Logistic Regression With Two Classes

Logistic regression with two classes models the class probabilities as:

$$P(G=1\,|\,X) \;\;=\;\; [1 \;+\; \exp(-(\beta_0+\beta_1X_1+\ldots+\beta_pX_p))]^{-1}$$

Looked at another way, with logistic regression, the "logit" of the probability is a linear function of the inputs:

$$\begin{aligned}\text{logit}(P(G=1\,|\,X)) \;&=\; \log\left(\frac{P(G=1\,|\,X)}{1-P(G=1\,|\,X)}\right)\\ &=\; \log\left(\frac{P(G=1\,|\,X)}{P(G=2\,|\,X)}\right)\\ &=\; \beta_0+\beta_1X_1+\ldots+\beta_pX_p\end{aligned}$$

This is also called the "log odds" of class 1 versus class 2.

Since logit is monotonically increasing, $\beta_0+\beta_1X_1+\ldots+\beta_pX_p$ is a linear discriminant function.

There's Nothing Magical About Logit

We'll see later how logits arise in connection with normal distributions, but typically there's no particular reason to insist on them.

Another possibility is *probit regression*, in which

$$P(G=1\,|\,X) \;\; = \;\; \Phi(\beta_0+\beta_1X_1+\ldots+\beta_pX_p)$$

where $\Phi(z)$ is the standard normal cumulative distribution function.

The functions $\Phi(z)$ and $[1+\exp(-z)]^{-1}$ look very similar, but Φ approaches 0 and 1 faster as $\beta_0+\beta_1X_1+\ldots+\beta_pX_p$ goes to $-\infty$ and $+\infty$.

Maximum Likelihood for Logistic Regression Models

We can use the *maximum likelihood* method to estimate the β_j.

Let x_i be the vector of inputs for training case i, and let y_i be a 0/1 representation of the response for train case i, with $y_i = 1$ when $g_i = 1$ and $y_i = 0$ when $g_i = 2$.

Write the probability of class 1 with inputs x_i and parameters β as

$$p(x_i;\beta) \;\; = \;\; [1 \; + \; \exp(-(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p))]^{-1}$$

We don't try to model the distribution of inputs, only of responses given inputs. We also assume that the training cases are independent. So the likelihood is just the product of probabilities for responses in training cases, which we can write as

$$L(\beta) \;\; = \;\; \prod_{i=1}^{N} p(x_i;\beta)^{y_i}\,(1 - p(x_i;\beta))^{1-y_i}$$

We find the $\hat\beta$ that maximizes this, and then make predictions for test cases using this. For a test case with inputs x,

$$P(\text{class } 1 \,|\, x) \;\; = \;\; p(x;\hat\beta)$$

Computation of the Maximum Likelihood Estimate

It's easier to deal with the log of the likelihood, which we can write as

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

The $\hat\beta$ that maximizes this is the same as the one that maximizes the likelihood.

There is no simple formula for $\hat\beta$, so numerical optimization methods must be used. Fortunately,

1) There is a unique maximum of $\ell(\beta)$, **unless** the classes can be perfectly separated by a hyperplane, as discussed later.

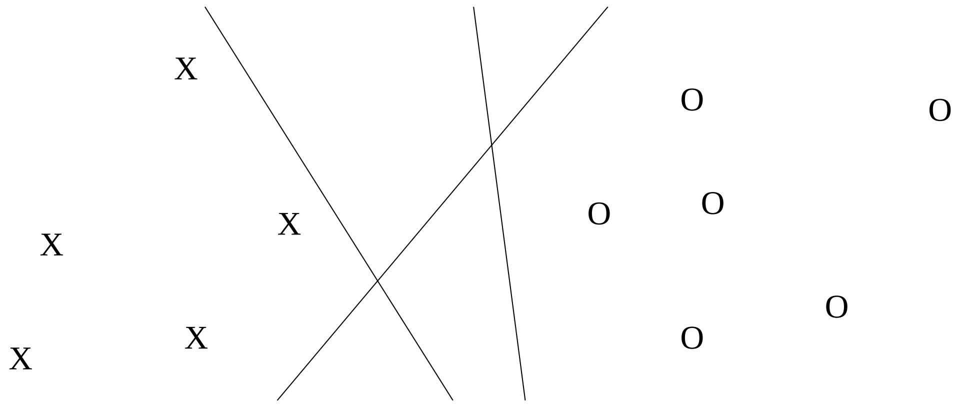
2) The log likelihood surface is convex.

3) The derivatives and second derivatives of $\ell(\beta)$ can easily be computed.

Efficient methods like Newton-Raphson iteration can be used to find the point where $\nabla\ell(\beta) = 0$, which will be where the maximum occurs. (Some care is needed, but this isn't a difficult numerical problem.)

Logistic Regression When Classes are Linearly Separable

A “problem” occurs when there is a hyperplane in the input space that perfectly separates the two classes. For example:



This shows the training cases in input space (with two inputs), with the classes distinguished as X and O. The classes can be perfectly separated by any of the lines shown.

With logistic regression, the hyperplane where the class probabilities are equal is given by

$$\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p \;\; = \;\; 0$$

Note that the *same* hyperplane is defined by $\beta' = a\beta$, for any positive a. Also the bigger we make a, the closer to 0 and 1 the probabilities $p(x; a\beta)$ will be.

. . . When Classes are Linearly Separable (Continued)

The consequence of classes being linearly separable are therefore:

- We can make the likelihood bigger and bigger by making the β_j be bigger and bigger, as long as these β_j define a hyperplane that perfectly separates the classes.
- There are many different hyperplanes that perfectly separate the data.

So there are many "maximum likelihood" solutions, all of which have β_j that are infinite (only their ratios matter).

Is this a problem? Actually, it's good news when the classes can be linearly separated! We just have to do something reasonable.

One possibility is to subtract a small penalty, $\lambda \sum_{j=1}^{p} \beta_j^2$, from the log likelihood. Even a tiny λ will stop the β_j from becoming infinite, and will produce a unique solution.

Logistic Regression With More Than Two Classes

This is also called the *multinomial logit* model in statistics, and the linear *softmax* model in machine learning. Class probabilities are given by

$$P(G = k \,|\, X) \;\; = \;\; \frac{\exp(\beta_{k0} + \beta_{k1}X_1 + \ldots + \beta_{kp}X_p)}{\sum_{\ell=1}^{K} \exp(\beta_{\ell 0} + \beta_{\ell 1}X_1 + \ldots + \beta_{\ell p}X_p)}$$

The book assumes that all $\beta_{Kj} = 0$, so that one of the exponential terms above becomes $\exp(0) = 1$. This makes the model "identifiable", which is convenient for talking about maximum likelihood, but makes the model asymmetric if one adds penalties on the β_{kj}.

With this model, the functions

$$\delta_k(x) \;\; = \;\; \beta_{k0} + \beta_{k1}x_1 + \ldots + \beta_{kp}x_p$$

form a set of linear discriminant functions.

We can find maximum likelihood estimates for this model too. Again, there is the issue of training data that can be perfectly separated, which can be handled by subtracting a small penalty from the log likelihood.