LOGISTIC REGRESSION

The classification problem can be written as follows:

$$\mathbf{Y} = \underset{y_k}{\operatorname{argmax}} \{\mathbf{P}(\mathbf{Y} = y_k \mid x_1,\ x_2,\ ...\ x_p)$$

"Generative" classifiers like the Naive Bayes classifier solve this problem using

$$\mathbf{P}(\mathbf{Y} = y_k \mid x_1,\ x_2,\ ...\ x_p) = \frac{\mathbf{P}(x_1,\ x_2,\ ...\ x_p|\mathbf{Y} = y_k) \times \mathbf{P}(\mathbf{Y} = y_k)}{\mathbf{P}(x_1,\ x_2,\ ...\ x_p)}$$

and modelling

$$\mathbf{P}(x_1,\ x_2,\ ...\ x_p|\mathbf{Y} = y_k)$$

On the other hand, "Discriminative" classifiers like Logistic Regression model

$$\mathbf{P}(\mathbf{Y} = y_k \mid x_1,\ x_2,\ ...\ x_p)$$

directly.

1 Probability Model

Assume the binary classification problem. Logistic Regression assumes:

$$\mathbf{P}(\mathbf{Y} = 1 \mid x_1,\ x_2,\ ...\ x_p) = \frac{e^{\beta_0+\beta_1 x_1+\ ...\ +\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\ ...\ +\beta_p x_p}}$$

To simplify this equation, the following notation is used,

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_p \end{pmatrix} \text{ and } \boldsymbol{x} = \begin{pmatrix} 0 \\ x_1 \\ . \\ . \\ . \\ x_p \end{pmatrix}$$

The equation now becomes,

$$\mathbf{P}(\mathbf{Y} = 1|\mathbf{X} = \boldsymbol{x}) = \frac{e^{\boldsymbol{\beta}^T \boldsymbol{x}}}{1 + e^{\boldsymbol{\beta}^T \boldsymbol{x}}}$$

The other half of this equation is,

$$\mathbf{P}(\mathbf{Y} = 0|\mathbf{X} = \boldsymbol{x}) = \frac{1}{1 + e^{\boldsymbol{\beta}^T \boldsymbol{x}}}$$

These half equations can be combined as follows,

$$\mathbf{P}(\mathbf{Y} = y|\mathbf{X} = \boldsymbol{x}) = \left(\frac{e^{\boldsymbol{\beta}^T \boldsymbol{x}}}{1 + e^{\boldsymbol{\beta}^T \boldsymbol{x}}}\right)^y \left(\frac{1}{1 + e^{\boldsymbol{\beta}^T \boldsymbol{x}}}\right)^{1-y}$$

2 Maximum Likelihood estimation

If the form of a Probability Density Function(PDF) is known and some independent and identically distributed(i.i.d.) observations are given, Maximum Likelihood Estimation(MLE) can be used to estimate the optimal parameters of the PDF.

For example, consider a random variable whose PDF is known to be Gaussian,

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

If some i.i.d. observations $x_1, x_2, \; ... \; x_n$ are given, likelihood can be written as follows,

$$\begin{aligned} l(\mu, \sigma^2) &= P(x_1, x_2, \; ... \; x_n|\mu, \sigma^2) \\ &= \prod_i P(x_i|\mu, \sigma^2) \end{aligned}$$

Log likelihood is more desirable to work with most of the times and expressed as follows,

$$
\begin{aligned}
L(\mu, \sigma^2) &= \sum_i log(P(x_i|\mu, \sigma^2)) \\
&= -\frac{n}{2} log(2\pi\sigma^2) - \sum_i \frac{(x_i - \mu)^2}{2\sigma^2}
\end{aligned}
$$

To maximize log likelihood, its derivative must be equated to zero.

$$
\begin{aligned}
\frac{\partial L(\mu, \sigma^2)}{\partial \mu} &= 0 \\
\frac{\partial}{\partial \mu}\left(-\frac{n}{2} log(2\pi\sigma^2) - \sum_i \frac{(x_i - \mu)^2}{2\sigma^2}\right) &= 0 \\
\sum_i (x_i - \mu) &= 0 \\
\mu = \frac{1}{n}\sum_i x_i
\end{aligned}
$$

So, in this case, MLE estimates μ to be average of the observations.

3 MLE application to Logistic Regression

$$
\begin{aligned}
L(\boldsymbol{\beta}) &= \sum_i log\left\{\left(\frac{e^{\boldsymbol{\beta}^T \boldsymbol{x_i}}}{1 + e^{\boldsymbol{\beta}^T \boldsymbol{x_i}}}\right)^{y_i}\left(\frac{1}{1 + e^{\boldsymbol{\beta}^T \boldsymbol{x_i}}}\right)^{1-y_i}\right\} \\
&= \sum_i \left(y_i \boldsymbol{\beta}^T \boldsymbol{x_i} - log(1 + e^{\boldsymbol{\beta}^T \boldsymbol{x_i}})\right)
\end{aligned}
$$

Taking the derivative and equating it to zero,

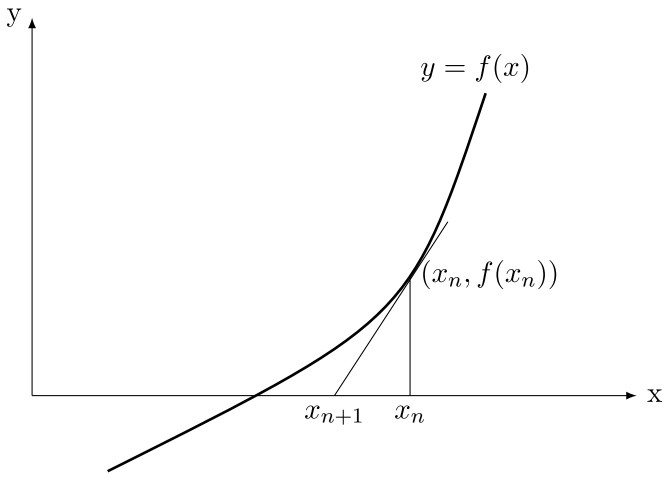
$$
\begin{aligned}
\frac{\partial \boldsymbol{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= 0 \\
\sum_i \left(y_i \boldsymbol{x_i} - \frac{e^{\boldsymbol{\beta}^T \boldsymbol{x_i}}}{1 + e^{\boldsymbol{\beta}^T \boldsymbol{x_i}}} \boldsymbol{x_i}\right) &= 0
\end{aligned}
$$

Unfortunately, no closed form solution exists in this case and hence this equation must be solved by numerical methods like the Newton-Raphson method. The objective function is always convex, so there are no problems of local maxima.

4 Newton-Raphson Method

Newton-Raphson method successively finds better approximations to the root of a real valued function.

4.1 Single varable



$$f(x + \Delta x) \approx f(x) + f'(x)\Delta x$$

Setting $f(x + \Delta x) = 0$ to find a better approximation of the root,

$$\begin{aligned} f(x) + f'(x)\Delta x &\approx 0 \\ \Delta x &\approx -\frac{f(x)}{f'(x)} \\ x_{n+1} - x_n &\approx -\frac{f(x)}{f'(x)} \end{aligned}$$

The method starts with an initial guess x_0 and then applies the above updation logic to come up with $x_1,\ x_2,$ ... until convergence.

4.2 Multiple Variables

$$f_1(x + \Delta x, y + \Delta y) \approx f_1(x, y) + \left(\tfrac{\partial f_1}{\partial x}, \tfrac{\partial f_1}{\partial y}\right)\begin{pmatrix}\Delta x \\ \Delta y\end{pmatrix}$$

$$f_2(x + \Delta x, y + \Delta y) \approx f_2(x, y) + \left(\tfrac{\partial f_2}{\partial x}, \tfrac{\partial f_2}{\partial y}\right)\begin{pmatrix}\Delta x \\ \Delta y\end{pmatrix}$$

These two equations can be written in matrix form as follows,

$$\begin{pmatrix} f_1(x+\Delta x, y+\Delta y) \\ f_2(x+\Delta x, y+\Delta y) \end{pmatrix} \approx \begin{pmatrix} f_1(x,y) \\ f_2(x,y) \end{pmatrix} + \begin{pmatrix} \frac{\partial f_1}{\partial x}, \frac{\partial f_1}{\partial y} \\ \frac{\partial f_2}{\partial x}, \frac{\partial f_2}{\partial y} \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix}$$

Setting the l.h.s to zero and solving,

$$\begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} \approx - \begin{pmatrix} \frac{\partial f_1}{\partial x}, \frac{\partial f_1}{\partial y} \\ \frac{\partial f_2}{\partial x}, \frac{\partial f_2}{\partial y} \end{pmatrix}^{-1} \begin{pmatrix} f_1(x,y) \\ f_2(x,y) \end{pmatrix}$$

The 2×2 matrix of partial derivatives is called Jacobian and denoted by J. The same idea can be extended to any number of variables.

4.3 Newton-Raphson application to Logistic Regression