Classification: Logistic Regression and Linear Discriminant Analysis

Michael

1 Intuition Behind Logistic Regression

In statistics 101 or econometrics 101, students are taught that they should be aware of the properties of independent variables. For instance, when we want to study the effects of different factors, such as education, gender, family background, etc., on individual's income level, we have to know that gender is dummy variable, education is categorical variables, etc. We don't care too much whether independent variables are continuous or not (e.g., dummy variable is not continuous) as long as dependent variable (Y) is continuous. Why? Because we need Y is continuous when we want to do differentiation to find the optimal solutions.

What if dependent variable Y now becomes discrete values, such as 1, 2, 3, 4, which they are not continuous? We need employ the logistic transformation to bring the model into a continuous domain. Through the logistic transformation, we *link our categorical values(like class A, B, C) with certain probability values ($\pi \in [0, 1]$).*

Now, we would like to have the probability π_i depend on a vector of observed covariates x_i (e.g., education, income). The simplest idea would be to let π_i be a linear function of the covariates, say

$$\pi_i = x_i'\beta \tag{1.1}$$

where β is a vector of regression coefficients. This model sometimes is also called the linear probability model. Once we get the estimated probability $\hat{\pi}$, if we can compare it with benchmark value, say if $\hat{\pi} > 0.6$ then observation with x_i characters can be classified as class A.

However, one problem with this model is that probability π_i on the left hand side has to be between 0 and 1, but the linear predictor $x_i'\beta$ on the right hand side can take any real value, so there is no guarantee that the predicted values will be in the correct range unless complex restrictions are imposed on the coefficients.

A simple solution to this problem is to transform the probability to remove the range restrictions, and model the transformation as a linear function of the covariates. We do this in two steps:

1. we move from the probability π to the odds:
$$odds_i = \frac{\pi_i}{1-\pi_i}$$
2. we take logarithms, calculating the *logit* or log-odds:
$$\eta_i = \log(odds) = \log(\frac{\pi_i}{1-\pi_i}), where\ \eta \in (-\infty, +\infty)$$

Now, Now, let's do some simple algebra and begin to use MLE. Solving for π_i in above equation, it gives

$$\pi_i = logit^{-1}(\eta_i) = \frac{e^{\eta_i}}{1+e^{\eta_i}} \tag{1.2}$$

We are now in a position to define the logistic regression model, by assuming that the logit of the odds ratio, rather than the probability itself, follows a linear model:

$$log(\frac{\pi_i}{1-\pi_i}) = x_i'\beta \quad \Rightarrow \frac{\pi_i}{1-\pi_i} = e^{x_i'\beta} \quad \Rightarrow \pi_i = F(x_i'\beta) = \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}} \tag{1.3}$$

In this case marginal effect can be obtained as

$$\frac{d\pi_i}{dx_{ij}} = \left[\frac{e^{x_i'\beta}}{(1+e^{x_i'\beta})^2}\right]\beta_j = [\frac{e^{x_i'\beta}}{(1+e^{x_i'\beta})}\frac{1}{(1+e^{x_i'\beta})}]\beta_j = \beta_j\pi_i(1-\pi_i) \tag{1.4}$$

2 Logistic Regression

Most of content in this section is taken from the notes by Ng (2014). As it has shown in (1.3), we can get a continuous function (or map) between covariates x and dependent variable probability π:

$$\pi = h_\theta(x) = \frac{e^{\theta' x}}{1+e^{\theta' x}} = \frac{1}{1+e^{-\theta' x}} \tag{2.1}$$

where the general format

$$g(z) = \frac{1}{1+e^{-z}}$$

is called the **logistic function**.

So, given the logistic regression model, how do we fit θ for it? As we are using the probability now, we need employ the maximum likelihood. Let us assume that

$$\begin{aligned} P(y=1|x;\theta) &= h_\theta(x) \\ P(y=0|x;\theta) &= 1-h_\theta(x) \end{aligned}$$

Note this can be written more compactly as

$$P(y|x;\theta) = (h_\theta(x))^y(1-h_\theta(x))^{1-y} \tag{2.2}$$

Assuming that the m training examples were generated independently, we can then write down the likelihood of the parameters as

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{m} P(y^i|x^i;\theta) \\ &= \prod_{i=1}^{m} (h_\theta(x))^y(1-h_\theta(x))^{1-y} \end{aligned}$$

How do we maximize the likelihood? We can still use gradient descent method. However, as we are doing maximum of likelihood rather than the minimum of cost function, so our update rule will become (we also call descent ascent not descent)

$$\theta = \theta + \alpha\Delta_\theta l(\theta)$$

where $\Delta_\theta l(\theta)$ is the just derivative of our likelihood function. Now, let's take one training example (x, y) and take derivative:

$$\begin{aligned}\frac{\partial L(\theta)}{\partial \theta} &= \left(y\frac{1}{g(\theta' x)} - (1-y)\frac{1}{1-g(\theta' x)}\right)\frac{\partial}{\partial \theta_j} g(\theta' x) \\ &= (y - h_\theta(x))x_j \\ &= (y - \frac{1}{1+e^{-\theta' x}})x_j\end{aligned}$$

This therefore gives us the *stochastic gradient ascent* rule

$$\theta_j = \theta_j + \alpha(y^i - h_\theta(x^i))x^i_j \tag{2.3}$$

where

$$h_\theta(x) = \frac{e^{\theta' x}}{1+e^{\theta' x}} = \frac{1}{1+e^{-\theta' x}}$$

If we have more than two categories to classify, we can still use logistic regression to model the classification. However, the method is not as efficient as linear discriminant analysis, which we will go through in the next section.

3 Linear Discriminant Analysis

To employ the logistic regression in section 2 we need know the class posteriors $P(y|x;\theta)$ for optimal classification. For instance, we are using Bernoulli distribution in equation (2.2). Now, suppose $f_k(x)$ is the class-conditional density of X in class $G = k$, and let π_k be the prior probability of class k, with $\sum_{k=1}^K \pi_k = 1$. A simple application of Bayes theorem gives us

$$P(G = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^K f_l(x)\pi_l} \tag{3.1}$$

We see that in terms of ability to classify, having the $f_k(x)$ is almost equivalent to having the quantity $P(G = k|X = x)$. Many techniques are based on models for the class densities:

- linear and quadratic discriminant analysis use Gaussian densities
- more flexible mixtures of Gaussians allow for nonlinear decision boundaries
- general nonparametric density estimates for each class density allow the most flexibility.

Remark: Andrew Ng wrote the brilliant notes for this section, please read it.

References

Ng, A. (2014). Cs229 machine learning: Lecture notes.