Logistic Regression

Advanced Methods for Data Analysis (36-402/36-608)

Spring 2014

1 Classification

1.1 Introduction to classification

- *Classification*, like regression, is a predictive task, but one in which the outcome takes only values across discrete categories; classification problems are very common (arguably just as or perhaps even more common than regression problems!)
- Examples:
  - Predicting whether a patient will develop breast cancer or remain healthy, given genetic information
  - Predicting whether or not a user will like a new product, based on user covariates and a history of his/her previous ratings
  - Predicting the region of Italy in which a brand of olive oil was made, based on its chemical composition
  - Predicting the next elected president, based on various social, political, and historical measurements
- Classification is fundamentally a different problem than regression, and so, we will need different tools. In this lecture we will learn one of the most common tools: logistic regression. You should know that there are many, many more methods beyond this one (just like there are many methods for estimating the regression function)

1.2 Why not just use least squares?

- Before we present logistic regression, we address the (reasonable) question: why not just use least squares?
- Consider a classification problem in which we are given samples (x_i, y_i), $i = 1, \ldots n$, and as usual $x_i \in \mathbb{R}^p$ denotes predictor measurements, and y_i discrete outcomes. In this classification problem, we will assume that y_i can take two values, which we will write (without a loss of generality) as 0 and 1. Then, to predict a future outcome Y from predictor measurement X, we might consider performing linear regression. I.e., we fit coefficients $\hat{\beta}$ according to the familiar criterion

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - \beta^T x_i)^2,$$

  and to predict the class of Y from X, we round $\hat{\beta}^T X$ to whichever class is closest, 0 or 1
- What is the problem with this, if any? For purely predictive purposes, this actually is not a crazy idea—it tends to give decent predictions. But there are two drawbacks:

1. We cannot use any of the well-established routines for statistical inference with least squares (e.g., confidence intervals, etc.), because these are based on a model in which the outcome is continuously distributed. At an even more basic level, it is hard to precisely interpret $\hat{\beta}$
2. We cannot use this method when the number of classes exceeds 2. If we were to simply code the response as $1, \dots K$ for a number of classes $K > 2$, then the ordering here would be arbitrary—but it actually matters[1]

2 Logistic regression

2.1 The logistic model

- Throughout this section we will assume that the outcome has two classes, for simplicity. (We return to the general K class setup at the end.) Logistic regression starts with different model setup than linear regression: instead of modeling Y as a function of X directly, we model the *probability* that Y is equal to class 1, given X. First, abbreviate $p(X) = \mathbb{P}(Y = 1|X)$. Then the *logistic model* is

$$p(X) = \frac{\exp(\beta^T X)}{1 + \exp(\beta^T X)} \tag{1}$$

  The function on the right-hand side above is called the *sigmoid* of $\beta^T X$. What does it look like for $\beta^T X$ large positive? For $\beta^T X$ negative? Plot it to gather some intuition (i.e., plot $e^a/(1 + e^a)$ as a function of a)

- Rearranged, the equation (1) says that

$$\log\Big(\frac{p(X)}{1 - p(X)}\Big) = \beta^T X. \tag{2}$$

  The left-hand side above is called the *log odds* or *logit* of $p(X)$, and is written as $\operatorname{logit} p(X)$. In general, $\operatorname{logit}(a) = \log(a/(1 - a))$

- Note that assuming (1) (or equivalently, (2)), is a modeling decision, just like it is a modeling decision to use linear regression

- Also note that, to include an intercept term of the form $\beta_0 + \beta^T X$, we just append a 1 to the vector X of predictors, as we do in linear regression

2.2 Interpreting coefficients

- How can we interpret the role of the coefficients $\beta = (\beta_1, \dots \beta_p) \in \mathbb{R}^p$ in (1) (i.e., in (2))? One nice feature of the logistic model is that it comes equipped with a useful interpretation for these coefficients

- Write

$$\frac{p(X)}{1 - p(X)} = e^{\beta^T X} = e^{\beta_1 X_1 + \dots + \beta_p X_p}.$$

  The left-hand side above is the odds of class 1 (conditional on X). We can see that increasing X_j by one unit, while keeping all other predictors fixed, multiplies the odds by e^{β_j}. This is because

$$e^{\beta_1 X_1 + \dots + \beta_j (X_j + 1) + \dots + \beta_p X_p} = e^{\beta_1 X_1 + \dots + \beta_j X_j + \dots + \beta_p X_p} \cdot e^{\beta_j}$$

[1] You might think we could get around this by modeling one class versus the rest with a binary coding, and performing K separate regressions, then using the strongest prediction at the end, given an input X. This is flawed too, however, as we would likely encounter a problem called *masking*. With this problem, there is some class j that never ends up being predicted in favor of the others, regardless of the input

- Equivalently, write
$$\log\Big(\frac{p(X)}{1-p(X)}\Big) = \beta^T X = \beta_1 X_1 + \ldots + \beta_p X_p.$$
Now increasing X_j by one unit, and keeping all other predictors fixed, changes the log odds by β_j

- It will help to get comfortable with the concept of odds, and log odds, if you haven't done so already in another class. Note that probabilities q close to 0 or 1 have odds $q/(1-q)$ close to 0 or ∞, respectively. And probabilities q close to 0 or 1 have log odds $\log(q/(1-q))$ close to $-\infty$ or ∞, respectively

2.3 Maximum likelihood estimation

- Given samples $(x_i, y_i) \in \mathbb{R}^p \times \{0,1\}$, $i = 1, \ldots n$, we let $p(x_i) = \mathbb{P}(y_i = 1 | x_i)$, and assume
$$\log\Big(\frac{p(x_i)}{1-p(x_i)}\Big) = \beta^T x_i, \quad i = 1, \ldots n$$

- To construct an estimate $\hat{\beta}$ of the coefficients, we will use the principle of *maximum likelihood*. I.e., assuming independence of the samples, the likelihood (conditional on x_i, $i = 1, \ldots n$) is
$$\begin{aligned} L(\beta) &= \prod_{i\,:\,y_i=1} p(x_i) \cdot \prod_{i\,:\,y_i=0} \big(1-p(x_i)\big) \\ &= \prod_{i=1}^n p(x_i)^{y_i} \big(1-p(x_i)\big)^{1-y_i}. \end{aligned}$$
We will choose $\hat{\beta}$ to maximize this likelihood criterion

- Note that maximizing a function is the same as maximizing the log of a function (because log is monotone increasing). Therefore $\hat{\beta}$ is equivalently chosen to maximize the log likelihood
$$\ell(\beta) = \sum_{i=1}^n y_i \log p(x_i) + (1-y_i) \log\big(1-p(x_i)\big).$$
It helps to rearrange this as
$$\begin{aligned} \ell(\beta) &= \sum_{i=1}^n y_i \big[\log p(x_i) - \log\big(1-p(x_i)\big)\big] + \log\big(1-p(x_i)\big) \\ &= \sum_{i=1}^n y_i \log\Big(\frac{p(x_i)}{1-p(x_i)}\Big) + \log\big(1-p(x_i)\big). \end{aligned}$$
Finally, plugging in for $\log(p(x_i)/(1-p(x_i))) = x_i^T\beta$ and using $1-p(x_i) = 1/(1+\exp(x_i^T\beta))$, $i = 1, \ldots n$,
$$\ell(\beta) = \sum_{i=1}^n y_i (x_i^T\beta) - \log\big(1+\exp(x_i^T\beta)\big). \tag{3}$$
You can see that, unlike the least squares criterion for regression, this criterion $\ell(\beta)$ does not have a closed-form expression for its maximizer (e.g., try taking its partial derivatives and setting them equal to zero). Hence we have to run an optimization algorithm to find $\hat{\beta}$

- Somewhat remarkably, we can maximize (3) by running repeated weighted least squares regressions! For those of you who have learned a little bit of optimization, this is actually just an instantiation of Newton's method. Applied to the criterion (3), we refer to it as *iteratively reweighted least squares* or IRLS
- In short: estimation of $\hat{\beta}$ in logistic regression is more involved than it is in linear regression, but it is possible to do so by iteratively using linear regression software

2.4 Decision boundary

- Suppose that we have formed the estimate $\hat{\beta}$ of the logistic coefficients, as discussed in the last section. To predict the outcome of a new input $x \in \mathbb{R}^p$, we form

$$\hat{p}(x) = \frac{\exp(\hat{\beta}^T x)}{1 + \exp(\hat{\beta}^T x)},$$

and then predict the associated class according

$$\hat{f}(x) = \begin{cases} 0 & \hat{p}(x) \leq 0.5 \\ 1 & \hat{p}(x) > 0.5 \end{cases}$$

- Equivalently, we can study the log odds

$$\operatorname{logit} \hat{p}(x) = \hat{\beta}^T x,$$

and predict the associated class using

$$\hat{f}(x) = \begin{cases} 0 & \hat{\beta}^T x \leq 0 \\ 1 & \hat{\beta}^T x > 0 \end{cases}$$

- The set of all $x \in \mathbb{R}^p$ such that

$$\hat{\beta}^T x = \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p = 0$$

is called the *decision boundary* between classes 0 and 1. On either side of this boundary, we would predict one class or the other

- Remembering the intercept, we would rewrite the decision boundary as

$$\hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p = 0. \tag{4}$$

This is a point when $p = 1$, it is a line when $p = 2$, and in general it is a $(p-1)$-dimensional subspace. We would therefore say that logistic regression has a *linear decision boundary*; this is because the equation (4) is linear in x

2.5 Inference

- A lot of the standard machinery for inference in linear regression carries over to logistic regression. Recall that we can solve for the logistic regression coefficients $\hat{\beta}$ by performing repeated weighted linear regressions; hence we can simply think of the logistic regession estimates $\hat{\beta}$ as the result of a single weighted linear regression—the last one in this sequence (upon convergence). Confidence intervals for $\hat{\beta}_j$, $j = 1, \ldots p$, and so forth, are then all obtained from this weighted linear regression perspective. We will not go into detail here, but such inferential tools are implemented in software, and it helps to be aware of where they come from

2.6 Multinomial regression

- With more than two classes, the story is similar. Now we use an extension of the logistic model called the *multinomial model*, which, given K classes for the outcome Y, takes the form

$$\begin{aligned}
\mathbb{P}(Y = 1|X) &= \frac{\exp(\beta_1^T X)}{1 + \sum_{j=1}^{K-1} \exp(\beta_j^T X)} \\
\mathbb{P}(Y = 2|X) &= \frac{\exp(\beta_2^T X)}{1 + \sum_{j=1}^{K-1} \exp(\beta_j^T X)} \\
&\vdots \\
\mathbb{P}(Y = K-1|X) &= \frac{\exp(\beta_{K-1}^T X)}{1 + \sum_{j=1}^{K-1} \exp(\beta_j^T X)} \\
\mathbb{P}(Y = K|X) &= \frac{1}{1 + \sum_{j=1}^{K-1} \exp(\beta_j^T X)}
\end{aligned}$$

- Equivalently, we can write this in log odds form as

$$\begin{aligned}
\log\Big(\frac{\mathbb{P}(Y = 1|X)}{\mathbb{P}(Y = K|X)}\Big) &= \beta_1^T X \\
\log\Big(\frac{\mathbb{P}(Y = 2|X)}{\mathbb{P}(Y = K|X)}\Big) &= \beta_2^T X \\
\log\Big(\frac{\mathbb{P}(Y = K-1|X)}{\mathbb{P}(Y = K|X)}\Big) &= \beta_{K-1}^T X
\end{aligned}$$

- The interpretation of coefficients is similar to before: increasing X_ℓ by one unit and keeping all other predictors fixed, $\beta_{j\ell}$ pertains to the change in $\log(\mathbb{P}(Y = j|X)/\mathbb{P}(Y = K-1|X))$

- Estimation proceeds by maximum likelihood, as before; inference is again drawn from treating the estimates as the result of a single weighted linear regression

- Finally, predictions are made at an input x by forming

$$\begin{aligned}
\hat{p}_1(x) &= \frac{\exp(\hat{\beta}_1^T x)}{1 + \sum_{j=1}^{K-1} \exp(\hat{\beta}_j^T x)} \\
\hat{p}_2(x) &= \frac{\exp(\hat{\beta}_2^T x)}{1 + \sum_{j=1}^{K-1} \exp(\hat{\beta}_j^T x)} \\
&\vdots \\
\hat{p}_{K-1}(x) &= \frac{\exp(\hat{\beta}_{K-1}^T x)}{1 + \sum_{j=1}^{K-1} \exp(\hat{\beta}_j^T x)} \\
\hat{p}_K(x) &= \frac{1}{1 + \sum_{j=1}^{K-1} \exp(\hat{\beta}_j^T x)},
\end{aligned}$$

  and then predicting the class according to

$$\hat{f}(x) = \underset{j=1,\dots K}{\operatorname{argmax}}\ \hat{p}_j(x)$$