

Advanced Metrics and Communicating Results

Ivan Corneillet

Data Scientist

Learning Objectives

After this lesson, you should be able to:

- Evaluate a model using advanced metrics such as confusion matrix and ROC/AUC curves
- Explain the trade-offs between the precision and recall of a model while articulating the cost of false positives vs. false negatives
- Describe the difference between visualization for presentations vs. exploratory data analysis
- Identify the components of a concise and convincing report and how they relate to specific audiences/stakeholders

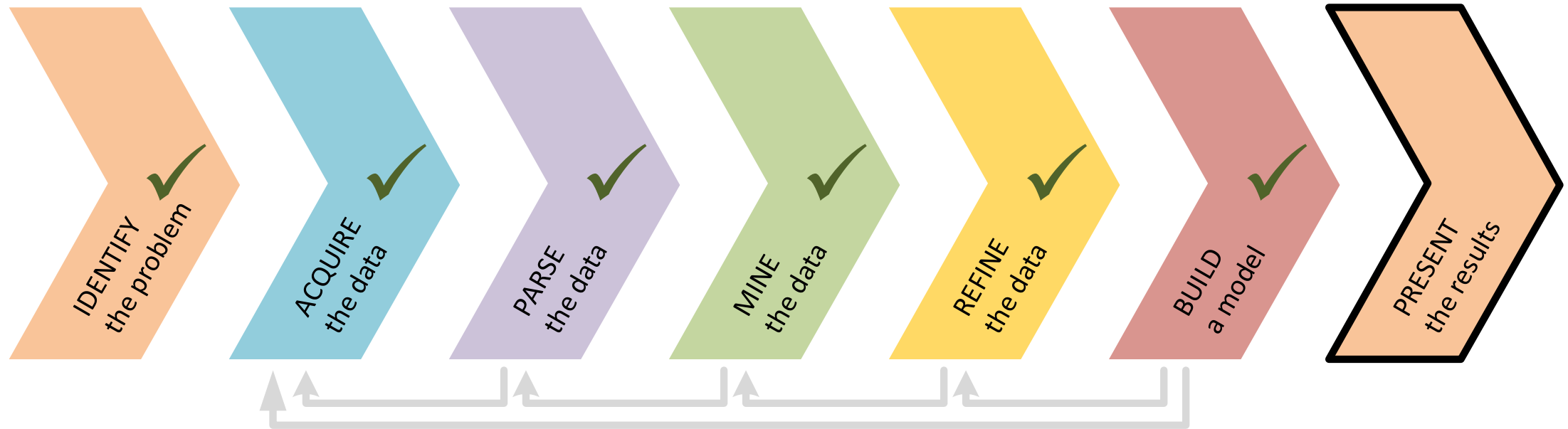
Outline

- Unit Project 4 (due today)
- Review (logistic regression)
- Advanced metrics
 - Confusion Matrix
 - True Positive, False Positive Rates, ROC, and AUC
 - Codealong for ROC/AUC
 - Precision, Recall, and F-score
 - Specificity and Prevalence
- Communicating Results
 - Showing our Work
 - Codealong to pretty up graphs
- Unit Project 4's Presentations
- Continue last session's lab (Part B)
- Review
- In-flight
 - Final Project 2 (due in 1.5 weeks)

Today we are wrapping Unit 2 – Foundation of Modeling

<i>Unit 1 – Research Design and Data Analysis</i>	<i>Research Design</i>	<i>Data Visualization in Pandas</i>	<i>Statistics</i>	<i>Exploratory Data Analysis in Pandas</i>
Unit 2 – Foundations of Modeling	Linear Regression <i>(sessions 6 and 10)</i>	Classification Models (KNN, Logistic Regression) <i>(sessions 8, 9, and 10)</i>	Evaluating Model Fit <i>(sessions 7, 10, and 11)</i>	Presenting Insights from Data Models <i>(session 11)</i>
<i>Unit 3 – Data Science in the Real World</i>	<i>Decision Trees and Random Forest</i>	<i>Time Series Data</i>	<i>Natural Language Processing</i>	<i>Databases</i>

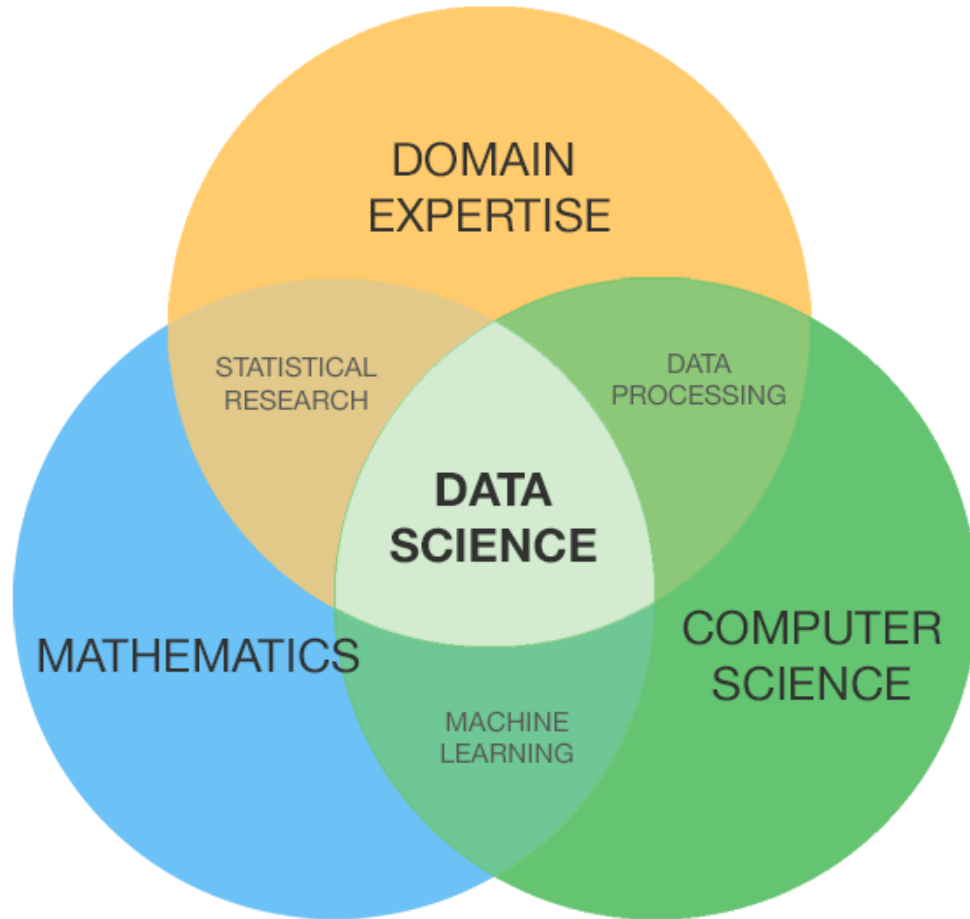
... as well as the first full pass of the Data Science Workflow



DS

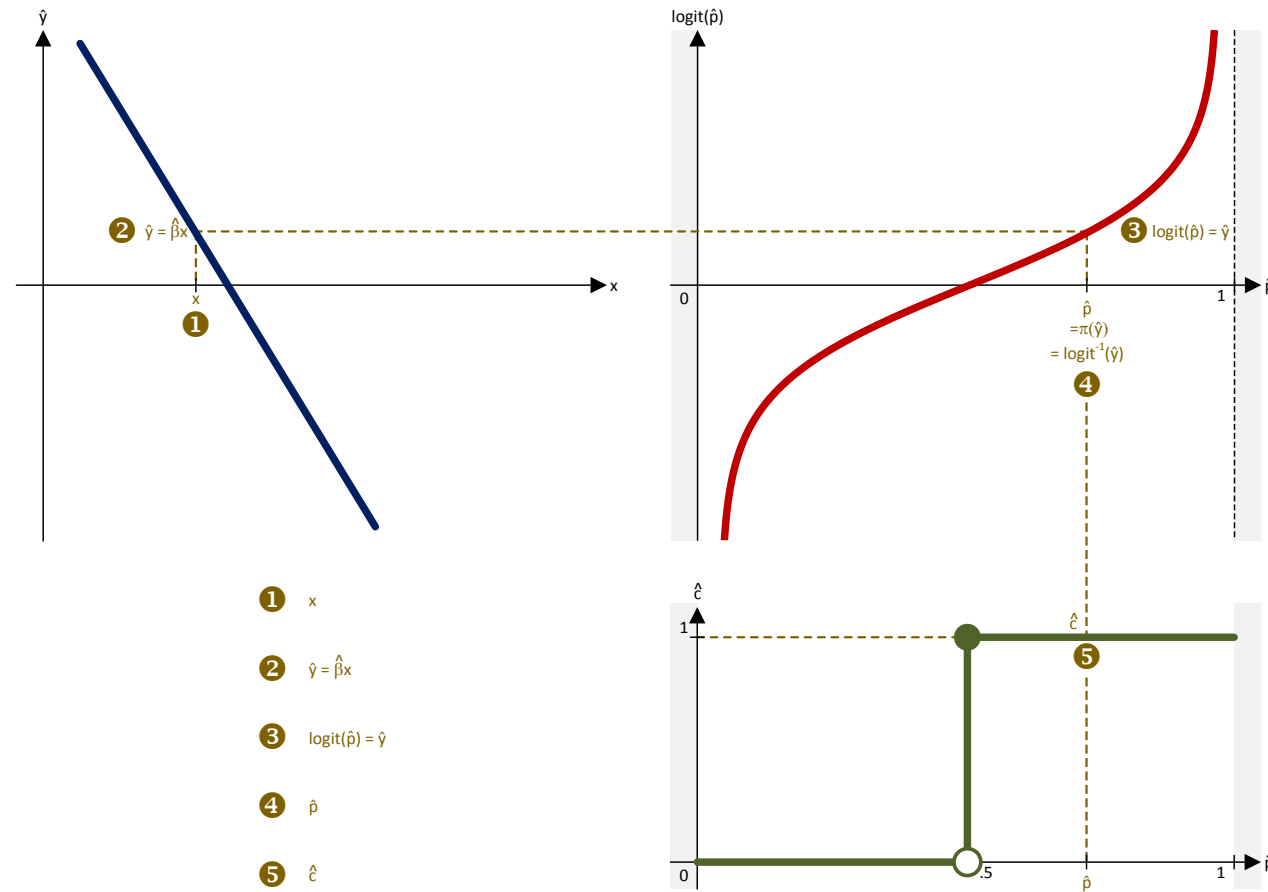
Review

The previous session was about practice,
practice, and practice...



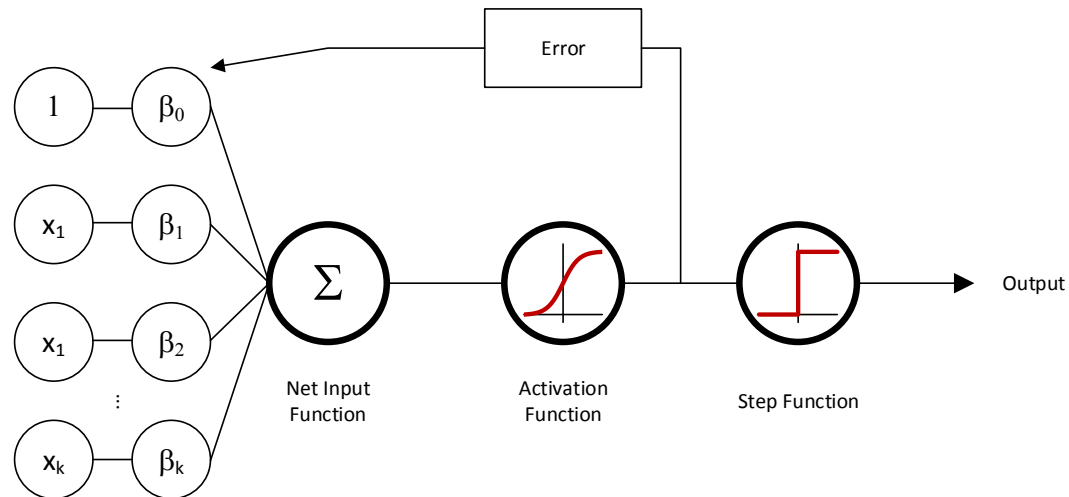
1 0 0 0 0
H O U R S

Logistic Regression and the \hat{y} , \hat{p} , and \hat{c} spaces

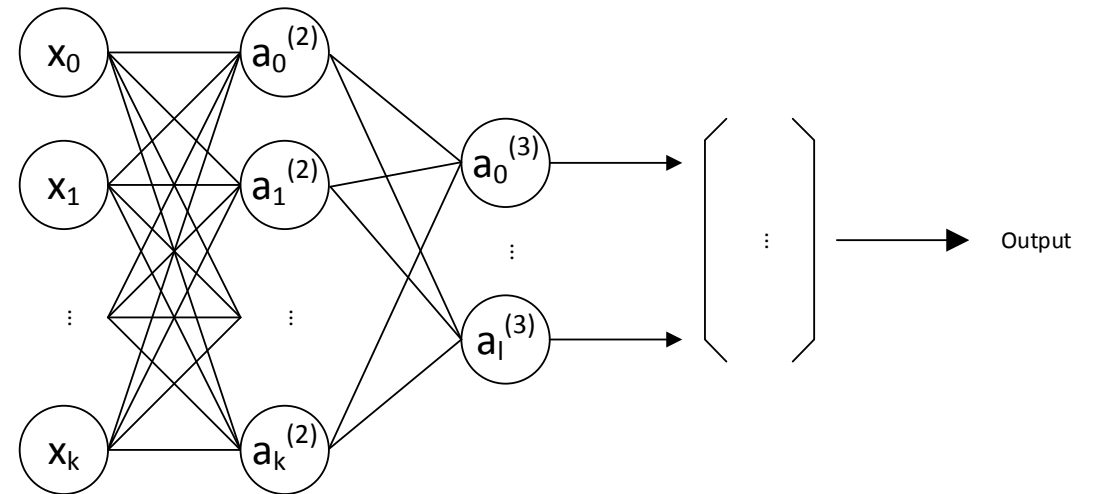


Opening up: neural networks are related to logistic regression; you can think of logistic regression as a one-layer neural network

One-layer neural network



Multi-layer neural network



DS

Pre-Work

Pre-Work

Before this lesson, you should already be able to:

- Create and interpret results from a binary classification problem
- Know what a decision line is in logistic regression



DS

Advanced Metrics

Advanced Metrics – Motivation

- Accuracy is only one of several metrics used when solving for a classification problem
 - E.g., if we know a prediction is 75% accurate, accuracy doesn't provide any insight into why the 25% was wrong. Was it wrong *equally* across all class labels? Did it just guess one class label for all predictions and 25% of the data was just the other label?
- It's important to look at other metrics to fully understand the problem

▸ Accuracy

- How many observations that we predicted were correct? This is a value we'd want to increase (like R^2)

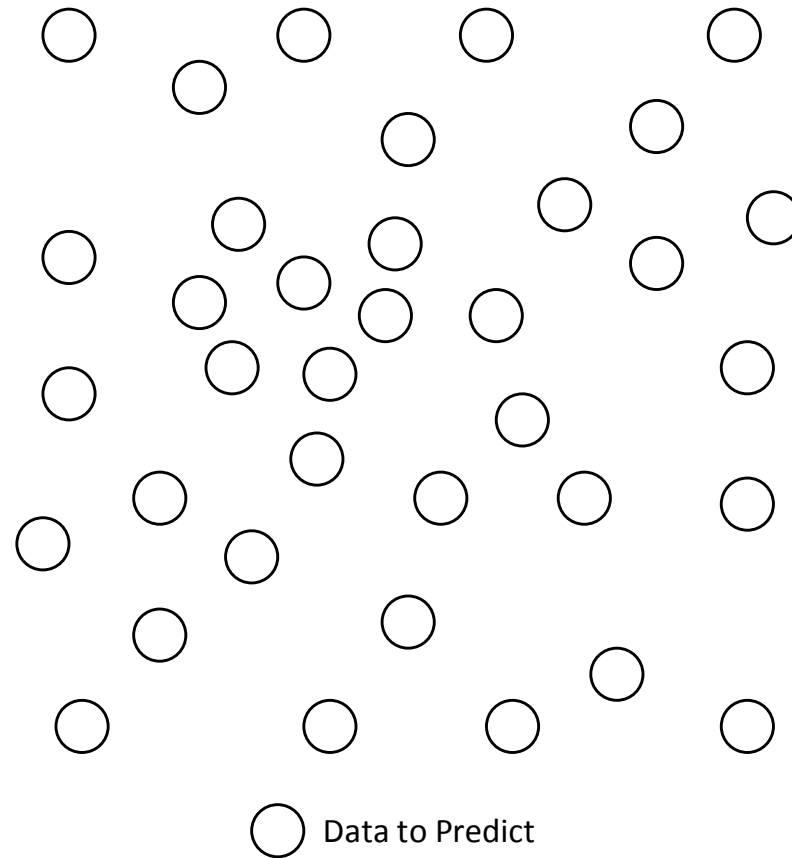
▸ Misclassification rate

- Directly opposite of accuracy
- Of all the observations we predicted, how many were incorrect? This is a value we'd want to decrease (like the mean squared error)

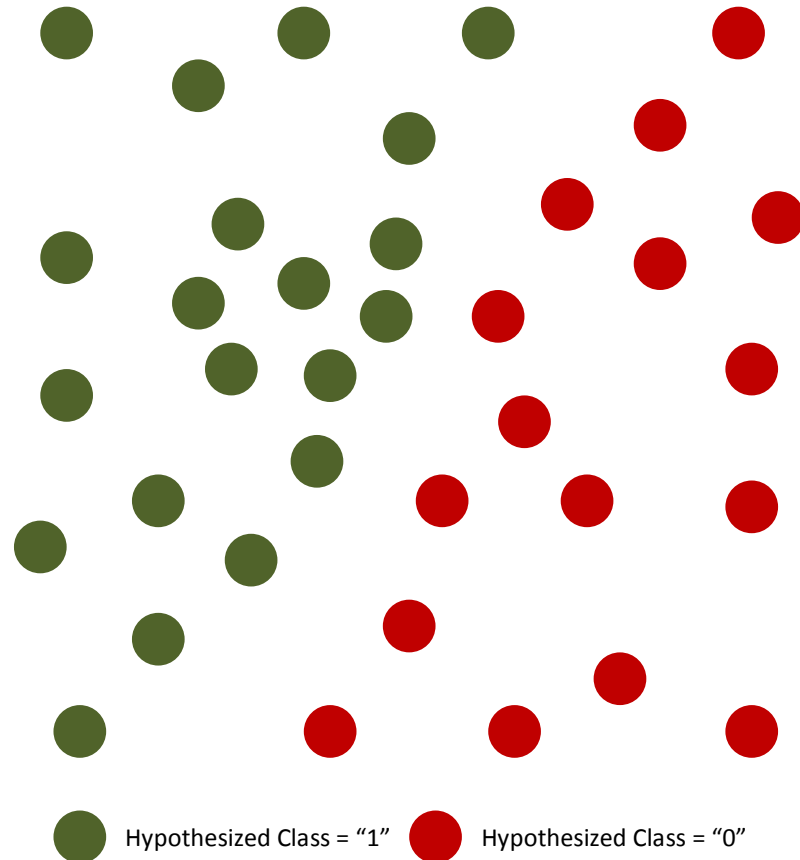
DS

Confusion Matrix

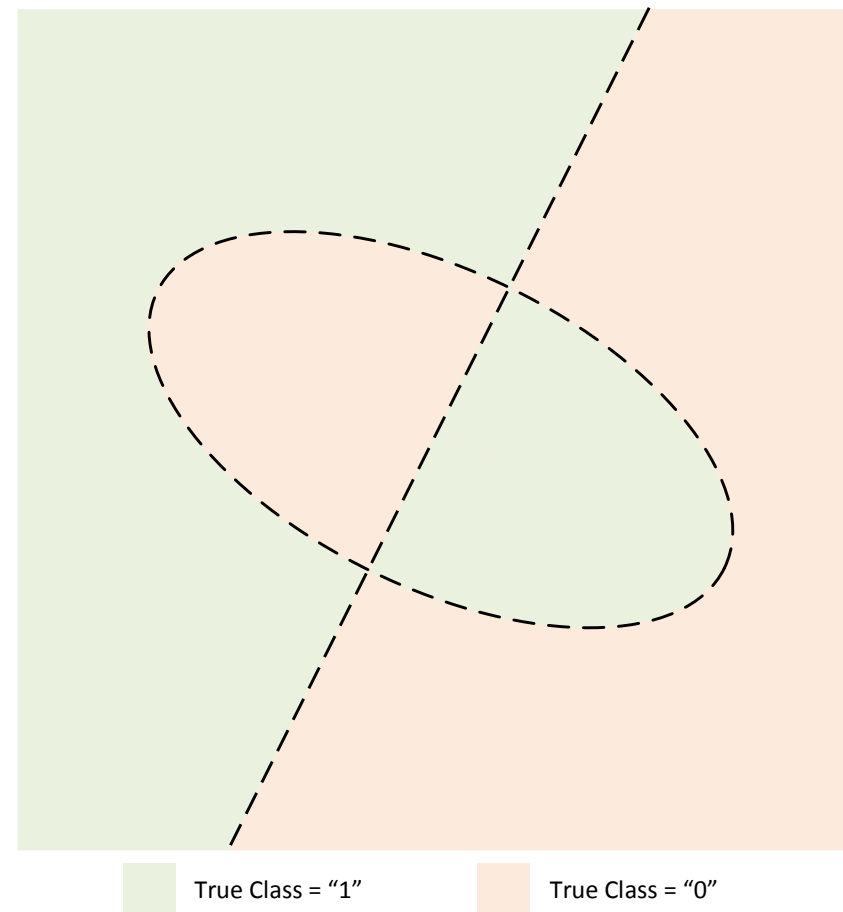
Stepping back: Let's say we want to classify this data



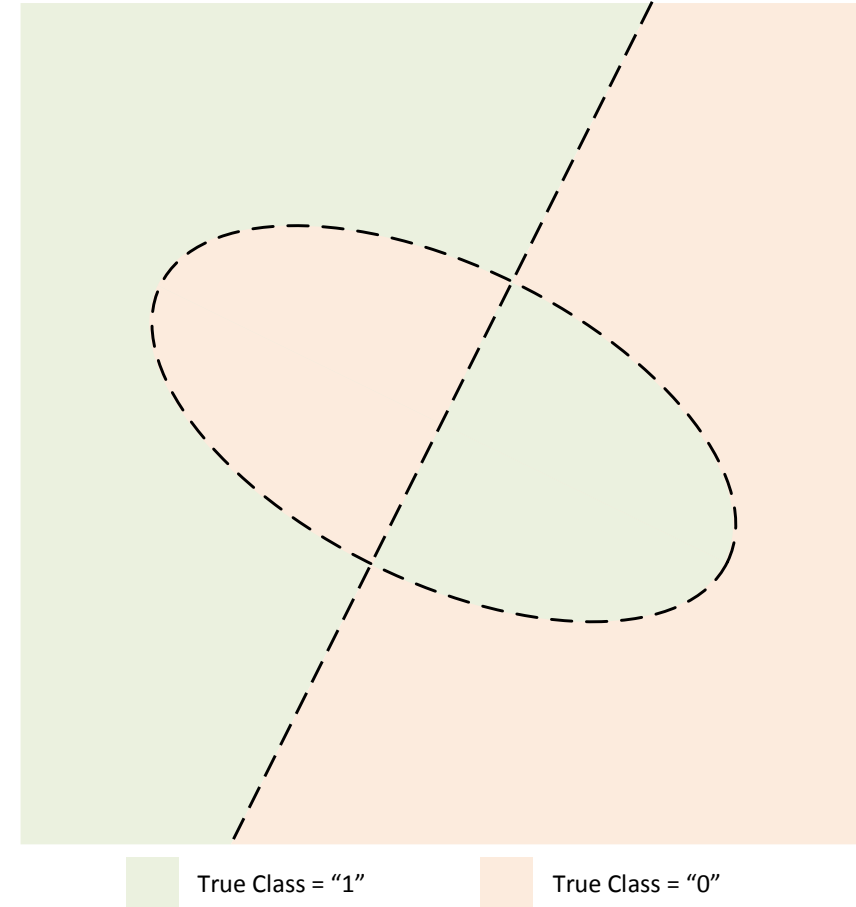
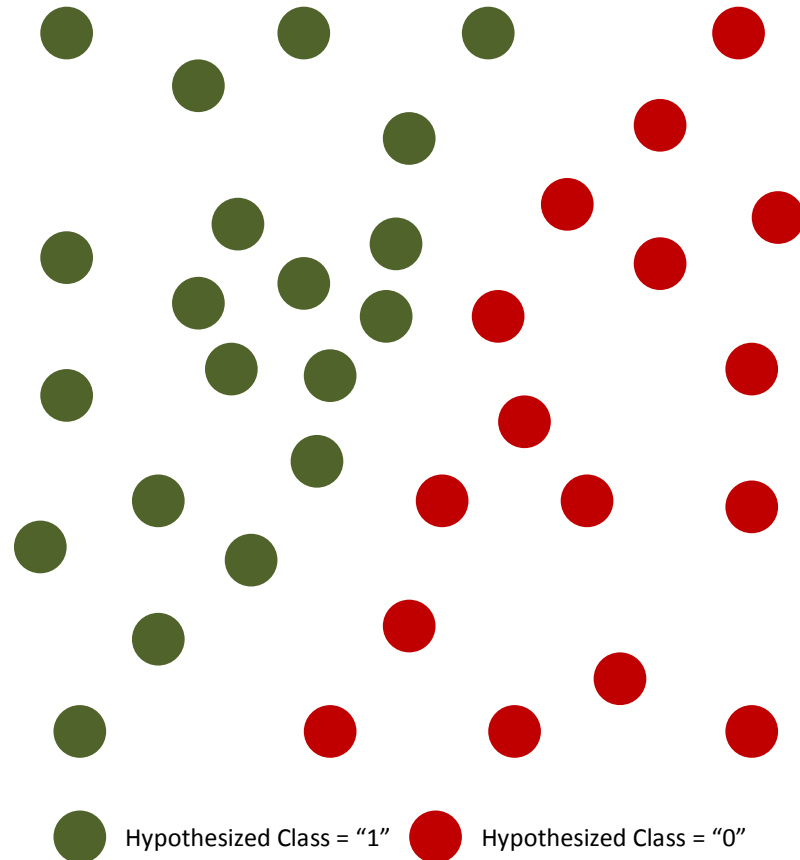
This is our prediction (the hypothesized classes)



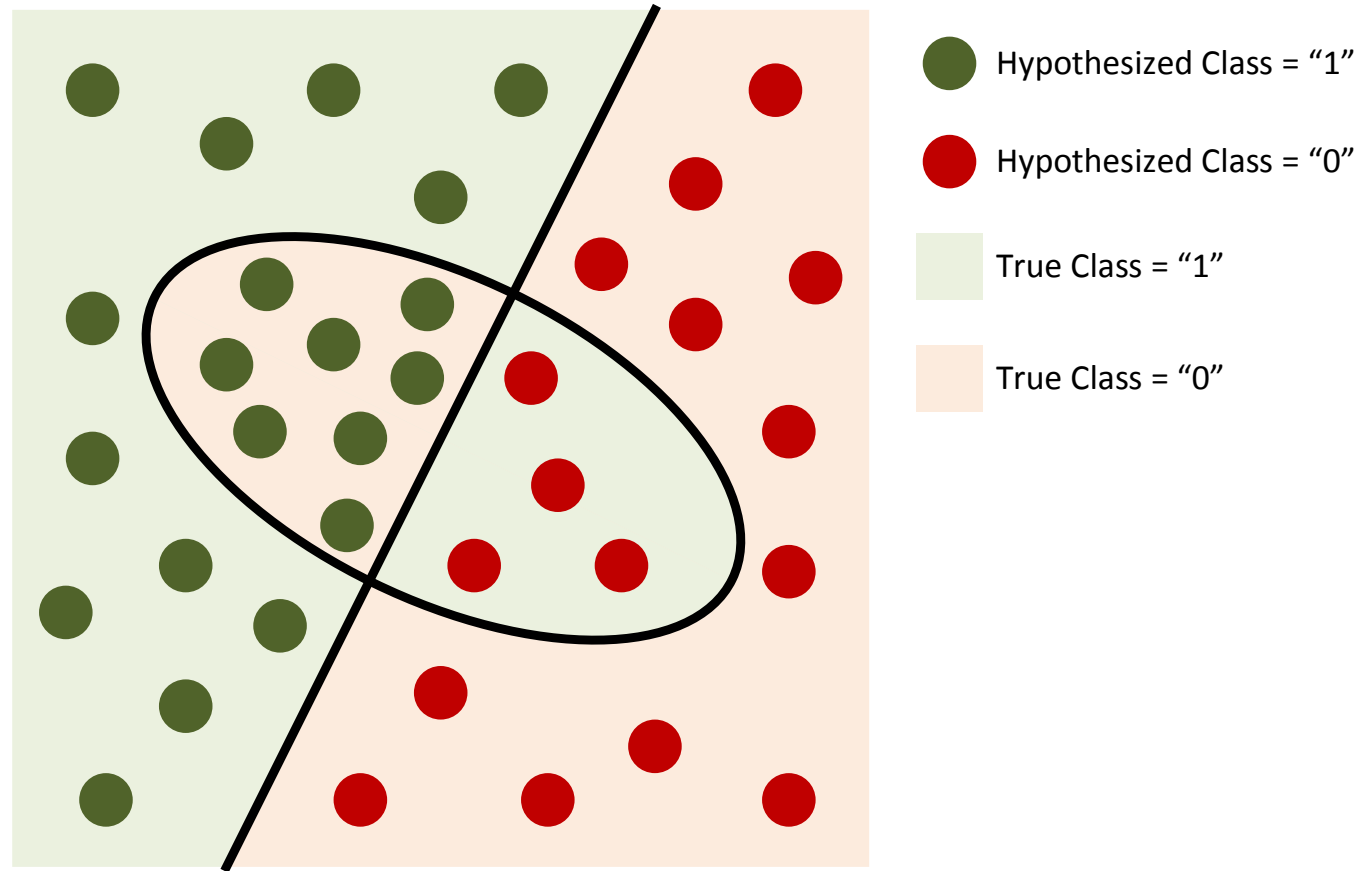
And this is the “true” (the true classes)



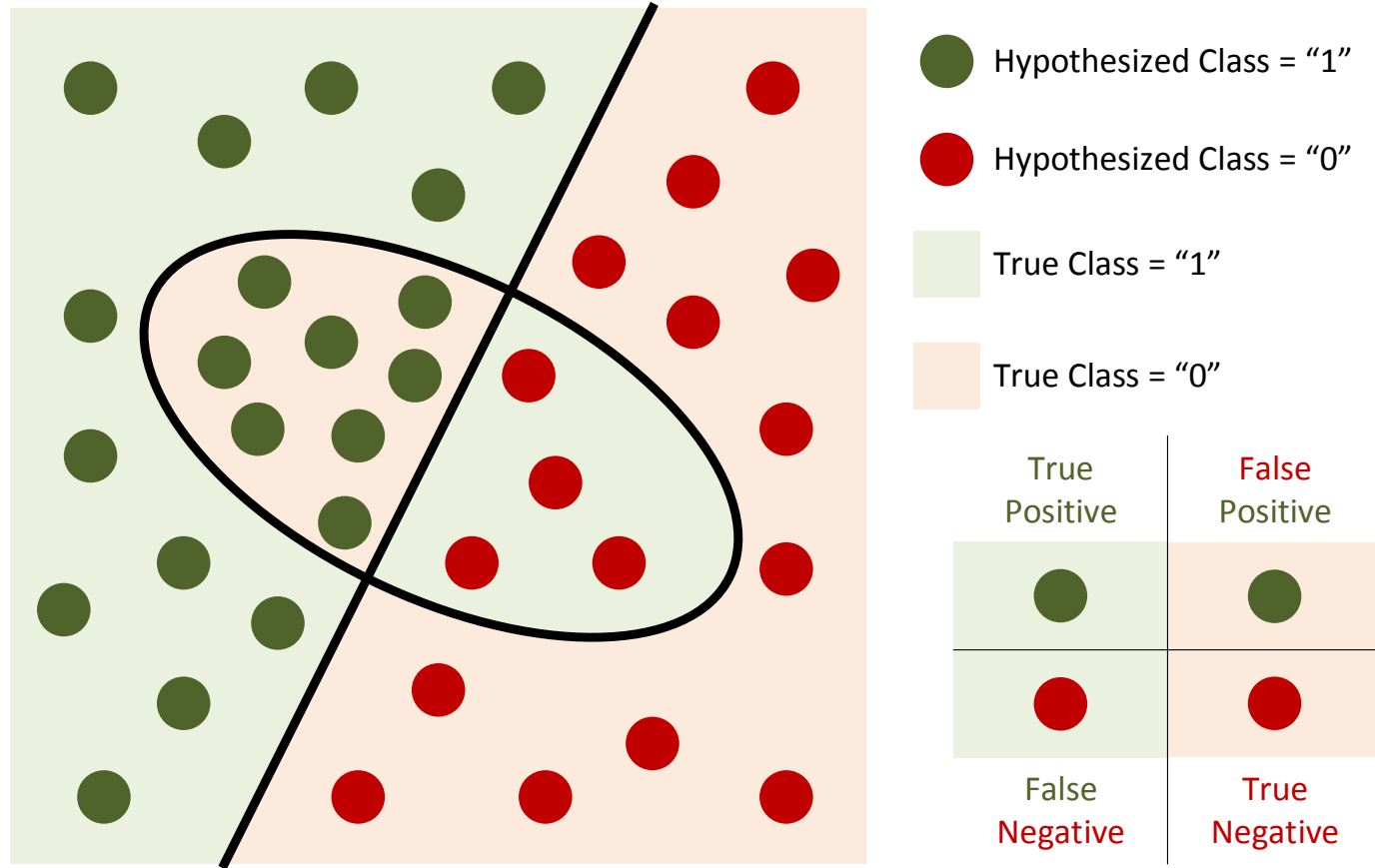
Hypothesized and true classes don't necessarily match







Putting hypothesized and true classes together, we get four possibilities



We can rearrange these four possibilities into a 2x2 table



Confusion Matrix (a.k.a., Contingency Table or Error Matrix)

		True Class	
		1	0
Hypothesized Class	1	 True Positives (TP)	 False Positives (FP) <i>(type I error)</i>
	0	 False Negatives (FN) <i>(type II error)</i>	 True Negatives (TN)
Total Columns		$P = TP + FN$	$N = FP + TN$

- A confusion matrix is a specific table layout that allows visualization of the performance of a supervised learning algorithm
- Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class
- The name stems from the fact that it makes it easy to see if the system is confusing two classes (i.e., commonly mislabeling one as another)

DS

Activity

Activity: Interpreting the confusion matrix



EXERCISE

ANSWER THE FOLLOWING QUESTIONS (20 minutes)

1. Use the variables defined in the confusion matrix (TP , FN , FP , TN , P , and N) to calculate the answers to the following questions:
 - a. Overall, how often is the classifier correct?
 - b. When the classifier predicts yes, how often is it correct?
 - c. How often does the yes condition actually occur in our sample?
 - d. When it's actually yes, how often does the classifier predict yes?
 - e. When it's actually no, how often does the classifier predict yes?
 - f. When it's actually no, how often does it predict no?
 - g. Overall, how often is the classifier wrong?

Activity: Interpreting the confusion matrix (cont.)



EXERCISE





ANSWER THE FOLLOWING QUESTIONS (20 minutes)

2. Given a medical exam that tests for cancer ($1 = \text{Cancer}$, $0 = \text{Cancer free}$), use the variables defined in the confusion matrix (TP , FN , FP , TN , P , and N) to calculate the answers to the following questions:
 - a. How often is it correct when it identify patients with cancer?
 - b. How often does it correctly identify patients without cancer?
 - c. How often does it trigger a “false alarm” by saying a patient has cancer when they actually don’t?
 - d. How often does it correctly identify patients with cancer?
3. When finished, share your answers with your table

DELIVERABLE





Answers to the above questions

Activity: Interpreting the confusion matrix (cont.)

		True Class	
		1	0
Hypothesized Class	1	 True Positives (TP)	 False Positives (FP) <i>(type I error)</i>
	0	 False Negatives (FN) <i>(type II error)</i>	 True Negatives (TN)
Total Columns		$P = TP + FN$	$N = FP + TN$

Question: Overall, how often is the classifier correct? Answer: $\frac{TP+TN}{P+N}$	When the classifier predicts yes, how often is it correct? Answer: $\frac{TP}{TP+FP}$
How often does the yes condition actually occur in our sample? Answer: $\frac{P}{P+N}$	When it's actually yes, how often does the classifier predict yes? Answer: $\frac{TP}{P}$
When it's actually no, how often does the classifier predict yes? Answer: $\frac{FP}{N}$	When it's actually no, how often does it predict no? Answer: $\frac{TN}{N}$
Overall, how often is the classifier wrong? Answer: $\frac{FP+FN}{P+N}$	

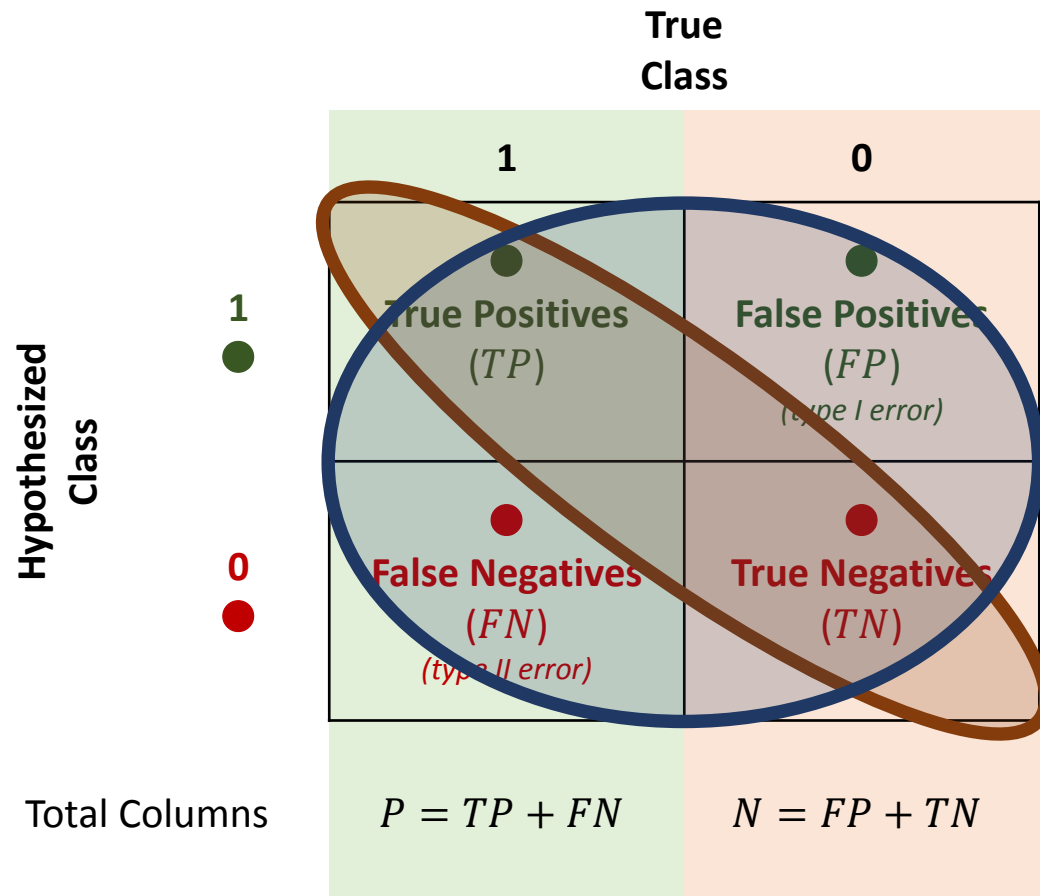
Activity: Interpreting the confusion matrix (cont.)

		True Class			
		Has Cancer	Doesn't have cancer		
Hypothesized Class	Predict Cancer	 True Positives (TP)	 False Positives (FP) <i>(type I error)</i>	<i>How often is it correct when it identify patients with cancer?</i> Answer: $\frac{TP}{TP+FP}$	<i>How often does it correctly identify patients without cancer?</i> Answer: $\frac{TN}{N}$
	Predict No Cancer	 False Negatives (FN) <i>(type II error)</i>	 True Negatives (TN)	<i>How often does it trigger a "false alarm" by saying a patient has cancer when they actually don't?</i> Answer: $\frac{FP}{N}$	<i>How often does it correctly identify patients with cancer?</i> Answer: $\frac{TP}{P}$
Total Columns		$P = TP + FN$	$N = FP + TN$	<i>How many patients have cancer?</i> Answer: $\frac{P}{P+N}$	

DS

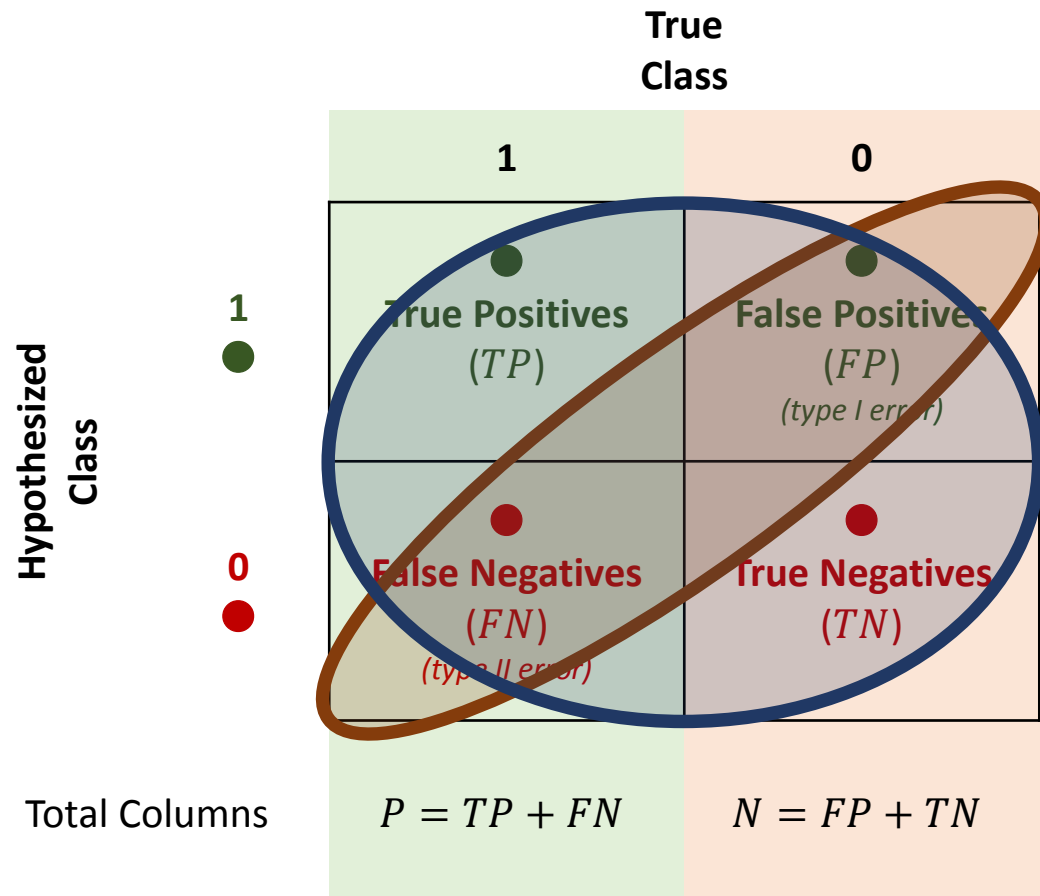
Accuracy and Misclassification Rate – Take 2

Accuracy Revisited, $\frac{TP+TN}{P+N}$ (Take 2)



► Overall, how often is the classifier correct?

Misclassification Rate, $\frac{FP+FN}{P+N}$ (Take 2)







► Overall, how often is the classifier wrong?

DS

True Positive, False Positive Rates, ROC, and AUC

True Positive Rate, $TPR = \frac{TP}{P}$

		True Class	
		1	0
Hypothesized Class	1	 True Positives (TP)	 False Positives (FP) (type I error)
	0	 False Negatives (FN) (type II error)	 True Negatives (TN)
Total Columns		$P = TP + FN$	$N = FP + TN$





▸ When it's actually yes, how often does the classifier predict yes?

▸ A.k.a., "Sensitivity"

▸ E.g., given a medical exam that tests for cancer, how often does it correctly identify patients with cancer?

▸ Likewise, this can be inverted: how often does a test *correctly* identify patients without cancer

False Positive Rate, $FPR = \frac{FP}{N}$

		True Class	
		1	0
Hypothesized Class	1	 True Positives (TP)	 False Positives (FP) <i>(type I error)</i>
	0	 False Negatives (FN) <i>(type II error)</i>	 True Negatives (TN)
Total Columns		$P = TP + FN$	$N = FP + TN$

‣ When it's actually no, how often does the classifier predict yes?

‣ A.k.a., “Fall-out”

‣ E.g., given a medical exam that tests for cancer, how often does it trigger a “false alarm” by saying a patient has cancer when they actually don't?

‣ Likewise, this can be also inverted: how often does a test *incorrectly* identify patients as being cancer-free when they might actually have cancer!

True Positive and False Positive Rates

- We can split up the accuracy of each label by using true positive and false positive rates. Using them, we can get a much clearer picture of where predictions begin to fall apart
- A good classifier would have a true positive rate approaching 1, and a false positive rate approaching 0. In a binary problem (say, predicting if someone smokes or not), it would accurately predict all of the smokers as smokers, and not accidentally predict any of the non-smokers as smokers

True Positive and False Positive Rates (cont.)

- We can split up the accuracy of each label by using true positive and false positive rates
- Using them, we can get a much clearer picture of where predictions begin to fall apart

Activity: Introduction to the ROC Space



EXERCISE

ANSWER THE FOLLOWING QUESTIONS (5 minutes)

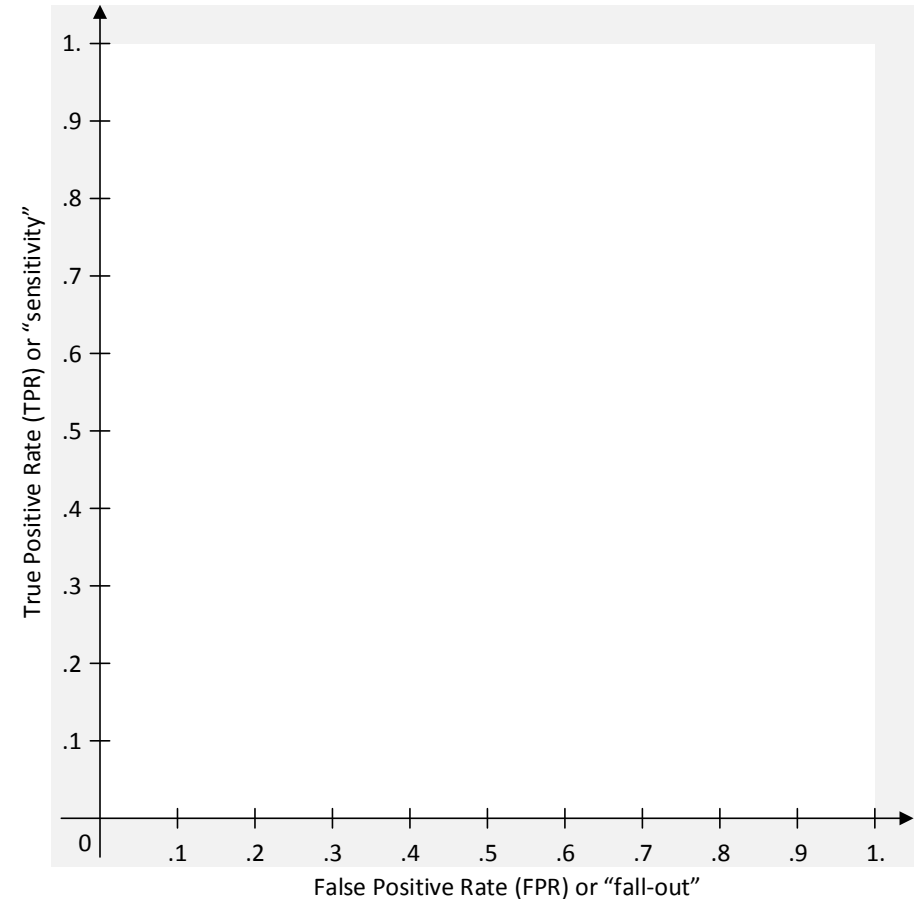
1. Calculate TPR and FPR for the four confusion matrices in the handout and place them in the ROC space (TPR as a function of FPR)
2. How would you classify these four cases as a function of their performance (e.g., better or worse)
3. What does the ROC space tells you?
4. When finished, share your answers with your table

DELIVERABLE

Answers to the above questions

ROC (receiver operating characteristic) curve (a.k.a., relative operating characteristic curve)

- An ROC curve plots the true positive rate (TPR) (or “sensitivity”) against the false positive rate (FPR) (or “fall-out”) at various threshold settings to illustrate the performance of a binary classifier system. The ROC curve is thus the sensitivity as a function of fall-out



Activity: Introduction to the ROC Space (cont.)

EXERCISE

A

$$TPR = \frac{63}{63 + 37} = .63$$

$$FPR = \frac{28}{28 + 72} = .28$$

B

$$TPR = \frac{77}{77 + 23} = .77$$

$$FPR = \frac{77}{77 + 23} = .77$$

C

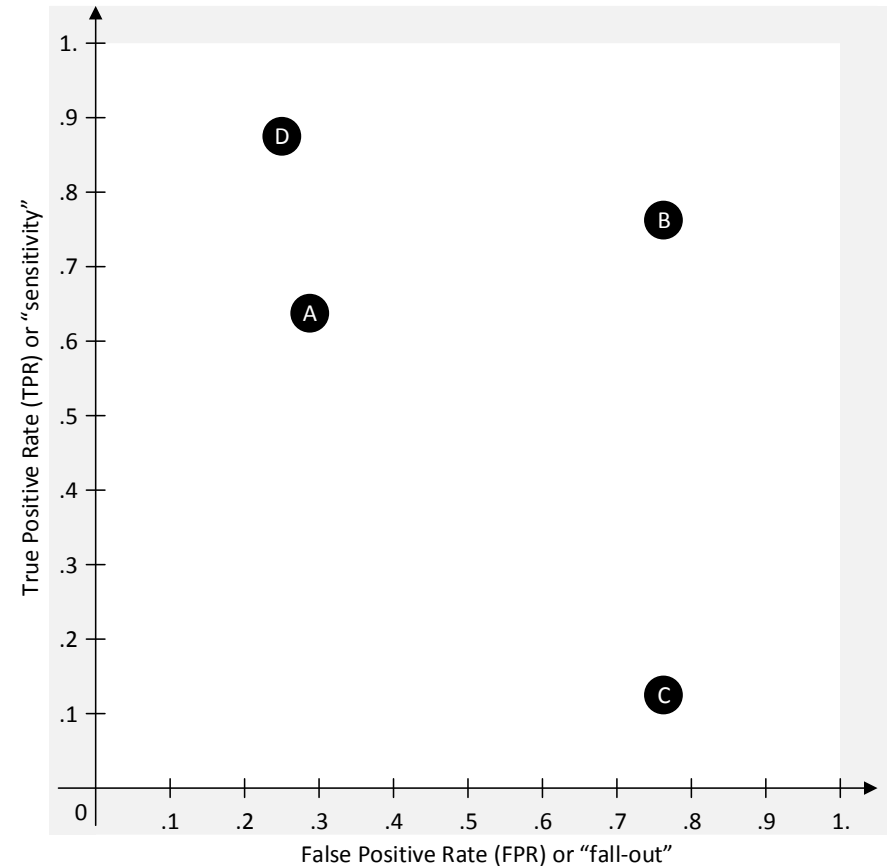
$$TPR = \frac{24}{24 + 76} = .24$$

$$FPR = \frac{88}{88 + 12} = .88$$

D

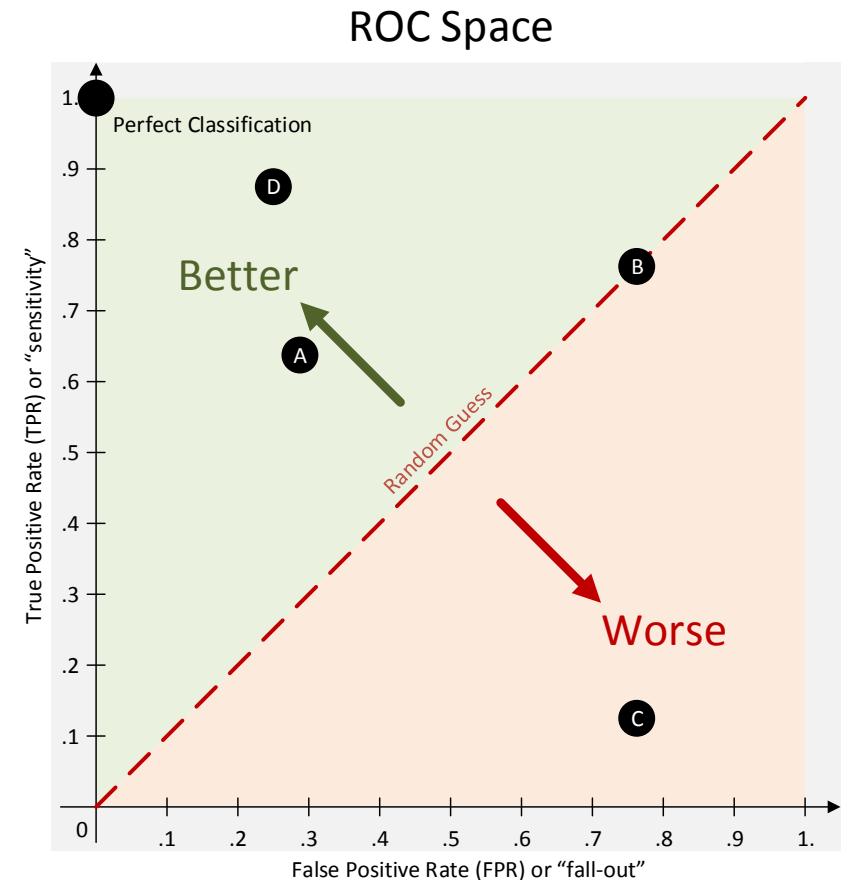
$$TPR = \frac{76}{76 + 24} = .76$$

$$FPR = \frac{12}{12 + 88} = .12$$



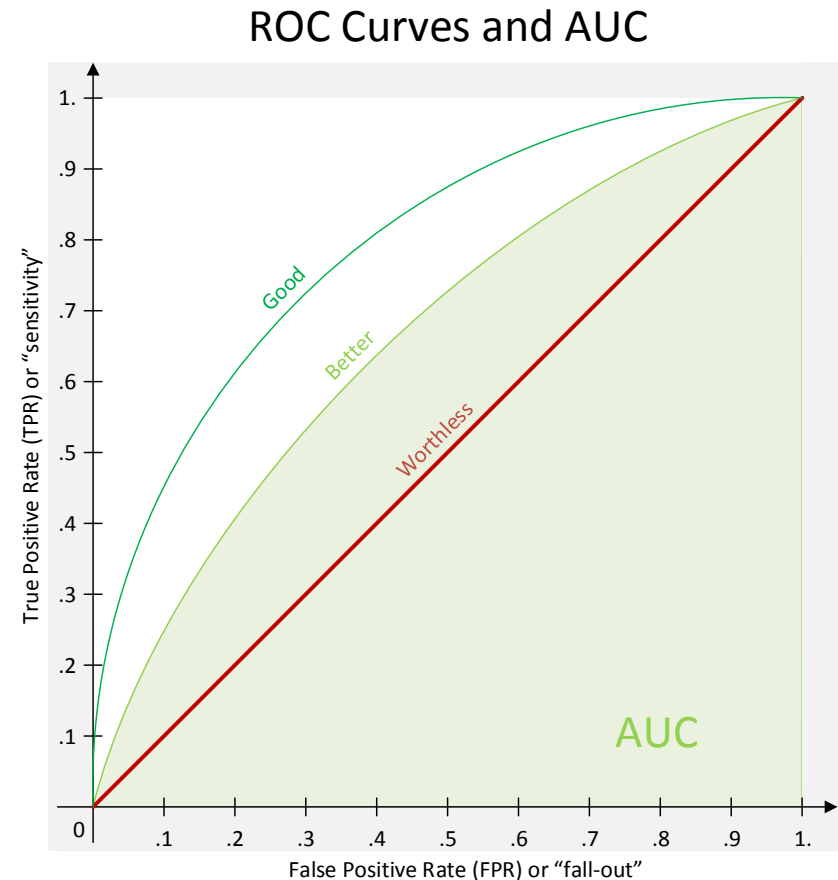
The ROC space demonstrates several things:

- It shows the tradeoff between sensitivity and fall-out (any increase in sensitivity will be accompanied by an increase in fallout)
 - The closer the points are in the left-hand border and then the top border of the ROC space, the more accurate the classifier is
 - The closer the points come to the 45-degree diagonal of the ROC space, the less accurate the classifier is



The ROC curves demonstrate several things:

- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the classifier is
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the classifier is
- The area under the curve (AUC) is a measure of classifier accuracy



DS





Codealong – Part A

ROC/AUC

DS

Precision, Recall, and F-score





$$\text{Precision, } \textit{precision} = \frac{TP}{TP+FP}$$

		True Class	
		1	0
Hypothesized Class	1	 True Positives (TP)	 False Positives (FP) <i>(type I error)</i>
	0	 False Negatives (FN) <i>(type II error)</i>	 True Negatives (TN)
Total Columns		$P = TP + FN$	$N = FP + TN$

▸ When the classifier predicts yes, how often is it correct?

- E.g., given a medical exam that tests for cancer, how often is it correct when it identifies patients with cancer?
- E.g., when searching on Google/Bing, precision is “how useful the search results are”

$$\text{Recall, } recall = \frac{TP}{P} = TPR$$

		True Class	
		1	0
Hypothesized Class	1	 True Positives (TP)	 False Positives (FP) (type I error)
	0	 False Negatives (FN) (type II error)	 True Negatives (TN)
Total Columns		$P = TP + FN$	$N = FP + TN$

- When it's actually yes, how often does the classifier predict yes?
 - A.k.a., True Positive Rate (TPR)
- E.g., given a medical exam that tests for cancer, how often does it correctly identify patients with cancer?
 - E.g., when searching on Google/Bing, recall is “how complete the results are”

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

- E.g., 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 (cont.)

- ▶ With *precision*, we're interested in producing 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?"
- ▶ With *recall*, we're interesting in seeing 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?"

Precision and recall (cont.)

- The key difference between the two is the attribution and value of an error:
 - Should our model be more picky in avoiding false positives (*precision*), or should it be more picky in avoiding false negatives (*recall*)?
- The answer should be determined by the problem you're trying to solve

F-score (a.k.a., F_1 -score, or F-measure)

- The F-score is the weighted average of precision and recall (the harmonic mean in fact), attempting to combine both measures into one





$$\frac{2}{F} = \frac{1}{precision} + \frac{1}{recall}$$

- The F-score reaches
 - Its best value at 1 (when both precision and recall are 1)
 - And its worst value at 0 (when both precision and recall are 0)

DS

Specificity and Prevalence

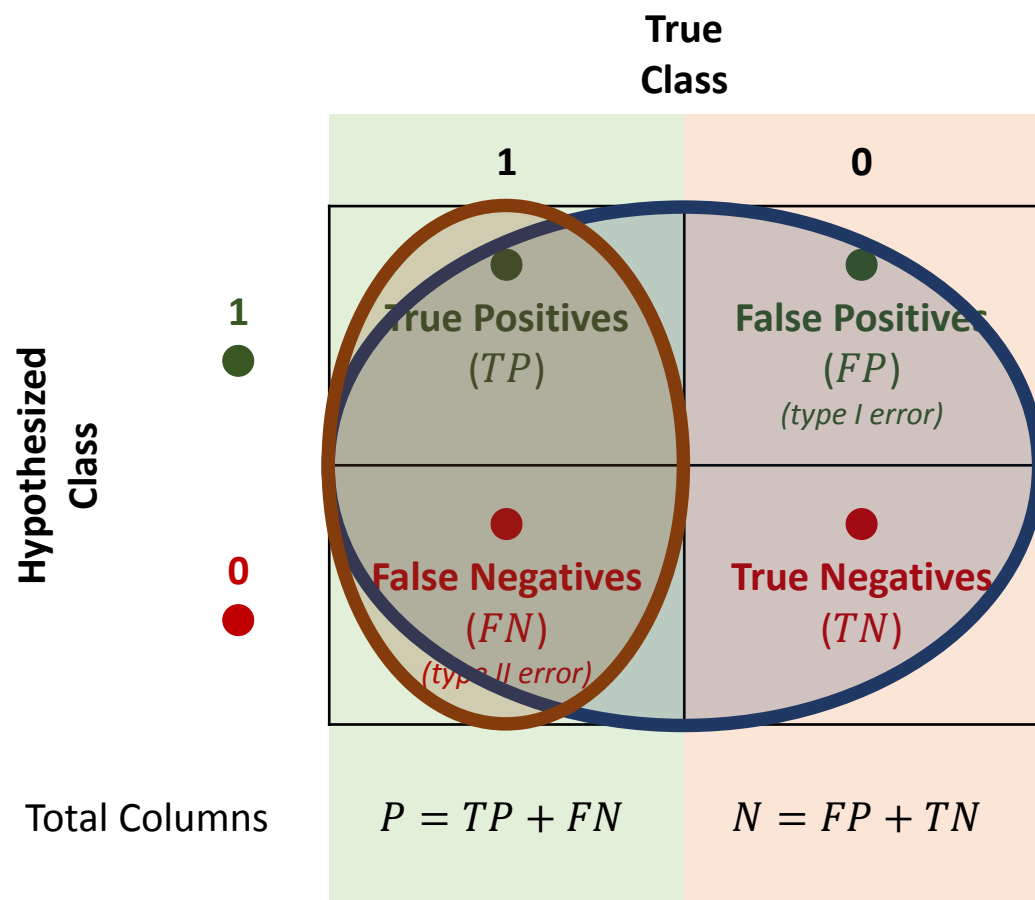
$$\text{Specificity, } \textit{specificity} = \frac{TN}{N} = 1 - FPR$$

		True Class	
		1	0
Hypothesized Class	1	 True Positives (TP)	 False Positives (FP) <i>(type I error)</i>
	0	 False Negatives (FN) <i>(type II error)</i>	 True Negatives (TN)
Total Columns		$P = TP + FN$	$N = FP + TN$

▸ When it's actually no, how often does it predict no?

▸ E.g., given a medical exam that tests for cancer, how often does it correctly identify patients without cancer?

$$\text{Prevalence, } prevalence = \frac{P}{P+N}$$







- ▶ How often does the yes condition actually occur in our sample?
- ▶ E.g., given a medical exam that tests for cancer, how many patients have cancer?

DS

Confusion Matrix (cont.)

Confusion Matrix

		True Class	
		1	0
Hypothesized Class	1	 True Positives (TP)	 False Positives (FP) <i>(type I error)</i>
	0	 False Negatives (FN) <i>(type II error)</i>	 True Negatives (TN)
Total Columns		$P = TP + FN$	$N = FP + TN$

Overall, how often is the classifier correct? $accuracy = \frac{TP + TN}{P + N}$	Overall, how often is the classifier wrong? $misclassification\ rate = \frac{FP + FN}{P + N}$
When it's actually yes, how often does the classifier predict yes? $TPR = \frac{TP}{P}$ (true positive rate) (also called sensitivity)	When it's actually no, how often does the classifier predict yes? $FPR = \frac{FP}{N}$ (false positive rate)
When the classifier predicts yes, how often is it correct? $precision = \frac{TP}{TP + FP}$	When it's actually yes, how often does the classifier predict yes? $recall = \frac{TP}{P}$ (TPR)
When it's actually no, how often does it predict no? $specificity = \frac{TN}{N} = 1 - FPR$	How often does the yes condition actually occur in our sample? $prevalence = \frac{P}{P + N}$



DS

Communicating Results

We built a model! Now what?

- We've built our model, but there is still a gap between our iPython notebook with its plots and figures and a slideshow needed to present our results
- Classes so far have focused on two core concepts:
 - Developing consistent practices
 - Interpreting metrics to evaluate and improve model performance
- But what does that mean to your audience?

We built a model! Now what? (cont.)

- Imagine how a non-technical audience might respond to the following statements:
 - “The predictive model I built has an accuracy of 80%”
 - “Logistic regression was optimized with L2 regularization”
 - “Gender was more important than age in the predictive model because it has a “larger coefficient”
 - “Here’s the AUC chart that shows how well the model did”

We built a model! Now what? (cont.)

- Who is your audience? Are they technical? What are their concerns?
 - In a business setting, you may be the only person who can interpret what you've built
- Some people may be familiar with basic visualization, but you will likely have to do a lot of “hand holding”
- You need to be able to efficiently explain your results in a way that makes sense to all stakeholders (technical or not)

We built a model! Now what? (cont.)

- Today, we'll focus on communicating results for “simpler” problems, but this applies to any type of model you may work with



DS

Showing our Work

Showing our Work

- We've spent a lot of time exploring our data and building a reasonable model that performs well
- However, if we look at our visuals, they are most likely:

- Statistically heavy:
most people don't
understand histograms

- Overly complicated:
scatter matrices
produce too much
information

- Poorly labeled: code
doesn't require adding
labels, so you may not
have added them

To convey important information to your audience, make sure your charts are simplified, easily interpretable, and clearly labeled

Simplified

- At most, you'll want to include figures that either explain a variable on its own or explain that variable's relationship against a target
- If your model used a data transformation (like natural log), just visualize the original data
- Remove unnecessary complexity

Easily interpretable

- Any stakeholder looking at a figure should be seeing the exact same thing you're seeing
 - A good test for this is to share the visual with others less familiar with the data and see if they come to the same conclusion
 - How long did it take them?

Clearly labeled

- Take the time to clearly label your axis, title your plot, and double check your scales – especially if the figures should be comparable
- If you're showing two graphs side by side, they should follow the same Y axis

When building visuals for another audience, ask yourself who, what, and how

Who

- Who is my target audience for the visual?

What

- What do they already know about this project?
What do they need to know?

How

- How does my project affect this audience? How might they interpret (or misinterpret) the data?

Visualizing Models over Variables

- One effective way to explain your model over particular variables is to plot the predicted values against the most explanatory variables
- E.g., in logistic regression, plotting the probability of a class against a variable can help explain the range of effect of the model

Visualizing Performance Against Baseline

- Another approach of visualization is the effect of your model against a baseline, or – even better – against previous models
- Plots like this will also be useful when talking to your peers – other data scientists or analysts who are familiar with your project and interested in the progress you've made

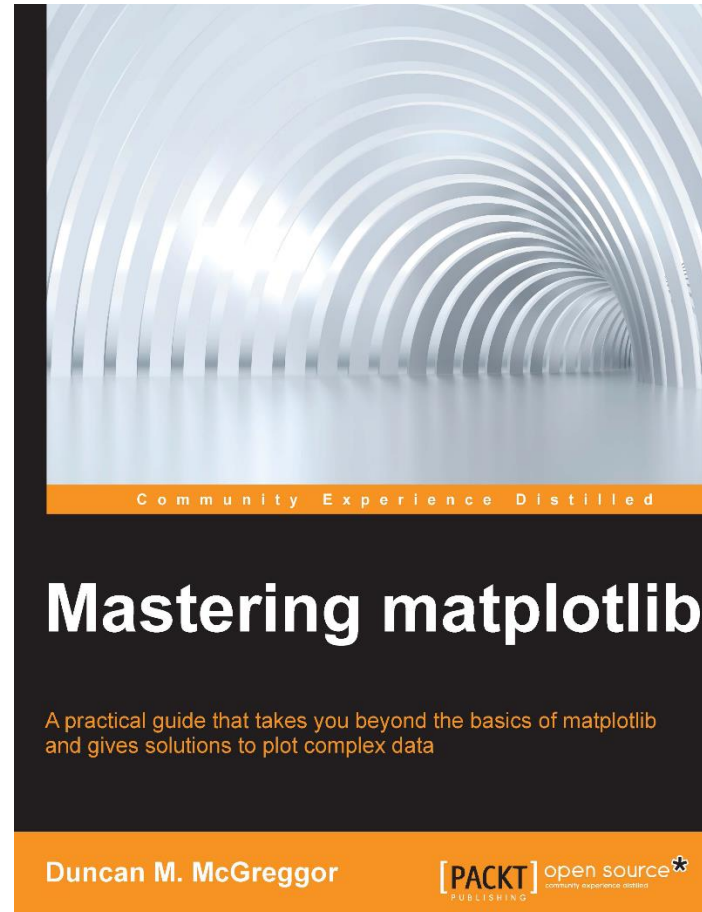


DS

Codealong – Part B

Prettying up Graphs

Going further



DS

Review

Review

- What do precision and recall mean? How are they similar and different to True Positive Rate and False Positive Rate?
- What are at least two very important details to consider when creating visuals for a project's stakeholders?
- Why would an AUC plot work well for a data science audience but not for a business audience? What would be a more effective visualization for that group?

Review (cont.)

You should now be able to:

- Evaluate a model using advanced metrics such as confusion matrix and ROC/AUC curves
- Explain the trade-offs between the precision and recall of a model while articulating the cost of false positives vs. false negatives
- Describe the difference between visualization for presentations vs. exploratory data analysis
- Identify the components of a concise, convincing report and how they relate to specific audiences/stakeholders

DS

Q & A



DS

Exit Ticket

Don't forget to fill out your exit ticket [here](#)