GENERAL ASSEMBLY

INTRODUCTION TO LOGISTIC REGRESSION

Chirayu Sariya

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

COURSE

PRE-WORK

PRE-WORK REVIEW

- Implement a linear model (LinearRegression) with sklearn
- Understand what a coefficient is
- Recall metrics such as accuracy and misclassification
- Recall the differences between L1 and L2 regularization

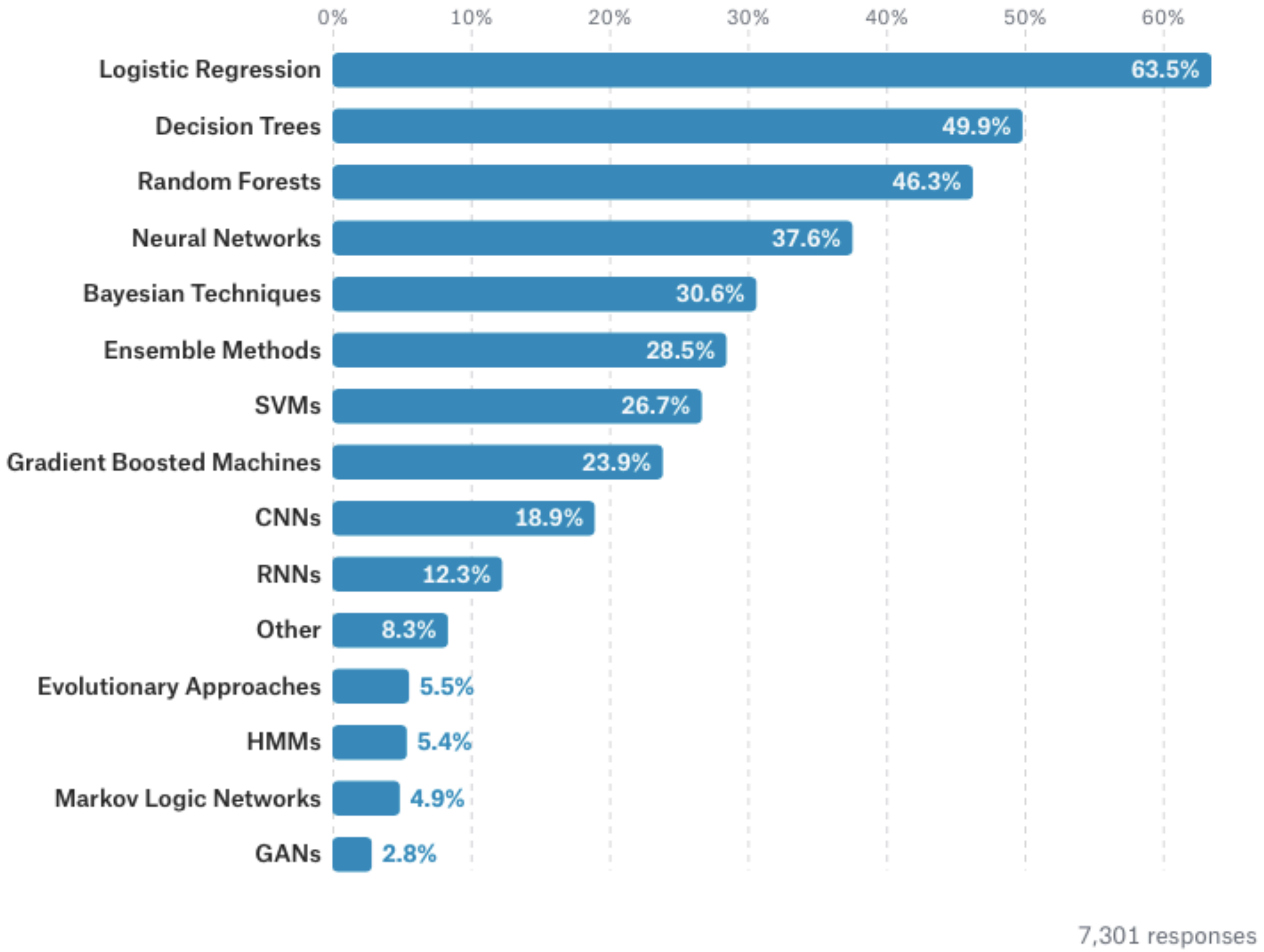
OPENING

INTRODUCTION TO LOGISTIC REGRESSION

WHO IS USING IT?

- Widely used data science method across industries. - Kaggle ML and Data Science survey.
- In some instances has even outperformed state-of-the-art models like Convolutional Neural Networks. - Winning model for Kaggle's Leaf Classification competition.
- Good news - its also fairly easy to implement!

*except Military and Security, where Neural Networks are used slightly more frequently.



LOGISTIC REGRESSION

LOGISTIC REGRESSION

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

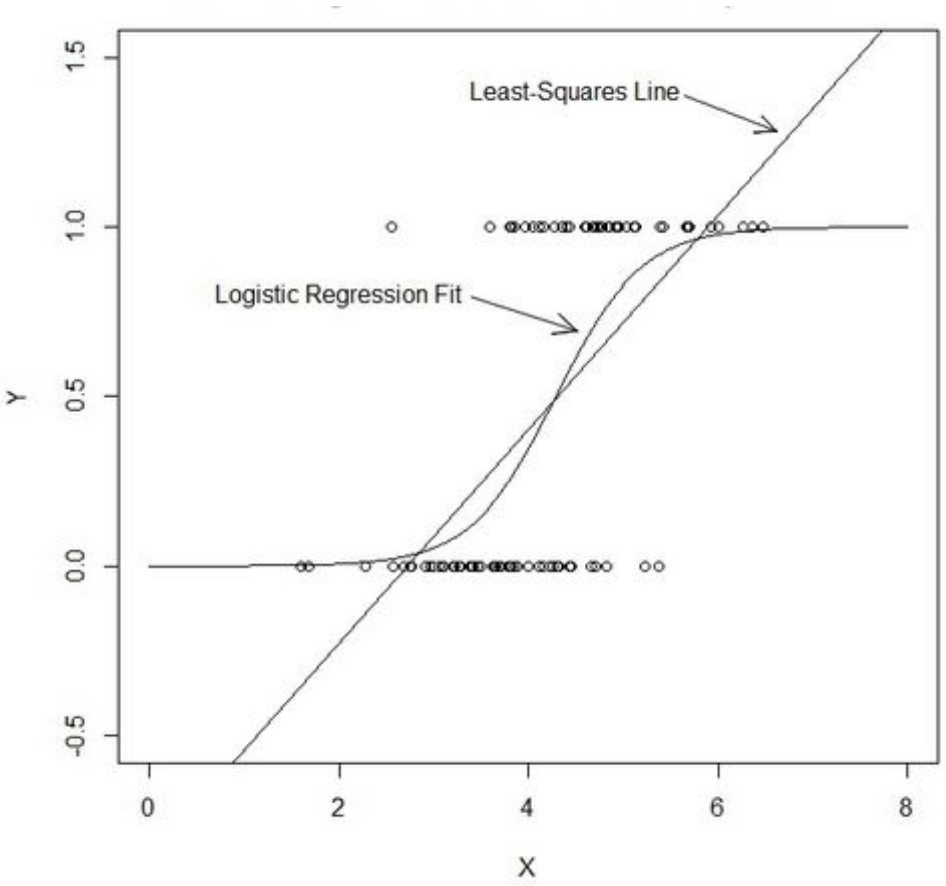
- That is, we can use a linear model, similar to Linear regression, in order to solve if an item *belongs* or *does not belong* to a class label.

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?

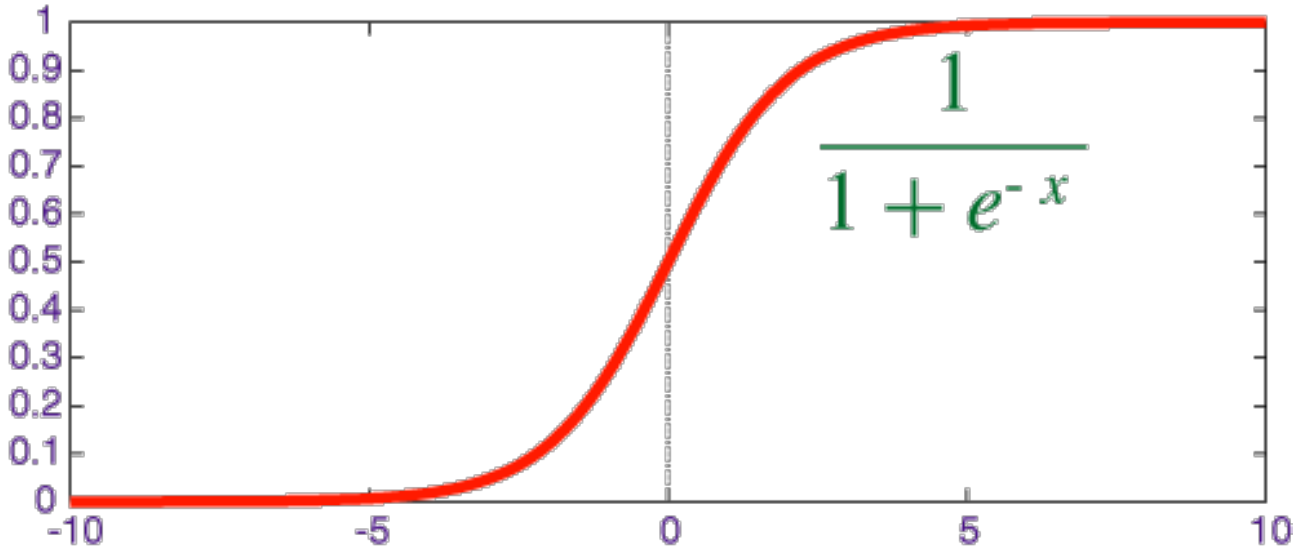
THE SIGMOID FUNCTION

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



THE SIGMOID FUNCTION

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



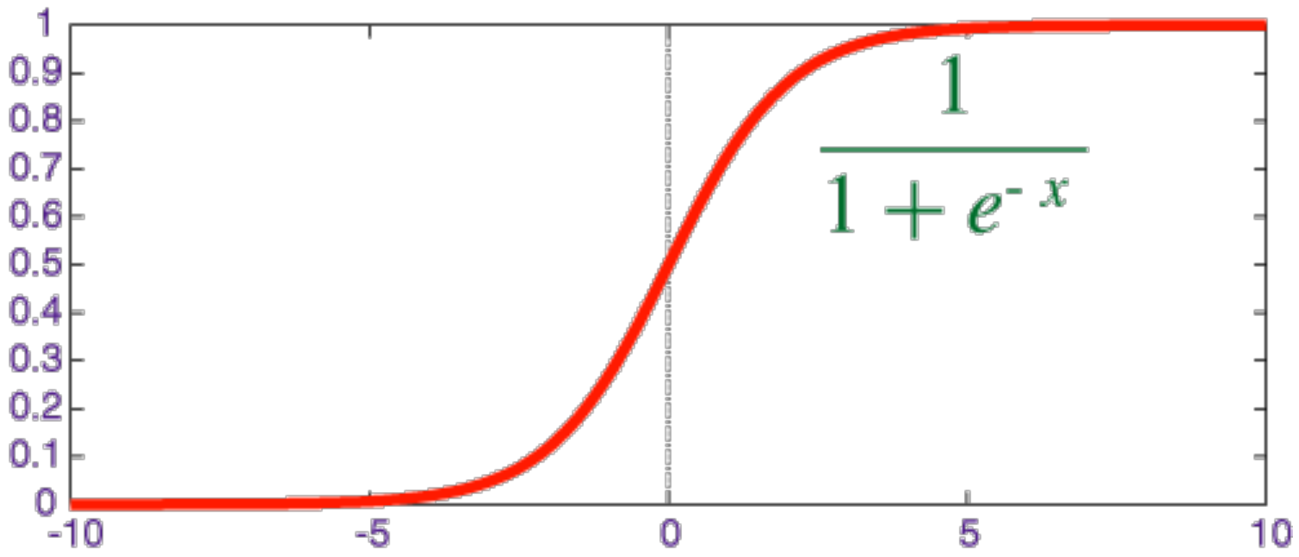
- Mathematically, it is defined as $f(x) = \dfrac{1}{1+\mathrm{e}^{-x}}$

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.



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?

ODDS, ODDS RATIO AND THE LOGIT FUNCTION

ODDS AND LOG-ODDS

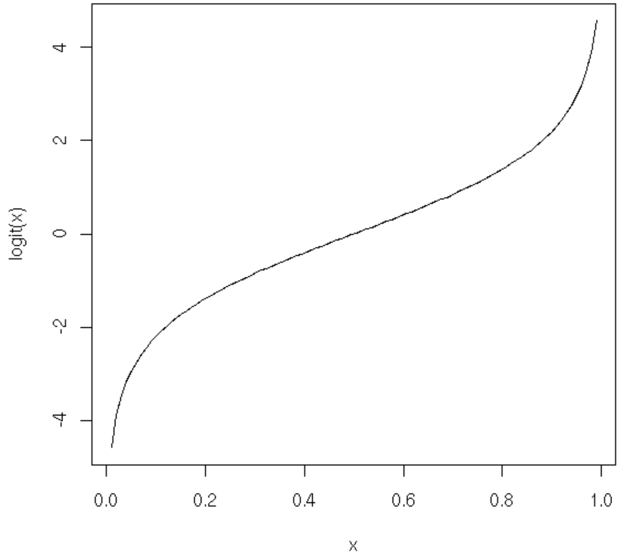
- Odds (like in gambling) are an expression of relative probabilities, generally quoted as the odds in favor.

- The odds (in favor) of an event or a proposition is the ratio of the probability that the event will happen to the probability that the event will not happen.

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

ODDS AND LOG-ODDS

- The log of odds is called the "*logit*" and it looks like the linear regression.
- The *logit* function is the inverse of the *sigmoid* function.
- Mathematically, the *logit* function is defined as $Ln\left(\frac{P}{1-P}\right)$
- This will act as our *link* function for logistic regression.



LINK FUNCTIONS

- Generalized linear models include a link function that relates the expected value of the response to the linear predictors in the model.
- A link function transforms the probabilities of the levels of a categorical response variable to a continuous scale that is unbounded.
- When you apply an appropriate link function to the probabilities, the numbers that result range from $-\infty$ to $+\infty$.

LINK FUNCTIONS

- The link function that best allows for this is the *logit* function, which is the inverse of the *sigmoid* function.
- We can now form a specific relationship between our linear predictors and the response variable.

ODDS AND LOG-ODDS

- For example, the logit value (log odds) of 0.2 (or odds of ~1.2:1):

$$0.2 = \ln(p / (1-p))$$

- Applying the sigmoid function, we would get the probability ~0.55.

$$1 / (1 + e^{-0.2})$$

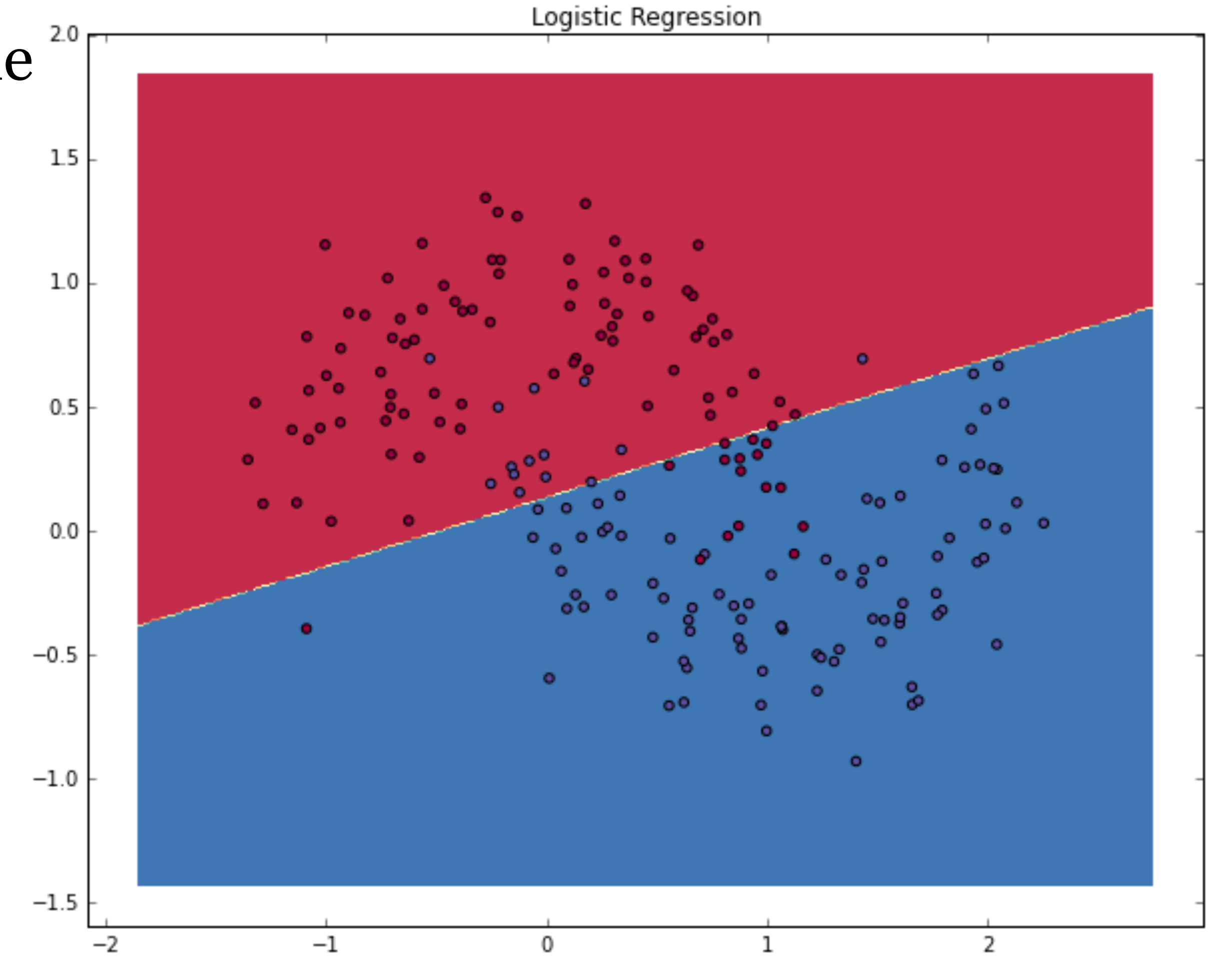
- To calculate this in python, we could use the following.

```
1 / (1 + numpy.exp(-0.2) )
```

ODDS AND LOG-ODDS

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

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



GUIDED PRACTICE

WAGER THOSE ODDS!

ACTIVITY: 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

ACTIVITY: 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

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

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→ real↓	*Class_pos*	*Class_neg*
Class_pos	TP	FN
Class_neg	FP	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?

predicted→ real↓	Class_pos	Class_neg
Class_pos	TP	FN
Class_neg	FP	TN

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

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?

predicted → real ↓	Class_pos	Class_neg
Class_pos	TP	FN
Class_neg	FP	TN

$$\text{FPR (1-specificity)} = \frac{\text{FP}}{\text{TN}+\text{FP}}$$

ADVANCED CLASSIFICATION METRICS

- These can also be inverted.
- How often does a test *correctly* identify patients without cancer? (True Negative Rate)
- How often does a test *incorrectly* identify patient as cancer-free? (False Negative Rate)

Name	Formula	Explanation
True Positive Rate (TP rate)	TP / (TP + FP)	The closer to 1, the better. TP rate = 1 when FP = 0. (No false positives)
True Negative Rate (TN rate)	TN / (TN + FN)	The closer to 1, the better. TN rate = 1 when FN = 0. (No false negatives)
False Positive Rate (FP rate)	FP / (FP + TN)	The closer to 0, the better. FP rate = 0 when FP = 0. (No false positives)
False Negative Rate (FN rate)	FN / (FN + TP)	The closer to 0, the better. FN rate = 0 when FN = 0. (No false negatives)

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.

ADVANCED CLASSIFICATION METRICS

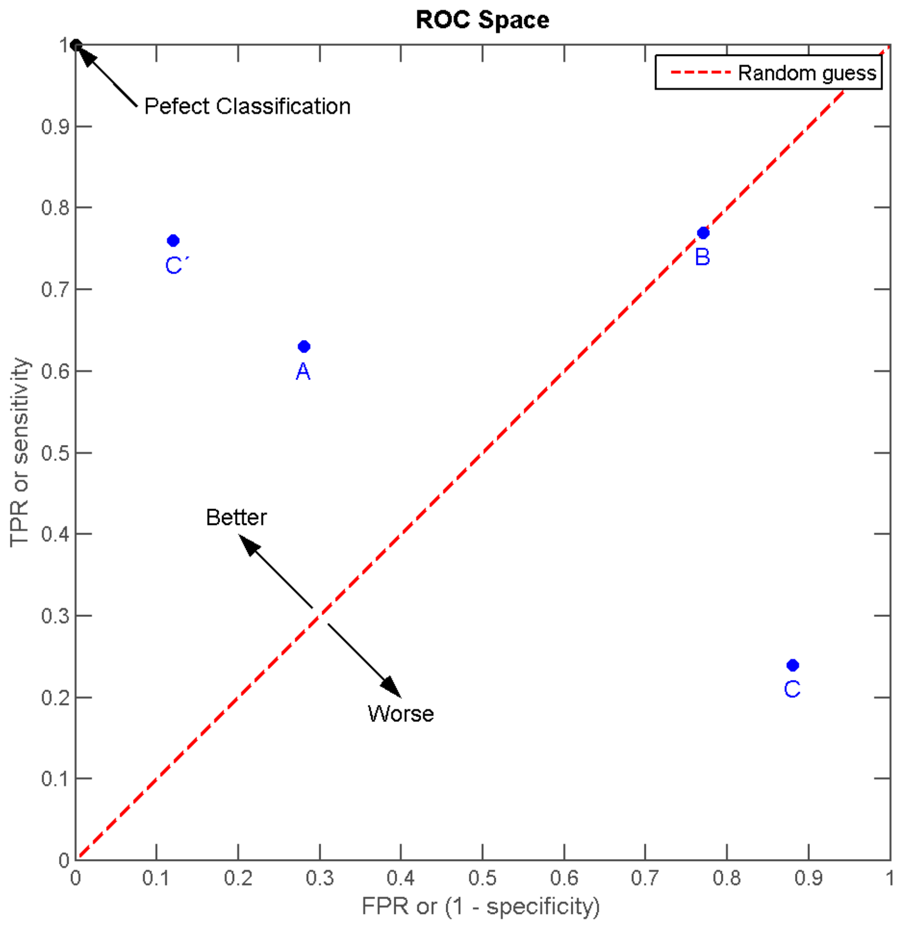
- We can vary the classification threshold for our model to get different predictions. But how do we know if a model is better overall than other model?
- We can compare the FPR and TPR of the models, but it can often be difficult to optimize two numbers at once.
- Logically, we like a single number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

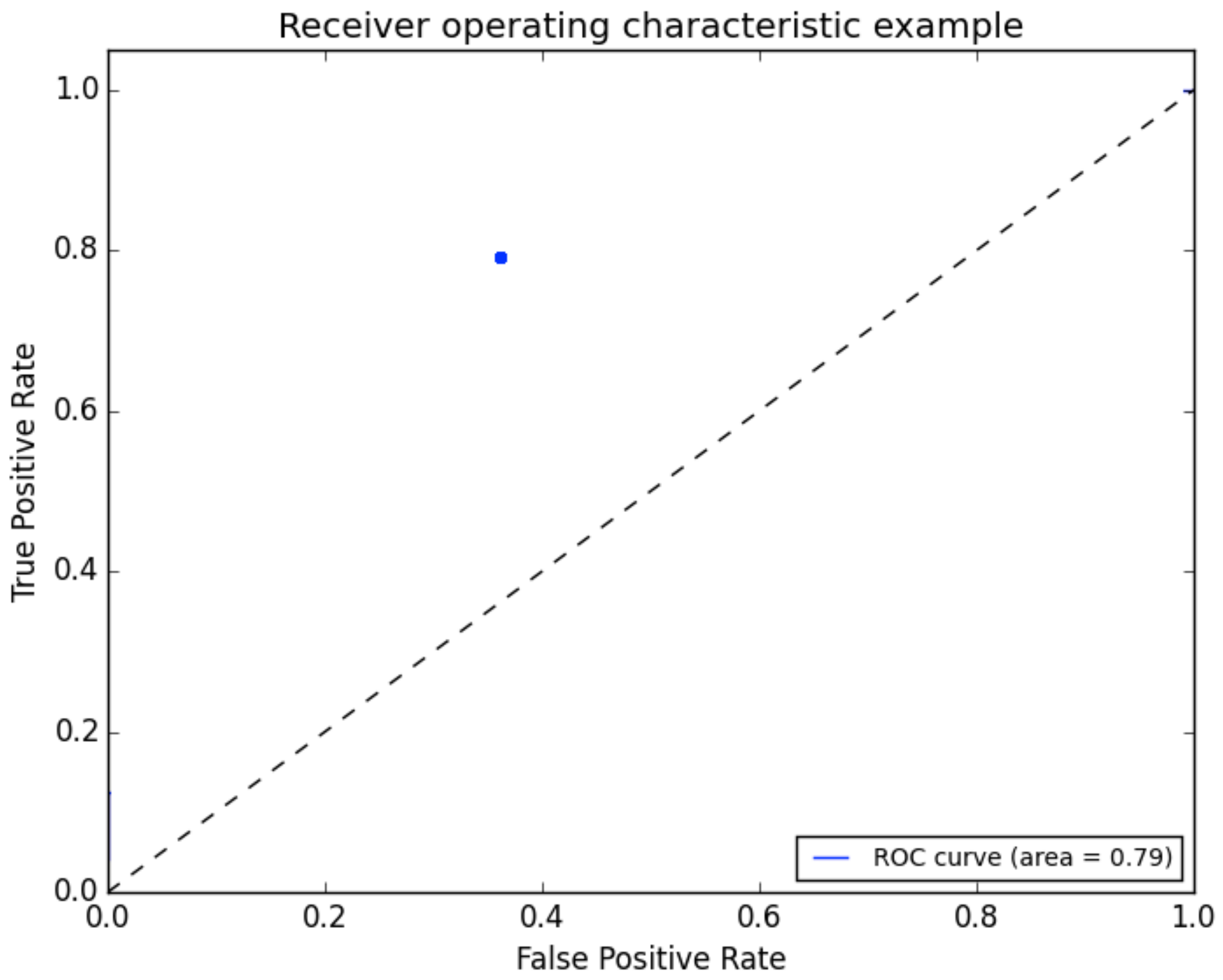
ADVANCED CLASSIFICATION METRICS

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



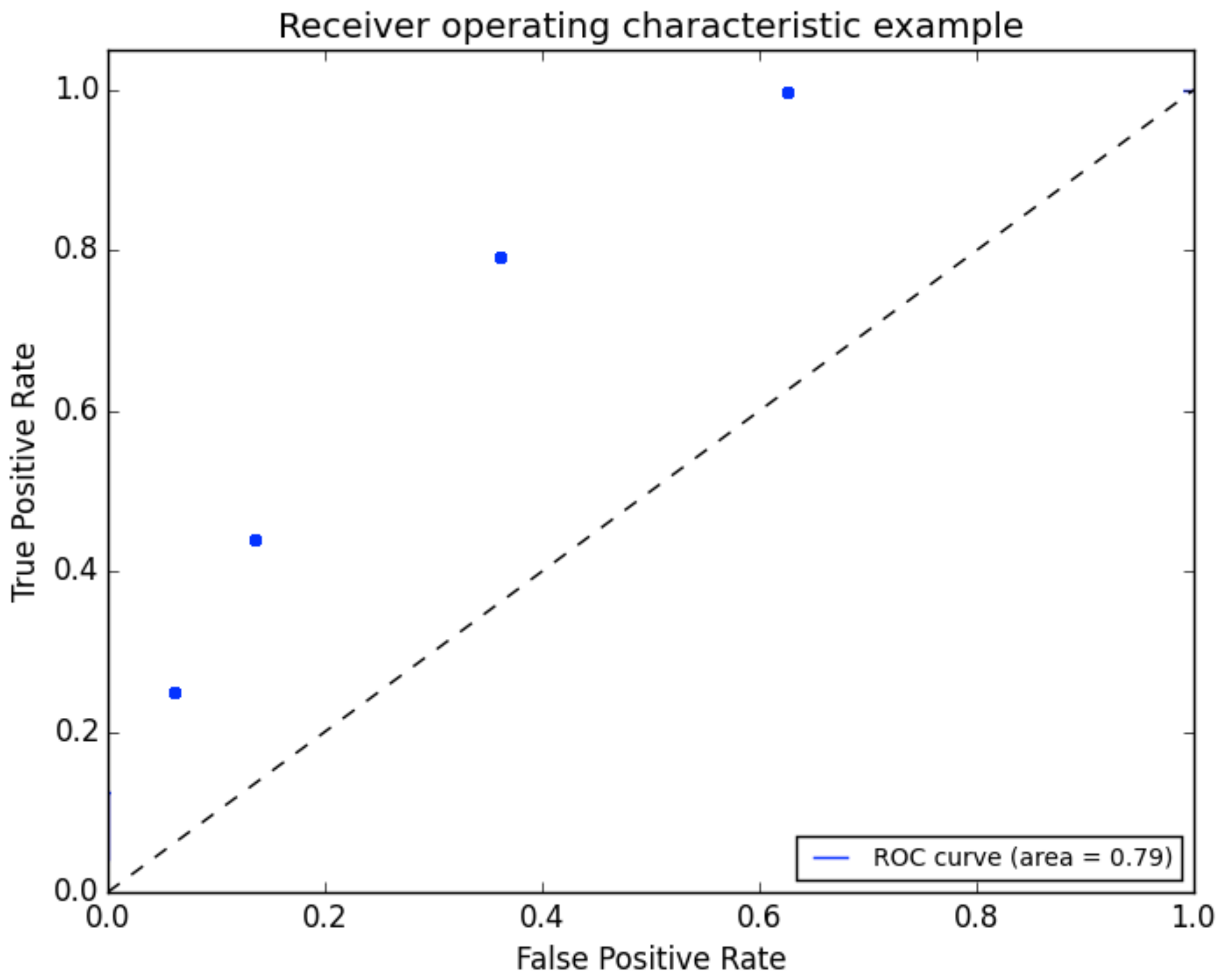
ADVANCED CLASSIFICATION METRICS

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



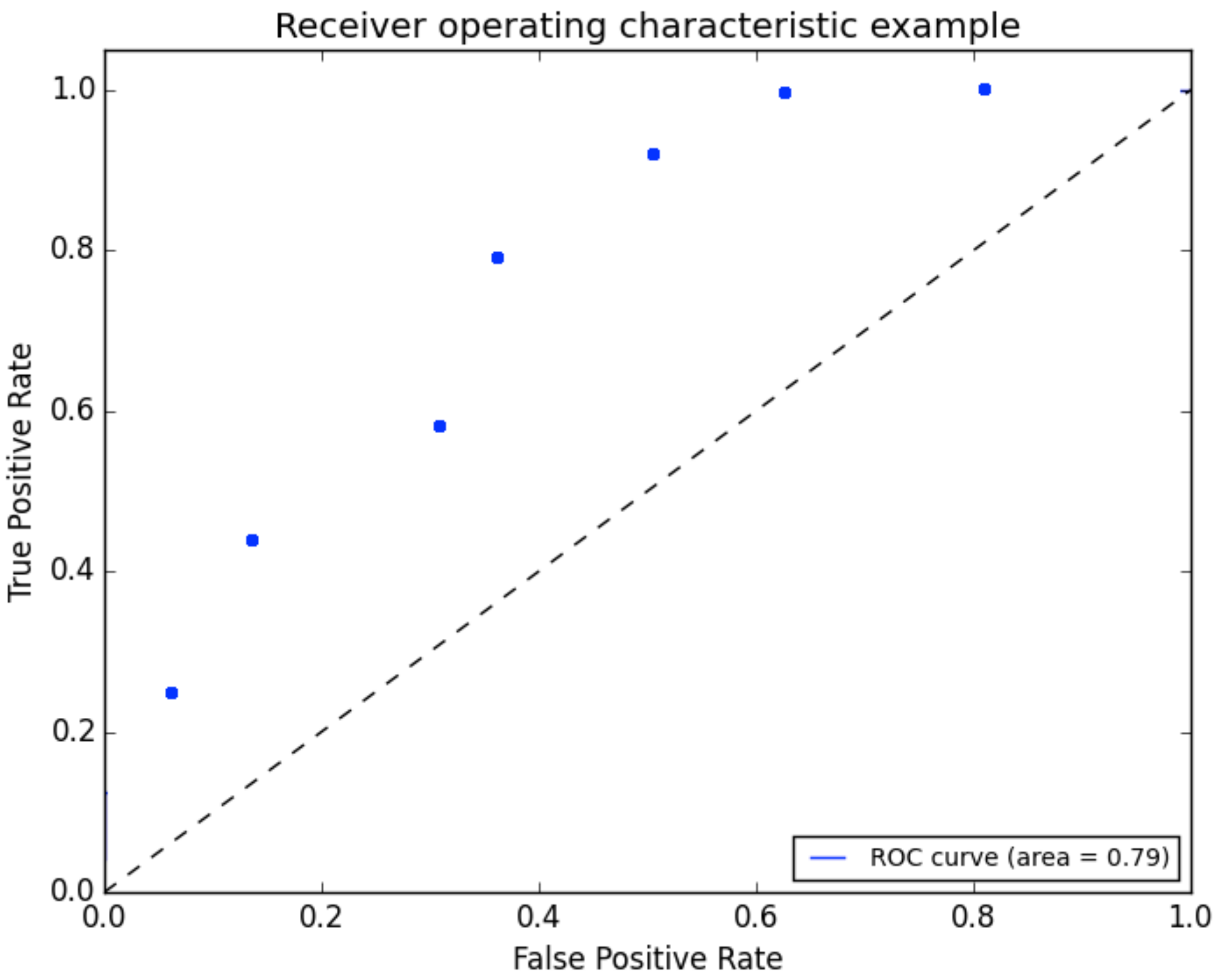
ADVANCED CLASSIFICATION METRICS

- We can continue adding pairs for different thresholds

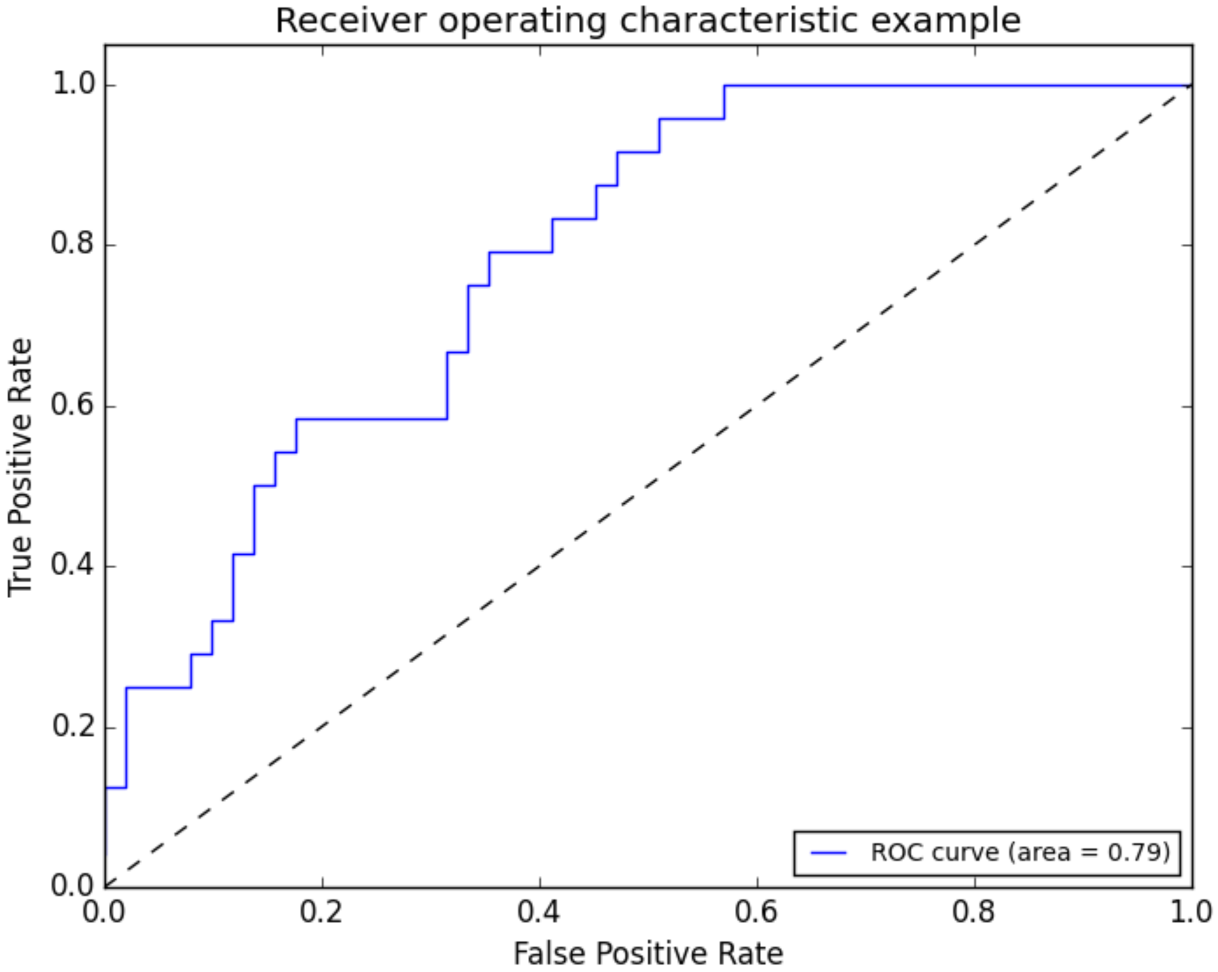


ADVANCED CLASSIFICATION METRICS

- We can continue adding pairs for different thresholds

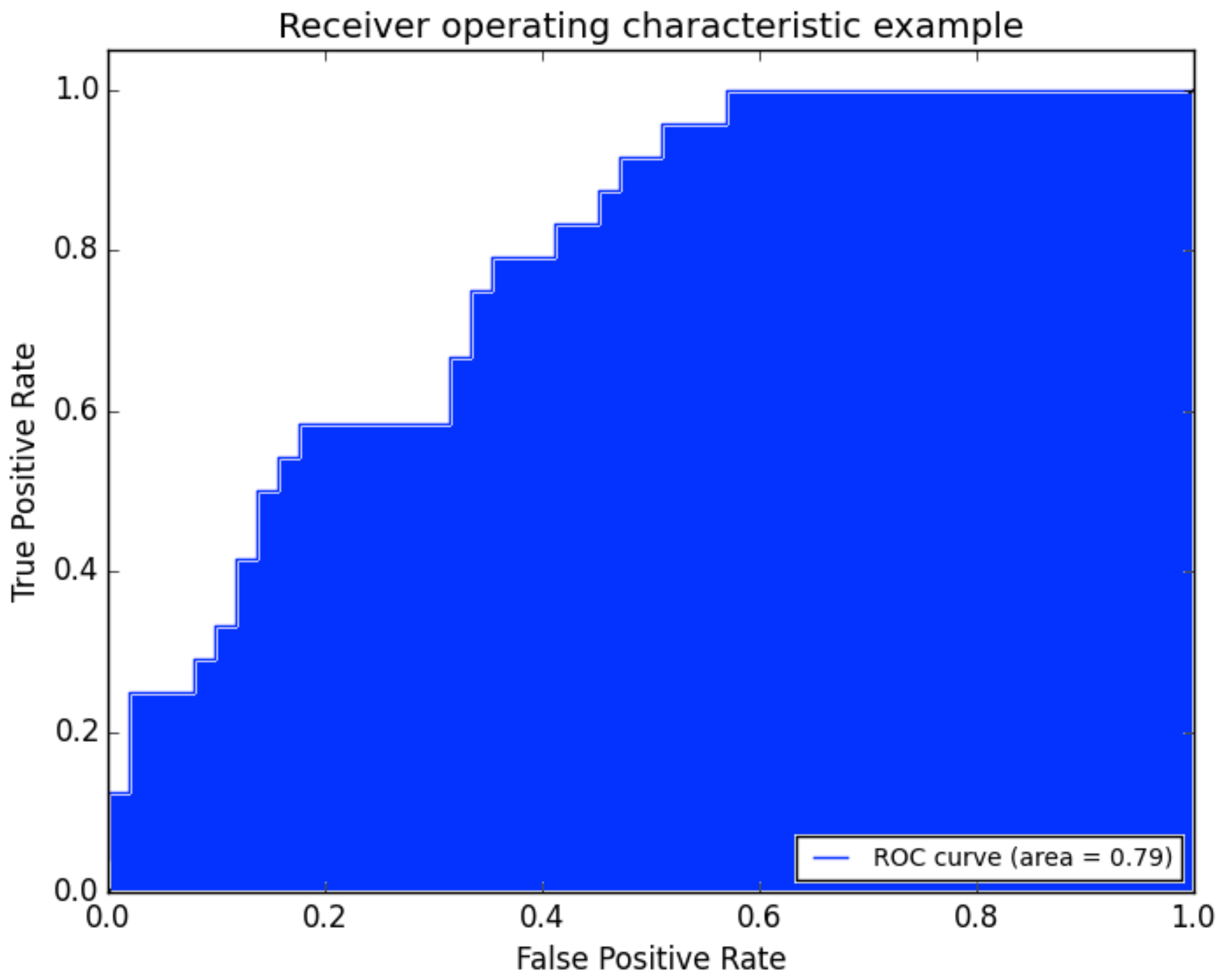


- Finally, we create a full curve that is described by TPR and FPR.



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.
- 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?

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

TOPIC REVIEW

REVIEW QUESTIONS

- What's the link function used in logistic regression?
- What kind of machine learning problems does logistic regression address?

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.

LESSON

Q & A

LESSON

EXIT TICKET

DON'T FORGET TO FILL OUT YOUR EXIT TICKET