Logistic Regression

Sean Sall

May 26th, 2016

Objectives

- Describe the motivation for logistic regression
- Understand how to fit a logistic model and interpret its coefficients
- Explain common classification metrics, and how they tie into the ROC curve

Agenda

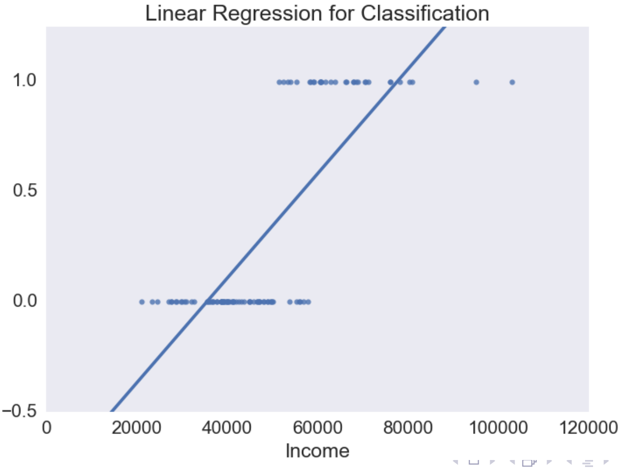
- Logistic regression motivation
- Logistic regression details
  - Sigmoid (logistic) function discussion
  - Solving through maximum likelihood estimation
  - Interpreting the results
- Classification metrics and the confusion matrix
  - Precision, Recall, Accuracy
  - Specificity, Sensitivity (recall)
  - True positive rate (recall), false positive rate
- ROC Curve

Logistic Regression - A Visual Motivation

Linear Review - Visual

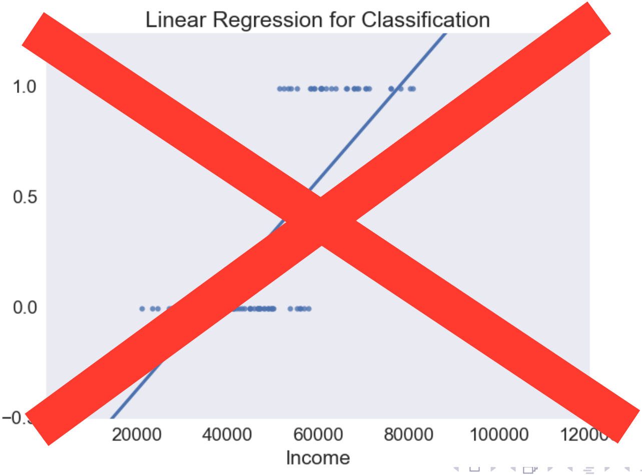
- With **linear regression**, we are modeling a **continuous** response and finding the linear function that gives the best fit

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



Linear Regression for Classification - Visual

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



A Model for Classification

- We need a model that:
  - Takes **continuous input** (e.g. -infinity to infinity)
  - Produces **output** between 0 and 1
  - Transitions from outputting 0 to outputting 1 quickly
  - Has interpretable coefficients (like our standard linear regression model)

Logistic Regression for Classification

- Enter **logistic regression**. . .

The Sigmoid function

- That general S-shaped curve is from the sigmoid family, who's general functional form is given by:

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

Logistic Regression - Motivation II

Linear Review - Underlying Assumptions

- In a standard **linear regression framework**, we assume that our response is **normally distributed**:

$$y_i \mid X \sim N(X\beta, \sigma^2)$$

- In a **classification setting**, that's **not** the case

Classification Setting - Obs. Distribution

- In a **binary classification setting**, the response is **binary**:

$$y_i \begin{cases} \textit{1, if event occurs} \\ \textit{0, if event doesn't occur} \end{cases}$$

- Each observation is drawn from a **Bernoulli distribution**:

$$y_i \mid X \sim Bernoulli(p)$$

- Our **standard** linear model won't work

A Model for Classification

- We need a model that:
  - Takes **continuous input** (e.g. -infinity to infinity)
  - Produces **output** between 0 and 1
  - Transitions from outputting 0 to outputting 1 quickly
  - Has interpretable coefficients (like our linear regression model)
    - Takes the mean response of our observations and *links* it to a linear combination of our inputs (e.g. $X\beta$)

Note: $X\beta = \beta_0 + X_1\beta_1 + X_2\beta_2 + ... + X_n\beta_n$

Logistic Regression for Classification

- Enter **logistic regression**...

$$p(y_i) = \frac{1}{1 + e^{-X\beta}}$$

Note: $p(y)$ denotes the probability of success for y. We can think of this as the mean of the response.

Logistic Regression - The Details

Logistic Regression

- Logistic regression fits a **logistic function** that we use to obtain the probability that an individual observation (y_i) is a success (typically denoted as a 1, where a failure is denoted by a 0).

$$p(y_i) = \frac{1}{1 + e^{-X\beta}}$$

- How do we get this function, though?

Logistic Regression - The Link Function

- The **link** function provides the relationship between a linear combination of our inputs ($X\beta$) and the mean of our response ($p(y_i)$)
- For logistic regression, we use the following link function:

$$\ln\left(\frac{p(y_i)}{1 - p(y_i)}\right) = X\beta$$

- See the appendix for a derivation of how to move from this to the logistic function that we use to predict the mean of our response

Logistic Regression - Solution 1

- The parameters of our logistic regression are estimated via **maximum likelihood**. We know that each individual observation follows a Bernoulli distribution:

$$y_i \mid X \sim Bernoulli(p)$$

- Given this, we can construct the likelihood of our β matrix as:

$$\mathcal{L}(\beta \mid y) = \prod_{i=1}^{N} p(y_i)^{y_i} + (1 - p(y_i))^{(1-y_i)}$$

- And from there, our log likelihood:

$$\ell = \sum_{i=1}^{N} y_i \log p(y_i) + (1 - y_i) log(1 - p(y_i))$$

Logistic Regression - Solution 2

- Unfortunately, there is no closed form solution (like in linear regression)
- As a result, iterative methods are typically used
  - These work with the first and/or second derivatives to try to take clever steps towards an optimal solution (defined by maximizing the likelihood function), often starting with a random guess
  - Tomorrow you'll be doing this through `stochastic gradient descent`, which is one of the most popular techniques

Logistic Regression - Interpretation Part 1

- Say we fit a logistic regression model with the outcome/response as whether or not a person works (yes/no, which is denoted with a 1/0) and only one predictor, `income`:

$$p(y_i) = \frac{1}{1 + e^{-(\beta_0 + X_1 \beta_{income})}}$$

- To actually interpret the coefficients, we need to go back to our original *link* function:

$$\ln\left(\frac{p(y_i)}{1 - p(y_i)}\right) = \beta_0 + X_1 \beta_{income}$$

Logistic Regression - Interpretation Part 2

- Typically, we'll take one step away from the raw *link* function, and then we'll build up our interpretation from here:

$$\frac{p(y_i)}{1 - p(y_i)} = e^{\beta_0 + X_1 \beta_{income}}$$

- We can modify that to this:

$$\frac{p(y_i)}{1 - p(y_i)} = e^{\beta_0} e^{X_1 \beta_{income}}$$

- Turns out that the left side of the above equation is known as the **odds** ratio. So, we interpret our results as follows:
  - For a one-unit increase in X_1, the odds increases by $e^{\beta_{income}}$

Logistic Regression - Interpretation Example

- Let's say in the context of our example (regressing whether or not somebody works on income), our β_1 is 0.00001. This means that a one-unit increase in income ($1) causes an $e^{(0.00001)}$ increase in the odds of somebody working.
  - $e^{(0.00001)} = 1.00001$
- This ultimately means that for each additional dollar that a person makes, we expect a 0.001% increase in the odds that they are working.
  - For an additional $1000 dollars that a person makes, we expect a 1% increase in the odds that they are working

Classification Metrics

Classification Metrics I

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

		Predicted	
		Positive	Negative
True	Positive	True Positives (TP)	False Negatives (FN)
	Negative	False Positives (FP)	True Negatives (TN)

Figure 5:Confusion Matrix

Classification Metrics II

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

$$\frac{\textit{True Positives}}{\textit{True Positives} + \textit{False Negatives}}$$

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

$$\frac{\textit{True Negatives}}{\textit{True Negatives} + \textit{False Positives}}$$

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

$$\frac{\textit{True Positives}}{\textit{True Positives} + \textit{False Positives}}$$

Classification Metrics III

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

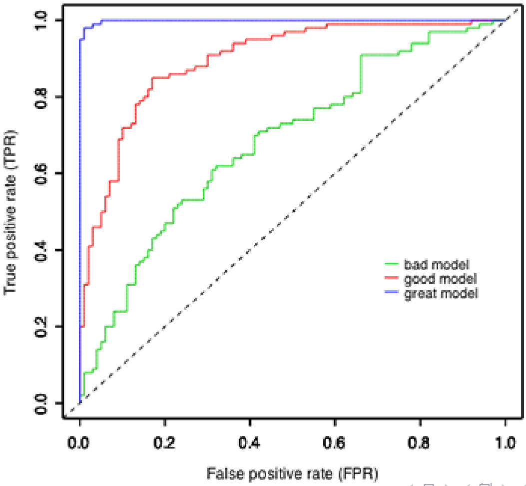
$$\frac{\textit{True Positives} + \textit{True Negatives}}{\textit{Total number of obs.}}$$

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

$$\frac{\textit{False Positives}}{\textit{False Positives} + \textit{True Negatives}}$$

ROC Curve I

- We use build a receiver operating curve (ROC) to visualize the performance of a given binary classifier:



ROC Curve II

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

Appendix

Logistic Regression - From link to probability I

1. $\ln\left(\frac{p(y_i)}{1-p(y_i)}\right) = X\beta$

2. $\frac{p(y_i)}{1-p(y_i)} = e^{X\beta}$

3. $p(y_i) = (1-p(y_i))e^{X\beta}$

4. $p(y_i) = e^{X\beta} - p(y_i)e^{X\beta}$

5. $p(y_i) + p(y_i)e^{X\beta} = e^{X\beta}$

Logistic Regression - From link to probability II

6. $p(y_i)(1 + e^{X\beta}) = e^{X\beta}$

7. $p(y_i) = \dfrac{e^{X\beta}}{1 + e^{X\beta}}$

8. $p(y_i) = \dfrac{\frac{e^{X\beta}}{e^{X\beta}}}{\frac{1+e^{X\beta}}{e^{X\beta}}}$

9. $p(y_i) = \dfrac{1}{1 + e^{-X\beta}}$