ROC

Agenda

Morning: **ROC**

- Regression vs. classification
- Logistic regression motivation
- Classification metrics and the confusion matrix
  - Precision, recall, accuracy
  - Specificity, sensitivity (recall)
  - True positive rate (recall), false positive rate
- Thresholding classification rules
  - ROC curve
- Pair Programming: Part 1 - ROC Curve

Logistic Regression - A Visual Motivation

Linear Regression Review - Visual

- With **linear regression**, we are modeling a **continuous response** and finding the linear function that gives the best fit

Figure 1:Linear Regression

Linear Regression for Classification - Visual

- What happens if we try linear regression for a **binary response** (such as a yes/no)?

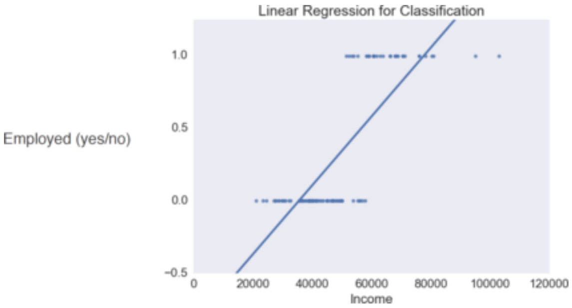

Figure 2:Linear regression for Classification

Linear Regression for Classification - Visual

- What happens if we try linear regression for a **binary response** (such as a yes/no)?

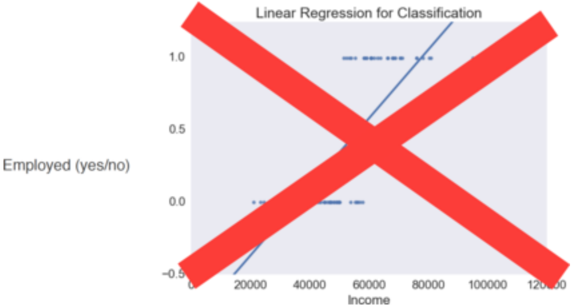

Figure 3:Linear regression for Classification

A Model for Classification

- We need a model that:
  - Takes continuous input (i.e., from $-\infty$ to ∞)
  - Produces output between 0 and 1
  - Transitions between 0 and 1 "without wasting much time"
  - Has interpretable coefficients (like our standard linear regression model)

Logistic Regression for Classification

- Enter **logistic regression**...

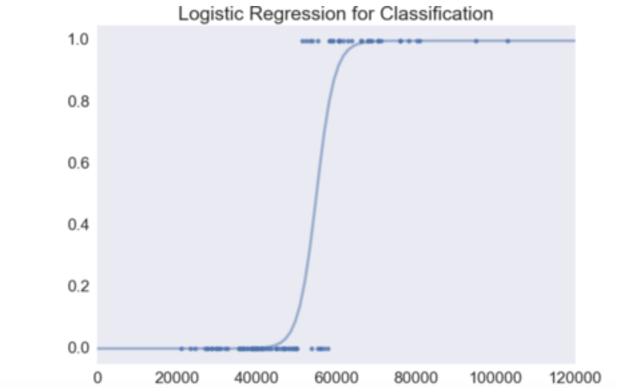

Figure 4:Logistic Regression

The Logistic Function

That general S-shaped curve is from the sigmoid family, and the logistic function that we use in logistic regression is from the sigmoid family

Its functional form is as follows:

$$S(t) = \frac{1}{1 + e^{-t}}$$

Classification Metrics

Logistic Regression Revisited

Think of sliding the purple/red lines along the sigmoid function

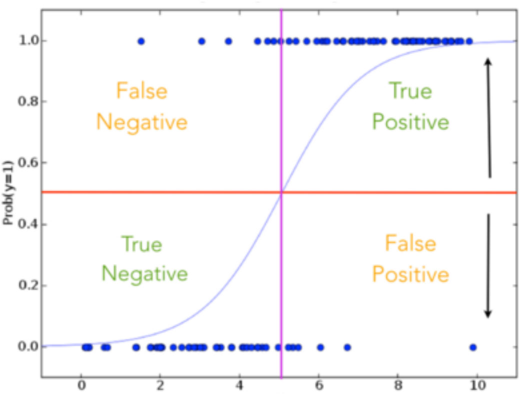

Figure 5:Logistic Regression Revisited

Classification Metrics

- We use the following metrics as a base by which to judge our model:

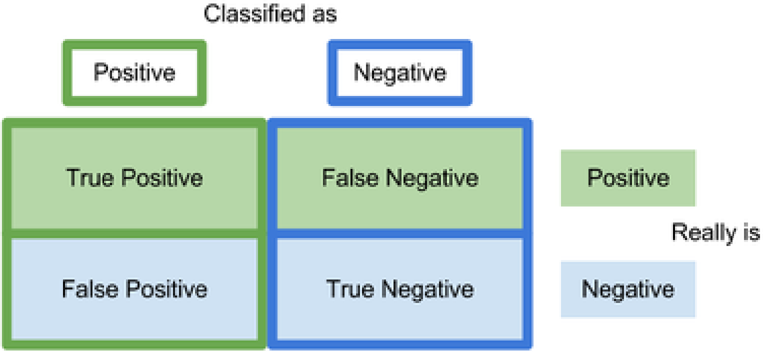

Figure 6:Confusion Matrix

Classification Metrics

- **Accuracy** - How many observations did I label correctly?

$$\frac{TP + TN}{P + N}$$

- **True Positive Rate (TPR), Recall, Sensitivity** - Of those observations that are actually positives, which ones did I label as positive?

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

- **False Positive Rate (FPR)** - Of those observations that are actually negatives, which ones did I label as positive?

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

Classification Metrics

- **Precision, Positive Predictive Value** - Of those observations that I labeled as positive, which ones are actually positive?

$$\frac{TP}{TP + FP}$$

- **True Negative Rate, Specificity** - Of those observations that are actually negative, which ones did I label as negative?

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

Classification Metrics

		predicted condition			
	total population	prediction positive	prediction negative	Prevalence $= \frac{\Sigma \text{ condition positive}}{\Sigma \text{ total population}}$	
true condition	condition positive	**True Positive (TP)**	**False Negative (FN)** (type II error)	True Positive Rate (TPR), Sensitivity, Recall, Probability of Detection $= \frac{\Sigma \text{ TP}}{\Sigma \text{ condition positive}}$	False Negative Rate (FNR), Miss Rate $= \frac{\Sigma \text{ FN}}{\Sigma \text{ condition positive}}$
	condition negative	**False Positive (FP)** (Type I error)	**True Negative (TN)**	False Positive Rate (FPR), Fall-out, Probability of False Alarm $= \frac{\Sigma \text{ FP}}{\Sigma \text{ condition negative}}$	True Negative Rate (TNR), Specificity (SPC) $= \frac{\Sigma \text{ TN}}{\Sigma \text{ condition negative}}$
	Accuracy $= \frac{\Sigma \text{ TP} + \Sigma \text{ TN}}{\Sigma \text{ total population}}$	Positive Predictive Value (PPV), Precision $= \frac{\Sigma \text{ TP}}{\Sigma \text{ prediction positive}}$	False Omission Rate (FOR) $= \frac{\Sigma \text{ FN}}{\Sigma \text{ prediction negative}}$	Positive Likelihood Ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic Odds Ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$
		False Discovery Rate (FDR) $= \frac{\Sigma \text{ FP}}{\Sigma \text{ prediction positive}}$	Negative Predictive Value (NPV) $= \frac{\Sigma \text{ TN}}{\Sigma \text{ prediction negative}}$	Negative Likelihood Ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$	

Figure 7:Confusion Matrix (Wikipedia)

Breakout: Pair Exercise, 10 mins

Why logistic and not just plain old linear?

- Discuss the problems with using standard linear regression for modeling binary response
- What shape does the logistic function take?
- Why is the logistic function a good, logical fit for binary classification? Compared to linear? What are the advantages?

Breakout: Pair Exercise, 10 mins

You built a fraud prediction model

- Label each square with one of TP, FP, FN, and TN
- How many total data points do you have? How many are fraudulent? How many aren't fraudulent?
- Calculate accuracy, precision and recall

	Predicted: Yes	Predicted: No
Actual: Yes	4	10
Actual: No	2	204

- Is the confusion matrix shown here representative of a good model?
- Which of the metrics you calculated above are most useful in determining how good the model is?
- What are cases where accuracy is useful? When do you need to be wary of using accuracy?

ROC Curve

ROC Curve

- Since logistic regression outputs **probabilities**, we can change our TPR and FPR by changing the **threshold** for positive classification

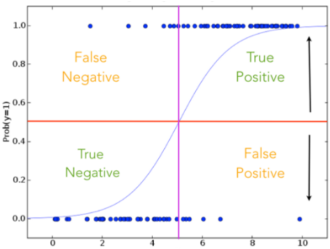

Figure 8:Probabilities and Threshold

- E.g., only say "Employed = yes" if the model gave a probability of being employed to be at least 0.75

ROC Curve

- A plot of the TPR vs. FPR at difference thresholds is called a ROC curve. It is used to visualize the performance of a given binary classifier:

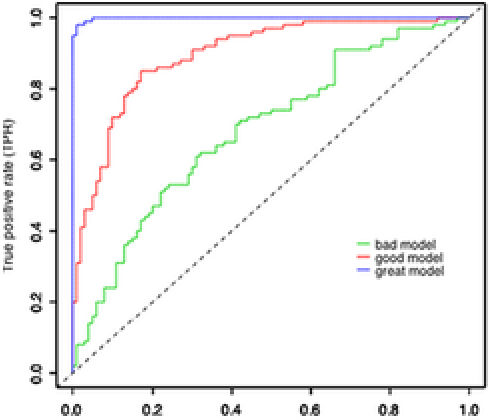

Figure 9:ROC Curve

Building the ROC Curve

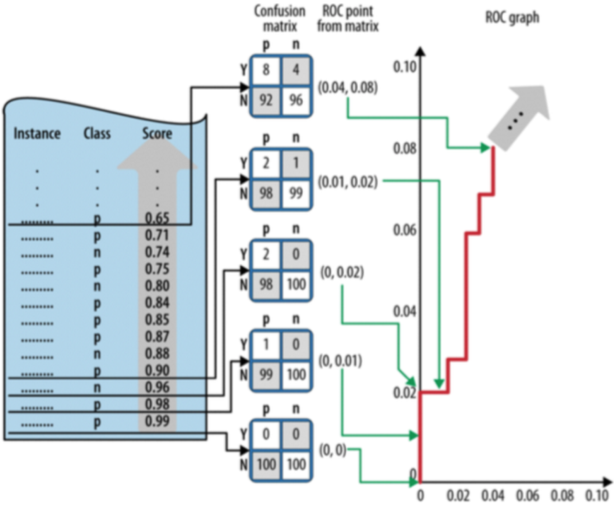

Figure 10:Building the ROC Curve

ROC Curve

- With the ROC curve, we can examine how the TPR changes as the FPR changes (or vice versa)
  - We can compare across curves to determine which model gives us a better TPR for a given FPR
  - We can also use the Area Under the Curve (AUC) to try to differentiate one model from another (greater area is typically better, but this also depends on what TPR/FPR you are willing to accept)
  - We can typically achieve the 45° line through random guessing (so we should always do better than this)

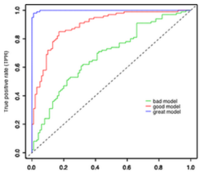

Figure 11:ROC Curve

Breakout: Pair, 5 mins

Assume we're dealing with predicting credit fraud. . .

- In this scenario, do you think you'd care more about optimizing TPR or FPR?
- What is a scenario where you'd care more about the other (TPR or FPR)?

Breakout: Pair, 5 mins

- Prompt: You have built 3 models to predict whether or not someone will default on a loan. You have 3,000 data points and these features: age, gender, city, FICO score, highest education completed

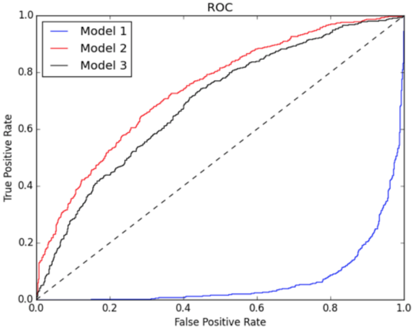

Figure 12:ROC Curve

Breakout: Pair, 5 mins

- Question: Which of the 3 ROC curves represents the model you should use?
- Question: How would you pick between 50 models? 100 models? 1,000 models?

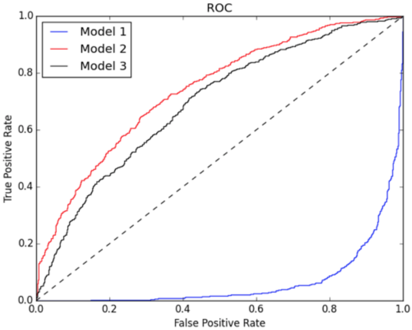

Figure 13:ROC Curve

Breakout: Pair, 15 mins

Construct a ROC curve only given the following predicted probabilities from a logistic regression and true labels

Predicted Probability	Actual fraud?
0.99	Fraud
0.84	Fraud
0.70	Fraud
0.70	Not Fraud
0.51	Fraud
0.22	Fraud
0.14	Not Fraud
0.05	Not Fraud