Logistic Regression

1. Model Setup

- Sample size: n
- Target variable: $\vec{Y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix}$, where $y^{(i)} \in \{0,1\} \rightarrow Y^{(i)} \sim Bernoulli(p_i)$, where $p_i = \Pr\{Y^{(i)} = 1\}$ (Y here is capitalized to reflect that it is a random variable)
- Input features: $X = \begin{bmatrix} \vec{x}^{(1)} \\ \vec{x}^{(2)} \\ \vdots \\ \vec{x}^{(n)} \end{bmatrix}$, where $\vec{x}^{(i)} = \begin{bmatrix} x_1^{(i)} & x_2^{(i)} & \dots & x_p^{(i)} \end{bmatrix}$, for $i \in \{1,2,\dots,n\}$
- Goal: For each i, estimate p_i given $\vec{x}^{(i)}$

For a given vector $\vec{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix}$, we assume the estimate of p_i, denoted as $\hat{p}(\vec{x}^{(i)}\vec{\beta})$, is calculated as follows:

$$\text{logit}\left(\hat{p}(\vec{x}^{(i)}\vec{\beta})\right) = \ln\left(\frac{\hat{p}(\vec{x}^{(i)}\vec{\beta})}{1-\hat{p}(\vec{x}^{(i)}\vec{\beta})}\right) = \vec{x}^{(i)}\vec{\beta} \Leftrightarrow \hat{p}(\vec{x}^{(i)}\vec{\beta}) = \frac{1}{1+e^{-\vec{x}^{(i)}\vec{\beta}}}$$

Note: the first term of the vector $\vec{x}^{(i)}$ is typically equal to one so that β_1 represents the intercept term

2. Parameter Estimation

2.1 Maximum Likelihood

Let $\mathcal{L}\left(\vec{\beta}|(X,\vec{Y})\right) = \Pr\{Y^{(1)} = y^{(1)}, Y^{(2)} = y^{(2)}, \dots, Y^{(n)} = y^{(n)}\}$

Assuming an independently and identically distributed sample, we have the following expressions:

$$\mathcal{L}\left(\vec{\beta}|(X,\vec{Y})\right) = \prod_{i=1}^{n} \Pr\{Y = y^{(i)}\} = \prod_{i=1}^{n} \Pr\{Y = 1\}^{y^{(i)}} \times (1 - \Pr\{Y = 1\})^{1-y^{(i)}}$$

$$= \prod_{i=1}^{n} \left(\hat{p}\left(\vec{x}^{(i)}\vec{\beta}\right)\right)^{y^{(i)}} \times \left(1 - \hat{p}\left(\vec{x}^{(i)}\vec{\beta}\right)\right)^{1-y^{(i)}}$$

$$= \prod_{i=1}^{n} \left(\frac{1}{1 + e^{-\vec{x}^{(i)}\vec{\beta}}}\right)^{y^{(i)}} \times \left(1 - \frac{1}{1 + e^{-\vec{x}^{(i)}\vec{\beta}}}\right)^{1-y^{(i)}}$$

$$= \prod_{i=1}^{n} \left(\frac{1}{1 + e^{-\vec{x}^{(i)}\vec{\beta}}}\right)^{y^{(i)}} \times \left(\frac{e^{-\vec{x}^{(i)}\vec{\beta}}}{1 + e^{-\vec{x}^{(i)}\vec{\beta}}}\right)^{1-y^{(i)}}$$

$$= \prod_{i=1}^{n} \left(\frac{e^{\vec{x}^{(i)}\vec{\beta}}}{1 + e^{\vec{x}^{(i)}\vec{\beta}}}\right)^{y^{(i)}} \times \left(\frac{1}{1 + e^{\vec{x}^{(i)}\vec{\beta}}}\right)^{1-y^{(i)}}$$

$$= \prod_{i=1}^{n} \frac{e^{y^{(i)}\vec{x}^{(i)}\vec{\beta}}}{1 + e^{\vec{x}^{(i)}\vec{\beta}}}$$

The goal is to maximize this function. This implies that if we maximize the natural logarithm of this function (which is easier to handle), we will achieve the same goal (since $\ln\left(f(.)\right)$ moves in the same direction as $f(.)$, for any function $f(.) > 0$).

Let $\ell\left(\vec{\beta}|(X,\vec{Y})\right) = \ln\left(\mathcal{L}\left(\vec{\beta}|(X,\vec{Y})\right)\right)$. Thus, we have the following expression:

$$\ell\left(\vec{\beta}|(X,\vec{Y})\right) = \sum_{i=1}^{n} y^{(i)}\vec{x}^{(i)}\vec{\beta} - \ln\left(1 + e^{\vec{x}^{(i)}\vec{\beta}}\right)$$

To obtain an optimal point, we can calculate the partial derivative of the log-likelihood function with respect to the parameter of interest and setting the derivative to zero:

For $j \in \{0,1,\dots,m\}$,

$$\frac{\partial \ell\left(\vec{\beta}|(X,\vec{Y})\right)}{\partial \beta_j} = \sum_{i=1}^{n} y^{(i)}x_j^{(i)} - \frac{x_j^{(i)}e^{\vec{x}^{(i)}\vec{\beta}}}{1 + e^{\vec{x}^{(i)}\vec{\beta}}} = \sum_{i=1}^{n} \left(y^{(i)} - \frac{e^{\vec{x}^{(i)}\vec{\beta}}}{1 + e^{\vec{x}^{(i)}\vec{\beta}}}\right) x_j^{(i)}$$

$$= \sum_{i=1}^{n} \left(y^{(i)} - \frac{1}{1 + e^{-\vec{x}^{(i)}\vec{\beta}}}\right) x_j^{(i)}$$

Thus, $\ell\left(\vec{\beta}|(X,\vec{Y})\right)$ is at an optimal value when varying β_j such that $\sum_{i=1}^{n}\left(y^{(i)} - \frac{1}{1+e^{-\vec{x}^{(i)}\vec{\beta}}}\right)x_j^{(i)} = 0$.

If we show that $\frac{\partial \ell^2\left(\vec{\beta}|(X,\vec{y})\right)}{\partial \beta_j^2} < 0$, we will establish that β_j's value that yields $\frac{\partial \ell\left(\vec{\beta}|(X,\vec{y})\right)}{\partial \beta_j} = 0$ maximizes $\ell\left(\vec{\beta}|(X,\vec{Y})\right)$. For $j \in \{1,2,\dots,m\}$,

$$\frac{\partial \ell^2\left(\vec{\beta}|(X,\vec{Y})\right)}{\partial \beta_j^2} = \frac{\partial}{\partial \beta_j}\left(\sum_{i=1}^{n}\left(y^{(i)} - \frac{1}{1+e^{-\vec{x}^{(i)}\vec{\beta}}}\right)x_j^{(i)}\right)$$

$$= \sum_{i=1}^{n}\left(-\frac{-1}{\left(1+e^{-\vec{x}^{(i)}\vec{\beta}}\right)^2} \times -x_j^{(i)}e^{-\vec{x}^{(i)}\vec{\beta}}\right)x_j^{(i)}$$

$$= -\sum_{i=1}^{n}\left(\frac{e^{-\vec{x}^{(i)}\vec{\beta}}}{\left(1+e^{-\vec{x}^{(i)}\vec{\beta}}\right)^2}\right)x_j^{(i)^2} < 0$$

To estimate $\vec{\beta}$, we will use the multi-variate Newton-Raphson procedure (we add a superscript $\{t\}$, where $t \in \{0,1,2,\dots\}$, to the variables and functions to denote that we are at iteration t):
Set $\vec{\beta}^{\{0\}}$ to random values.

$t \geq 0$: $\vec{\beta}^{\{t+1\}} = \vec{\beta}^{\{t\}} - \left[\frac{\partial \ell^2\left(\vec{\beta}^{\{t\}}|X,\vec{Y}\right)}{\partial \vec{\beta}^{\{t\}^2}}\right]^{-1} \times \frac{\partial \ell\left(\vec{\beta}^{\{t\}}|X,\vec{Y}\right)}{\partial \vec{\beta}^{\{t\}}}$

We derive $\frac{\partial \ell\left(\vec{\beta}^{\{t\}}|(X,\vec{Y})\right)}{\partial \vec{\beta}^{\{t\}}}$ and $\frac{\partial \ell^2\left(\vec{\beta}^{\{t\}}|(X,\vec{Y})\right)}{\partial \vec{\beta}^{\{t\}^2}}$ explicitly (note: when differentiating by vector, we use the so-called numerator layout notation; source: https://en.wikipedia.org/wiki/Matrix_calculus#Layout_conventions):

$$\frac{\partial \ell\left(\vec{\beta}^{\{t\}}|(X,\vec{Y})\right)}{\partial \vec{\beta}^{\{t\}}} = \left[\frac{\partial \ell\left(\vec{\beta}^{\{t\}}|(X,\vec{Y})\right)}{\partial \beta_1^{\{t\}}} \quad \frac{\partial \ell\left(\vec{\beta}^{\{t\}}|(X,\vec{Y})\right)}{\partial \beta_2^{\{t\}}} \quad \dots \quad \frac{\partial \ell\left(\vec{\beta}^{\{t\}}|(X,\vec{Y})\right)}{\partial \beta_m^{\{t\}}}\right]$$

$$= \left[\sum_{i=1}^{n}\left(y^{(i)} - \frac{1}{1+e^{-\vec{x}^{(i)}\vec{\beta}^{\{t\}}}}\right)x_1^{(i)} \quad \sum_{i=1}^{n}\left(y^{(i)} - \frac{1}{1+e^{-\vec{x}^{(i)}\vec{\beta}^{\{t\}}}}\right)x_2^{(i)} \quad \dots \quad \sum_{i=1}^{n}\left(y^{(i)} - \frac{1}{1+e^{-\vec{x}^{(i)}\vec{\beta}^{\{t\}}}}\right)x_m^{(i)}\right]$$

$$= X^T\left(\vec{y} - \vec{p}\left(X,\vec{\beta}^{\{t\}}\right)\right)$$

Where $\vec{p}\left(X,\vec{\beta}^{\{t\}}\right) = \begin{bmatrix}\hat{p}\left(\vec{x}^{(1)}\vec{\beta}^{\{t\}}\right)\\ \hat{p}\left(\vec{x}^{(2)}\vec{\beta}^{\{t\}}\right)\\ \vdots \\ \hat{p}\left(\vec{x}^{(n)}\vec{\beta}^{\{t\}}\right)\end{bmatrix}$

$$\frac{\partial \ell^2\left(\vec{\beta}^{\{t\}}|(X,\vec{Y})\right)}{\partial \vec{\beta}^{\{t\}^2}} = \frac{\partial \ell^2\left(\vec{\beta}^{\{t\}}|(X,\vec{Y})\right)}{\partial \vec{\beta}^{\{t\}}\partial \vec{\beta}^{\{t\}^T}}$$

$$= \frac{\partial}{\partial \vec{\beta}^{\{t\}^T}}\left(\frac{\partial \ell\left(\vec{\beta}^{\{t\}}|(X,\vec{Y})\right)}{\partial \vec{\beta}^{\{t\}}}\right)$$

$$= \frac{\partial}{\partial \vec{\beta}^{\{t\}^T}}\left[\frac{\partial \ell\left(\vec{\beta}^{\{t\}}|(X,\vec{Y})\right)}{\partial \beta_1^{\{t\}}} \quad \frac{\partial \ell\left(\vec{\beta}^{\{t\}}|(X,\vec{Y})\right)}{\partial \beta_2^{\{t\}}} \quad \dots \quad \frac{\partial \ell\left(\vec{\beta}^{\{t\}}|(X,\vec{Y})\right)}{\partial \beta_m^{\{t\}}}\right]$$

$$= \begin{bmatrix} \frac{\partial}{\partial \beta_1^{\{t\}}}\left(\frac{\partial \ell\left(\vec{\beta}^{\{t\}}|(X,\vec{Y})\right)}{\partial \beta_1^{\{t\}}}\right) & \frac{\partial}{\partial \beta_1^{\{t\}}}\left(\frac{\partial \ell\left(\vec{\beta}^{\{t\}}|(X,\vec{Y})\right)}{\partial \beta_2^{\{t\}}}\right) & \cdots & \frac{\partial}{\partial \beta_1^{\{t\}}}\left(\frac{\partial \ell\left(\vec{\beta}^{\{t\}}|(X,\vec{Y})\right)}{\partial \beta_m^{\{t\}}}\right) \\ \frac{\partial}{\partial \beta_2^{\{t\}}}\left(\frac{\partial \ell\left(\vec{\beta}^{\{t\}}|(X,\vec{Y})\right)}{\partial \beta_1^{\{t\}}}\right) & \frac{\partial}{\partial \beta_2^{\{t\}}}\left(\frac{\partial \ell\left(\vec{\beta}^{\{t\}}|(X,\vec{Y})\right)}{\partial \beta_2^{\{t\}}}\right) & \vdots & \frac{\partial}{\partial \beta_2^{\{t\}}}\left(\frac{\partial \ell\left(\vec{\beta}^{\{t\}}|(X,\vec{Y})\right)}{\partial \beta_m^{\{t\}}}\right) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial}{\partial \beta_m^{\{t\}}}\left(\frac{\partial \ell\left(\vec{\beta}^{\{t\}}|(X,\vec{Y})\right)}{\partial \beta_1^{\{t\}}}\right) & \frac{\partial}{\partial \beta_m^{\{t\}}}\left(\frac{\partial \ell\left(\vec{\beta}^{\{t\}}|(X,\vec{Y})\right)}{\partial \beta_2^{\{t\}}}\right) & \cdots & \frac{\partial}{\partial \beta_m^{\{t\}}}\left(\frac{\partial \ell\left(\vec{\beta}^{\{t\}}|(X,\vec{Y})\right)}{\partial \beta_m^{\{t\}}}\right) \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial \ell^2\left(\vec{\beta}^{\{t\}}|(X,\vec{Y})\right)}{\partial \beta_1^{\{t\}}\partial \beta_1^{\{t\}}} & \frac{\partial \ell^2\left(\vec{\beta}^{\{t\}}|(X,\vec{Y})\right)}{\partial \beta_1^{\{t\}}\partial \beta_2^{\{t\}}} & \cdots & \frac{\partial \ell^2\left(\vec{\beta}^{\{t\}}|(X,\vec{Y})\right)}{\partial \beta_1^{\{t\}}\partial \beta_m^{\{t\}}} \\ \frac{\partial \ell^2\left(\vec{\beta}^{\{t\}}|X,\vec{Y}\right)}{\partial \beta_2^{\{t\}}\partial \beta_1^{\{t\}}} & \frac{\partial \ell^2\left(\vec{\beta}^{\{t\}}|X,\vec{Y}\right)}{\partial \beta_2^{\{t\}}\partial \beta_2^{\{t\}}} & \vdots & \frac{\partial \ell^2\left(\vec{\beta}^{\{t\}}|(X,\vec{Y})\right)}{\partial \beta_2^{\{t\}}\partial \beta_m^{\{t\}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \ell^2\left(\vec{\beta}^{\{t\}}|(X,\vec{Y})\right)}{\partial \beta_m^{\{t\}}\partial \beta_1^{\{t\}}} & \frac{\partial \ell^2\left(\vec{\beta}^{\{t\}}|(X,\vec{Y})\right)}{\partial \beta_m^{\{t\}}\partial \beta_2^{\{t\}}} & \cdots & \frac{\partial \ell^2\left(\vec{\beta}^{\{t\}}|(X,\vec{Y})\right)}{\partial \beta_m^{\{t\}}\partial \beta_m^{\{t\}}} \end{bmatrix}, \text{ where}$$

$$\frac{\partial \ell^2\left(\vec{\beta}^{\{t\}}|(X,\vec{Y})\right)}{\partial \beta_j^{\{t\}}\partial \beta_k^{\{t\}}} = \frac{\partial}{\partial \beta_k^{\{t\}}}\left(\frac{\partial \ell\left(\vec{\beta}^{\{t\}}|(X,\vec{Y})\right)}{\partial \beta_j^{\{t\}}}\right) = \frac{\partial}{\partial \beta_k^{\{t\}}}\left(\sum_{i=1}^{n}\left(y^{(i)} - \frac{1}{1+e^{-\vec{x}^{(i)}\vec{\beta}^{\{t\}}}}\right)x_j^{(i)}\right)$$

$$= \sum_{i=1}^{n}\left(-\frac{-1}{\left(1+e^{-\vec{x}^{(i)}\vec{\beta}^{\{t\}}}\right)^2} \times -e^{-\vec{x}^{(i)}\vec{\beta}^{\{t\}}}\right)x_j^{(i)}x_k^{(i)}$$

$$= -\sum_{i=1}^{n}\left(\frac{1}{1+e^{-\vec{x}^{(i)}\vec{\beta}^{\{t\}}}} \times \frac{e^{-\vec{x}^{(i)}\vec{\beta}^{\{t\}}}}{1+e^{-\vec{x}^{(i)}\vec{\beta}^{\{t\}}}}\right)x_j^{(i)}x_k^{(i)}$$

$$= -\sum_{i=1}^{n}\left(\frac{1}{1+e^{-\vec{x}^{(i)}\vec{\beta}^{\{t\}}}} \times \left(1 - \frac{1}{1+e^{-\vec{x}^{(i)}\vec{\beta}^{\{t\}}}}\right)\right)x_j^{(i)}x_k^{(i)}$$

$$= -\sum_{i=1}^{n}\hat{p}\left(\vec{x}^{(i)}\vec{\beta}^{\{t\}}\right) \times \left(1 - \hat{p}\left(\vec{x}^{(i)}\vec{\beta}^{\{t\}}\right)\right) \times x_j^{(i)} \times x_k^{(i)}$$

In matrix form, letting $\vec{1}$ be a column vector of ones of dimension n,

$$\frac{\partial \ell^2\left(\vec{\beta}^{\{t\}}|(X,\vec{Y})\right)}{\partial \vec{\beta}^{\{t\}^2}} = -\sum_{i=1}^{n}\hat{p}\left(\vec{x}^{(i)}\vec{\beta}^{\{t\}}\right) \times \left(1 - \hat{p}\left(\vec{x}^{(i)}\vec{\beta}^{\{t\}}\right)\right) \times \vec{x}^{(i)^T} \times \vec{x}^{(i)}$$

$$= -\sum_{i=1}^{n}\hat{p}\left(\vec{x}^{(i)}\vec{\beta}^{\{t\}}\right) \times \vec{x}^{(i)^T} \times \left(1 - \hat{p}\left(\vec{x}^{(i)}\vec{\beta}^{\{n\}}\right)\right) \times \vec{x}^{(i)}$$

$$= -\sum_{i=1}^{n}\left(\hat{p}\left(\vec{x}^{(i)}\vec{\beta}^{\{t\}}\right)^T \times \vec{x}^{(i)^T}\right) \times \left(1 - \hat{p}\left(\vec{x}^{(i)}\vec{\beta}^{\{n\}}\right)\right) \times \vec{x}^{(i)}$$

$$= -\sum_{i=1}^{n} \left(\hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}^{\{t\}}\right) \times \vec{x}^{(i)}\right)^T \times \left(1 - \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}^{\{n\}}\right)\right) \times \vec{x}^{(i)}$$

$$= -\sum_{i=1}^{n} \left(\hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}^{\{t\}}\right) \times \vec{x}^{(i)}\right)^T \times \left(\left(1 - \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}^{\{t\}}\right)\right) \times \vec{x}^{(i)}\right)$$

$$= -\left(\vec{p}\left(X, \vec{\beta}^{\{t\}}\right) \circ X\right)^T \times \left(\left(\vec{1} - \vec{p}\left(X, \vec{\beta}^{\{t\}}\right)\right) \circ X\right)$$

Therefore, we have the following result for $t \geq 0$:

$$\vec{\beta}^{\{t+1\}} = \vec{\beta}^{\{t\}} - \left[\frac{\partial \ell^2\left(\vec{\beta}^{\{t\}} | (X, \vec{y})\right)}{\partial \vec{\beta}^{\{t\}^2}}\right]^{-1} \times \frac{\partial \ell\left(\vec{\beta}^{\{t\}} | (X, \vec{y})\right)}{\partial \vec{\beta}^{\{t\}}}$$

$$\Leftrightarrow$$

$$\boxed{\vec{\beta}^{\{t+1\}} = \vec{\beta}^{\{t\}} + \left[\left(\vec{p}\left(X, \vec{\beta}^{\{t\}}\right) \circ X\right)^T \times \left(\left(\vec{1} - \vec{p}\left(X, \vec{\beta}^{\{t\}}\right)\right) \circ X\right)\right]^{-1} \times X^T \left(\vec{Y} - \vec{p}\left(X, \vec{\beta}^{\{t\}}\right)\right)}$$

2.2 Gradient Descent

2.2.1 Cross-entropy Loss Function

We choose the following loss function to minimize:

$$L\left(\hat{\vec{\beta}}\right) = -\sum_{i=1}^{n} y^{(i)} \ln\left(\hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)\right) + \left(1 - y^{(i)}\right) \ln\left(1 - \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)\right)$$

We assume that X does not have a row of all zeros.

We begin by taking the first derivative and second derivatives with respect to $\hat{\beta}_j$ for $j = 1, 2, \ldots, m$ and following through with the algebra:

$$\frac{\partial L\left(\hat{\vec{\beta}}\right)}{\partial \hat{\beta}_j} = \frac{\partial\left(-\sum_{i=1}^{n} y^{(i)} \ln\left(\hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)\right) + \left(1 - y^{(i)}\right) \ln\left(1 - \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)\right)\right)}{\partial \hat{\beta}_j}$$

$$= -\sum_{i=1}^{n} y^{(i)} \frac{\partial \ln\left(\hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)\right)}{\partial \hat{\beta}_j} + \left(1 - y^{(i)}\right) \frac{\partial \ln\left(1 - \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)\right)}{\partial \hat{\beta}_j}$$

$$= -\sum_{i=1}^{n} y^{(i)} \times \frac{1}{\hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)} \times \frac{\partial \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)}{\partial \hat{\beta}_j} + \left(1 - y^{(i)}\right) \times \frac{1}{1 - \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)} \times \frac{\partial\left(1 - \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)\right)}{\partial \hat{\beta}_j}$$

$$= -\sum_{i=1}^{n} \frac{y^{(i)}}{\hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)} \times \frac{\partial \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)}{\partial \hat{\beta}_j} + \frac{\left(1 - y^{(i)}\right)}{1 - \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)} \times \frac{-\partial \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)}{\partial \hat{\beta}_j}$$

Since $\hat{p}\left(\vec{x}^{(i)}\vec{\beta}\right) = \frac{1}{1+e^{-\vec{x}^{(i)}\vec{\beta}}}$, therefore $\frac{\partial \hat{p}\left(\vec{x}^{(i)}\vec{\beta}\right)}{\partial \hat{\beta}_j} = \frac{\partial\left(\frac{1}{1+e^{-\vec{x}^{(i)}\vec{\beta}}}\right)}{\partial \hat{\beta}_j} = -\frac{1}{\left(1+e^{-\vec{x}^{(i)}\vec{\beta}}\right)^2} \times \left(e^{-\vec{x}^{(i)}\vec{\beta}}\right) \times -x_j^{(i)} = \left(\frac{1}{1+e^{-\vec{x}^{(i)}\vec{\beta}}}\right) \times \left(\frac{e^{-\vec{x}^{(i)}\vec{\beta}}}{1+e^{-\vec{x}^{(i)}\vec{\beta}}}\right) \times x_j^{(i)} = \left(\frac{1}{1+e^{-\vec{x}^{(i)}\vec{\beta}}}\right) \times \left(1 - \frac{1}{1+e^{-\vec{x}^{(i)}\vec{\beta}}}\right) \times x_j^{(i)} = \hat{p}\left(\vec{x}^{(i)}\vec{\beta}\right) \times \left(1 - \hat{p}\left(\vec{x}^{(i)}\vec{\beta}\right)\right) \times x_j^{(i)}$

Therefore,

$$\frac{\partial L\left(\hat{\vec{\beta}}\right)}{\partial \hat{\beta}_j} = -\sum_{i=1}^{n} \frac{y^{(i)}}{\hat{p}\left(\vec{x}^{(i)}\vec{\beta}\right)} \times \left(\hat{p}\left(\vec{x}^{(i)}\vec{\beta}\right) \times \left(1 - \hat{p}\left(\vec{x}^{(i)}\vec{\beta}\right)\right) \times x_j^{(i)}\right) + \frac{\left(1-y^{(i)}\right)}{1-\hat{p}\left(\vec{x}^{(i)}\vec{\beta}\right)} \times -\left(\hat{p}\left(\vec{x}^{(i)}\vec{\beta}\right) \times \left(1 - \hat{p}\left(\vec{x}^{(i)}\vec{\beta}\right)\right) \times x_j^{(i)}\right)$$

$$= -\sum_{i=1}^{n} y^{(i)} \times \left(\left(1 - \hat{p}\left(\vec{x}^{(i)}\vec{\beta}\right)\right) \times x_j^{(i)}\right) + \left(1-y^{(i)}\right) \times -\left(\hat{p}\left(\vec{x}^{(i)}\vec{\beta}\right) \times x_j^{(i)}\right)$$

$$= \sum_{i=1}^{n} -y^{(i)} \times \left(\left(1 - \hat{p}\left(\vec{x}^{(i)}\vec{\beta}\right)\right) \times x_j^{(i)}\right) + \left(1-y^{(i)}\right) \times \left(\hat{p}\left(\vec{x}^{(i)}\vec{\beta}\right) \times x_j^{(i)}\right)$$

$$= \sum_{i=1}^{n} -y^{(i)} x_j^{(i)} + y^{(i)} \hat{p}\left(\vec{x}^{(i)}\vec{\beta}\right) \times x_j^{(i)} + \hat{p}\left(\vec{x}^{(i)}\vec{\beta}\right) \times x_j^{(i)} - y^{(i)} \times \hat{p}\left(\vec{x}^{(i)}\vec{\beta}\right) \times x_j^{(i)}$$

$$= \sum_{i=1}^{n} -y^{(i)} x_j^{(i)} + \hat{p}\left(\vec{x}^{(i)}\vec{\beta}\right) \times x_j^{(i)}$$

$$= -\sum_{i=1}^{n} \left(y^{(i)} - \hat{p}\left(\vec{x}^{(i)}\vec{\beta}\right)\right) x_j^{(i)}$$

If we can show that $\frac{\partial L^2\left(\hat{\vec{\beta}}\right)}{\partial \hat{\beta}_j^2} > 0$, we will establish that $L\left(\hat{\vec{\beta}}\right)$ is a strictly convex function.

$$\frac{\partial L^2\left(\hat{\vec{\beta}}\right)}{\partial \hat{\beta}_j^2} = \frac{\partial}{\partial \hat{\beta}_j}\left(-\sum_{i=1}^{n} \left(y^{(i)} - \hat{p}\left(\vec{x}^{(i)}\vec{\beta}\right)\right) x_j^{(i)}\right)$$

$$= \sum_{i=1}^{n} \left(0 + \frac{\partial \hat{p}\left(\vec{x}^{(i)}\vec{\beta}\right)}{\partial \hat{\beta}_j}\right) x_j^{(i)}$$

$$= \sum_{i=1}^{n} \frac{\partial \hat{p}\left(\vec{x}^{(i)}\vec{\beta}\right)}{\partial \hat{\beta}_j} x_j^{(i)}$$

$$= \sum_{i=1}^{n} \left(\hat{p}\left(\vec{x}^{(i)}\vec{\beta}\right) \times \left(1 - \hat{p}\left(\vec{x}^{(i)}\vec{\beta}\right)\right) \times x_j^{(i)}\right) x_j^{(i)}$$

$$= \sum_{i=1}^{n} \hat{p}\left(\vec{x}^{(i)}\vec{\beta}\right) \times \left(1 - \hat{p}\left(\vec{x}^{(i)}\vec{\beta}\right)\right) \times {x_j^{(i)}}^2$$

$$> 0$$

Therefore, $L\left(\hat{\vec{\beta}}\right)$ is strictly convex function with a global minimum.

Let $\hat{\beta}_j^{\{t\}}$ represent the estimate of β_j for $j = 1,2,\dots,m$, at the tth iteration. Given initial values at $t = 0$, a learning rate η, we have the following expression for each j:

$$\boxed{\hat{\beta}_J^{\{t+1\}} = \hat{\beta}_J^{\{t\}} - \eta \times \frac{\partial L\left(\hat{\vec{\beta}}^{\{t\}}\right)}{\partial \hat{\beta}_J^{\{t\}}}}$$

Using the results and notation given in above, we proceed as follows:

$$\hat{\beta}_J^{\{t+1\}} = \hat{\beta}_J^{\{t\}} - \eta \times \left(-\sum_{i=1}^{n}\left(y^{(i)} - \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}^{\{t\}}\right)\right)x_j^{(i)}\right)$$

$$= \hat{\beta}_J^{\{t\}} + \eta \times \sum_{i=1}^{n}\left(y^{(i)} - \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}^{\{t\}}\right)\right)x_j^{(i)}$$

In matrix form, we have the following results:

$$\vec{\beta}^{\{t+1\}} = \vec{\beta}^{\{t\}} + \eta \times \sum_{i=1}^{n}\left(y^{(i)} - \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}^{\{t\}}\right)\right)\vec{x}^{\,(i)^T}$$

$$= \vec{\beta}^{\{t\}} + \eta \times \sum_{i=1}^{n}\vec{x}^{\,(i)^T}\left(y^{(i)} - \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}^{\{t\}}\right)\right)$$

$$= \vec{\beta}^{\{t\}} + \eta \times X^T \times \left(\vec{Y} - \vec{p}\left(X, \vec{\beta}^{\{t\}}\right)\right)$$

Thus, the final expression we can use can be given as follows:

$$\boxed{\vec{\beta}^{\{t+1\}} = \vec{\beta}^{\{t\}} + \eta X^T\left(\vec{Y} - \vec{p}\left(X, \vec{\beta}^{\{t\}}\right)\right)}$$

Note that the derivation obtained here and the Newton-Raphson derivation are very similar. In the gradient descent formula, if we let $\eta = \left[\left(\vec{p}\left(X, \vec{\beta}^{\{t\}}\right) \circ X\right)^T \times \left(\left(\vec{1} - \vec{p}\left(X, \vec{\beta}^{\{t\}}\right)\right) \circ X\right)\right]^{-1}$, then we obtain the Newton-Raphson formula.

2.2.2 Sum of Squared Error Loss Function

We choose the following loss function to minimize:

$$L\left(\hat{\vec{\beta}}\right) = \left(\vec{Y} - \vec{p}\left(X, \vec{\beta}^{\{t\}}\right)\right)^T \times \left(\vec{Y} - \vec{p}\left(X, \vec{\beta}^{\{t\}}\right)\right)$$

We assume that X does not have a row of all zeros.

We begin by taking the first derivative and second derivatives with respect to $\hat{\beta}_j$ for $j = 1,2,\dots,m$ and following through with the algebra:

$$\frac{\partial L\left(\hat{\vec{\beta}}\right)}{\partial \hat{\beta}_j} = \frac{\partial\left(\left(\vec{Y} - \vec{p}(X,\vec{\beta})\right)^T \times \left(\vec{Y} - \vec{p}(X,\vec{\beta})\right)\right)}{\partial \hat{\beta}_j}$$

$$= \frac{\partial\left(\sum_{i=1}^{n}\left(y^{(i)} - \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)\right)^2\right)}{\partial \hat{\beta}_j}$$

$$= \sum_{i=1}^{n} \frac{\partial\left(y^{(i)} - \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)\right)^2}{\partial \hat{\beta}_j}$$

$$= \sum_{i=1}^{n} 2 \times \left(y^{(i)} - \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)\right) \times \frac{\partial\left(-\hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)\right)}{\partial \hat{\beta}_j}$$

$$= \sum_{i=1}^{n} 2\left(y^{(i)} - \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)\right) \times \frac{-\partial \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)}{\partial \hat{\beta}_j}$$

$$= \sum_{i=1}^{n} 2\left(y^{(i)} - \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)\right) \times \frac{-\partial \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)}{\partial \hat{\beta}_j}$$

$$= -\sum_{i=1}^{n} 2\left(y^{(i)} - \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)\right) \times \left(\hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right) \times \left(1 - \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)\right) \times x_j^{(i)}\right)$$

$$= -2\sum_{i=1}^{n} y^{(i)}\hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)x_j^{(i)} - y^{(i)}\hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)^2 x_j^{(i)} - \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)^2 x_j^{(i)} + \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)^3 x_j^{(i)}$$

$$= -2\sum_{i=1}^{n} \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)^3 x_j^{(i)} - \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)^2\left\{y^{(i)}x_j^{(i)} + x_j^{(i)}\right\} + \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)y^{(i)}x_j^{(i)}$$

If we can show that $\frac{\partial L^2\left(\hat{\vec{\beta}}\right)}{\partial \hat{\beta}_j^2} > 0$, we will establish that $L\left(\hat{\vec{\beta}}\right)$ is a strictly convex function.

$$\frac{\partial L^2\left(\hat{\vec{\beta}}\right)}{\partial \hat{\beta}_j^2} = \frac{\partial}{\partial \hat{\beta}_j}\left(-2\sum_{i=1}^{n} \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)^3 x_j^{(i)} - \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)^2\left\{y^{(i)}x_j^{(i)} + x_j^{(i)}\right\} + \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)y^{(i)}x_j^{(i)}\right)$$

$$= -2\sum_{i=1}^{n} 3\hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)^2 \frac{\partial \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)}{\partial \hat{\beta}_j} x_j^{(i)} - 2\hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)\left\{y^{(i)}x_j^{(i)} + x_j^{(i)}\right\}\frac{\partial \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)}{\partial \hat{\beta}_j} + y^{(i)}x_j^{(i)}\frac{\partial \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)}{\partial \hat{\beta}_j}$$

$$= -2\sum_{i=1}^{n} 3\hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)^2\left(\hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)\times\left(1-\hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)\right)\times x_j^{(i)}\right)x_j^{(i)}$$
$$- 2\hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)\left\{y^{(i)}x_j^{(i)}+x_j^{(i)}\right\}\left(\hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)\times\left(1-\hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)\right)\times x_j^{(i)}\right)$$
$$+ y^{(i)}x_j^{(i)}\left(\hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)\times\left(1-\hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)\right)\times x_j^{(i)}\right)$$

$$= -2\sum_{i=1}^{n} 3\hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)^3 x_j^{(i)^2} - 3\hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)^4 x_j^{(i)^2} - 2\hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)^2\left\{y^{(i)}x_j^{(i)^2}+x_j^{(i)^2}\right\}$$
$$+ 2\hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)^3\left\{y^{(i)}x_j^{(i)^2}+x_j^{(i)^2}\right\} + \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)y^{(i)}x_j^{(i)^2} - \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)^2 y^{(i)}x_j^{(i)^2}$$

$$= \sum_{i=1}^{n} 6\hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)^4 x_j^{(i)^2} - 2\hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)^3\left(5x_j^{(i)^2} - 2y^{(i)}x_j^{(i)^2}\right) + 2\hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)^2\left(2x_j^{(i)^2} - 3y^{(i)}x_j^{(i)^2}\right)$$
$$- 2\hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)y^{(i)}x_j^{(i)^2}$$

The sign of this expression ($\frac{\partial L^2\left(\hat{\vec{\beta}}\right)}{\partial\hat{\beta}_j^2}$) will ultimately be determined by the values of $\hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}\right)$ and $y^{(i)}$, and therefore, we cannot guarantee a strictly convex function, implying that we cannot guarantee a global minimum. Also note that finding $\hat{\beta}_j$ by plugging $\frac{\partial L\left(\hat{\vec{\beta}}\right)}{\partial\hat{\beta}_j} = 0$ will likely yield several possible values further suggesting the existence of more than one optimum point. Therefore, a potential remedy is to run the model under several random initializations and pick the initialization set that yields the lowest the loss function.

Let $\hat{\beta}_j^{\{t\}}$ represent the estimate of β_j for $j = 1,2,\dots,n$, at the tth iteration. Given initial values at $t = 0$, a learning rate η, we have the following expression for each j:

$$\boxed{\hat{\beta}_j^{\{t+1\}} = \hat{\beta}_j^{\{t\}} - \eta\times\frac{\partial L\left(\hat{\vec{\beta}}^{\{t\}}\right)}{\partial\hat{\beta}_j^{\{t\}}}}$$

Using the results and notation given in above, we proceed as follows:

$$\hat{\beta}_j^{\{t+1\}} = \hat{\beta}_j^{\{t\}} - \eta$$
$$\times\left(-2\sum_{i=1}^{n}\hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}^{\{t\}}\right)^3 x_j^{(i)} - \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}^{\{t\}}\right)^2\left\{y^{(i)}x_j^{(i)}+x_j^{(i)}\right\}\right.$$
$$\left.+\,\hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}^{\{t\}}\right)y^{(i)}x_j^{(i)}\right)$$

$$= \hat{\beta}_j^{\{t\}} + 2\eta\times\sum_{i=1}^{n}\hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}^{\{t\}}\right)^3 x_j^{(i)} - \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}^{\{t\}}\right)^2\left\{y^{(i)}x_j^{(i)}+x_j^{(i)}\right\} + \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}^{\{t\}}\right)y^{(i)}x_j^{(i)}$$

If all η to absorb the constant 2, we can simply replace 2η with η, and get

$$\widehat{\beta_j}^{\{t+1\}} = \widehat{\beta_j}^{\{t\}} + \eta \times \sum_{i=1}^{n} \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}^{\{t\}}\right)^3 x_j^{(i)} - \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}^{\{t\}}\right)^2 \left\{y^{(i)}x_j^{(i)} + x_j^{(i)}\right\} + \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}^{\{t\}}\right) y^{(i)} x_j^{(i)}$$

$$= \widehat{\beta_j}^{\{t\}} + \eta \times \sum_{i=1}^{n} \left(\hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}^{\{t\}}\right)^3 - \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}^{\{t\}}\right)^2 \left\{y^{(i)} + 1\right\} + \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}^{\{t\}}\right) y^{(i)}\right) x_j^{(i)}$$

In matrix form, we have the following results:

$$\vec{\beta}^{\{t+1\}} = \vec{\beta}^{\{t\}} + \eta \times \sum_{i=1}^{n} \left(\hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}^{\{t\}}\right)^3 - \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}^{\{t\}}\right)^2 \left\{y^{(i)} + 1\right\} + \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}^{\{t\}}\right) y^{(i)}\right) \vec{x}^{\,(i)^T}$$

$$= \vec{\beta}^{\{t\}} + \eta \times \sum_{i=1}^{n} \vec{x}^{\,(i)^T} \left(\hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}^{\{t\}}\right)^3 - \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}^{\{t\}}\right)^2 \left\{y^{(i)} + 1\right\} + \hat{p}\left(\vec{x}^{\,(i)}\vec{\beta}^{\{t\}}\right) y^{(i)}\right)$$

$$= \vec{\beta}^{\{t\}} + \eta \times X^T \times \left(\vec{p}\left(X, \vec{\beta}^{\{t\}}\right) \circ \left(\left(\vec{1} - \vec{p}\left(X, \vec{\beta}^{\{t\}}\right)\right) \circ \vec{Y} - \left(\vec{1} - \vec{p}\left(X, \vec{\beta}^{\{t\}}\right)\right) \circ \vec{p}\left(X, \vec{\beta}^{\{t\}}\right)\right)\right)$$

$$= \vec{\beta}^{\{t\}} + \eta \times X^T \times \left(\vec{p}\left(X, \vec{\beta}^{\{t\}}\right) \circ \left(\vec{1} - \vec{p}\left(X, \vec{\beta}^{\{t\}}\right)\right) \circ \left(\vec{Y} - \vec{p}\left(X, \vec{\beta}^{\{t\}}\right)\right)\right)$$

$$= \vec{\beta}^{\{t\}} + \eta \times X^T \times \left(\left(\vec{Y} - \vec{p}\left(X, \vec{\beta}^{\{t\}}\right)\right) \circ \vec{p}\left(X, \vec{\beta}^{\{t\}}\right) \circ \left(\vec{1} - \vec{p}\left(X, \vec{\beta}^{\{t\}}\right)\right)\right)$$

Thus, the final expression we can use can be given as follows:

$$\boxed{\vec{\beta}^{\{t+1\}} = \vec{\beta}^{\{t\}} + \eta X^T \left(\left(\vec{Y} - \vec{p}\left(X, \vec{\beta}^{\{t\}}\right)\right) \circ \vec{p}\left(X, \vec{\beta}^{\{t\}}\right) \circ \left(\vec{1} - \vec{p}\left(X, \vec{\beta}^{\{t\}}\right)\right)\right)}$$

2.3 Standardization Comment

Due to the use of exponentiation in the terms above, it is possible to run into numerical instability issues if the values of X are either large or slow. Thus, it is recommended to standardize the training dataset for each feature (with the exception of the intercept term). For example, for each column vector of the matrix X, denoted as $X[:, j]$, with $j = 1,2, \ldots, m$, we can standardize via the following formula:

$$X[:, j] = \frac{X[:, j] - \mu(X[:, j])}{std(X[:, j])},$$

where $\mu(X[:, j]) = \frac{1}{n}\sum_{i=1}^{n} x_{i,j}$, and $std(X[:, j]) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{i,j} - \mu(X[:, j])\right)^2}$.