Supplementary Material

Mathematical working of Logistic Regression

- **Problem and data**

The problem demonstrated is a binary classification problem with two predictor variables $X = (X_1, X_2)$ and two classes $Y = 1,2$.

- **Construction of decision boundary**

The motivation behind Logistic Regression is to use Linear Regression to model the posterior probabilities of the two classes.

$$\text{i.e.}\quad P(Y = 1 \mid X = x) = \beta_0 + \beta^T x$$

But, $0 \leq P(X) \leq 1$, for any random variable X and $\sum P(X) = 1$.

To satisfy these conditions, the **logistic function** is used to model the probabilities. Therefore,

$$P(Y = 1 \mid X = x) = \frac{e^{\beta_0 + \beta^T x}}{1 + e^{\beta_0 + \beta^T x}},$$

$$P(Y = 2 \mid X = x) = \frac{1}{1 + e^{\beta_0 + \beta^T x}}$$

So $P(X)$ belongs to [0,1] and clearly sum to 1.

Calculating the logit function for the data we get,

$$\frac{P(Y = 1 \mid X = x)}{P(Y = 2 \mid X = x)} = e^{\beta_0 + \beta^T x}$$

$$or, \qquad \log\left(\frac{P(Y = 1 \mid X = x)}{P(Y = 2 \mid X = x)}\right) = \beta_0 + \beta^T x$$

The linear logit function forms the decision boundary of any Logistic Regression model.

Note: For K > 2 classes, the model can be specified in terms of K-1 logit transformations. In that case, the parameter set $\theta = \{\beta_{10},\ {\beta_1}^T, \dots, \beta_{(K-1)0},\ {\beta_{(K-1)}}^T\}$ must be estimated.

Logistic Regression is widely used in applications of a Binary classification problem, where only a single linear function is formed and only two parameters have to be estimated.

Note: The logit function models a linear regression with the predictor variables. But there doesn't exist a linear relationship of $P(X)$ with the predictor variables.

Estimating the parameters of Logistic Regression

In Logistic Regression, we estimate the regression coefficients (β_0, β) using the method of Maximum likelihood.
The basic intuition behind using the method is to estimate β_0, β such that the conditional Probability for all the observations of Class K = 0 is a number close to 0 and for Class K =1 is a number close to 1.

Given N observations (x_i, y_i), $i = 1,2,3, \dots, N$ and two classes y_i = 1,2.
The estimates of $\beta' = \{\beta_0, \beta\}$ can be obtained by maximizing the likelihood function,

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

where $p(x_i;\beta') = P(Y = 1 \mid X = x_i ; \beta')$.

The log-likelihood can be written as,

$$l(\beta') = \sum_{i=1}^{N} \left\{ y_i \log\left(p(x_i;\beta')\right) + (1 - y_i) \log\left(1 - p(x_i;\beta')\right) \right\}$$
$$= \sum_{i=1}^{N} \{ y_i \, (\beta')^T x_i - \log\left(1 + e^{(\beta')^T x_i}\right) \}$$

The ML estimates $(\hat{\beta}_0, \hat{\beta})$ are obtained by maximizing $l(\beta')$ w.r.t β_0, β.

References

- The Elements of Statistical Learning
- An Introduction to Statistical Learning with Applications in R
- Modern Multivariate Statistical Techniques