4 Classifications

4.1 Logistic Regression

4.1.1 Logistic Model

logistic function:

$$p(x_i) = \frac{e^{\beta^T x_i}}{1 + e^{\beta^T x_i}} = \frac{1}{1 + e^{-\beta^T x_i}}, \quad 0 \leq p(x_i) \leq 1, \quad -\infty < x_i < \infty \tag{4.1}$$

The logistic function will always produce an S-shaped curve.

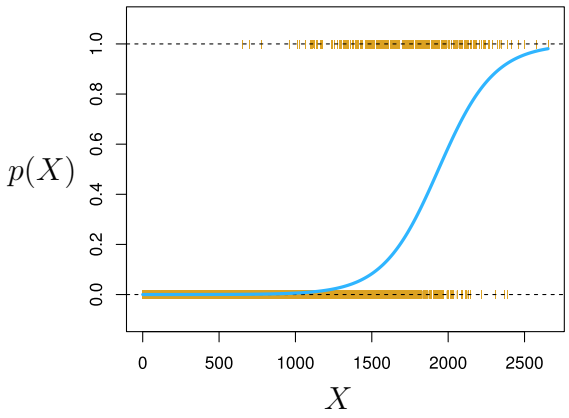

Figure 1

With a bit of manipulation, we will get a odds function:

$$\frac{p(x_i)}{1 - p(x_i)} = e^{\beta^T x_i}, \quad 0 < \frac{p(x_i)}{1 - p(x_i)} < \infty \tag{4.2}$$

The notation $p(x)$ can be interpreted as $\Pr(\mathrm{Y} = 1|\mathrm{X} = \mathrm{x})$, which means the probability of Y is True given x. The equation (4.2) imply the ratio of success compare to fail. For example, average nine of ten people go to school. That means 9 times of people go to school compare to those who don't. So:

$$p(x) = \Pr(\mathrm{y} = 1|\mathrm{x}) = 9/10$$
$$\frac{p(x)}{1 - p(x)} = \frac{9/10}{1 - 9/10} = 9$$

By taking the logarithm of the odds function, we get

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta^T x_i \tag{4.3}$$

The left-hand side is called the log-odds or logit. We see that the logistic regression model has a logit that is linear in X.

4.1.2 Estimating the Regression Coefficients

To fit logistic model (4.1), we use Maximum Likelihood to estimate the coefficients:

Definition 1 (Likelihood Function)**.**
$X_1, X_2, \ldots, X_n$ 爲一組樣本大小爲n之隨機樣本, 記爲$X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} f_{X_i}(x_i;\theta)$, 定義母體參數 θ 之 *likelihood function:*

$$L(\theta) = f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i;\theta)$$

可解讀爲在不同的母體參數值θ之下, 抽到這組觀測的隨機樣本 $(X_1, X_2, \ldots, X_n)$ 之可能性

Maximum Likelihood Estimator, $\hat{\theta}_{\text{MLE}}$:
找一個 θ 使得拿到這組隨機樣本 $(X_1, X_2, \ldots, X_n)$ 的可能性爲最大, 亦即, 找一個 θ 使得概似函數 $L(\theta)$ 爲最大:

$$\arg\max_{\theta} L(\theta), \text{求解 } \hat{\theta}_{\text{MLE}}$$

Unlike linear regression, we can no longer write down the MLE in closed form. Instead, we need to use an optimization algorithm to compute it. For this, we need to derive the gradient and Hessian. The likelihood function of logistic model:

$$L(\beta) = \prod_{i=1}^{n} p(x_i;\beta)^{y_i}(1-p(x_i;\beta))^{1-y_i}, \text{ where } R_{y_i} = \{0,1\} \tag{4.4}$$

The log-likelihood of logistic model:

$$\begin{aligned}
\ell(\beta) &= \sum_{i=1}^{n} y_i \log p(x_i;\beta) + (1-y_i)\log(1-p(x_i;\beta)) \\
&= \sum_{i=1}^{n} y_i \log p(x_i;\beta) + \log(1-p(x_i;\beta)) - y_i \log(1-p(x_i;\beta)) \\
&= \sum_{i=1}^{n} y_i \log \frac{p(x_i;\beta)}{1-p(x_i;\beta)} + \log(1-p(x_i;\beta)) \\
&= \sum_{i=1}^{n} y_i \beta^T x_i + \log(\frac{1}{1+e^{\beta^T x_i}}) \\
&= \sum_{i=1}^{n} \{y_i \beta^T x_i - \log(1+e^{\beta^T x_i})\}
\end{aligned}$$

First order partial differential of the log-likelihood function (gradient descent):

$$\begin{aligned}
\frac{\partial \ell(\beta)}{\partial \beta} &= \sum_{i=1}^{n} x_i (y_i - \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}) \\
&= \sum_{i=1}^{n} x_i (y_i - p(x_i;\beta))
\end{aligned}$$

Second order partial differential of the log-likelihood function (the s.o.c can transfer to a Hessian matrix):

$$\begin{aligned}
\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^T} &= \sum_{i=1}^{n} x_i (-x_i \cdot \frac{e^{-\beta^T x_i}}{1+e^{-\beta^T x_i}} \cdot \frac{1}{1+e^{-\beta^T x_i}}) \\
&= -\sum_{i=1}^{n} x_i x_i^T p(x_i;\beta)(1-p(x_i;\beta))
\end{aligned}$$

Starting with β^{old}, a single Newton update is

$$\beta^{new} = \beta^{old} - \left(\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^T}\right)^{-1} \frac{\partial \ell(\beta)}{\partial \beta}$$

Now we will try to simplify by vectorizing the model. Let $\mathbf{y}$ denote the vector of y_i, $\mathbf{X}$ the $N \times (p+1)$ matrix of x_i (p predictors with one intercept), $\mathbf{p}$ the vector of fitted probabilities with ith element $p(x_i;\boldsymbol{\beta}^{\text{old}})$ and $\mathbf{W}$ a $N \times N$ diagonal matrix of weights with ith diagonal element $p(x_i;\boldsymbol{\beta}^{\text{old}})(1-p(x_i;\boldsymbol{\beta}^{\text{old}}))$. The structure of the notation and the model:

$$\mathbf{y}_{N\times 1} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}, \mathbf{X}_{N\times(p+1)} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & x_{N2} & \cdots & x_{Np} \end{pmatrix},$$

$$\mathbf{p}_{N\times 1} = \begin{pmatrix} p(x_1;\boldsymbol{\beta}^{\text{old}}) \\ p(x_2;\boldsymbol{\beta}^{\text{old}}) \\ \vdots \\ p(x_N;\boldsymbol{\beta}^{\text{old}}) \end{pmatrix}, \boldsymbol{\beta}^{\text{old}}_{(p+1)\times 1} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix},$$

$$\mathbf{W}_{N\times N} = \begin{pmatrix} p(x_1;\boldsymbol{\beta}^{\text{old}})(1-p(x_1;\boldsymbol{\beta}^{\text{old}})) & 0 & \cdots & 0 \\ 0 & p(x_2;\boldsymbol{\beta}^{\text{old}})(1-p(x_2;\boldsymbol{\beta}^{\text{old}})) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p(x_N;\boldsymbol{\beta}^{\text{old}})(1-p(x_N;\boldsymbol{\beta}^{\text{old}})) \end{pmatrix}$$

$$\begin{aligned} \frac{\partial \ell(\beta)}{\partial \beta} &= \mathbf{X}^T(\mathbf{y}-\mathbf{p}) \\ \frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^T} &= -\mathbf{X}^T\mathbf{W}\mathbf{X} \\ \boldsymbol{\beta}^{\text{new}} &= \boldsymbol{\beta}^{\text{old}} + \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T(\mathbf{y}-\mathbf{p}) \\ &= \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\left(\mathbf{X}\boldsymbol{\beta}^{\text{old}} + \mathbf{W}^{-1}(\mathbf{y}-\mathbf{p})\right) \\ &= \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\mathbf{z} \end{aligned}$$

Response of weighted least squares step, sometimes known as adjusted response:

$$\mathbf{z} = \mathbf{X}\boldsymbol{\beta}^{\text{old}} + \mathbf{W}^{-1}(\mathbf{y}-\mathbf{p}) \tag{4.5}$$

The equations above solved repeatedly, since at each iteration $\mathbf{p}$ changes, and hence so does $\mathbf{W}$ and $\mathbf{z}$. This algorithm is referred to as iteratively reweighed least squares (IRLS).

Algorithm 1 Iteratively reweighted least squares (IRLS)

1: $\boldsymbol{\beta}^{\text{new}} = \begin{pmatrix} 1 \\ \boldsymbol{\beta}_p \end{pmatrix}$;
2: **repeat**
3: $\quad \mathbf{p} = p(\mathbf{X}; \boldsymbol{\beta}^{\text{new}})$
4: $\quad \mathbf{W}_{N \times N} = \text{diag}(\mathbf{p}(1 - \mathbf{p}))$
5: $\quad \mathbf{z} = \mathbf{X}\boldsymbol{\beta}^{\text{new}} + \mathbf{w}^{-1}(\mathbf{y} - \mathbf{p})$
6: $\quad \boldsymbol{\beta}^{\text{new}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{z}$
7: **until** calculate loss function to check converged;

4.1.3 Discussion

1. How to calculate the fitness of logistic model?
2. How to check the estimated coefficients is significant or not?
3. How to perform the hypothesis testing for logistic regression?
4. What is the assumptions of logistic regression model?

4.2 Linear Discriminant Analysis (LDA)

Difference between Linear Discriminant Analysis and Logistic Analysis:

- when the classes are well-separated, the parameter estimates for the logistic regression model are surprisingly unstable. LDA does not suffer from this problem
- If n is small and the distribution of the predictors X is approximately normal in each of the classes, LDA is more stable than logistic regression
- LDA is preferable when we have more than two response classes.
- Logistic regression can take both qualitative and quantitative as predictor variables. However LDA can only take quantitative predictor variables, due to the assumption of multivariate Gaussian Distribution.

- LDA assumes that the classes have a common covariance matrix (Continuous random variable X)

4.2.1 Bayes Theorem for Classification

Classify an observation into one of K classes, where $K \geq 2$.
Notations:

- Y: The qualitative response variable with K different categories, $R_Y = 1, \ldots, K$
- π_k: The prior probability that a randomly chosen observation comes from the kth class
- $f_k(X)$: the density function of X for an observation comes form the kth class

Note:

$$\hat{\pi}_k = n_k/n = \hat{Pr}(Y = k)$$
$$f_k(X) = Pr(X = x|Y = k)$$

From the law of total probability, the above notations can be stated as below:

$$Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{\ell=1}^{K} \pi_\ell f_\ell(x)} \tag{4.6}$$

The abbreviation for $Pr(Y = k|X = x)$ will be denoted as $p_k(X)$ which is the posterior probability that an observation $X = x$ belongs to the kth class. In the following section, we have to make assumptions and approximates $f_k(X)$ to build a classifier that approximates the Bayes classifier.

Definition 2 (The law of total probability)**.**
設 $A_1, A_2, \ldots, A_n$ 爲樣本空間S中之一組分割, 則對於樣本空間S之任意事件B而言,

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

4.2.2 Linear Discriminant Analysis for p=1

In the assumption of the Linear Discriminant Analysis, we assume the independent variable X follow Gaussian distribution. So the $f_k(x)$ takes the form

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{1}{2\sigma_k^2}(x-\mu_k)^2} \tag{4.7}$$

LDA also assume homoscedasticity for $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_K^2$, so the simplify version of $f_k(x)$

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu_k)^2} \tag{4.8}$$

By plugging (4.8) into (4.6), we get

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu_k)^2}}{\sum_{\ell=1}^{K} \pi_\ell \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu_\ell)^2}} \tag{4.9}$$

In order to estimate the parameters from (4.8), we apply MAP,

$$\begin{aligned} \hat{p}_k &= \arg\max_k P(Y = k|X = x) \\ &= \arg\max_k \pi_k f_k(x) \end{aligned} \tag{4.10}$$

The objective of LDA is to find a k that maximizes posterior probability($\hat{p}_k$) among all Kth posterior probabilities. From the above model, in order to meet this objective, we will try to find a k which will maximize the conditional probability ($\pi_k f_k(x)$). Since, the maximum of $\pi_k f_k(x)$, imply the largest probability among all Kth posterior probabilities. In order to simplify (4.10), we take the log of (4.10)

$$\begin{aligned} \arg\max_k \delta_k(x) &= \arg\max_k \left(\log\left(\pi_k f_k(x)\right) \right) \\ &= \arg\max_k \left(\log \pi_k + x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \underline{\log \frac{1}{\sqrt{2\pi}\sigma} - \frac{x^2}{2\sigma^2}} \right) \\ &= \arg\max_k \left(\log \pi_k + x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \cancel{\text{fixed constant}} \right) \end{aligned} \tag{4.11}$$

In practice we don't know the parameters of the Gaussian distributions(unknown μ_k, σ^2), and will need to estimate them using our training data:

- $\hat{\pi}_k = n_k/n$
- $\hat{\mu}_k = \sum_{i:y_i=k} x_i/n_k$
- $\hat{\sigma}^2 = \sum_{k=1}^{K}(n_k - 1)S_k/(n-k)$
- $S_k = \sum_{i:y_i=k}(x_i - \hat{\mu}_k)^2/(n_k - 1)$

The discriminant function,

$$\arg\max_k \hat{\delta}_k(x) = \arg\max_k \left(\left(\log \hat{\pi_k} - \frac{\hat{\mu_k}^2}{2\hat{\sigma^2}} \right) + x\frac{\hat{\mu}_k}{\hat{\sigma}^2} \right) \tag{4.12}$$

The reason linear in the name of LDA, is because the discriminant functions are linear functions of x

4.2.3 Linear Discriminant Analysis for $p > 1$

Assume that $X = \begin{pmatrix} X_1 & X_1 & \dots & X_p \end{pmatrix}$ drown from a multivariate Gaussian distribution, with a class-specific mean vector and a common covariance matrix. To indicate that a p-dimensional random variable X has a multivariate Gaussian Distribution, we write

$$X \sim N(\mu, \mathbf{\Sigma})$$
$$E(X) = \sum_{i=1}^{p} X_i/p = \mu$$
$$Cov(X) = \mathbf{\Sigma}$$

$\mathbf{\Sigma}$ is the $p \times p$ covariance matrix of X. The multivariate Gaussian Distribution density is defined as

$$f(x) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \mathbf{\Sigma}^{-1}(x-\mu)}$$

Using the same method as describe for $p = 1$, we get the optimization function for $p > 1$ version

$$\begin{aligned}
\arg\max_k \delta_k(x) &= \log\big(\pi_k f_k(x)\big) \\
&= \arg\max_k \left(\log \pi_k + \log\big(\frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}}\big) - \frac{1}{2}(x-\mu_k)^T\mathbf{\Sigma}^{-1}(x-\mu_k) \right) \\
&= \arg\max_k \left(\log \pi_k + x^T\mathbf{\Sigma}^{-1}\mu_k - \frac{1}{2}\mu_k^T\mathbf{\Sigma}^{-1}\mu_k + \cancel{\text{constant}} \right) \\
&= \arg\max_k \left(\log \pi_k + x^T\mathbf{\Sigma}^{-1}\mu_k - \frac{1}{2}\mu_k^T\mathbf{\Sigma}^{-1}\mu_k \right)
\end{aligned}$$

The parameters $(\hat{\pi}_k, \hat{\mu}_k)$ are estimated as before except $\mathbf{\Sigma}$. Let S_k denote the $p \times p$ covariance matrix for class k. Then the sample covariance matrix $\mathbf{\Sigma}$ can be estimated by substituting in the pooled covariance matrix S :

$$S = \frac{\sum_{k=1}^{K}(n_k - 1)S_k}{n - K}$$

The discriminant function for $p > 1$

$$\arg\max_k \hat{\delta}_k(x) = \arg\max_k \left(\log \hat{\pi}_k + x^T S^{-1}\hat{\mu}_k - \frac{1}{2}\hat{\mu}_k^T S^{-1}\hat{\mu}_k \right)$$

Example:

Two predictors with 11 observations and $k = 2$:

$$[x^T, y] = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \\ x_9 \\ x_{10} \\ x_{11} \end{array} \begin{array}{c} \begin{array}{ccc} X_1 & X_2 & y \end{array} \\ \left(\begin{array}{cc|c} 1 & 2 & 1 \\ 2 & 3 & 1 \\ 3 & 3 & 1 \\ 4 & 5 & 1 \\ 5 & 5 & 1 \\ 1 & 0 & 2 \\ 2 & 1 & 2 \\ 3 & 1 & 2 \\ 3 & 2 & 2 \\ 5 & 3 & 2 \\ 6 & 5 & 2 \end{array}\right) \end{array}$$

$$c_1 = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{array} \begin{array}{c} \begin{array}{ccc} X_1 & X_2 & y \end{array} \\ \left(\begin{array}{cc|c} 1 & 2 & 1 \\ 2 & 3 & 1 \\ 3 & 3 & 1 \\ 4 & 5 & 1 \\ 5 & 5 & 1 \end{array}\right) \end{array} \qquad c_2 = \begin{array}{c} \\ x_6 \\ x_7 \\ x_8 \\ x_9 \\ x_{10} \\ x_{11} \end{array} \begin{array}{c} \begin{array}{ccc} X_1 & X_2 & y \end{array} \\ \left(\begin{array}{cc|c} 1 & 0 & 2 \\ 2 & 1 & 2 \\ 3 & 1 & 2 \\ 3 & 2 & 2 \\ 5 & 3 & 2 \\ 6 & 5 & 2 \end{array}\right) \end{array}$$

$$\mu_1 = \begin{bmatrix} 3 & 3.6 \end{bmatrix} \qquad \mu_2 = \begin{bmatrix} 3.3 & 2 \end{bmatrix}$$

$$S_1 = \begin{bmatrix} \sigma_{11}^{k=1} = 2.5 & \sigma_{12}^{k=1} = 2 \\ \sigma_{21}^{k=1} = 2 & \sigma_{22}^{k=1} = 1.8 \end{bmatrix} \qquad S_2 = \begin{bmatrix} \sigma_{11}^{k=2} = 3.47 & \sigma_{12}^{k=2} = 3.2 \\ \sigma_{21}^{k=2} = 3.2 & \sigma_{22}^{k=2} = 3.2 \end{bmatrix}$$

$$S = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n - K} = \frac{\begin{bmatrix} 10 & 8 \\ 8 & 7.2 \end{bmatrix} + \begin{bmatrix} 17.3 & 16 \\ 16 & 16 \end{bmatrix}}{11 - 2} = \frac{\begin{bmatrix} 27.3 & 24 \\ 24 & 23.2 \end{bmatrix}}{11 - 2}$$

4.2.4 Discussion

1. What is the purpose of partial differential on Newton's update?
2. What is the definition of Maximum a posteriori estimation(MAP)?
3. What is the Bayes decision boundary when $K = 2$? What is the Bayes decision boundary when $K > 2$?
4. What is Fisher linear discriminant function ?

Reference

張翔 (2012), 《提綱契領學統計》, 四版, 鼎茂圖書

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). Springer, Berlin: Springer series in statistics.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 6). New York: springer.

Wasserman, L. (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.