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,

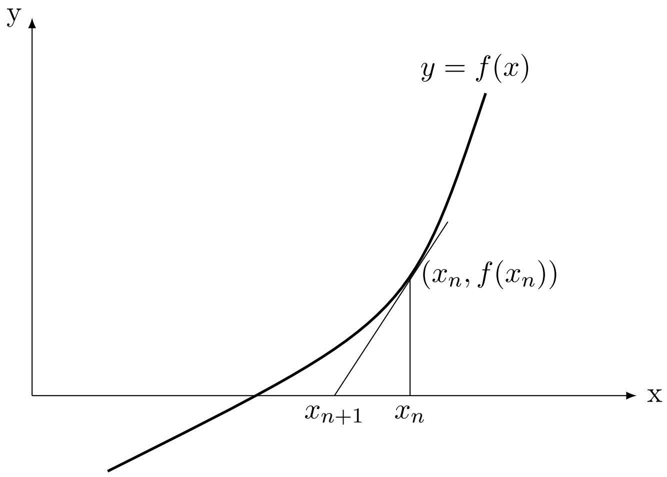
$$
\begin{aligned}
\frac{\partial \boldsymbol{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= 0 \\
\sum_i \left(y_i \boldsymbol{x_i} - \frac{e^{\boldsymbol{\beta}^T \boldsymbol{x_i}}}{1 + e^{\boldsymbol{\beta}^T \boldsymbol{x_i}}} \boldsymbol{x_i}\right) &= 0
\end{aligned}
$$

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.

5 Newton-Raphson application to Logistic Regression

To find optimal parameter β for Logistic Regression using MLE, the roots of the following equation must be found,

$$\frac{\partial \boldsymbol{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0$$

The Newton-Raphson update rule for this has the form,

$$\Delta\boldsymbol{\beta} = -\left(\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right)^{-1} \frac{\partial \boldsymbol{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$$

$$\Delta\boldsymbol{\beta} = \left[\sum_i \left(\frac{e^{\boldsymbol{\beta}^T\boldsymbol{x_i}}}{1+e^{\boldsymbol{\beta}^T\boldsymbol{x_i}}}\right)\left(\frac{1}{1+e^{\boldsymbol{\beta}^T\boldsymbol{x_i}}}\right)\boldsymbol{x_i}\boldsymbol{x_i}^T\right]^{-1}\left[\sum_i\left(y_i\boldsymbol{x_i} - \frac{e^{\boldsymbol{\beta}^T\boldsymbol{x_i}}}{1+e^{\boldsymbol{\beta}^T\boldsymbol{x_i}}}\boldsymbol{x_i}\right)\right]$$

Using the following notation

$$\boldsymbol{X} = \left(\boldsymbol{x_1}, \boldsymbol{x_2}, .., \boldsymbol{x_N}\right)$$
$$P_i = \frac{e^{\boldsymbol{\beta}^T\boldsymbol{x_i}}}{1+e^{\boldsymbol{\beta}^T\boldsymbol{x_i}}}$$
$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_N \end{pmatrix} \boldsymbol{P} = \begin{pmatrix} P_1 \\ P_2 \\ . \\ . \\ . \\ P_N \end{pmatrix}$$

the update rule can be rewritten as

$$\Delta\boldsymbol{\beta} = [\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^T]^{-1}\boldsymbol{X}(\boldsymbol{Y} - \boldsymbol{P})$$

where $\boldsymbol{W}$ is a $N \times N$ diagonal matrix of $P_i(1 - P_i)$s with N as total number of $\boldsymbol{x_i}^T, y_i$ pairs. This equation bears a surprising resemblance with the OLS solution of the problem of linear regression. In fact $\Delta\boldsymbol{\beta}$ is solutions to the following linear regression problem.

$$(\boldsymbol{Y} - \boldsymbol{P}) = \boldsymbol{W}\boldsymbol{X}^T\Delta\boldsymbol{\beta}$$

Hence, the update in each iteration of Newton-Raphson method for Logistic Regression is the solution of a weighted least square linear regression problem where the response is the difference between the observed response and estimated probability of the Logistic Regression model.

6 Interpretation

6.1 Coefficients

Rewriting the base equations for Logistic Regression

$$\begin{aligned}\mathbf{P}(\mathbf{Y} = 1 \mid x_1,\ x_2,\ ...\ x_p) &= \frac{e^{\beta_0 + \beta_1 x_1 +\ ...\ + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 +\ ...\ + \beta_p x_p}} \\ \mathbf{P}(\mathbf{Y} = 0 \mid x_1,\ x_2,\ ...\ x_p) &= \frac{1}{1 + e^{\beta_0 + \beta_1 x_1 +\ ...\ + \beta_p x_p}}\end{aligned}$$

Dividing both equations

$$\frac{\mathbf{P}(\mathbf{Y} = 1 \mid x_1,\ x_2,\ ...\ x_p)}{\mathbf{P}(\mathbf{Y} = 0 \mid x_1,\ x_2,\ ...\ x_p)} = e^{\beta_0 + \beta_1 x_1 +\ ...\ + \beta_p x_p}$$

The left side of the equation: the ratio of the probability of an event occurring to the probability of it not occurring is called the “odds” for the event. Hence, the odds of the event $\mathbf{Y} = 1$ is equal to e raised to a linear function of the input features. Every other thing remaining equal, if one of the input features is increased by one unit, the odds are multiplied by e raised to the coefficient of the input features. Positive coefficients have a multiplicative effect on the odds, but negative coefficients have a divisive effect.

6.2 Coordinate Freedom

Rewriting the MLE solution equation again,

$$\sum_i \left(y_i \boldsymbol{x_i} - \frac{e^{\boldsymbol{\beta}^T \boldsymbol{x_i}}}{1 + e^{\boldsymbol{\beta}^T \boldsymbol{x_i}}} \boldsymbol{x_i} \right) = 0$$

or

$$\sum_i (y_i \boldsymbol{x_i} - P_i \boldsymbol{x_i}) = 0$$

or

$$\boldsymbol{X}(\boldsymbol{Y} - \boldsymbol{P}) = 0$$

If $\boldsymbol{P}$ satisfies $\boldsymbol{X}(\boldsymbol{Y} - \boldsymbol{P}) = 0$, then it will also satisfy $\boldsymbol{M}\boldsymbol{X}(\boldsymbol{Y} - \boldsymbol{P}) = 0$ for non singular matrix $\boldsymbol{M}$. Since $\boldsymbol{M}\boldsymbol{X}$ is related to $\boldsymbol{X}$ by linear transformations, it is clear that linear combinations of input features like rescaling and combination will not have any effect on the probabilities estimated by a Logistic Regression model. Hence, Logistic regression is a coordinate free model.

6.3 Marginal Probability Preservation

Since the Logistic Regression model demands,

$$\sum_i (y_i \boldsymbol{x_i} - P_i \boldsymbol{x_i}) = 0$$

For every feature, the sum of the input feature values corresponding to positive response will be equal to the weighted sum of the all the input feature values where the probabilities estimated by the model serve as the weights. Hence, Logistic Regression preserves marginal probabilities of the input.