Lecture Outline

Module 2: Classification

Why not Linear Regression?

Binary Response & Logistic Regression

Estimating the Simple Logistic Model

Classification using the Logistic Model

Extending the Logistic Model

Multiple Logistic Regression

Classification Boundaries

Module 2: Classification

Classification

Up to this point, the methods we have seen have centered around modeling and the prediction of a quantitative response variable (ex, # taxi pickups, # bike rentals, etc…). Linear 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.

Simple Classification Example

Given a dataset $\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_N, y_N)\}$, 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:

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

Binary Response & Logistic Regression

Logistic Regression

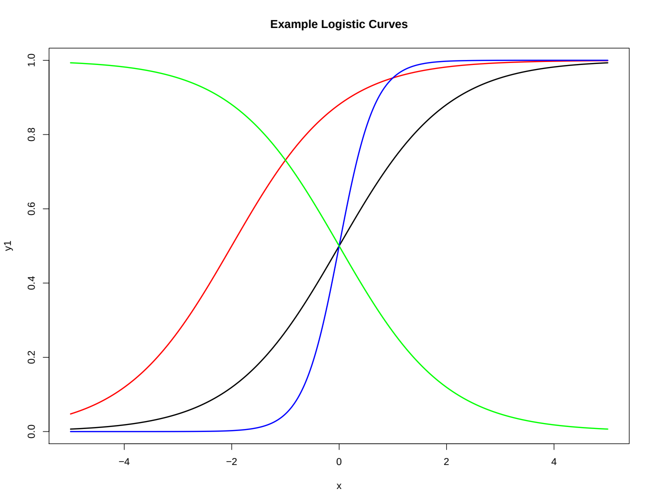
Logistic Regression addresses the problem of estimating a probability, $P(y=1)$, to be outside the range of $[0,1]$. The logistic regression model uses a function, called the *logistic* function, to model $P(y=1)$:

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

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^{β_1} change in the odds that $Y=1$.

Classification using 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?

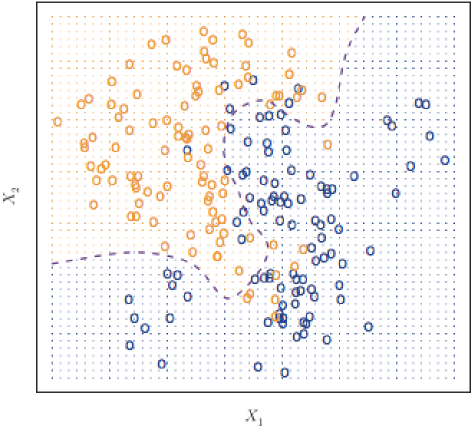
The Bayes classifier is the one that minimizes the overall classification error rate. That is, it minimizes:

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

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

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?

Multiple Logistic Regression

Earlier we saw the general form of *simple* logistic regression, meaning when there is just one predictor used in the model. What was the model statement (in terms of linear predictors)?

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

Multiple logistic regression is a generalization to multiple predictors. More specifically we can define a multiple logistic regression model to predict $P(Y=1)$ as such:

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right)=\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_p X_p$$

where there are p predictors: $X=(X_1,X_2,...,X_p)$.
Note: statisticians are often lazy and use the notation log to mean ln (the text does this). We will write $\log_{10}$ if this is what we mean.

Interpreting Multiple Logistic Regression: an Example

Let's get back to the NFL data. We are attempting to predict whether a play results in a TD based on location (yard line) and whether the play was a pass. The simultaneous effect of these two predictors can be brought into one model.
Recall from earlier we had the following estimated models:

$$\log\left(\frac{P(\widehat{Y=1})}{1-P(\widehat{Y=1})}\right) = -7.425 + 0.0626 \cdot X_{yard}$$

$$\log\left(\frac{P(\widehat{Y=1})}{1-P(\widehat{Y=1})}\right) = -4.061 + 1.106 \cdot X_{pass}$$

The results for the multiple logistic regression model are on the next slide.

Lecture Outline

Logistic Regression: a Brief Review

Classification Boundaries

Regularization in Logistic Regression

Multinomial Logistic Regression

Bayes Theorem and Misclassification Rates

ROC Curves

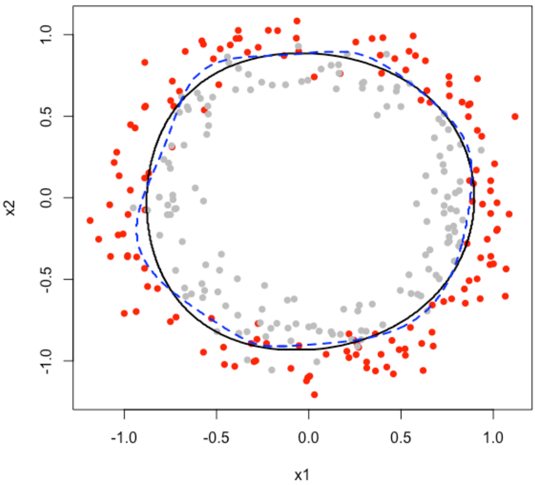
Classification

Recall that we could attempt to purely classify each observation based on whether the estimated $P(Y = 1)$ from the model was greater than 0.5.

When dealing with ‘well-separated’ data, logistic regression can work well in performing classification.

We saw a 2-D plot last time which had two predictors, X_1 and X_2 and depicted the classes as different colors. A similar one is shown on the next slide.

2D Classification in Logistic Regression: an Example



2D Classification in Logistic Regression: an Example

Would a logistic regression model perform well in classifying the observations in this example?

What would be a good logistic regression model to classify these points?

Based on these predictors, two separate logistic regression model were considered that were based on different ordered polynomials of X_1 and X_2 and their interactions. The ‘circles’ represent the boundary for classification.

How can the classification boundary be calculated for a logistic regression?

2D Classification in Logistic Regression: an Example

In the previous plot, which classification boundary performs better? How can you tell? How would you make this determination in an actual data example?

We could determine the misclassification rates in left out validation or test set(s)

Logistic Regression for predicting more than 2 Classes

There are several extensions to standard logistic regression when the response variable Y has more than 2 categories. The two most common are :

1. ordinal logistic regression
2. multinomial logistic regression.

Ordinal logistic regression is used when the categories have a specific hierarchy (like class year: Freshman, Sophomore, Junior, Senior; or a 7-point rating scale from strongly disagree to strongly agree).

Multinomial logistic regression is used when the categories have no inherent order (like eye color: blue, green, brown, hazel, et...).

Multinomial Logistic Regression

The most common approach to estimating a nominal (not-ordinal) categorical variable that has more than 2 classes. The first approach sets one of the categories in the response variable as the *reference* group, and then fits separate logistic regression models to predict the other cases based off of the reference group. For example we could attempt to predict a student's concentration:

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

from predictors x_1 number of psets per week and x_2 how much time spent in Lamont Library.

Multinomial Logistic Regression (cont.)

We could select the $y = 3$ case as the reference group (other concentration), and then fit two separate models: a model to predict $y = 1$ (CS) from $y = 3$ (others) and a separate model to predict $y = 2$ (Stat) from $y = 3$ (others).

Ignoring interactions, how many parameters would need to be estimated?

How could these models be used to estimate the probability of an individual falling in each concentration?

One vs. Rest (ovr) Logistic Regression (cont.)

The default multiclass logistic regression model is called the ’One vs. Rest’ approach.

If there are 3 classes, then 3 separate logistic regressions are fit, where the probability of each category is predicted over the rest of the categories combined. So for the concentration example, 3 models would be fit:

1. a first model would be fit to predict CS from (Stat and Others) combined
2. a second model would be fit to predict Stat from (CS and Others) combined
3. a third model would be fit to predict Others from (CS and Stat) combined

An example to predict play call from the NFL data follows...

Classification for more than 2 Categories

When there are more than 2 categories in the response variable, then there is no guarantee that $P(Y = k) \geq 0.5$ for any one category. So any classifier based on logistic regression will instead have to select the group with the largest estimated probability.

The classification boundaries are then much more difficult to determine. We will not get into the algorithm for drawing these in this class.

Bayes Theorem and Misclassification Rates

Bayes' Theorem

We defined conditional probability as:

$$P(B|A) = \frac{P(B \cap A)}{P(A)}$$

And using the fact that $P(B \cap A) = P(A|B)P(B)$ we get Bayes' Theroem:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Another version of Bayes' Theorem is found by substituting in the Law of Total Probability (LOTP) into the denominator:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^C)P(B^C)}$$

Where have we seen Bayes' Theorem before? Why do we care?

Diagnostic Testing

In the diagnostic testing paradigm, one cares about whether the results of a test (like a classification test) matches truth (the true class that observation belongs to). The simplest version of this is trying to detect disease ($D+$ vs. $D-$) based on a diagnostic test ($T+$ vs. $T-$).

Medical examples of this include various screening tests: breast cancer screening through (i) self-examination and (ii) mammographies, prostate cancer screening through (iii) PSA tests, and Colo-rectal cancer through (iv) colonoscopies.

These tests are a little controversial because of poor predictive probability of the tests.

Diagnostic Testing (cont.)

Bayes' theorem can be rewritten for diagnostic tests:

$$P(D+|T+) = \frac{P(T+|D+)P(D+)}{P(T+|D+)P(D+) + P(T+|D-)P(D-)}$$

These probability quantities can then be defined as:

- *Sensitivity*: $P(T+|D+)$
- *Specificity*: $P(T-|D-)$
- *Prevalence*: $P(D+)$
- *Positive Predictive Value*: $P(D+|T+)$
- *Negative Predictive Value*: $P(D-|T-)$

How do positive and negative predictive values relate? Be careful…

Error in Classification

There are 2 major types of error in classification problems based on a binary outcome. They are:

- False positives: incorrectly predicting $\hat{Y} = 1$ when it truly is in $Y = 0$.
- False negative: incorrectly predicting $\hat{Y} = 0$ when it truly is in $Y = 1$.

The results of a classification algorithm are often summarized in two ways: a confusion table, sometimes called a contingency table, or a 2x2 table (more generally kxk table) and an receiver operating characteristics (ROC) curve.

Confusion table

When a classification algorithm (like logistic regression) is used, the results can be summarize in a kxk table as such:

		True Republican Status	
		Yes	No
Predicted	Yes	487	288
Republican	No	218	314

The table above was a classification based on a logistic regression model to predict political party (Dem. vs. Rep.) based on 3 predictors: $X_1 =$ whether respondent believes abortion is legal, $X_1 =$ income (logged) and $X_3 =$ years of education.

What are the false positive and false negative rates for this classifier?

Bayes' Classifier Choice

A classifier's error rates can be tuned to modify this table. How?

The choice of the Bayes' classifier level will modify the characteristics of this table.

If we thought is was more important to predict republicans correctly (lower false positive rate), what could we do for our Bayes' classifier level?

We could classify instead based on:

$$\hat{P}(Y = 1) < \pi$$

and we could choose π to be some level other than 0.5. Let's see what the table looks like if π were 0.28 or 0.52 instead (why such strange numbers?).

Other Confusion table

Based on $\pi = 0.28$:

		True Republican Status	
		Yes	No
Predicted	Yes	247	528
Republican	No	80	452

What has improved? What has worsened?

Other Confusion table

Based on $\pi = 0.28$:

		True Republican Status	
		Yes	No
Predicted	Yes	247	528
Republican	No	80	452

What has improved? What has worsened?

Based on $\pi = 0.52$:

		True Republican Status	
		Yes	No
Predicted	Yes	627	148
Republican	No	388	144

Which should we choose? Why?

ROC Curves

ROC Curves

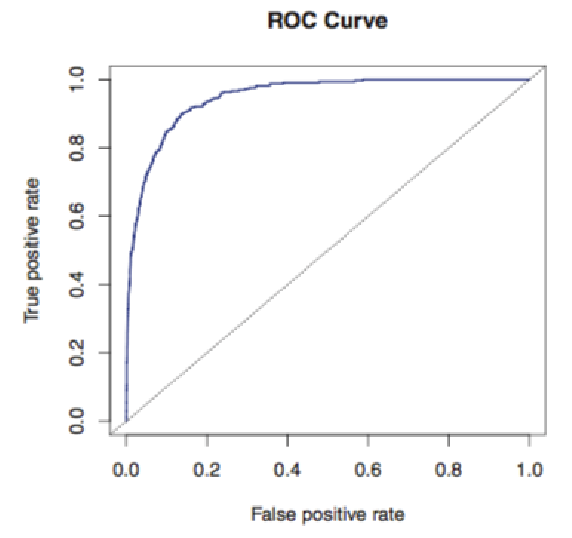
The ROC curve illustrates the trade-off for all possible thresholds chosen for the two types of error (or correct classification).

The vertical axis displays the true positive predictive value and the horizontal axis depicts the true negative predictive value.

What is the shape of an ideal ROC curve?

See next slide for an example.

ROC Curve Example



ROC Curve for measuring classifier preformance

The overall performance of a classifier, calculated over all possible thresholds, is given by the area under the ROC curve ('AUC').

An ideal ROC curve will hug the top left corner, so the larger the AUC the better the classifier.

What is the worst case scenario for AUC? What is the best case? What is AUC if we independently just flip a coin to perform classification?

This AUC then can be use to compare various approaches to classification: Logistic regression, LDA (to come), kNN, etc...