BEFORE WE START

Mid-Course Feedback

BEFORE WE START

Self-evaluation survey

https://goo.gl/forms/IqgaPBXVH36S3bXG2

REVIEW

FROM LAST LESSON

- Difference between normal equation (OLS) & gradient descent
- Define and explain the gradient descent algorithm
- Define class label and classification
- Understand what a coefficient is
- Recall metrics such as accuracy and misclassification

LEARNING OBJECTIVES

- Define and apply the KNN algorithm
- Dive into the mechanisms of Logistic Regression:
  - Difference with the linear regression algorithm
  - Loss function (also called Cost function)
  - Evaluation of your classification model

INTRODUCTION

WHAT IS K NEAREST NEIGHBORS?

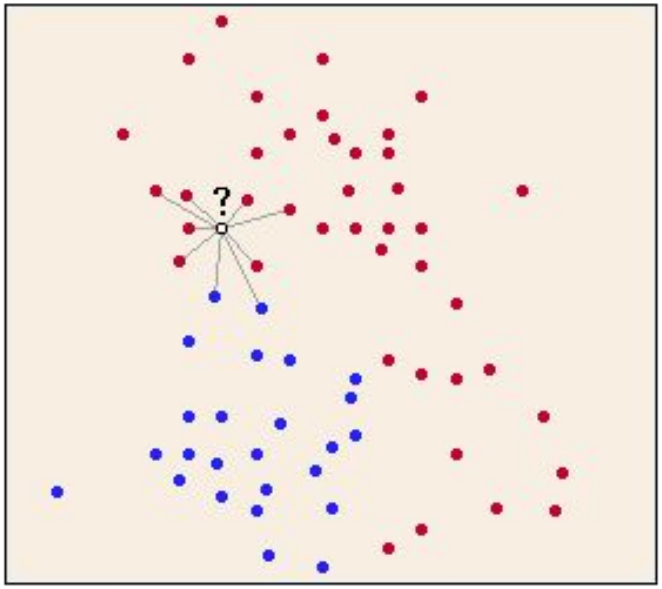
WHAT IS K NEAREST NEIGHBORS?

- **K Nearest Neighbors (KNN)** is an algorithm that makes a prediction based upon the closest data points.

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

WHAT IS K NEAREST NEIGHBORS?

- KNN uses distance to predict a class label
- Distance is used as a measure of similarity between observations



WHAT IS K NEAREST NEIGHBORS?

- Suppose we want to determine your favorite type of music
- **Friends share similar traits and interests** (e.g. music, sports teams, hobbies, etc). We could ask your five closest friends what their favorite type of music is and take the majority vote
- This is the intuition behind **KNN**: we look for **things similar** to (or close to) our new observation and identify shared traits. We can use this information to make an **educated guess** about a trait of our new observation

ACTIVITY: KNOWLEDGE CHECK

ANSWER THE FOLLOWING QUESTIONS

1. In what other tasks do we use a heuristic similar to K Nearest Neighbors?

DELIVERABLE

Answers to the above questions

DEMO

KNN IN ACTION

WHAT HAPPENS IN TIES?

- What happens if two classes get the same number of votes?
- This could happen in binary classification if we use an even number for *k*. This could also happen if there are multiple class labels.
- In sklearn, it will choose the class that it first saw in the *training set*.

DEMO

- http://localhost:8888/notebooks/GA/DS_PartTime/DS-course-materials/lessons/lesson-08/code/starter-code/starter-code-8.ipynb

WHAT HAPPENS IN TIES?

- ***weight***, taking into account the distance between the point and its neighbors.

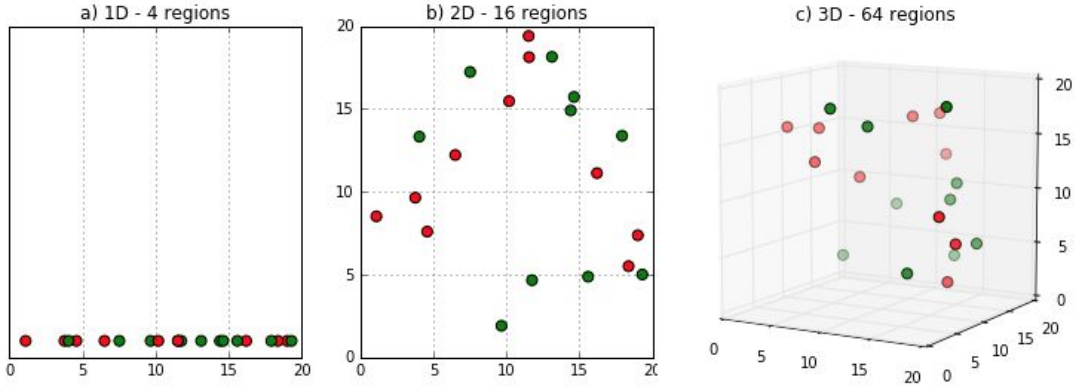
- This can be done in sklearn by changing the `weights` parameter to `"distance"`.

THE CURSE OF DIMENSIONALITY?

- Since KNN works with distance, higher dimensionality of data (i.e. more features) requires ***significantly*** **more samples** in order to have the same predictive power.

- Consider this: with more dimensions, all points slowly start averaging out to be equally distant. This causes significant issues for KNN.

- Keep the feature space limited and KNN will do well. Exclude extraneous features when using KNN.



INDEPENDENT PRACTICE

SOLVING FOR K

ACTIVITY: SOLVING FOR K

EXERCISE

DIRECTIONS (35 minutes)

One of the primary challenges of KNN is solving for k - how many neighbors do we use?

The **smallest** k we can use is 1. However, using only one neighbor will probably perform poorly.

The largest k we can use is n-1 (every other point in the data set). However, this would result in always choosing the largest class in the sample. This would also perform poorly.

Use the lesson 8 starter code and the iris data set to answer the following questions:

1. What is the accuracy for k=1?
2. What is the accuracy for k=n-1?
3. Using cross validation, what value of k optimizes model accuracy. Create a plot with *k* as the x-axis and *accuracy* as the y-axis (called a "fit chart") to help find the answer.

DELIVERABLE

Answers to the above questions

ACTIVITY: SOLVING FOR K

EXERCISE

DIRECTIONS

Bonus Questions:

1. By default, the KNN classifier in sklearn uses the *Minkowski metric* for distance.
   a. What *type* of data does this metric work best for?
   b. What *type* of data does this distance metric not work for?
   c. You can read about distance metrics in the sklearn documentation.

2. It is possible to use KNN as a regression estimator. Determine the following:
   a. Steps that KNN Regression would follow
   b. How it predicts a regression value

DELIVERABLE

Answers to the above questions

BACK TO THE LOGISTIC REGRESSION

INTRODUCTION TO LOGISTIC REGRESSION

EXERCISE

ANSWER THE FOLLOWING QUESTIONS

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?

2. 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)?

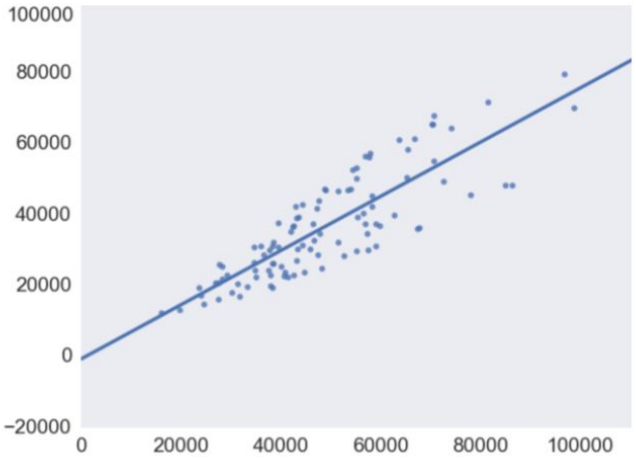
DELIVERABLE

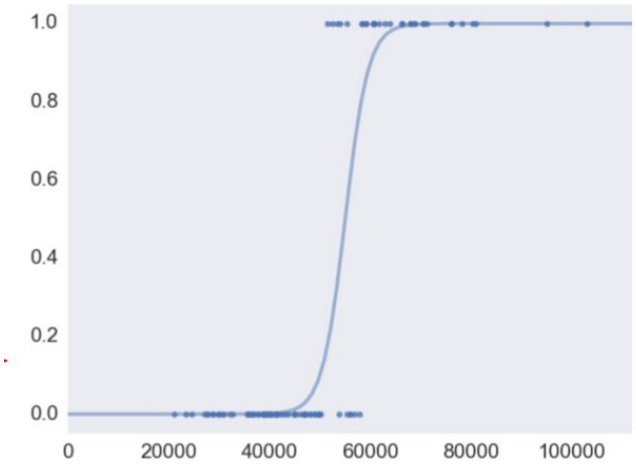
Answers to the above questions

LOGISTIC REGRESSION

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

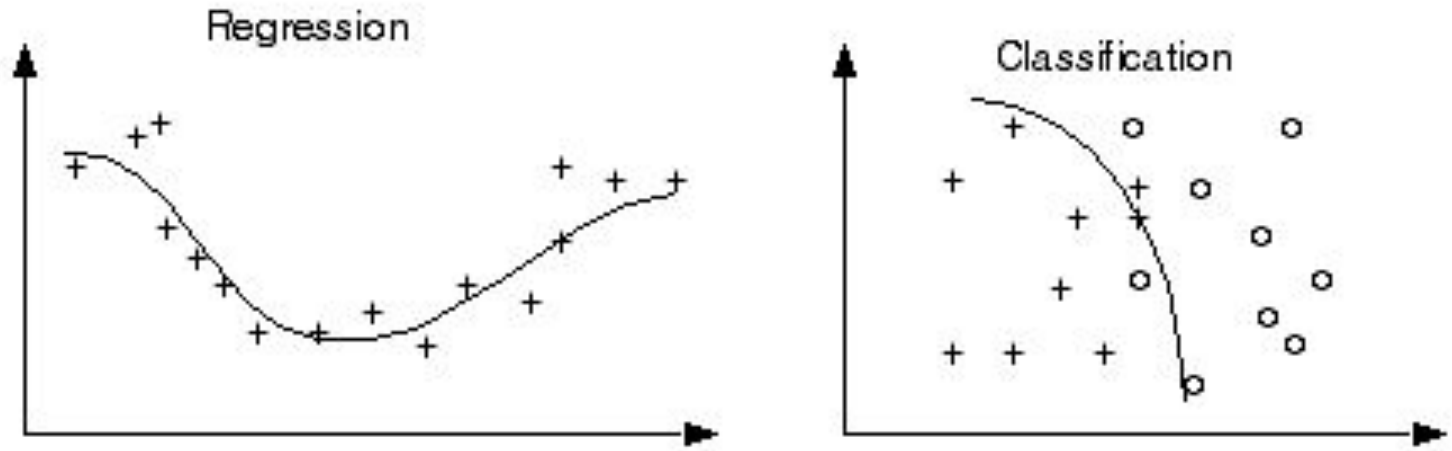
- Linear model to solve if an item *belongs* or *does not belong* to a class label.





LOGISTIC REGRESSION

- Regression results can have a value range from -∞ to ∞.
- Classification is used to find out to which group an observation belongs

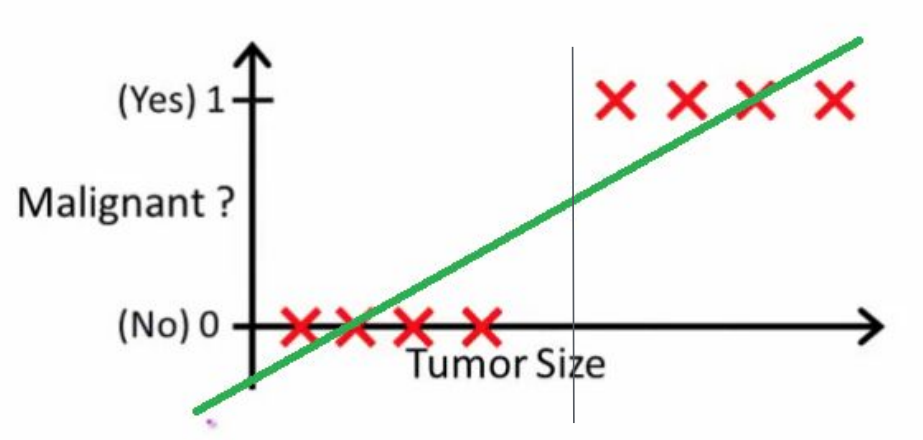


LOGISTIC REGRESSION

- Most classification problems are binary (0 or 1)
- **Can someone think about why linear regression** to solve classification does not work?

LOGISTIC REGRESSION

- One approach is predicting the **probability** that an observation belongs to a certain class
- Linear regression: Θ'*X ~ Y



LOGISTIC REGRESSION

- Linear Regression can be >1 and <0 - **WE DON'T WANT THAT**
- The Logistic Regression prevent it

LINK FUNCTIONS AND THE SIGMOID FUNCTION

- OLS allows for *other* models using a ***link function***

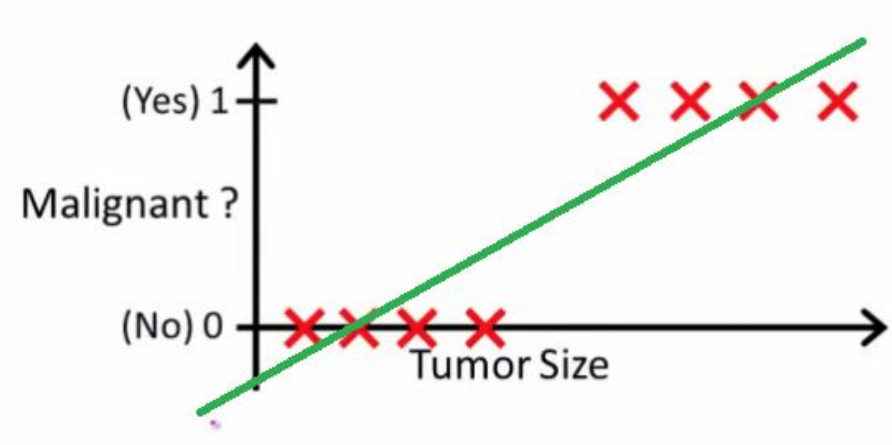
- What the H*** is a Sigmoid?

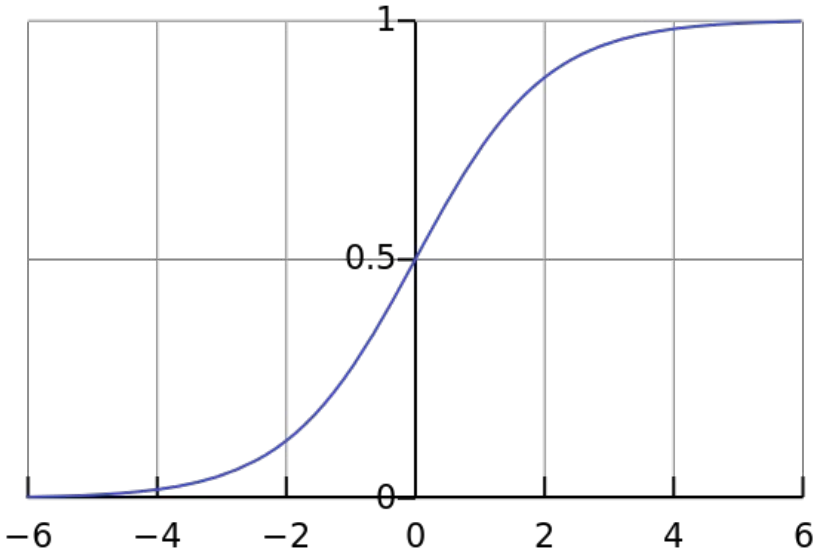
$$f(x) = \frac{1}{1 + e^{-(x)}}$$

- What's the link with Linear regression?

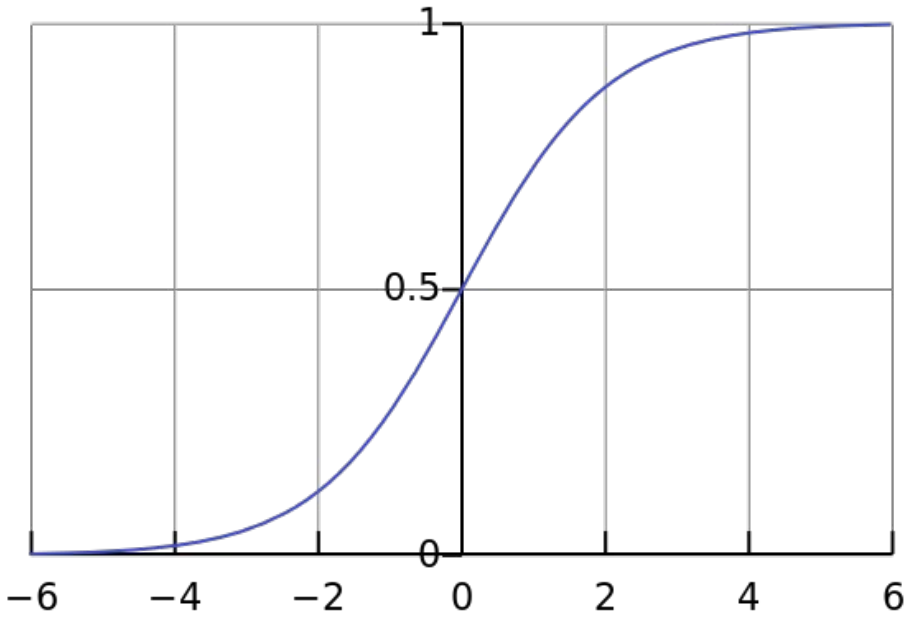
LINK FUNCTIONS AND THE SIGMOID FUNCTION

- For classification, we need a distribution associated with categories:
- Given all observations, what is the probability of an observation to be 0 or 1?
- In the Tumor example?





LINK FUNCTIONS AND THE SIGMOID FUNCTION



DEMO

PLOTTING A SIGMOID FUNCTION

PLOTTING A SIGMOID FUNCTION

http://localhost:8888/notebooks/GA/DS_PartTime/DS-course-materials/lessons/lesson-09/Cost-fct-reg.ipynb

INTRODUCTION

LOGISTIC REGRESSION: Cost Function

COST FUNCTION

Cost function

Linear regression: $J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{2} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$

- The sigmoid function is not convex

COST FUNCTION

Logistic regression cost function

$$\text{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

- NOW THE COST FUNCTION IS CONVEX!!! YEAHHHH

$$\text{Cost} = 0 \text{ if } y = 1, h_\theta(x) = 1$$
$$\text{But as} \quad h_\theta(x) \to 0$$
$$Cost \to \infty$$

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

CODE ALONG

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 model using multiple features with either sklearn or sm. Interpreting the odds, which features have the most impact on admission rate? Which features have the least?
3. What is the accuracy of your model?

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.

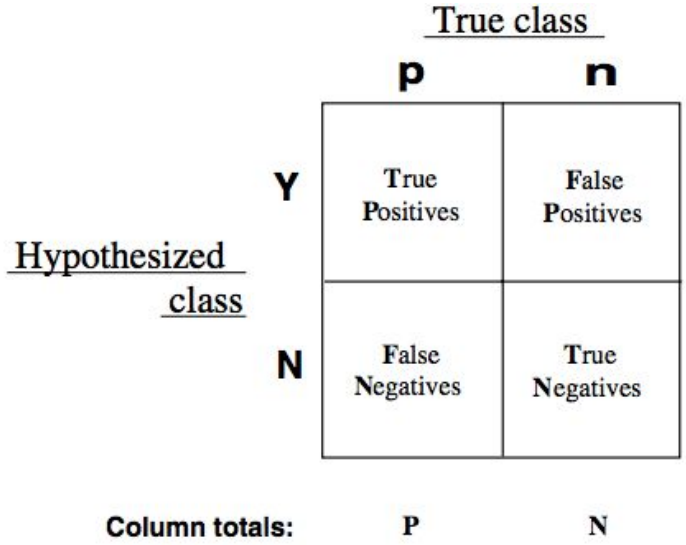
ADVANCED CLASSIFICATION METRICS

- 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*.

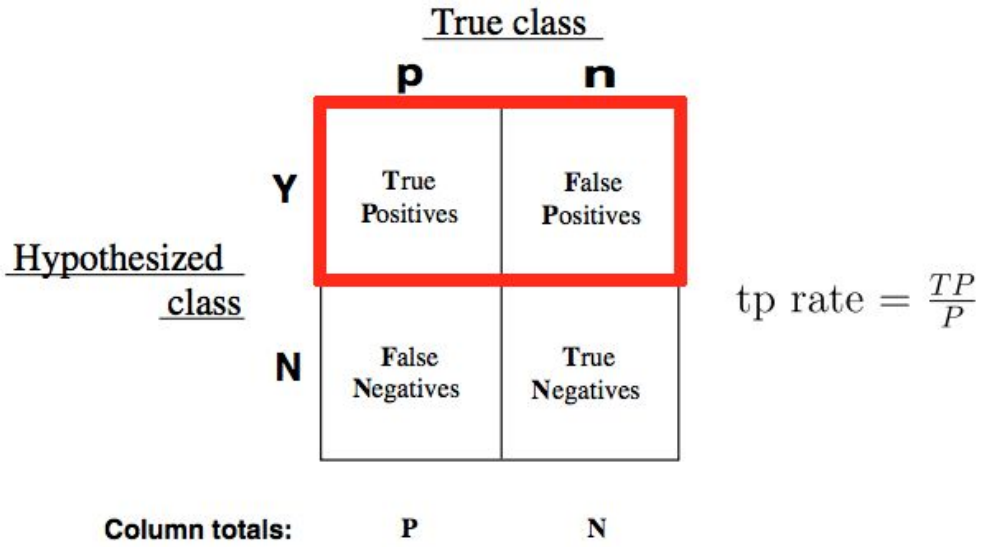
- For each label, we can put it into the category of a true positive, false positive, true negative, or false negative.



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?”

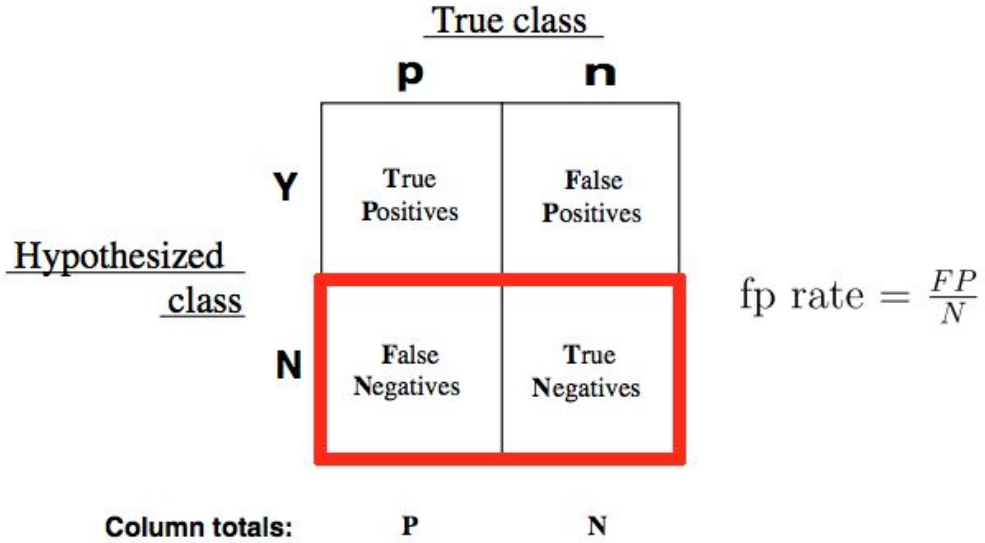
- For example, given a medical exam that tests for cancer, how often does it correctly identify patients with cancer?



ADVANCED CLASSIFICATION METRICS

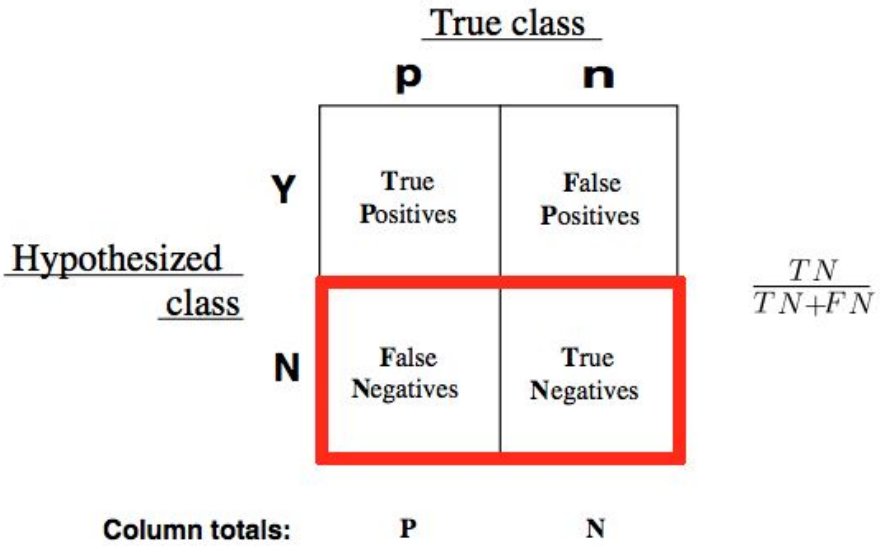
- False Positive Rate (FPR) asks, “Out of all items not belonging to a class label, how many were predicted as belonging to that target class label?”

- 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?



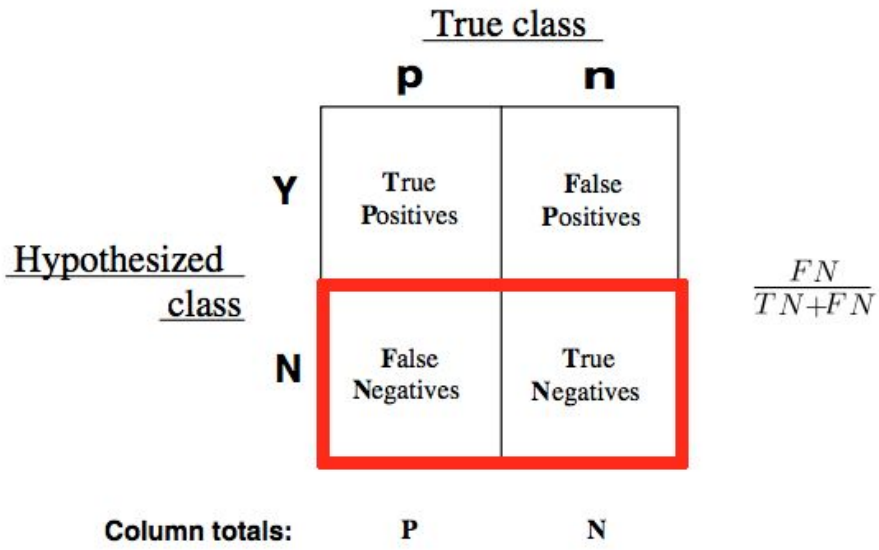
ADVANCED CLASSIFICATION METRICS

- These can also be inverted.
- How often does a test *correctly* identify patients without cancer?



ADVANCED CLASSIFICATION METRICS

- How often does a test *incorrectly* identify patient as cancer-free?



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.

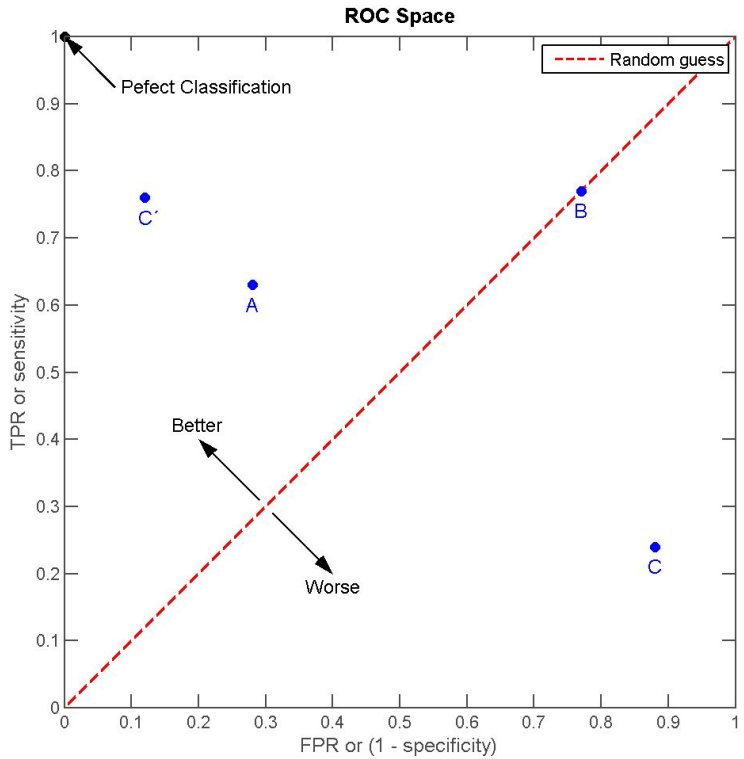
- In our smoking problem, this model would accurately predict ***all* of the smokers as smokers** and **not accidentally predict any of the nonsmokers as smokers**.

ADVANCED CLASSIFICATION METRICS

- Receiver Operation Characteristic (ROC) curve help optimizing 2 parameters
- 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.

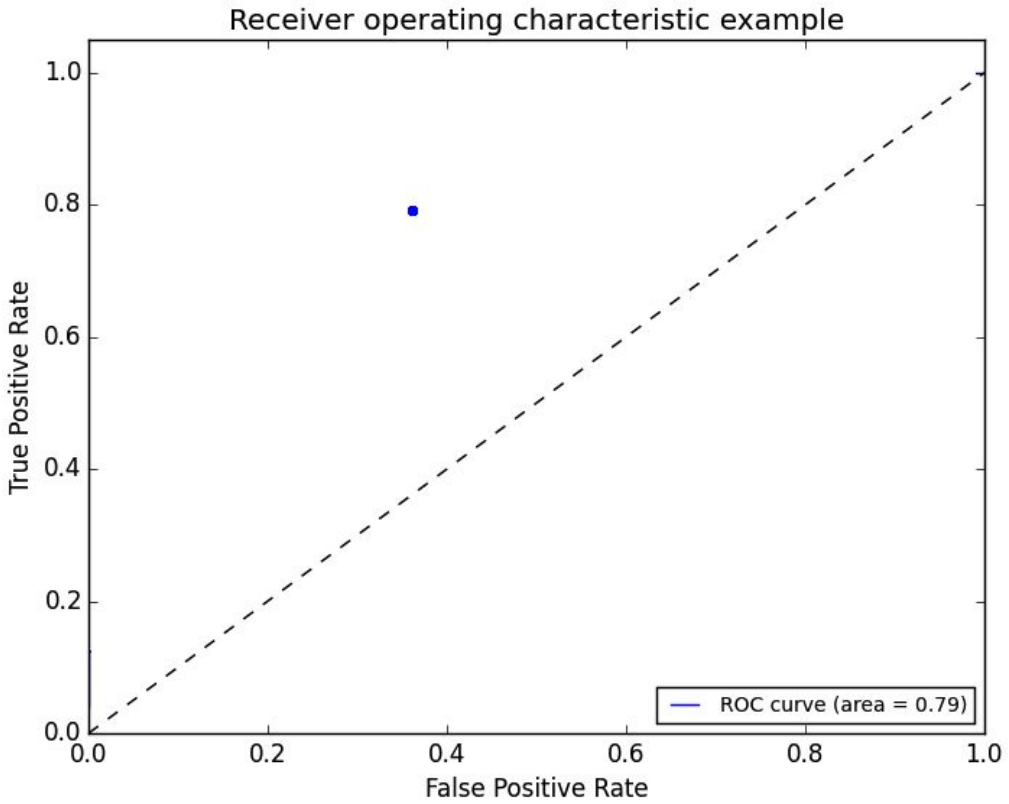
ADVANCED CLASSIFICATION METRICS

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



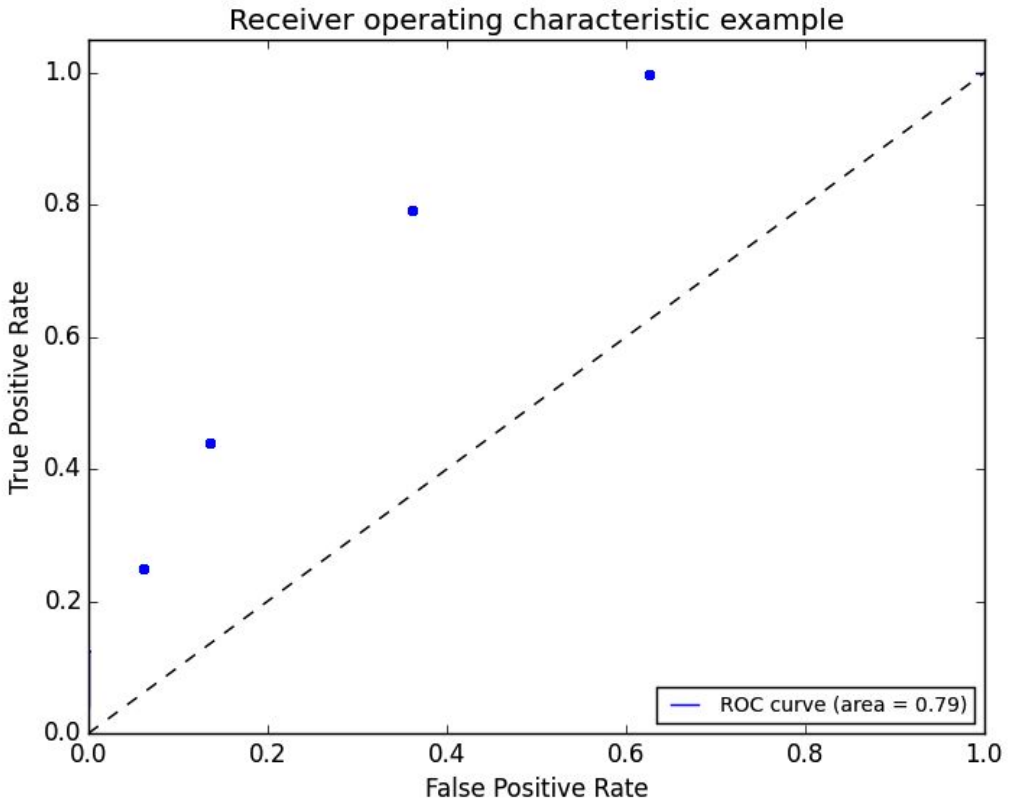
ADVANCED CLASSIFICATION METRICS

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



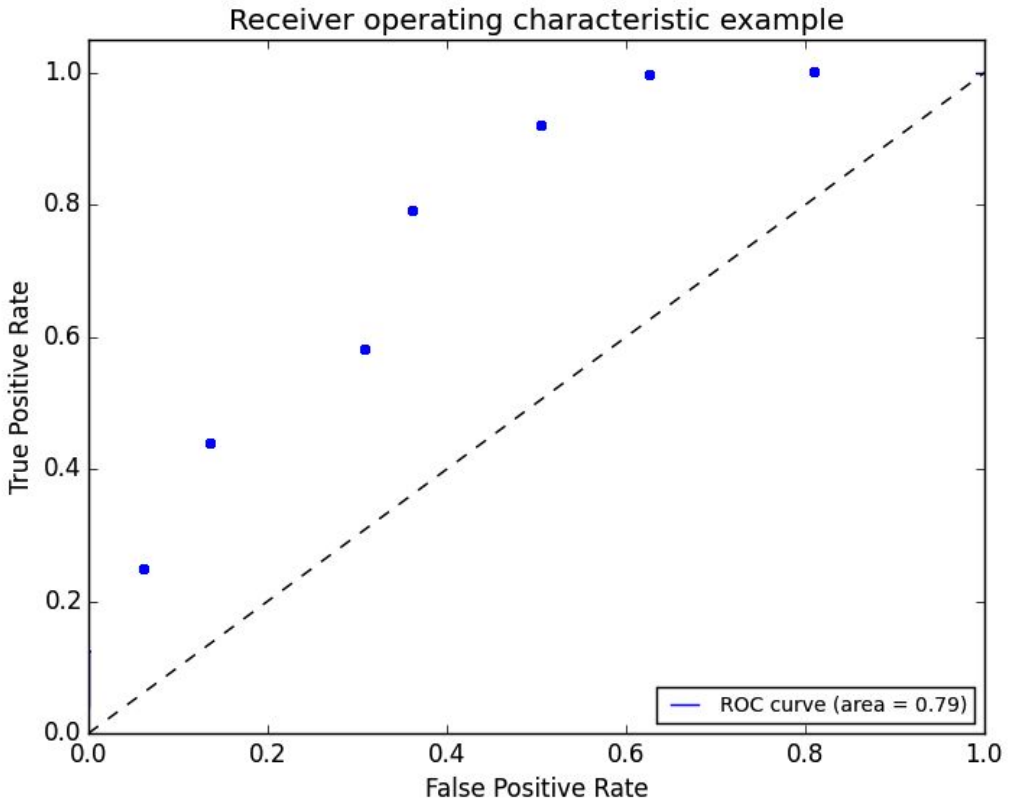
ADVANCED CLASSIFICATION METRICS

- We can continue adding pairs for different thresholds



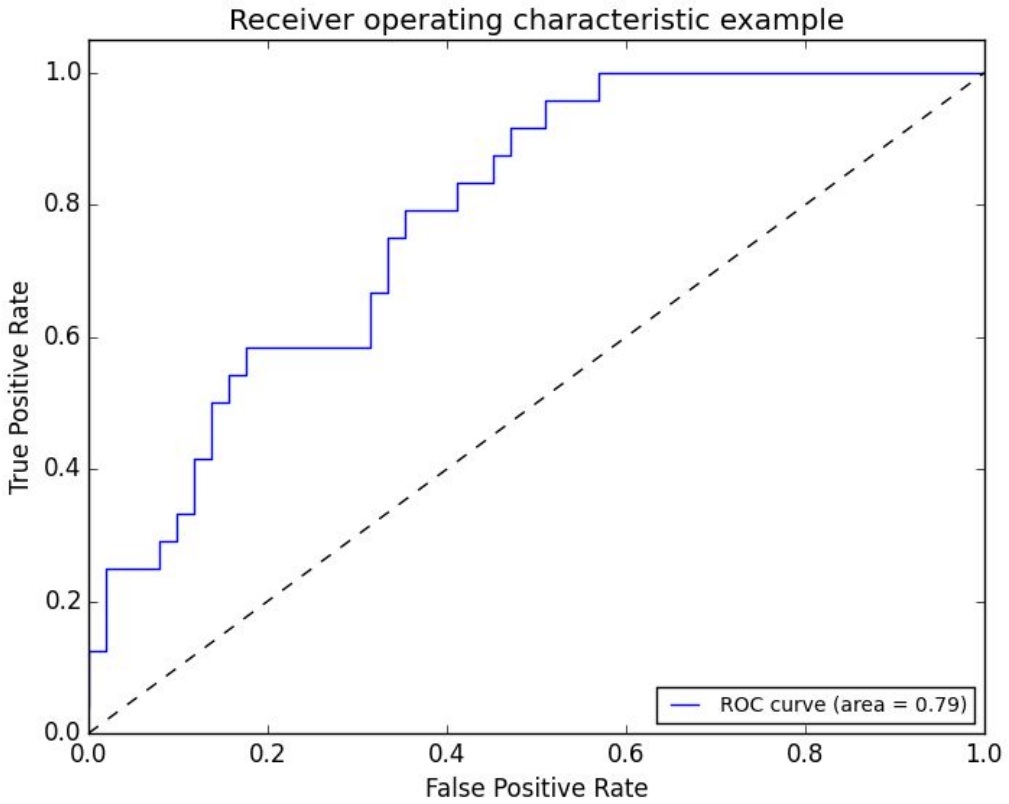
ADVANCED CLASSIFICATION METRICS

- We can continue adding pairs for different thresholds



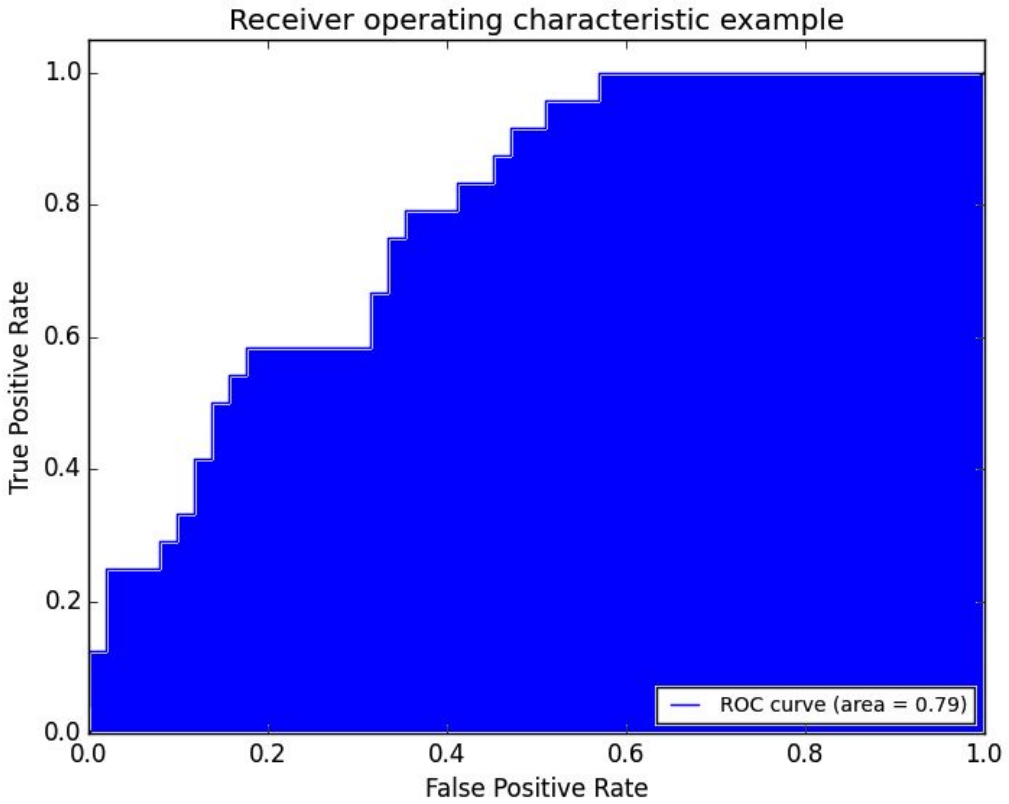
ADVANCED CLASSIFICATION METRICS

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



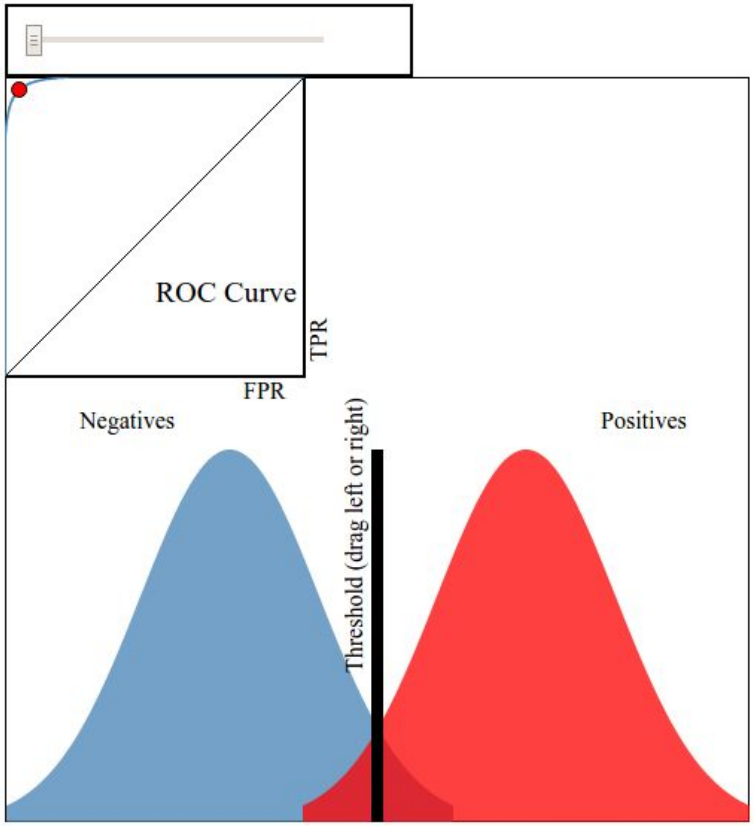
ADVANCED CLASSIFICATION METRICS

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



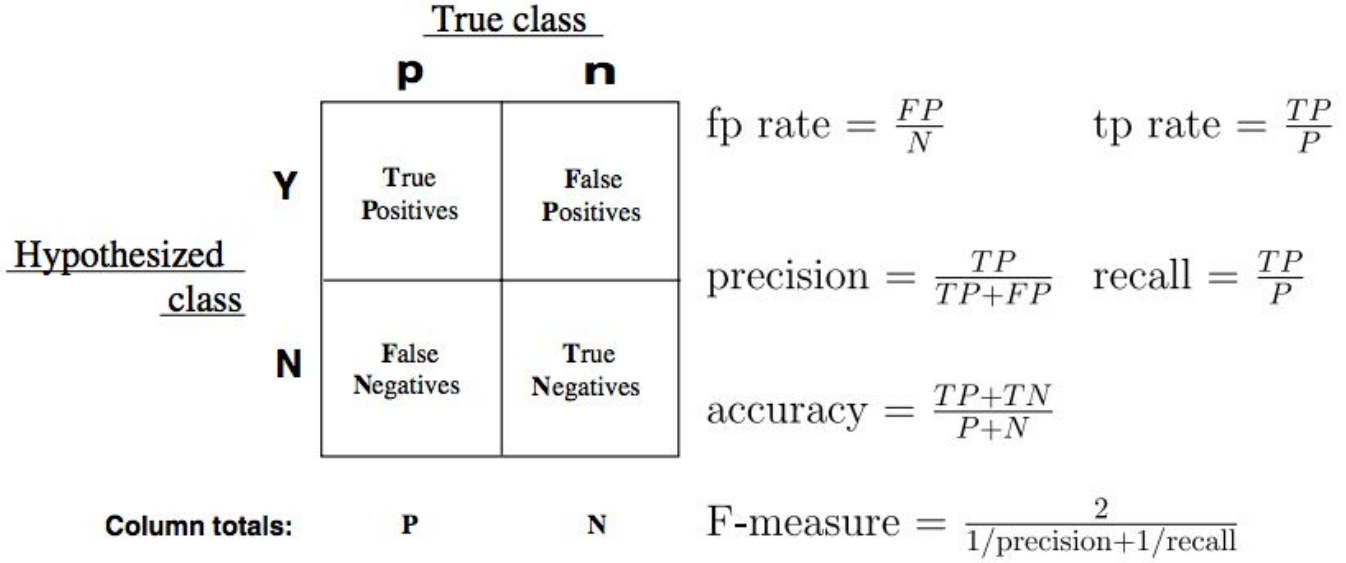
ADVANCED CLASSIFICATION METRICS

- This interactive visualization can help practice visualizing ROC curves.



ADVANCED CLASSIFICATION METRICS

- 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?

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.
4. What's an AUC of 0.5

DELIVERABLE

Answers for each example

INDEPENDENT PRACTICE

EVALUATING LOGISTIC REGRESSION WITH ALTERNATIVE METRICS

ACTIVITY: EVALUATING LOGISTIC REGRESSION

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

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?

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 do the *coefficients* in a logistic regression represent? How does the interpretation differ from ordinary least squares? How is it similar?
- What would an AUC of 0.5 represent for a model? What about an AUC of 0.9?

REVIEW TOPICS

- Define and apply the KNN algorithm
- Dive into the mechanisms of Logistic Regression:
  - Difference with the linear regression algorithm
  - Loss function (also called Cost function)
  - Evaluation of your classification model

COURSE

BEFORE NEXT CLASS

BEFORE NEXT CLASS

DUE DATE

- Unit Project 3: start modelling assignment

LESSON

Q & A