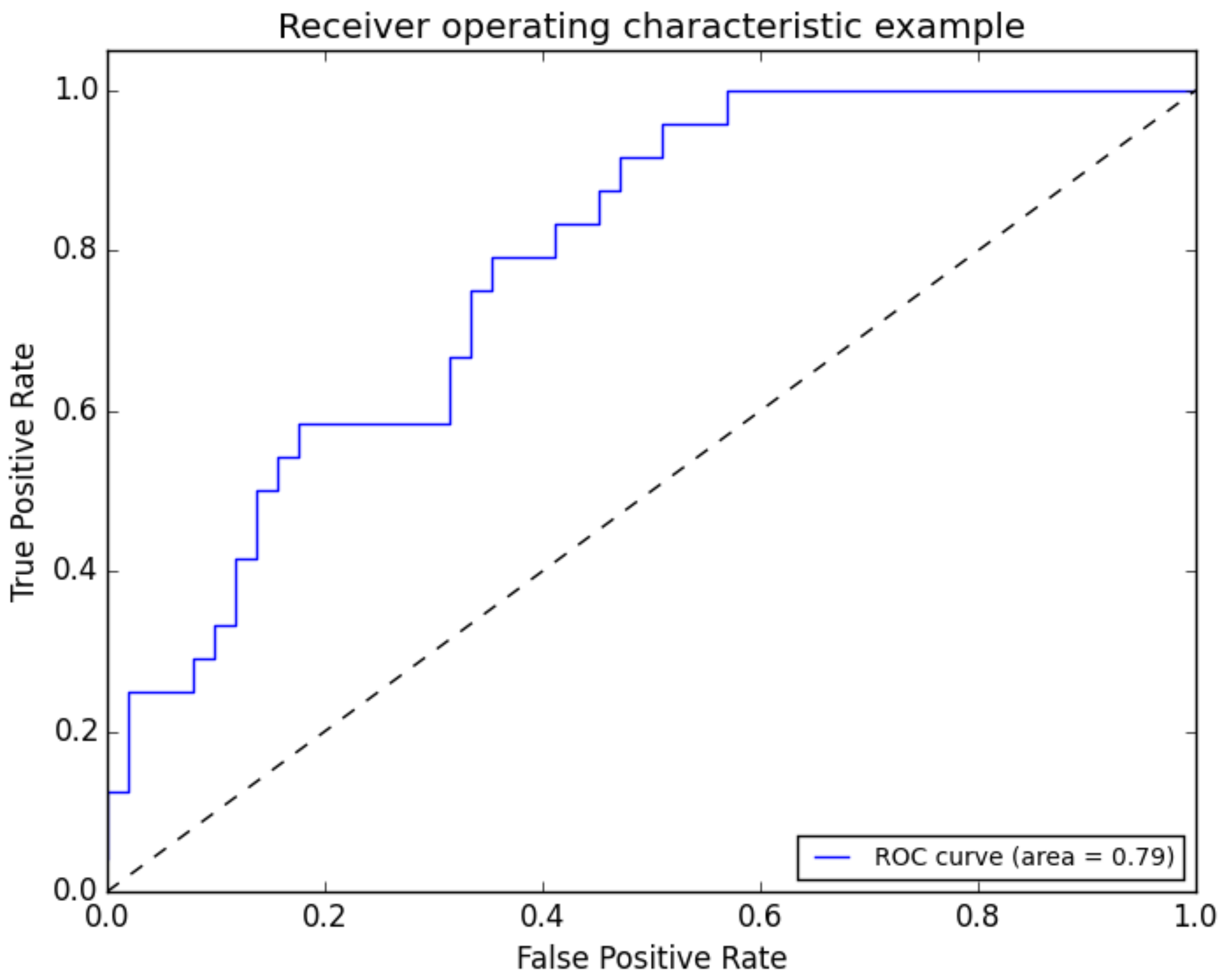
ADVANCED CLASSIFICATION METRICS

- We can continue adding pairs for different thresholds



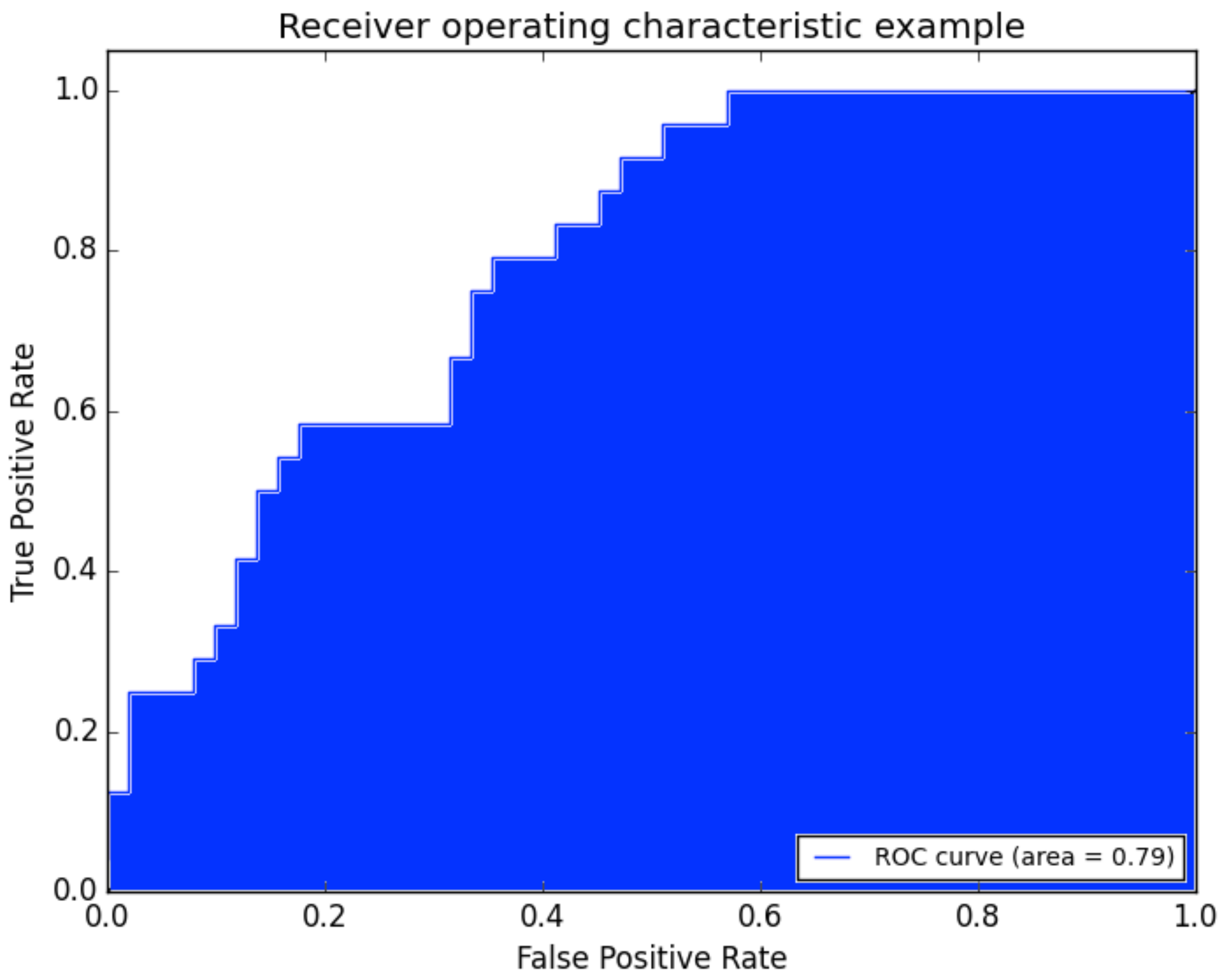
ADVANCED CLASSIFICATION METRICS

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



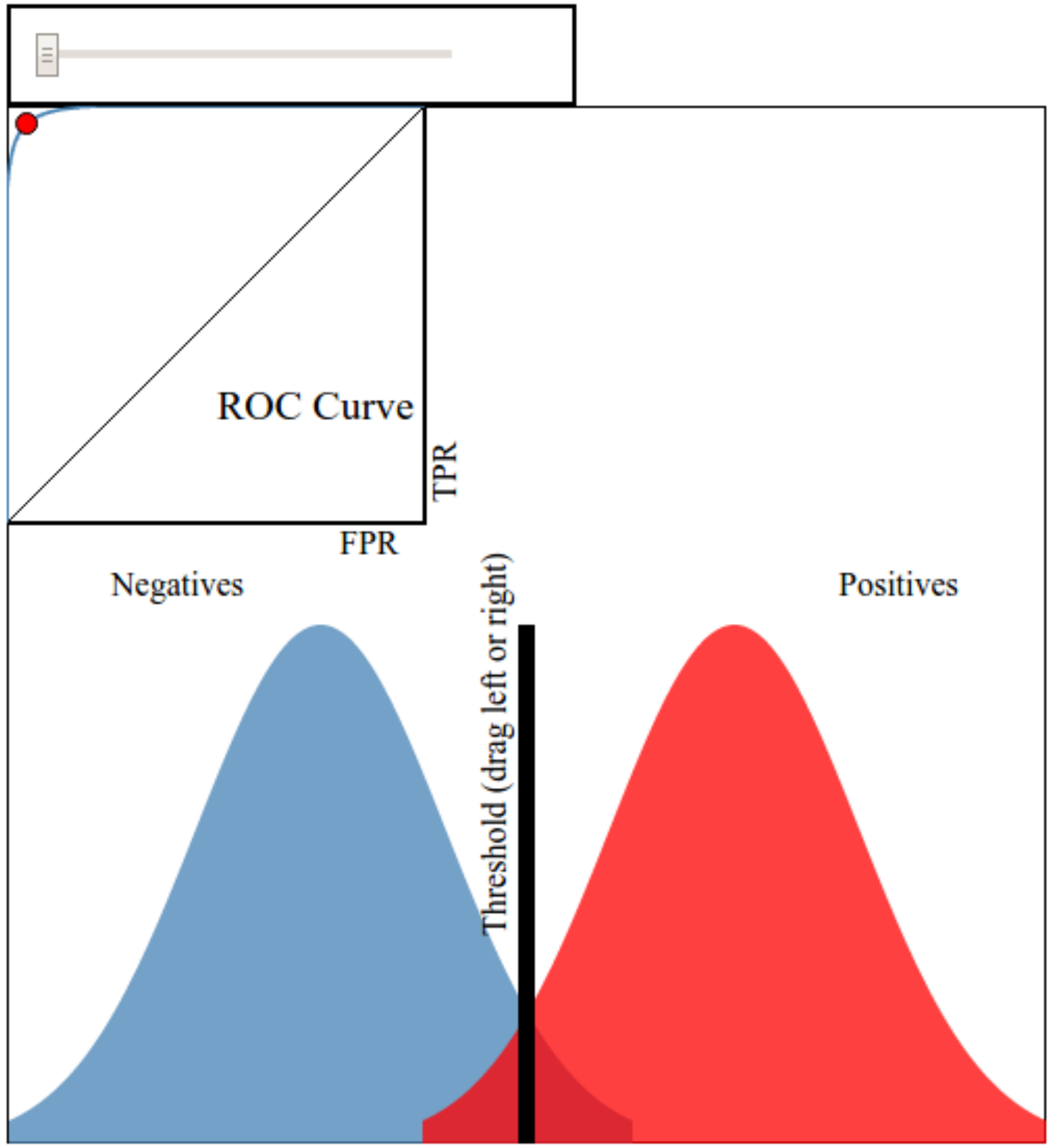
ADVANCED CLASSIFICATION METRICS

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



ADVANCED CLASSIFICATION METRICS

- This [interactive visualization] can help practice visualizing ROC curves.

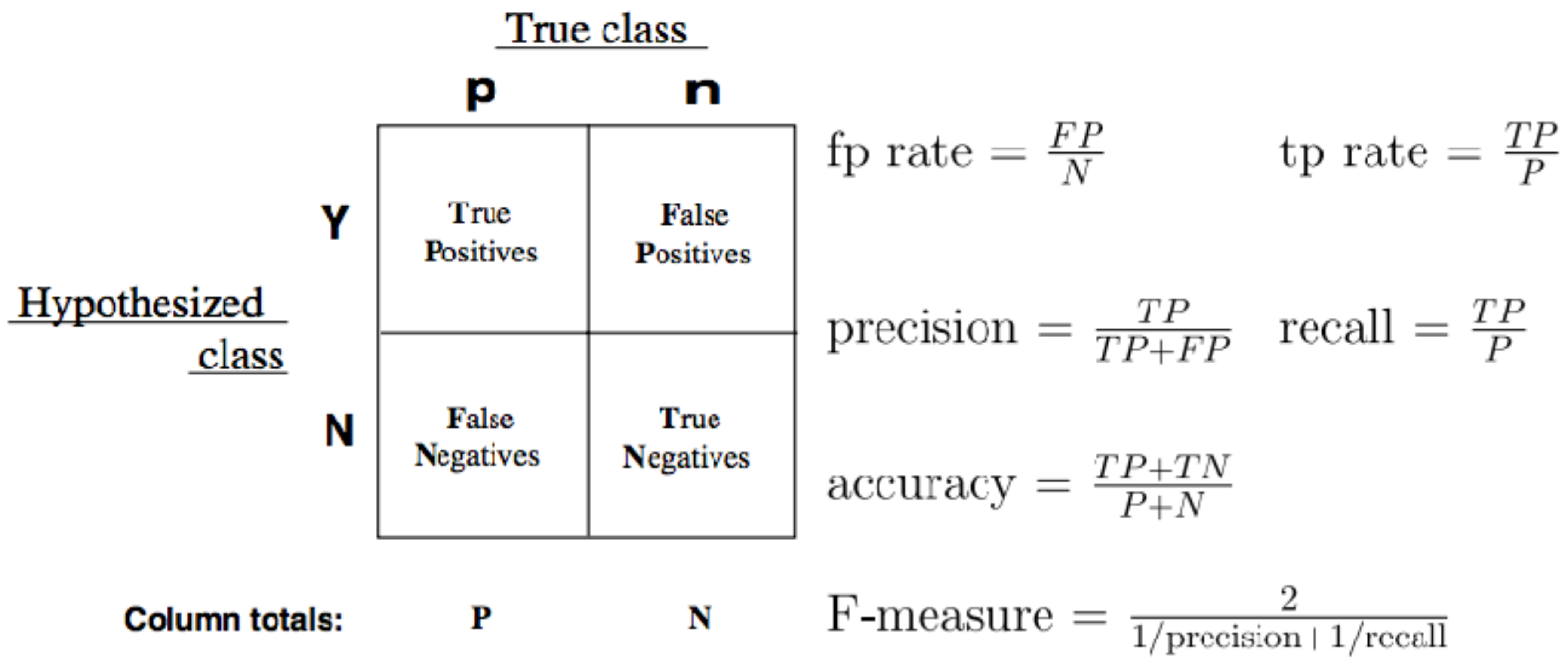


ADVANCED CLASSIFICATION METRICS

- If we have a TPR of 1 (all positives are marked positive) and FPR of 0 (all negatives are not marked positive), we'd have an AUC of 1. This means everything was accurately predicted.
- If we have a TPR of 0 (all positives are not marked positive) and an FPR of 1 (all negatives are marked positive), we'd have an AUC of 0. This means nothing was predicted accurately.
- An AUC of 0.5 would suggest randomness (somewhat) and is an excellent benchmark to use for comparing predictions (i.e. is my AUC above 0.5?).

ADVANCED CLASSIFICATION METRICS

- There are several other common metrics that are similar to TPR and FPR.



- Sklearn has all of the metrics located on [one convenient page].

GUIDED PRACTICE

WHICH METRIC SHOULD I USE?

ACTIVITY: WHICH METRIC SHOULD I USE?

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.

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.