Lecture #11: Classification & Logistic Regression

CS 109A, STAT 121A, AC 209A: Data Science

Weiwei Pan, Pavlos Protopapas, **Kevin Rader**

Fall 2016

Harvard University

Announcements

- **Midterm**: will be graded this week. Results on Multiple Choice are promising :)
- Don't worry about some since Canvas isn't smart, unfortunately. Score could [only] increase.
- **Project**: Milestone #2 due on Wednesday, but will be open until Saturday. **Optional** to customize your project based on your requests.
- Expect an email from your project TF today!
- **Survey**: Results, we are incorporating suggestions, thank you for your feedback!
- **OHs** Update: Pavlos: Wed 4-5pm, Weiwei 5-6pm (updating on Canvas).
- **HW Grading**: A 4 is good ("A-")!

Quiz Time

Time for Quiz...

Quiz Time

Time for Quiz...**NOT!!!**

Lecture Outline

Classification [revisited]

Why Linear Regression Fails

Binary Response & Logistic Regression

Estimating the Simple Logistic Model

Classification using the Logistic Model

Extending the Logistic Model

Classification

Up to this point, the methods we have seen have centered around modeling and the prediction of a quantitative response variable (ex, # Uber pickups, location of a meteorite, etc...). Regression (and Ridge, LASSO, etc...) perform well under these situations

When the response variable is categorical, then the problem is no longer called a regression problem (from the machine learning perspective) but is instead labeled as a *classification* problem.

The goal is to attempt to classify each observation into a category (aka, class or cluster) defined by Y, based on a set of predictor variables (aka, features), X.

Lecture Outline

Classification [revisited]

Why Linear Regression Fails

Binary Response & Logistic Regression

Estimating the Simple Logistic Model

Classification using the Logistic Model

Extending the Logistic Model

Simple Classification Example

Given a dataset $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_N, y_N)\}$, where the y are categorical (sometimes referred to as *qualitative*), we would like to be able to predict which category y takes on given x. Linear regression does not work well, or is not appropriate at all, in this setting. A categorical variable y could be encoded to be quantitative. For example, if Y represents concentration of Harvard undergrads, then y could take on the values:

$$y = \begin{cases} 1 & \textit{if } \text{Computer Science (CS)} \\ 2 & \textit{if } \text{Statistics} \\ 3 & \text{otherwise} \end{cases}.$$

Simple Classification Example (cont.)

A linear regression could be used to predict y from $\mathbf{x}$. What would be wrong with such a model?

Simple Classification Example (cont.)

A linear regression could be used to predict y from $\mathbf{x}$. What would be wrong with such a model?

The model would imply a specific ordering of the outcome, and would treat a one-unit change in y equivalent. The jump from $y = 1$ to $y = 2$ (CS to Statistics) should not be interpreted as the same as a jump from $y = 2$ to $y = 3$ (Statistics to everyone else).

Similarly, the response variable could be reordered such that $y = 1$ represents Statistics and $y = 2$ represents CS, and then the model estimates and predictions would be fundamentally different.

If the categorical response variable was *ordinal* (had a natural ordering...like class year, Freshman, Sophomore, etc...), then a linear regression model would make some sense but is still not ideal.

Even Simpler Classification Problem: Binary Response

The simplest form of classification is when the response variable Y has only two categories, and then an ordering of the categories is natural. For example, an upperclassmen Harvard student could be categorized as (note, the $y = 0$ category is a "catch-all" so it would involve both River House students and those who live in other situations: off campus, etc...):

$$y = \begin{cases} 1 & \textit{if} \text{ lives in the Quad} \\ 0 & \text{otherwise} \end{cases} .$$

Linear regression could be used to predict y directly from a set of covariates (like sex, whether an athlete or not, concentration, GPA, etc...), and if $\hat{y} \geq 0.5$, we could predict the student lives in the Quad and predict other houses if $\hat{y} < 0.5$.

Even Simpler Classification Example (cont.)

What could go wrong with this linear regression model?

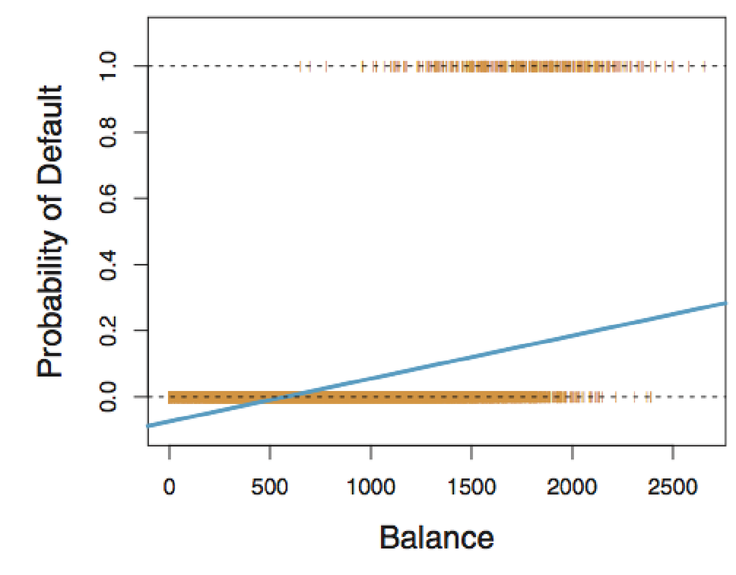
Even Simpler Classification Example (cont.)

What could go wrong with this linear regression model?

The main issue is you could get non-sensical values for $\hat{y}$. Since this is modeling $P(y = 1)$, values for $\hat{y}$ below 0 and above 1 would be at odds with the natural measure for y, and linear regression can lead to this issue.

A picture is worth a thousand words...

Why linear regression fails



Lecture Outline

Classification [revisited]

Why Linear Regression Fails

Binary Response & Logistic Regression

Estimating the Simple Logistic Model

Classification using the Logistic Model

Extending the Logistic Model

Logistic Regression

Logistic Regression addresses the problem of estimating a probability, $P(y = 1)$, to be outside the range of $[0, 1]$. The logistic regression model uses a function, called the *logistic* function, to model $P(y = 1)$:

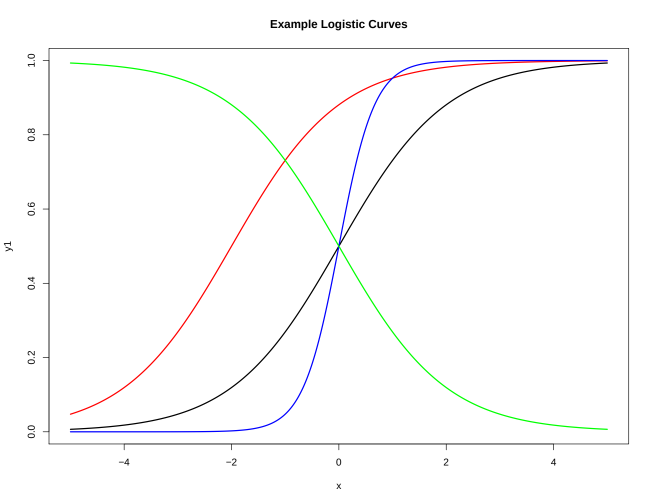
$$P(Y = 1) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}.$$

As a result the model will predict $P(Y = 1)$ with an S-shaped curve, as seen in a future slide, which is the general shape of the logistic function. β_0 shifts the curve right or left and β_1 controls how steep the S-shaped curve is.

Note: if β_1 is positive, then the predicted $P(Y = 1)$ goes from zero for small values of X to one for large values of X and if β_1 is negative, then $P(Y = 1)$ has the opposite association.

Logistic Regression(cont.)

Below are four different logistic models with different values for β_0 and β_1: $\beta_0 = 0, \beta_1 = 1$ is in black, $\beta_0 = 2, \beta_1 = 1$ is in red, $\beta_0 = 0, \beta_1 = 3$ is in blue, and $\beta_0 = 0, \beta_1 = -1$ is in green.



Logistic Regression(cont.)

With a little bit of algebraic work, the logistic model can be rewritten as:

$$\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right)=\beta_0+\beta_1 X.$$

The value inside the natural log function, $\frac{P(Y=1)}{1-P(Y=1)}$, is called the *odds*, thus logistic regression is said to model the *log-odds* with a linear function of the predictors or features, X. This gives us the natural interpretation of the estimates similar to linear regression: a one unit change in X is associated with a β_1 change in the log-odds of $Y=1$; or better yet, a one unit change in X is associated with an e_1^β change in the odds that $Y=1$.

Lecture Outline

Classification [revisited]

Why Linear Regression Fails

Binary Response & Logistic Regression

Estimating the Simple Logistic Model

Classification using the Logistic Model

Extending the Logistic Model

Estimation in Logistic Regression

Unlike in linear regression where there exists a closed-form solution to finding the estimates, $\hat{\beta}_j$'s, for the true parameters, logistic regression estimates cannot be calculated through simple matrix multiplication.

In linear regression what loss function was used to determine the parameter estimates?

Estimation in Logistic Regression

Unlike in linear regression where there exists a closed-form solution to finding the estimates, $\hat{\beta}_j$'s, for the true parameters, logistic regression estimates cannot be calculated through simple matrix multiplication.

In linear regression what loss function was used to determine the parameter estimates? What was the probabilistic perspective on linear regression?

Estimation in Logistic Regression

Unlike in linear regression where there exists a closed-form solution to finding the estimates, $\hat{\beta}_j$'s, for the true parameters, logistic regression estimates cannot be calculated through simple matrix multiplication.

In linear regression what loss function was used to determine the parameter estimates? What was the probabilistic perspective on linear regression? Logistic Regression also has a likelihood based approach to estimating parameter coefficients.

Logistic Regression's Likelihood

What are the possible values for the response variable, Y? What distribution defines this type of variable?

Logistic Regression's Likelihood

What are the possible values for the response variable, Y? What distribution defines this type of variable?

A Bernoulli random variable is a discrete random variable defined as one that takes on the values 0 and 1, where $P(Y = 1) = p$. This can be written as $Y \sim \text{Bern}(p)$.

What is the PMF of Y?

Logistic Regression's Likelihood

What are the possible values for the response variable, Y? What distribution defines this type of variable?

A Bernoulli random variable is a discrete random variable defined as one that takes on the values 0 and 1, where $P(Y = 1) = p$. This can be written as $Y \sim \text{Bern}(p)$.

What is the PMF of Y?

$$P(Y = y) = p^y (1 - p)^{1-y}$$

In logistic regression, we say that the parameter p_i depends on the predictor X through the logistic function: $p_i = \frac{e^{\beta X_i}}{1+e^{\beta X_i}}$. Thus not every p_i is the same for each individual.

Logistic Regression's Likelihood (cont.)

Given the observations are independent, what is the likelihood function for p?

Logistic Regression's Likelihood (cont.)

Given the observations are independent, what is the likelihood function for p?

$$\begin{aligned} L(p|Y) &= \prod P(Y_i = y_i) = \prod p_i^{y_i}(1-p_i)^{1-y_i} \\ &= \prod \left(\frac{e^{\beta X_i}}{1+e^{\beta X_i}}\right)^{y_i} \left(1 - \frac{e^{\beta X_i}}{1+e^{\beta X_i}}\right)^{1-y_i} \end{aligned}$$

How do we maximize this?

Logistic Regression's Likelihood (cont.)

Given the observations are independent, what is the likelihood function for p?

$$\begin{aligned} L(p|Y) &= \prod P(Y_i = y_i) = \prod p_i^{y_i}(1-p_i)^{1-y_i} \\ &= \prod \left(\frac{e^{\beta X_i}}{1+e^{\beta X_i}}\right)^{y_i}\left(1-\frac{e^{\beta X_i}}{1+e^{\beta X_i}}\right)^{1-y_i} \end{aligned}$$

How do we maximize this? Take the log and differentiate!

But jeeze does this look messy! It will not necessarily have a closed form solution? So how do we determine the parameter estimates? Through an iterative approach (Newton-Raphson).

NFL TD Data

We'd like to predict whether or not a play from scrimmage (aka, regular play) in the NFL resulted in an offensive touchdown. And we'd like to make this prediction, for now, just based on distance from the goal line.

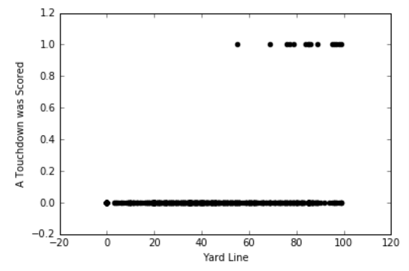
How should we visualize these data?

NFL TD Data

We'd like to predict whether or not a play from scrimmage (aka, regular play) in the NFL resulted in an offensive touchdown. And we'd like to make this prediction, for now, just based on distance from the goal line.

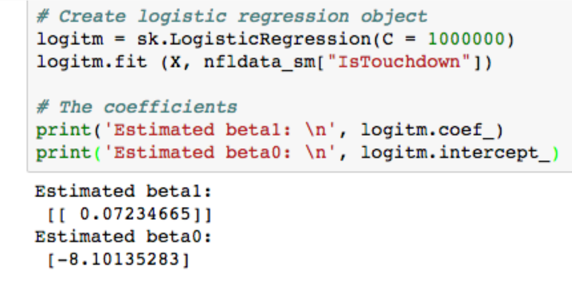
How should we visualize these data?

We start by visualizing the data via a scatterplot (to illustrate the logistic fit):



NFL TD Data: logistic estimation

There are various ways to fit a logistic model to this data set in Python. The most straightforward in `sklearn` is via `linear_model.LogisticRegression`. A little bit of preprocessing work may need to be done first.



```
# Create logistic regression object
logitm = sk.LogisticRegression(C = 1000000)
logitm.fit (X, nfldata_sm["IsTouchdown"])

# The coefficients
print('Estimated beta1: \n', logitm.coef_)
print('Estimated beta0: \n', logitm.intercept_)
```

```
Estimated beta1:
 [[ 0.07234665]]
Estimated beta0:
 [-8.10135283]
```

Use this output to answer a few questions (on the next slide)...

NFL TD Data: Answer some questions

1. Write down the logistic regression model.

NFL TD Data: Answer some questions

1. Write down the logistic regression model.
2. Interpret $\hat{\beta}_1$.

NFL TD Data: Answer some questions

1. Write down the logistic regression model.
2. Interpret $\hat{\beta}_1$.
3. Estimate the probability of scoring a touchdown for a play from the 10 yard line.

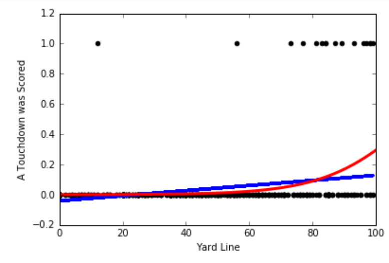
NFL TD Data: Answer some questions

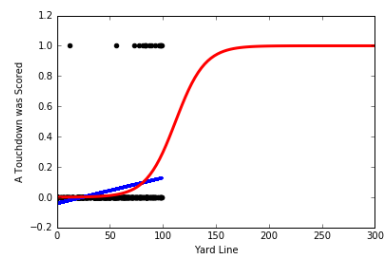
1. Write down the logistic regression model.
2. Interpret $\hat{\beta}_1$.
3. Estimate the probability of scoring a touchdown for a play from the 10 yard line.
4. If we were to use this model purely for classification, how would we do so? See any issues?

NFL TD Data: Solutions

NFL TD Data: curve plot

The probabilities can be calculated/predicted directly using the `predict_proba` command based on your `sklearn` model.





Special case: when the predictor is binary

Just like in linear regression, when the predictor, X, is binary, the interpretation of the model simplifies (and there is a quick closed form solution).

In this case, what are the interpretations of β_0 and β_1?

For the NFL data, let X be the indicator that the play called was a pass. What is the interpretation of the coefficient estimates in this case?

The observed percentage of pass plays that result in a TD is 7.28% while it is just 0.34% for non-passes. Calculate the estimates for β_0 and β_1 if the indicator for TD was predicted from the indicator for pass play.

Predict TD from Pass Play: Solutions

Statistical Inference in Logistic Regression

The uncertainty of the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ can be quantified and used to calculate both confidence intervals and hypothesis tests.

The estimate for the standard errors of these estimates, likelihood-based, is based on a quantity called Fisher's Information (beyond the scope of this class), which is related to the curvature of the function.

Due to the nature of the underlying Bernoulli distribution, if you estimate the underlying proportion p_i, you get the variance for free! Because of this, the inferences will be based on the normal approximation (and not t-distribution based).

Of course, you could always bootstrap the results to perform these inferences as well.

Lecture Outline

Classification [revisited]

Why Linear Regression Fails

Binary Response & Logistic Regression

Estimating the Simple Logistic Model

Classification using the Logistic Model

Extending the Logistic Model

Using Logistic Regression for Classification

How can we use a logistic regression model to perform classification?

That is, how can we predict when $Y = 1$ vs. when $Y = 0$?

We mentioned before, we can classify all observations for which $\hat{P}(Y = 1) \geq 0.5$ to be in the group associated with $Y = 1$ and then classify all observations for which $\hat{P}(Y = 1) < 0.5$ to be in the group associated with $Y = 0$.

Using such an approach is called the standard *Bayes classifier*. The Bayes classifier takes the approach that assigns each observation to the most likely class, given its predictor values.

Bayes classifier details

When will this Bayes classifier be a good one? When will it be a poor one?

Bayes classifier details

When will this Bayes classifier be a good one? When will it be a poor one?

The Bayes classifier is the one that minimizes the overall classification error rate. That is, it minimizes:

$$\frac{1}{n}\sum I\left(y_i = \hat{y}_i\right)$$

Is this a good Loss function to minimize? Why or why not?

Bayes classifier details (cont.)

The Bayes classifier may be a poor indicator within a group. Think about the NFL scatter plot...

This has potential to be a good classifier if the predicted probabilities are on both sides of 0 and 1.

How do we extend this classifier if Y has more than two categories?

Lecture Outline

Classification [revisited]

Why Linear Regression Fails

Binary Response & Logistic Regression

Estimating the Simple Logistic Model

Classification using the Logistic Model

Extending the Logistic Model

Model Diagnostics in Logistic Regression

In linear regression, when is the model appropriate (aka, what are the assumptions)?

In logistic regression, when is the model appropriate?

Model Diagnostics in Logistic Regression

In linear regression, when is the model appropriate (aka, what are the assumptions)?

In logistic regression, when is the model appropriate?

We don't have to worry about the distribution of the resisduals (we get that for free). What we do have to worry about is how Y 'links' to X in its relationship. More specifically, we assume the 'S'-shaped (aka, sigmoidal) curve follows the logistic function.

How could we check this?

Alternatives to logistic regression

Why was the logistic function chosen to model how a binary response variable can be predicted from a quantitative predictor?

Alternatives to logistic regression

Why was the logistic function chosen to model how a binary response variable can be predicted from a quantitative predictor?

Because it it takes as inputs values in $(0, 1)$ and outputs values $(-\infty, \infty)$ so that the estimation of β is unbounded.

This is not the only function that does this. Any suggestions?

Alternatives to logistic regression

Why was the logistic function chosen to model how a binary response variable can be predicted from a quantitative predictor?

Because it it takes as inputs values in $(0, 1)$ and outputs values $(-\infty, \infty)$ so that the estimation of β is unbounded.

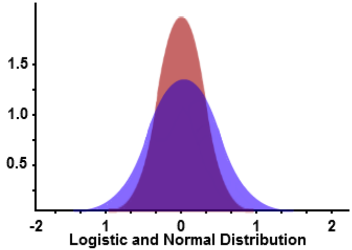
This is not the only function that does this. Any suggestions?

Any *inverse CDF* function for an unbounded continuous distribution can work as the 'link' between the observed values for Y and how it relates 'linearly' to the predictors.

So what are possible other choices? What differences do they have? Why is logistic regression preferred?

Logistic vs. Normal pdf

The choice of link function determines the shape of the ‘S’ shape. Let’s compare the pdf’s for the Logistic and Normal distributions (called a ’probit’ model...econometricians love these):

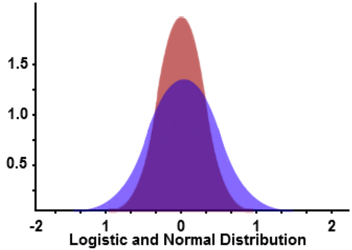


In above graph, blue color represents logistic distribution ; while red denotes normal distribution.

So what?

Logistic vs. Normal pdf

The choice of link function determines the shape of the ‘S’ shape. Let’s compare the pdf’s for the Logistic and Normal distributions (called a ’probit’ model...econometricians love these):



In above graph, blue color represents logistic distribution ; while red denotes normal distribution.

So what?

Choosing a distribution with longer tails will make for a shape that asymptotes more slowly (likely a good thing for model fitting).

Multiple logistic regression

It is simple to illustrate examples in logistic regression when there is just one predictors variable. But the approach 'easily' generalizes to the situation where there are multiple predictors.

A lot of the same details as linear regression apply to logistic regression. Interactions can be considered. Multicollinearity is a concern. So is overfitting. Etc...

So how do we correct for such problems?

Multiple logistic regression

It is simple to illustrate examples in logistic regression when there is just one predictors variable. But the approach ‘easily’ generalizes to the situation where there are multiple predictors.

A lot of the same details as linear regression apply to logistic regression. Interactions can be considered. Multicollinearity is a concern. So is overfitting. Etc...

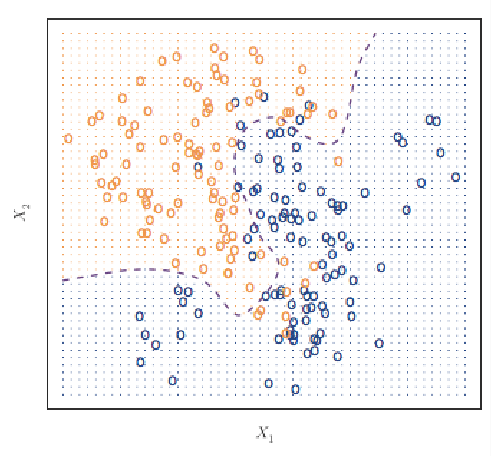
So how do we correct for such problems?

Regularization and checking though train, test, and cross-validation!

We will get into the details of this, along with other extensions of logistic regression, in the next lecture.

Classifier with two predictors

How can we estimate a classifier, based on logistic regression, for the following plot?



How else can we calculate a classifier from these data?