Classification

Moses Marsh

galvanize

- What is classification?
- What is the difference between Linear Regression and Logistic Regression?
- What are the metrics for evaluating a classifier?
- How is a ROC Curve constructed?

Classification: Examples

galvanize

- Identifying spam emails
- Predicting if borrowers will default on their loans
- Determining whether someone has a disease
- Determining if an animal is a dog or a horse (credit: Michael Jancsy)

- These are all examples of ***binary classification***

What is classification?

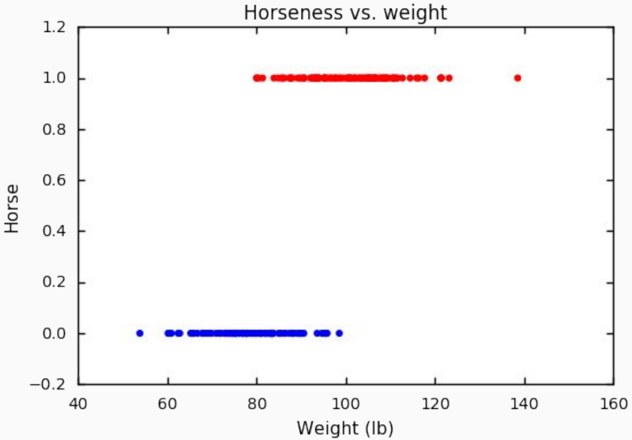
- Review: in ***regression***, the target y is ***numerical*** and ***unbounded***
- In ***classification***, the target y is ***categorical***: it is a ***finite set*** of ***class labels***

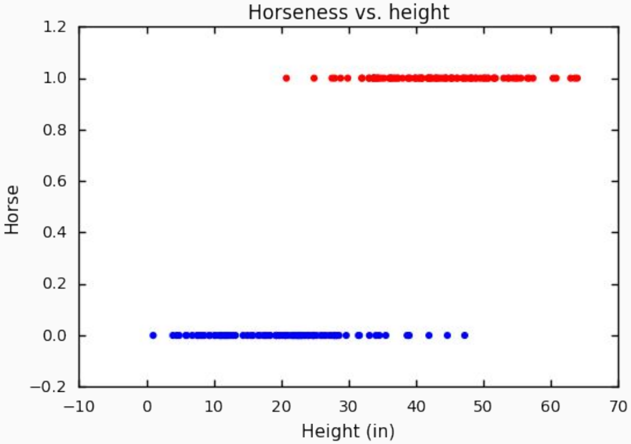
Classification: Mathematical Description

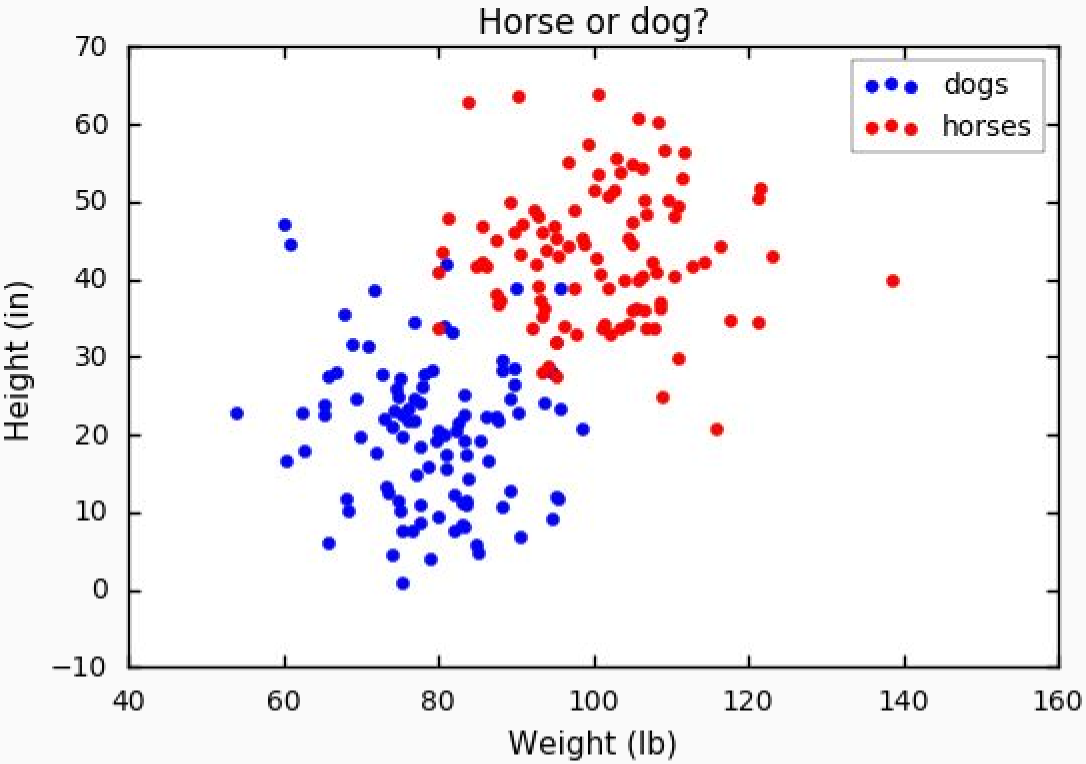
- A ***classifier model*** maps between feature space and a finite set of class labels
- A ***binary classifier*** maps onto `{0, 1}`
- Example: predicting college admission based on academic performance
  - Features
    - GPA: real number in the range `[0, 4]`
    - SAT score: integer in the range `[600, 2400]`
  - Target
    - Not admitted: `{0}`
    - Admitted: `{1}`
  - Our model is some function that takes GPA and SAT scores as input, then outputs either a zero or a one

$$f(GPA, SAT) \rightarrow \{0, 1\}$$

Binary Classification Example: Horse vs Dog

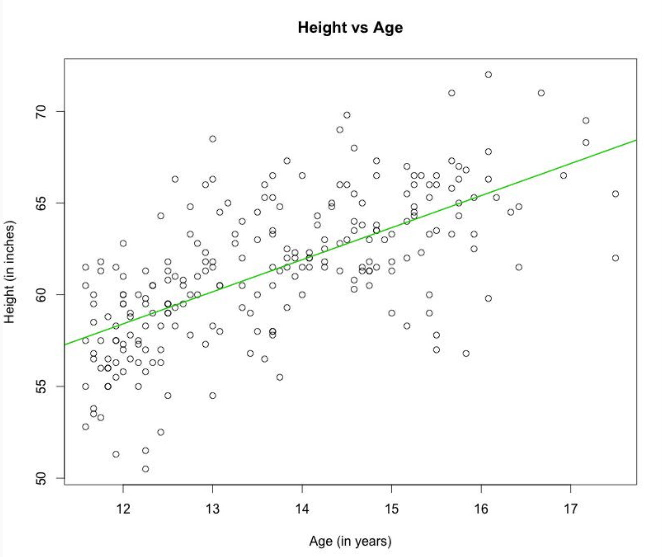




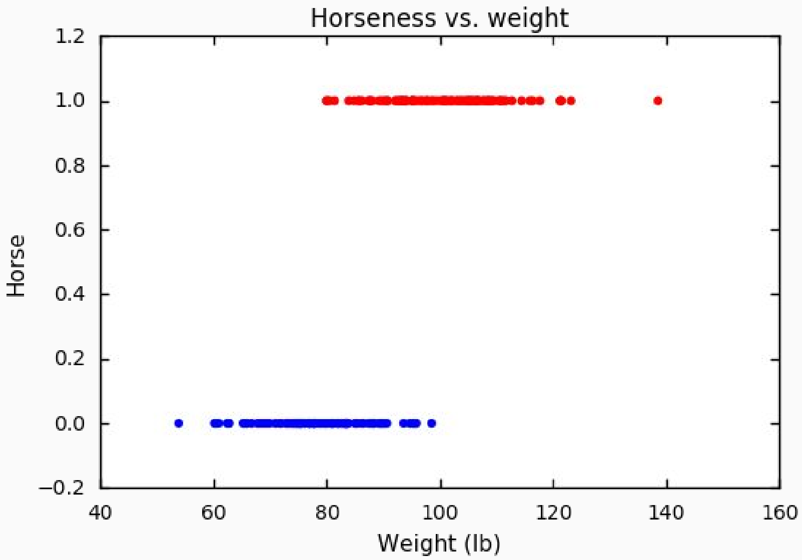


Regression vs. Classification

Regression



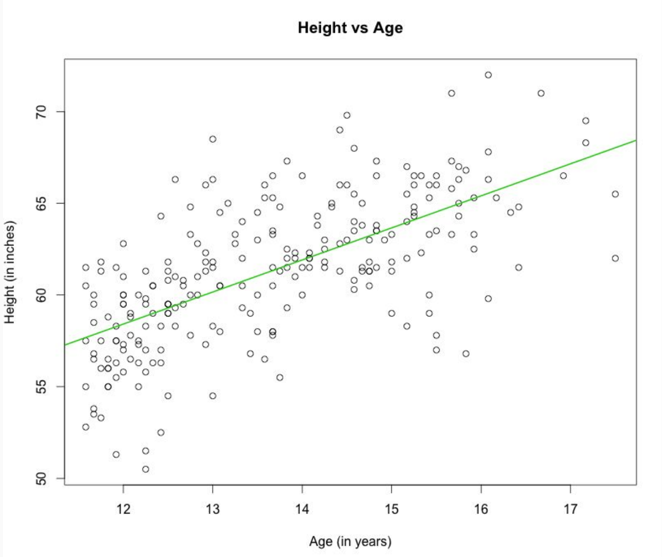
Classification



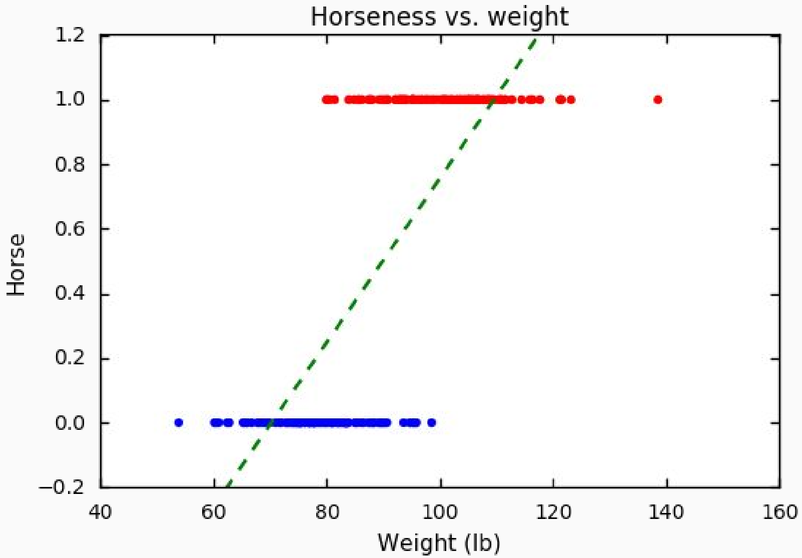
How should we model this?

Regression vs. Classification

Regression



Classification

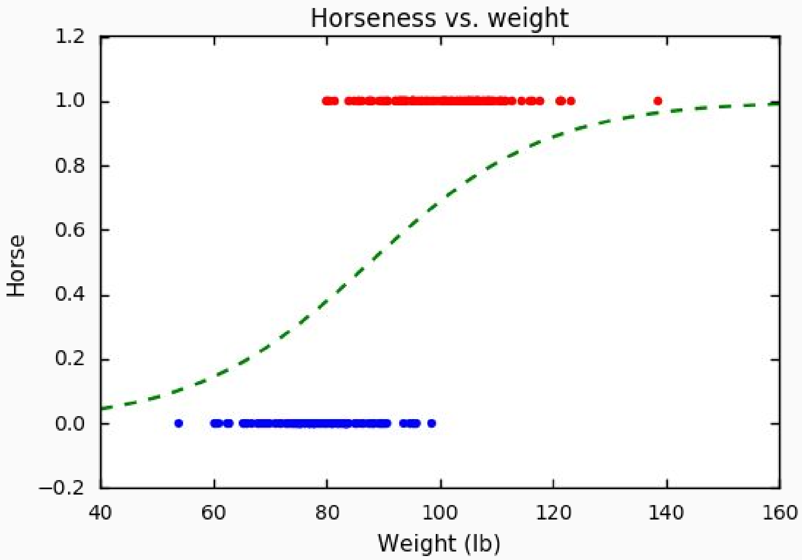


Throwing linear regression at it: not super effective

Predict the probability!

Logistic Regression

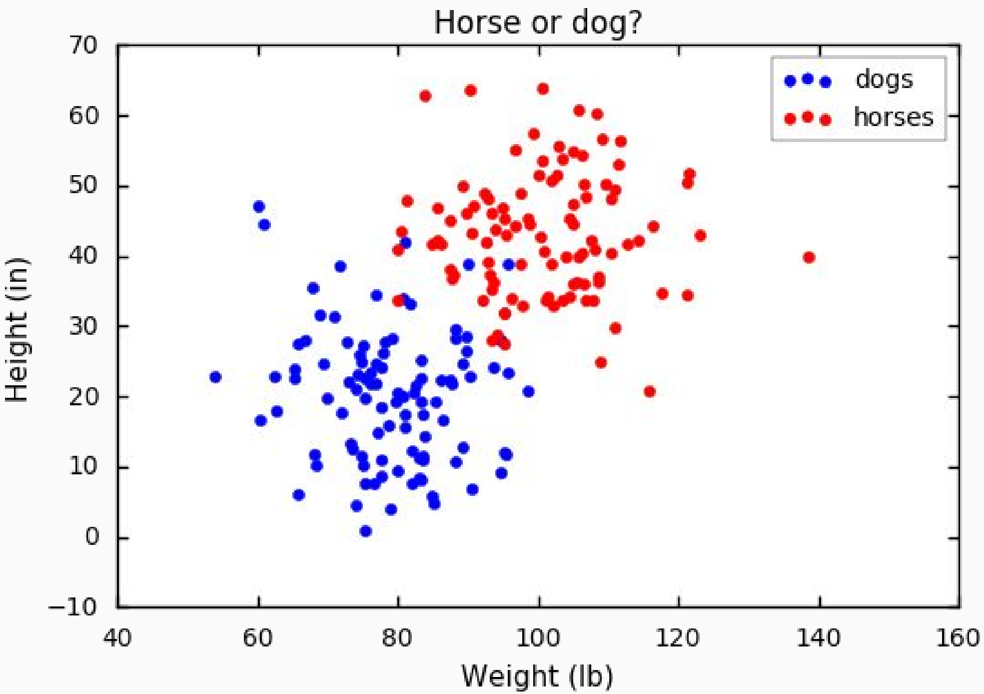
- Instead, we predict a ***probability*** of being a horse
- Then we simply round the probability to `0` or `1` to classify
- The ***sigmoid shape*** of the curve ensured predictions in the range `(0,1)`
  - (more math this afternoon!)

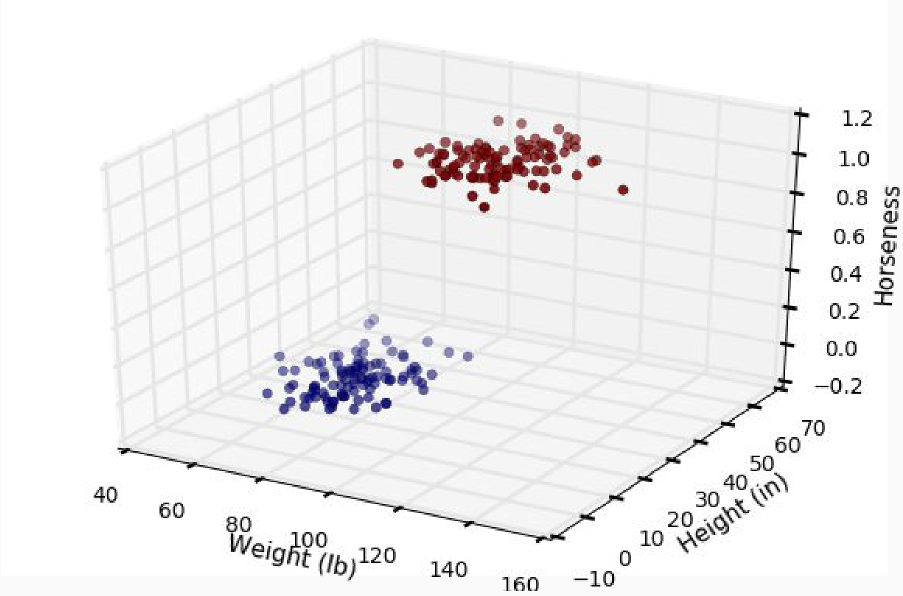


Why logistic and not just plain old linear?

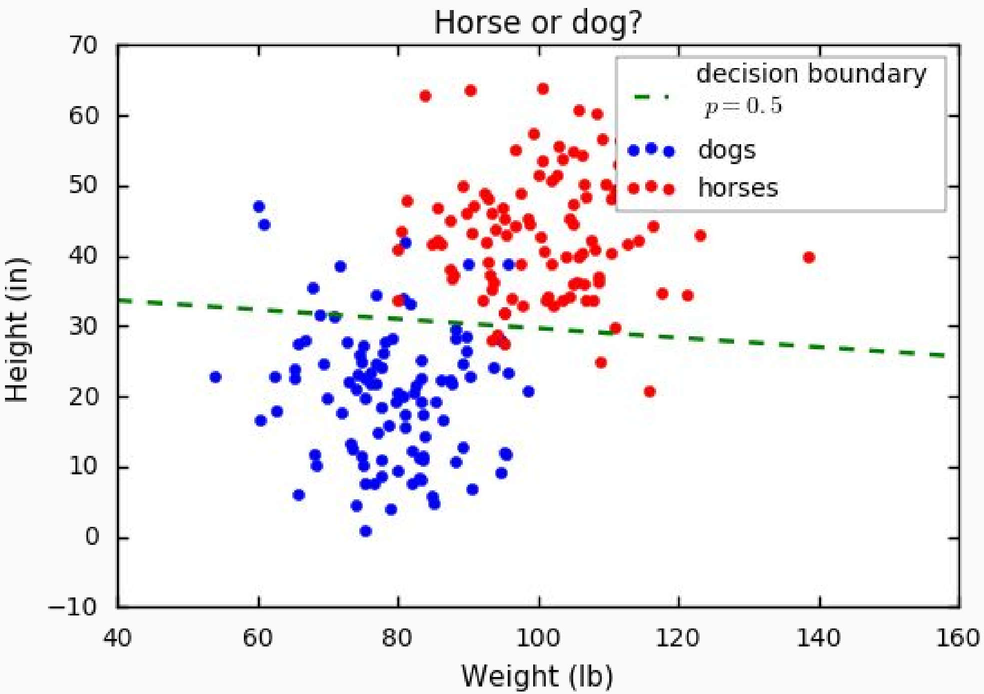
1. Discuss the problems with using standard linear regression for modeling binary response.
2. What shape does the logistic function take?
3. Why is the logistic function a good, logical fit for binary classification? Compared to linear? What advantages?

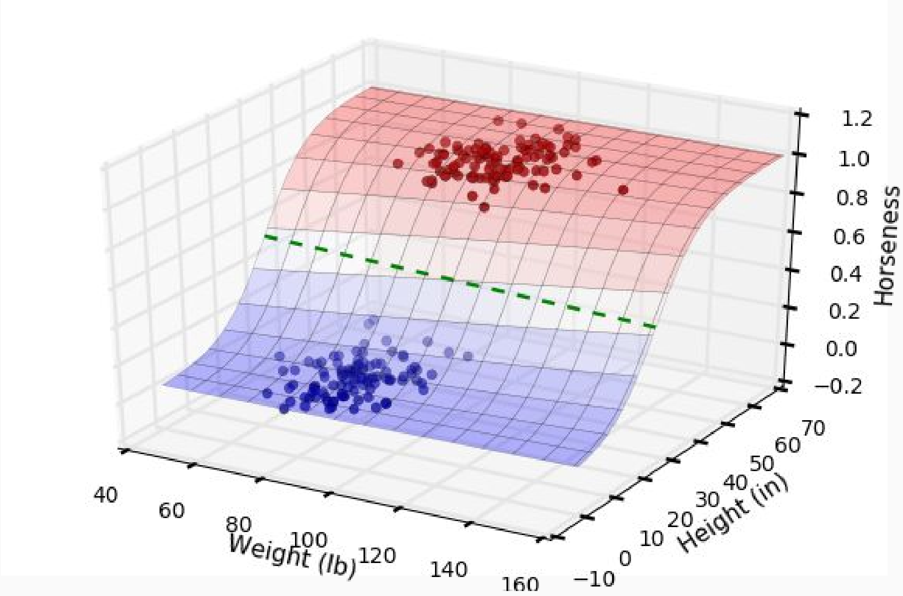
Binary Classification: Horse vs Dog





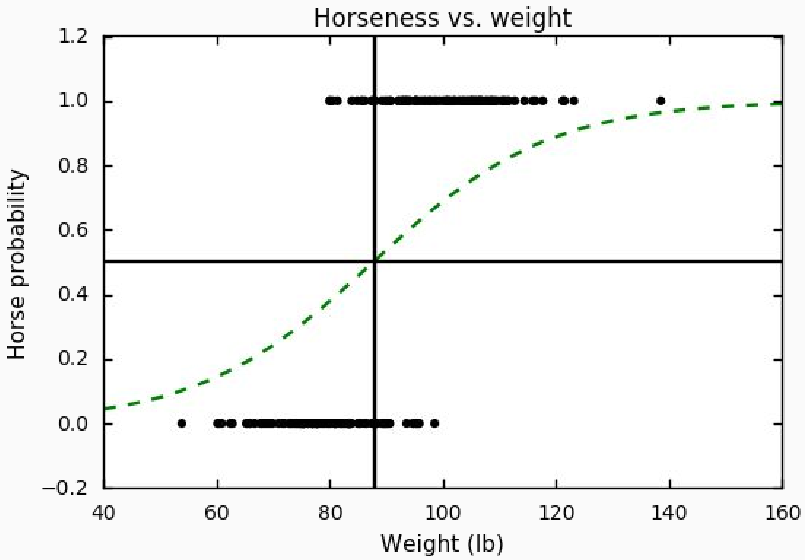
Binary Classification: Horse vs Dog





So you have a class probability for each data point. Now what?

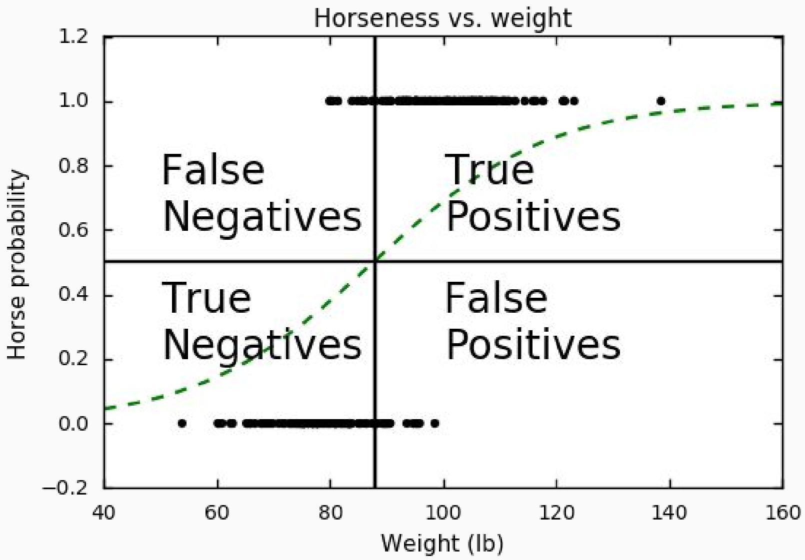
How do you evaluate your predictions against the true class labels?



Evaluating a classifier: metrics

Confusion matrix

	Predicted Negative	Predicted Positive
Actually Positive	False Negatives	True Positives
Actually Negative	True Negatives	False Positives



Evaluating a classifier: metrics

Confusion matrix

	Predicted Negative	Predicted Positive
Actually Positive	False Negatives	True Positives
Actually Negative	True Negatives	False Positives

n = # of data points

Accuracy: fraction of data correctly classified
`(TP + TN) / n`

True Positive Rate (aka Sensitivity, Recall):
fraction of actual positives that were labeled positive
`(TP) / (TP + FN)`

True Negative Rate (Specificity):
fraction of actual negatives that were labeled negative
`(TN) / (TN + FP)`

Precision: fraction of labeled positive points that were actually positive
`(TP) / (TP + FP)`

False Positive Rate
`(FP) / (TN + FP)`

False Negative Rate
`(FN) / (TP + FN)`

Evaluating a classifier: F-Score

The **F-Score** is a weighted harmonic mean of ***precision*** and ***recall***

$$F = \frac{1}{\alpha \frac{1}{precision} + (1-\alpha)\frac{1}{recall}}$$

The F1-Score (aka “balanced F-Score) has $\alpha = 0.5$

$$F_1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}}$$

You have built a credit card fraud prediction model (Pair exercise, 10 min)

- Label each square with one of TP, FP, FN, 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

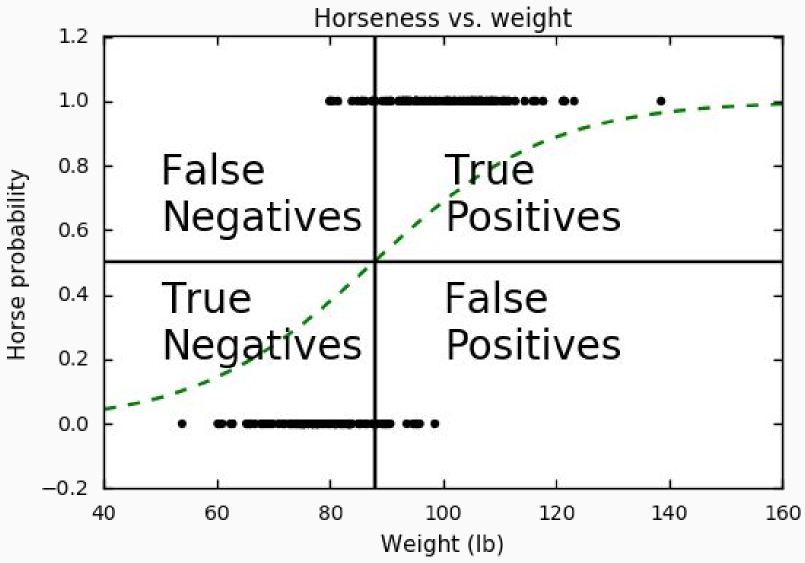
- Is the confusion matrix shown here representative of a good model?
- Which of the metrics you calculated above are most useful in determining how good the model is?
- What are cases where accuracy is useful? When do you need to be wary of using accuracy?

Evaluating a classifier: ROC Curves

Since Logistic Regression outputs ***probabilities***, we can change our TP & FP rates by changing the ***threshold*** for positive classification

e.g., only say “horse!” if the model gave a probability of at least 0.7

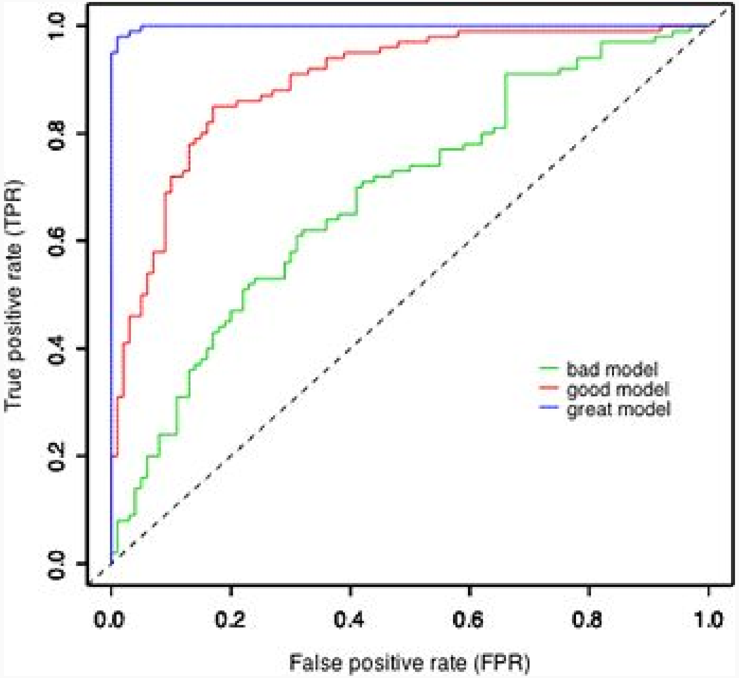
A plot of the TPR vs FPR at difference thresholds is called a ***ROC plot***



[Fun gif]

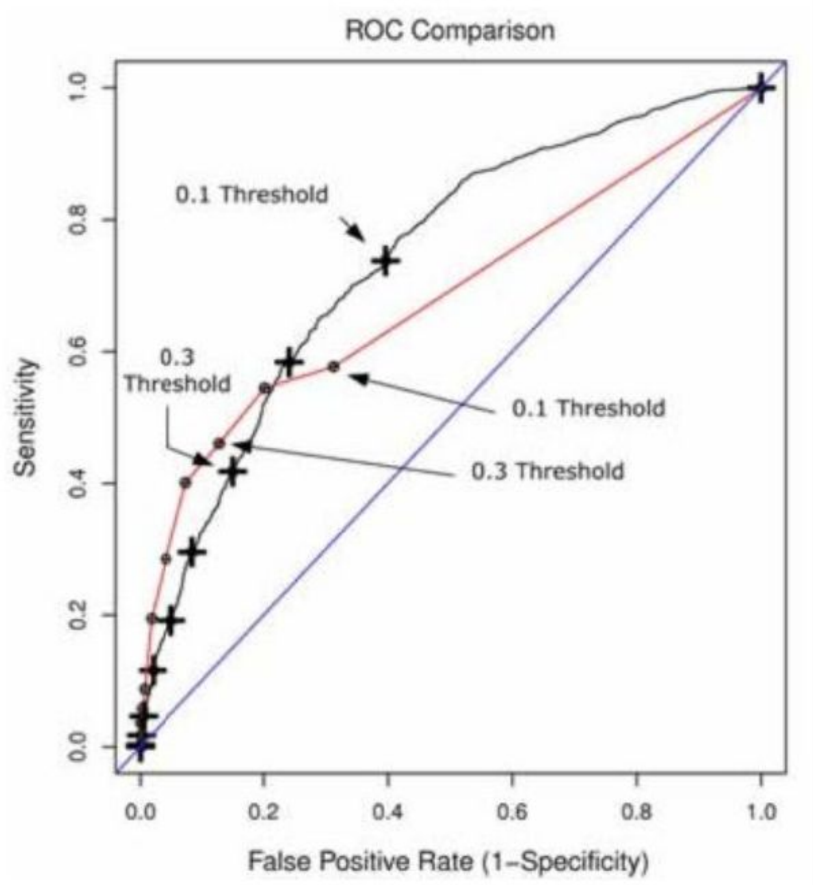
Interpreting ROC curves

- If classifier A's ROC curve is strictly greater than classifier B, then A is preferred
- The ***AUC*** (area under curve) is useful for comparing ROC curves.
    - It equals the probability that the model will rank a randomly chosen positive observation higher than a randomly chosen negative observation



Interpreting ROC curves

- If two classifier's ROC curves intersect, then the choice depends on the relative importance of sensitivity and specificity



Assume we’re dealing with predicting credit fraud… (Pair discussion, 5 minutes)

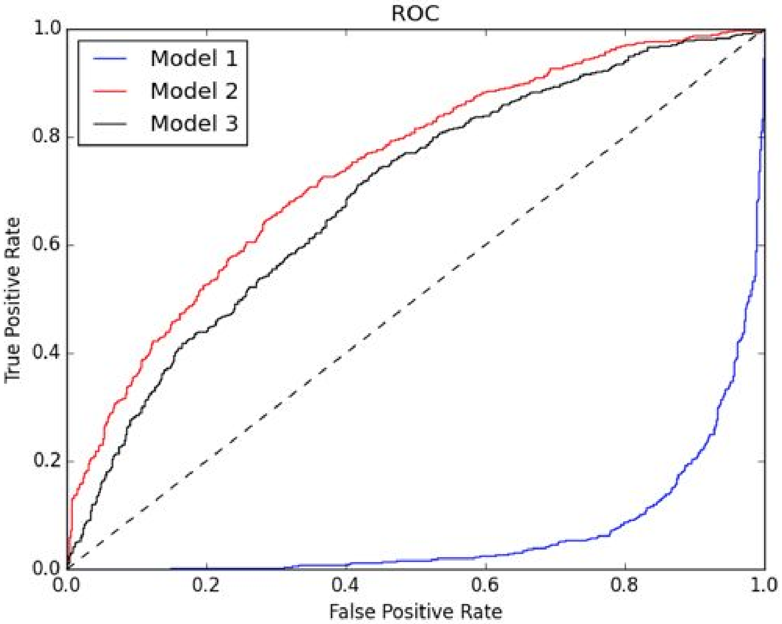
1. In this scenario, do you think you’d care more about optimizing TPR or FPR?
2. What is a scenario where you’d care more about the other (TPR or FPR)?

(Pair discussion, 5 minutes)

Prompt: You have built 3 models to predict whether or not someone will default on a loan. You have 3000 data points and these features: age, gender, city, FICO score, highest education completed

Question: Which of the 3 ROC curves represents the model you should use?

Question: How would you pick between 50 models? 100 models? 1000 models?



(Class exercise)
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

Logistic Regression

Moses Marsh

galvanize

- How does Logistic Regression produce probabilities?
- How is Linear Regression involved?
- How are its parameters found?
- What do its parameters mean?

The Logit Function

- Recall our problem:
  - we know linear regression $y = \beta_0 + \beta \cdot x = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$
  - Its predictions are unbounded $y \in (-\infty, \infty)$
  - We want to predict a probability $p \in (0, 1)$
  - So we need a function to do this transformation
- Punchline: the logit function

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta \cdot x)}}$$

The Logit Function: let's break it down

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta \cdot x)}}$$

- Let $\theta = \beta_0 + \beta \cdot x$ (the linear regression piece)
- Then $p = \frac{1}{1 + e^{-\theta}}$
- And hey presto: $\theta = \ln\left(\frac{p}{1-p}\right)$ this is the ***log of the odds***

The Log-Odds: let's break it down

$$\theta = \ln\left(\frac{p}{1-p}\right)$$

- What is probability? $p \sim \frac{\#successes}{\#trials}$ $\quad p \in [0, 1]$
- **Odds** are defined as $d = \frac{p}{1-p}$ $\quad d \in [0, \infty)$

  $$d \sim \frac{\#successes}{\#failures}$$

- Hence log-odds: $\theta = \ln(d) = \ln(\frac{p}{1-p})$ $\quad \theta \in (-\infty, \infty)$

galvanize

- To recap:
  - We use linear regression to model the log of the odds

$$\theta = \beta_0 + \beta \cdot x \qquad \theta = \ln\left(\frac{p}{1-p}\right)$$

  - Then we convert that to a probability

$$p = \frac{1}{1+e^{-\theta}} \qquad p = \frac{1}{1+e^{-(\beta_0+\beta\cdot x)}}$$

- The function used to estimate p is called the ***hypothesis function***

$$h_\beta(x_i) = \frac{1}{1+e^{-(\beta_0+\beta\cdot x)}}$$

- Now we can map a feature vector x to a probability! CLASSIFICATION!

Finding the coefficients

- Unlike linear regression, there is no analytic solution for the set of parameters that produces the best fit.
- So we use Maximum Likelihood Estimation (MLE)!

$$\hat{\beta} = \arg\max_{\beta} P(X|\beta)$$

Finding the coefficients

$$\hat{\beta} = \arg\max_{\beta} P(X|\beta)$$

- Likelihood of an observation given the model:

$$p(y_i|x_i;\beta) = h_\beta(x_i)^{y_i}(1 - h_\beta(x_i))^{1-y_i}$$

- Assuming each observation is independent:

$$P(Y|X;\beta) = \prod_{i=1}^{n} h_\beta(x_i)^{y_i}(1 - h_\beta(x_i))^{1-y_i}$$

- Choose the coefficients that maximize this expression.
- In practice, we maximize the log likelihood:

$$\ln P(Y|X;\beta) = \sum_{i=1}^{n}(y_i \ln h_\beta(x_i) + (1 - y_i)\ln(1 - h_\beta(x_i)))$$

- Note that a **mismatch** between the hypothesis and the true value for a singly point **negatively** impacts the log likelihood

Interpreting the coefficients

- What do all these betas mean?
- Consider the case of only one feature

$$\theta(x) = \beta_0 + \beta_1 x$$
$$\theta_0 = \beta_0$$
$$\theta_1 = \beta_0 + \beta_1$$

- A one-unit increase in x corresponds to a one-unit increase in the ***log odds***
- What about the ***odds***?

$$d_0 = e^{\theta_0} = e^{\beta_0}$$
$$d_1 = e^{\theta_1} = e^{\beta_0 + \beta_1}$$
$$\frac{d_1}{d_0} = e^{\beta_1}$$ this is the ***odds ratio***

Interpreting the coefficients

$$\frac{d_1}{d_0} = e^{\beta_1}$$

- A one-unit increase in x ***multiplied*** the odds by e^{β_1}
- What does this do to the ***probability*** if β_1 is:
    - Positive?
    - Negative?
    - Zero?

Interpreting the coefficients

$$\frac{d_1}{d_0} = e^{\beta_1}$$

- A one-unit increase in x ***multiplied*** the odds by e^{β_1}
- What does this do to the ***probability*** if β_1 is:
    - Positive?
    - Negative?
    - Zero?
- You can show that the new probability isn't the prettiest function of the old probability and β_1

$$p_1 = \frac{1}{1 + (\frac{1}{p_0} - 1)e^{-\beta_1}}$$

- But that's OK! Odds are very interpretable! "Ratio of successes to failures"

Understanding your chances (Pair, 3 mins)

1. State what each of the following terms are:

   Probability, Odds, Log-Odds, Odds Ratio

2. Give an example to demonstrate what each of the 4 terms are

Interpret the results from this logistic regression model

1. What are my current chances of getting into grad school?
2. How would my chances change if I increased my GPA by 100 pts?
3. What score would I need on the GRE's to increase my chances to 95%?

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                  admit   No. Observations:                  400
Model:                          Logit   Df Residuals:                      397
Method:                           MLE   Df Model:                            2
Date:                Fri, 02 Dec 2016   Pseudo R-squ.:                 0.03927
Time:                        16:43:29   Log-Likelihood:                -240.17
converged:                       True   LL-Null:                       -249.99
                                        LLR p-value:                 5.456e-05
==============================================================================
                 coef    std err          z      P>|z|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const         -4.9494      1.075     -4.604      0.000        -7.057    -2.842
gre            0.0027      0.001      2.544      0.011         0.001     0.005
gpa            0.7547      0.320      2.361      0.018         0.128     1.381
==============================================================================
```

Model 1 and 2 are from the same dataset. Explain what you see. (individual, 2 mins, then pair, 5)

Dep. Variable:	Survived	**No. Observations:**	712
Model:	Logit	**Df Residuals:**	709
Method:	MLE	**Df Model:**	2
Date:	Tue, 22 Nov 2016	**Pseudo R-squ.:**	0.2528
Time:	15:27:35	**Log-Likelihood:**	-359.02
converged:	True	**LL-Null:**	-480.45
		LLR p-value:	1.825e-53

	coef	std err	z	P>|z|	[95.0% Conf. Int.]
Intercept	0.6590	0.167	3.935	0.000	0.331 0.987
Sex[T.male]	-2.3711	0.189	-12.524	0.000	-2.742 -2.000
Fare	0.0121	0.003	4.595	0.000	0.007 0.017

Model 1

Dep. Variable:	Survived	**No. Observations:**	712
Model:	Logit	**Df Residuals:**	708
Method:	MLE	**Df Model:**	3
Date:	Tue, 06 Dec 2016	**Pseudo R-squ.:**	0.3013
Time:	08:33:07	**Log-Likelihood:**	-335.70
converged:	True	**LL-Null:**	-480.45
		LLR p-value:	1.852e-62

	coef	std err	z	P>|z|	[95.0% Conf. Int.]
Intercept	3.1335	0.399	7.863	0.000	2.352 3.915
Sex[T.male]	-2.5536	0.204	-12.528	0.000	-2.953 -2.154
Fare	0.0019	0.002	0.850	0.395	-0.002 0.006
Pclass	-0.9283	0.137	-6.788	0.000	-1.196 -0.660

Model 2