Chapter 1

Machine Learning - By Dorival Pedroso

Note: This chapter does *not* use the summation convention on repeated indices.

1.1 Linear Regression

Given $m = nSamples$ data points and $n = nFeatures$ features, the matrix $\boldsymbol{X}$ organises the data along rows such that

$$\boldsymbol{X} = \begin{bmatrix} 1 & X_{00} & X_{01} & \cdots & X_{0n} \\ 1 & X_{10} & X_{11} & \cdots & X_{1n} \\ 1 & X_{20} & X_{21} & \cdots & X_{2n} \\ 1 & X_{30} & X_{31} & \cdots & X_{3n} \\ 1 & X_{40} & X_{41} & \cdots & X_{4n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{\mu 0} & X_{\mu 1} & \cdots & X_{\mu n} \end{bmatrix}_{(nSamples \times nFeatures+1)} \tag{1.1}$$

where the columns from the second column correspond to each feature and $\mu = m - 1$. For example, X_{ij} is the value of data point i and feature j.

The vector of parameters is expressed as

$$\boldsymbol{\theta} = \begin{Bmatrix} \theta_0 \\ \theta_1 \\ \cdots \\ \theta_n \end{Bmatrix} \tag{1.2}$$

Thus, a linear regression applied to data point i results in

$$\ell_i(\boldsymbol{\theta}) = \sum_{j=0}^{n} X_{ij}\,\theta_j \qquad \text{or} \qquad \boldsymbol{\ell}(\boldsymbol{\theta}) = \boldsymbol{X}\,\boldsymbol{\theta} \tag{1.3}$$

and

$$\frac{\partial \ell_i}{\partial \theta_j} = X_{ij} \qquad \text{or} \qquad \frac{\mathrm{d}\boldsymbol{\ell}}{\mathrm{d}\boldsymbol{\theta}} = \boldsymbol{X} \tag{1.4}$$

An error vector is defined by

$$\boldsymbol{e}(\boldsymbol{\theta}) = \boldsymbol{\ell}(\boldsymbol{\theta}) - \boldsymbol{y} \tag{1.5}$$

and the cost function by

$$C(\boldsymbol{\theta}) = \frac{1}{2\,m}\sum_{i=0}^{\mu}(\ell_i - y_i)^2 = \frac{1}{2\,m}\,\boldsymbol{e}^T\boldsymbol{e} \tag{1.6}$$

thus

$$\frac{\partial C}{\partial \theta_j} = \frac{1}{m}\sum_{i=0}^{\mu}(\ell_i - y_i)\,\frac{\partial \ell_i}{\partial \theta_j} = \frac{1}{m}\sum_{i=0}^{\mu} e_i\,X_{ij} \tag{1.7}$$

or

$$\frac{\mathrm{d}C}{\mathrm{d}\boldsymbol{\theta}} = \frac{1}{m}\boldsymbol{X}^T\boldsymbol{e} = \frac{1}{m}\boldsymbol{X}^T[\boldsymbol{\ell}(\boldsymbol{\theta}) - \boldsymbol{y}] \tag{1.8}$$

The minimum cost corresponds to

$$\frac{\mathrm{d}C}{\mathrm{d}\boldsymbol{\theta}} = 0 \qquad \text{or} \qquad \boldsymbol{X}^T\boldsymbol{X}\,\boldsymbol{\theta} - \boldsymbol{X}^T\boldsymbol{y} = 0 \tag{1.9}$$

Therefore,

$$\boldsymbol{\theta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\,\boldsymbol{X}^T\boldsymbol{y} \tag{1.10}$$

The gradient-descent update is

$$\boldsymbol{\theta} := \boldsymbol{\theta} - \alpha\,\frac{\mathrm{d}C}{\mathrm{d}\boldsymbol{\theta}} \tag{1.11}$$

1.1.1 External bias parameter plus regularization

The previous θ_0 is now called b and the $\boldsymbol{\theta}$ contains one less component with the other indices being decreased by one. $\boldsymbol{X}$ also contains one less column. The data matrix now is

$$\boldsymbol{X} = \begin{bmatrix} X_{00} & X_{01} & \cdots & X_{0\nu} \\ X_{10} & X_{11} & \cdots & X_{1\nu} \\ X_{20} & X_{21} & \cdots & X_{2\nu} \\ X_{30} & X_{31} & \cdots & X_{3\nu} \\ X_{40} & X_{41} & \cdots & X_{4\nu} \\ \vdots & \vdots & \ddots & \vdots \\ X_{\mu 0} & X_{\mu 1} & \cdots & X_{\mu\nu} \end{bmatrix}_{(nSamples \times nFeatures)} \tag{1.12}$$

where $\nu = n - 1$ (number of features minus one) with $\mu = m - 1$ still being the number of samples minus one.

Let's define the vector "one" of length equal to the number of samples as

$$o_i = 1 \qquad \text{thus} \qquad \boldsymbol{o} = \left[1\ 1\ 1\ \dots\ 1\right]^T \tag{1.13}$$

Thus, the linear model can be written as

$$\ell_i(\boldsymbol{\theta}, b) = b\, o_i + \sum_{j=0}^{\nu} X_{ij}\, \theta_j \tag{1.14}$$

In vector notation

$$\boxed{\boldsymbol{\ell}(\boldsymbol{\theta}, b) = b\, \boldsymbol{o} + \boldsymbol{X}\, \boldsymbol{\theta}}$$

We also define an error vector with all sample data as

$$e_i(\boldsymbol{\theta}, b) = \ell_i(\boldsymbol{\theta}, b) - y_i \tag{1.15}$$

or

$$\boxed{\boldsymbol{e} = b\, \boldsymbol{o} + \boldsymbol{X}\, \boldsymbol{\theta} - \boldsymbol{y}}$$

thus

$$\frac{\partial e_i}{\partial \theta_j} = \frac{\partial \ell_i}{\partial \theta_j} = X_{ij} \tag{1.16}$$

and

$$\frac{\partial e_i}{\partial b} = \frac{\partial \ell_i}{\partial b} = o_i \tag{1.17}$$

The regularized cost function is given by (with $l_i = l_i(\boldsymbol{\theta}, b)$)

$$C(\boldsymbol{\theta}, b) = \frac{1}{2\, m} \sum_{i=0}^{\mu} (\ell_i - y_i)^2 + \frac{\lambda}{2\, m} \sum_{i=0}^{\mu} \theta_i^2 \tag{1.18}$$

or, in vector notation,

$$\boxed{C(\boldsymbol{\theta}, b) = \frac{1}{2\, m} \left(\boldsymbol{e}^T \boldsymbol{e} + \lambda\, \boldsymbol{\theta}^T \boldsymbol{\theta} \right)}$$

The gradient of C is calculated by two parts. The first part is

$$\frac{\partial C}{\partial \theta_j} = \frac{1}{m} \sum_{i=0}^{\mu} (\ell_i - y_i)\, \frac{\partial \ell_i}{\partial \theta_j} + \frac{\lambda}{m} \sum_{i=0}^{\mu} \theta_i\, \delta_{ij} \tag{1.19}$$

or

$$\frac{\partial C}{\partial \theta_j} = \frac{1}{m} \sum_{i=0}^{\mu} e_i \, X_{ij} + \frac{\lambda}{m} \theta_j \tag{1.20}$$

similarly

$$\frac{\partial C}{\partial \theta_j} = \frac{1}{m} \sum_{i=0}^{\mu} X_{ji}^T \, e_i + \frac{\lambda}{m} \theta_j \tag{1.21}$$

or, in vector notation,

$$\boxed{\frac{\partial C}{\partial \boldsymbol{\theta}} = \frac{1}{m} \boldsymbol{X}^T \boldsymbol{e} + \frac{\lambda}{m} \boldsymbol{\theta}}$$

The second part is

$$\frac{\partial C}{\partial b} = \frac{1}{m} \sum_{i=0}^{\mu} (\ell_i - y_i) \, \frac{\partial \ell_i}{\partial b} = \frac{1}{m} \sum_{i=0}^{\mu} \, e_i \, o_i \tag{1.22}$$

or,

$$\boxed{\frac{\partial C}{\partial b} = \frac{1}{m} \boldsymbol{o}^T \boldsymbol{e}}$$

Let's define the vector

$$s_j = \sum_{i=0}^{\mu} o_i \, X_{ij} \equiv \mathrm{sum} \left(\mathrm{cols} \left(\boldsymbol{X} \right) \right) \tag{1.23}$$

and the scalar

$$t = \sum_{i=0}^{\mu} o_i \, y_i \equiv \mathrm{sum} \left(\mathrm{cols} \left(\boldsymbol{y} \right) \right) \tag{1.24}$$

In vector notation

$$\boldsymbol{s} = \boldsymbol{X}^T \boldsymbol{o} = \begin{Bmatrix} \sum_i^{\mu} X_{i0} \\ \sum_i^{\mu} X_{i1} \\ \cdots \\ \sum_i^{\mu} X_{i\nu} \end{Bmatrix} \tag{1.25}$$

and

$$t = \boldsymbol{o}^T \boldsymbol{y} \tag{1.26}$$

Note that

$$\boldsymbol{s}^T = (\boldsymbol{X}^T \boldsymbol{o})^T = \boldsymbol{o}^T \boldsymbol{X} = \left[\sum_i^{\mu} X_{i0} \;\; \sum_i^{\mu} X_{i1} \;\; \cdots \;\; \sum_i^{\mu} X_{i\nu} \right] \tag{1.27}$$

Let's further define the vector

$$\boldsymbol{a} = \boldsymbol{X}^T \boldsymbol{y} \tag{1.28}$$

and the following matrices

$$\boldsymbol{A} = \boldsymbol{X}^T \boldsymbol{X} \tag{1.29}$$

and

$$B_{ij} = \frac{1}{m}\, s_i\, s_j \qquad \text{hence} \qquad \boldsymbol{B} = \frac{1}{m}\, \boldsymbol{s}\, \boldsymbol{s}^T \tag{1.30}$$

Note that these three quantities $\boldsymbol{a}$, $\boldsymbol{A}$ and $\boldsymbol{B}$ can be directly computed from the input data.

By expanding the gradient expressions, we get

$$\begin{aligned} \frac{\partial C}{\partial \boldsymbol{\theta}} &= \frac{1}{m}\left(b\, \boldsymbol{X}^T \boldsymbol{o} + \boldsymbol{X}^T \boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{X}^T \boldsymbol{y} + \lambda\, \boldsymbol{\theta}\right) \\ &= \frac{1}{m}\left(b\, \boldsymbol{s} + \boldsymbol{A}\, \boldsymbol{\theta} - \boldsymbol{a} + \lambda\, \boldsymbol{\theta}\right) \end{aligned} \tag{1.31}$$

and

$$\begin{aligned} \frac{\partial C}{\partial b} &= \frac{1}{m}\left(b\, \boldsymbol{o}^T \boldsymbol{o} + \boldsymbol{o}^T \boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{o}^T\, \boldsymbol{y}\right) \\ &= \frac{1}{m}\left(b\, m + \boldsymbol{s}^T \boldsymbol{\theta} - t\right) \end{aligned} \tag{1.32}$$

The minimum is found by zeroing both partial derivatives and solving the following system for $\boldsymbol{\theta}$ and b

$$\begin{aligned} b\, \boldsymbol{s} + \boldsymbol{A}\, \boldsymbol{\theta} - \boldsymbol{a} + \lambda\, \boldsymbol{\theta} &= 0 \\ b\, m + \boldsymbol{s}^T \boldsymbol{\theta} - t &= 0 \end{aligned} \tag{1.33}$$

From the second equation we have

$$\boxed{b = \frac{t}{m} - \frac{1}{m}\boldsymbol{s}^T \boldsymbol{\theta}}$$

By substituting this result into the first equation, we obtain

$$\begin{aligned} \left(\frac{t}{m} - \frac{1}{m}\boldsymbol{s}^T \boldsymbol{\theta}\right)\, \boldsymbol{s} + (\boldsymbol{A} + \lambda \boldsymbol{I})\, \boldsymbol{\theta} &= \boldsymbol{a} \\ -\frac{1}{m}\boldsymbol{s}\, (\boldsymbol{s}^T \boldsymbol{\theta}) + (\boldsymbol{A} + \lambda \boldsymbol{I})\, \boldsymbol{\theta} &= \boldsymbol{a} - \frac{t}{m}\boldsymbol{s} \\ -\frac{1}{m}(\boldsymbol{s}\, \boldsymbol{s}^T)\, \boldsymbol{\theta} + (\boldsymbol{A} + \lambda \boldsymbol{I})\, \boldsymbol{\theta} &= \boldsymbol{a} - \frac{t}{m}\boldsymbol{s} \end{aligned} \tag{1.34}$$

Therefore, the linear system to be solved is

$$\boxed{(\boldsymbol{A} - \boldsymbol{B} + \lambda \boldsymbol{I})\ \boldsymbol{\theta} = \boldsymbol{r}}$$

where

$$\boxed{\boldsymbol{r} = \boldsymbol{a} - \frac{t}{m}\boldsymbol{s}}$$

After the computation of $\boldsymbol{\theta}$, b can be easily found.

1.2 Logistic Regression

The Logistic function is given by

$$g(z) = \frac{1}{1 + \mathrm{e}^{-z}} \tag{1.35}$$

and the Logistic regression applied to each data point i is

$$h_i(\boldsymbol{\theta}) = g(\ell_i(\boldsymbol{\theta})) \qquad \text{or} \qquad h_i(\boldsymbol{\theta}) = \frac{1}{1 + \mathrm{e}^{-\sum_j X_{ij}\,\theta_j}} \tag{1.36}$$

where the summation is indicated in Eq. (1.3).

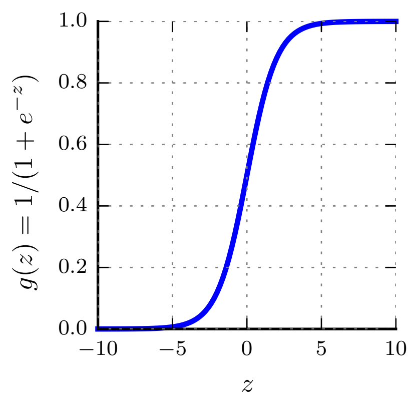


Fig. 1.1: Logistic function

Let's define p_i as

$$p_i(\boldsymbol{\theta}) = 1 + \mathrm{e}^{-\ell_i(\boldsymbol{\theta})} \qquad \text{hence} \qquad h_i = (p_i)^{-1} \tag{1.37}$$

Thus (considering Eq. 1.4)

$$\frac{\partial p_i}{\partial \theta_j} = -\mathrm{e}^{-\ell_i} \frac{\partial \ell_i}{\partial \theta_j} = -\mathrm{e}^{-\ell_i} X_{ij} \tag{1.38}$$

Let's define q_i as

$$q_i(\boldsymbol{\theta}) = \log\left[p_i(\boldsymbol{\theta})\right] \tag{1.39}$$

Thus

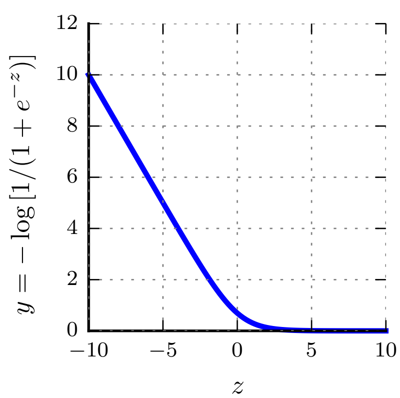
$$\frac{\partial q_i}{\partial \theta_j} = \frac{1}{p_i} \frac{\partial p_i}{\partial \theta_j} = \frac{-\mathrm{e}^{-\ell_i}}{1 + \mathrm{e}^{-\ell_i}} X_{ij} \tag{1.40}$$

Note that

$$\log h_i = \log\left(\frac{1}{p_i}\right) = -\log p_i = -q_i \tag{1.41}$$

Note also that

$$\log\left(1 - h_i\right) = \log\left(\frac{p_i}{p_i} - \frac{1}{p_i}\right) = \underbrace{\log\left(p_i - 1\right)}_{-\ell_i} - \log p_i = -\ell_i - q_i \tag{1.42}$$



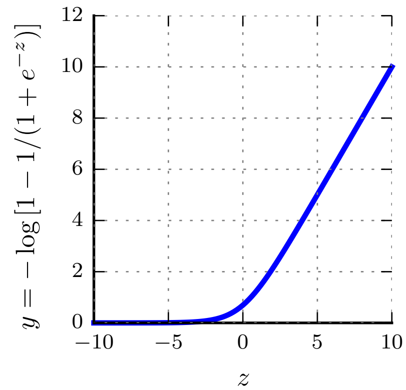
The cost function is defined as

$$C(\boldsymbol{\theta}) = \frac{1}{m} \sum_{i=0}^{\mu} c_i(\boldsymbol{\theta}) \tag{1.43}$$

where

$$\begin{aligned} c_i(\boldsymbol{\theta}) &= -y_i \log\left[h_i(\boldsymbol{\theta})\right] - (1 - y_i) \log\left[1 - h_i(\boldsymbol{\theta})\right] \\ &= y_i\, q_i(\boldsymbol{\theta}) + (1 - y_i)\left[\ell_i(\boldsymbol{\theta}) + q_i(\boldsymbol{\theta})\right] \\ &= \cancel{y_i\, q_i} + \ell_i + q_i - y_i\, \ell_i - \cancel{y_i\, q_i} \end{aligned} \tag{1.44}$$

or



$$c_i(\boldsymbol{\theta}) = q_i(\boldsymbol{\theta}) + (1 - y_i)\,\ell_i(\boldsymbol{\theta}) \tag{1.45}$$

The cost function can hence be written as

$$C(\boldsymbol{\theta}) = \underbrace{\frac{1}{m}\sum_{i=0}^{\mu} q_i(\boldsymbol{\theta})}_{s_q} + \sum_{i=0}^{\mu} \underbrace{\frac{1 - y_i}{m}}_{\bar{y}_i}\,\ell_i(\boldsymbol{\theta}) \tag{1.46}$$

or

$$C(\boldsymbol{\theta}) = s_q + \bar{\boldsymbol{y}}^T \boldsymbol{\ell} \tag{1.47}$$

The derivative of c_i is

$$\begin{aligned}
\frac{\partial c_i}{\partial \theta_j} &= \frac{\partial q_i}{\partial \theta_j} + (1 - y_i)\,\frac{\partial \ell_i}{\partial \theta_j} \\
&= \left(\frac{-\mathrm{e}^{-\ell_i}}{1 + \mathrm{e}^{-\ell_i}} + 1 - y_i\right) X_{ij} \\
&= \left(\frac{-\mathrm{e}^{-\ell_i}}{1 + \mathrm{e}^{-\ell_i}} + \frac{1 + \mathrm{e}^{-\ell_i}}{1 + \mathrm{e}^{-\ell_i}} - y_i\right) X_{ij} \\
&= \left(\frac{1}{1 + \mathrm{e}^{-\ell_i}} - y_i\right) X_{ij}
\end{aligned} \tag{1.48}$$

or

$$\frac{\partial c_i}{\partial \theta_j} = (h_i(\boldsymbol{\theta}) - y_i)\,X_{ij} \tag{1.49}$$

The derivative of the cost function is

$$\frac{\partial C}{\partial \theta_j} = \frac{1}{m}\sum_{i=0}^{\mu}(h_i(\boldsymbol{\theta}) - y_i)\,X_{ij} \tag{1.50}$$

Therefore

$$\frac{\mathrm{d}C}{\mathrm{d}\boldsymbol{\theta}} = \frac{1}{m} (\boldsymbol{h} - \boldsymbol{y}) \boldsymbol{X} \tag{1.51}$$

or

$$\frac{\mathrm{d}C}{\mathrm{d}\boldsymbol{\theta}} = \frac{1}{m} \boldsymbol{X}^T (\boldsymbol{h} - \boldsymbol{y}) \tag{1.52}$$

1.2.1 External bias parameter plus regularization

The logistic regression model with $\boldsymbol{\theta}$ and b not in the same vector is

$$h_i(\boldsymbol{\theta}, b) = g(\ell_i(\boldsymbol{\theta}, b)) \tag{1.53}$$

where ℓ_i is given by Eq. (1.14).

The auxiliary vectors are

$$p_i(\boldsymbol{\theta}, b) = 1 + \mathrm{e}^{-\ell_i(\boldsymbol{\theta}, b)} \tag{1.54}$$

and

$$q_i(\boldsymbol{\theta}, b) = \log\left[p_i(\boldsymbol{\theta}, b)\right] \tag{1.55}$$

Thus, the cost is still given by the same expression (Eq. 1.43) with

$$c_i(\boldsymbol{\theta}, b) = q_i(\boldsymbol{\theta}, b) + (1 - y_i)\, \ell_i(\boldsymbol{\theta}, b) \tag{1.56}$$

Therefore, now adding regularization, we obtain

$$\boxed{C(\boldsymbol{\theta}, b) = s_q(\boldsymbol{\theta}, b) + \bar{\boldsymbol{y}}^T \boldsymbol{\ell}(\boldsymbol{\theta}, b) + \frac{\lambda}{2\, m} \boldsymbol{\theta}^T \boldsymbol{\theta}}$$

with

$$s_q(\boldsymbol{\theta}, b) = \frac{1}{m} \operatorname{sum}(\boldsymbol{q}) \qquad \text{and} \qquad \bar{y}_i = \frac{1 - y_i}{m} \tag{1.57}$$

The derivative of p_i with respect to θ_j is (no change)

$$\frac{\partial p_i}{\partial \theta_j} = -\mathrm{e}^{-\ell_i} \frac{\partial \ell_i}{\partial \theta_j} = -\mathrm{e}^{-\ell_i} X_{ij} \tag{1.58}$$

and the derivative of q_i with respect to θ_j is (no change)

$$\frac{\partial q_i}{\partial \theta_j} = \frac{1}{p_i} \frac{\partial p_i}{\partial \theta_j} = \frac{-\mathrm{e}^{-\ell_i}}{1 + \mathrm{e}^{-\ell_i}} X_{ij} = -\frac{1}{1 + \mathrm{e}^{\ell_i}} X_{ij} \tag{1.59}$$

The derivative of c_i with respect to θ_j is (no change)

$$\frac{\partial c_i}{\partial \theta_j} = (h_i(\boldsymbol{\theta}, b) - y_i)\, X_{ij} \tag{1.60}$$

Therefore, the partial derivative of the cost function with respect to $\boldsymbol{\theta}$ is (no change, except for the regularization term)

$$\boxed{\frac{\partial C}{\partial \boldsymbol{\theta}} = \frac{1}{m} \boldsymbol{X}^T \left(\boldsymbol{h}(\boldsymbol{\theta}, b) - \boldsymbol{y}\right) + \frac{\lambda}{m}\boldsymbol{\theta}}$$

The derivative of p_i with respect to b is

$$\frac{\partial p_i}{\partial b} = -\mathrm{e}^{-\ell_i} \frac{\partial \ell_i}{\partial b} = -\mathrm{e}^{-\ell_i}\, o_i \tag{1.61}$$

and the derivative of q_i with respect to b is

$$\frac{\partial q_i}{\partial b} = \frac{1}{p_i}\frac{\partial p_i}{\partial b} = \frac{-\mathrm{e}^{-\ell_i}}{1+\mathrm{e}^{-\ell_i}}\, o_i = -\frac{1}{1+\mathrm{e}^{\ell_i}}\, o_i \tag{1.62}$$

The derivative of c_i with respect to b is

$$\begin{aligned}
\frac{\partial c_i}{\partial b} &= \frac{\partial q_i}{\partial b} + (1 - y_i)\frac{\partial \ell_i}{\partial b} \\
&= \left(\frac{-\mathrm{e}^{-\ell_i}}{1+\mathrm{e}^{-\ell_i}} + 1 - y_i\right) o_i \\
&= \left(\frac{-\mathrm{e}^{-\ell_i}}{1+\mathrm{e}^{-\ell_i}} + \frac{1+\mathrm{e}^{-\ell_i}}{1+\mathrm{e}^{-\ell_i}} - y_i\right) o_i \\
&= \left(\frac{1}{1+\mathrm{e}^{-\ell_i}} - y_i\right) o_i
\end{aligned} \tag{1.63}$$

hence

$$\frac{\partial c_i}{\partial b} = (h_i(\boldsymbol{\theta}, b) - y_i)\, o_i \tag{1.64}$$

Therefore, the partial derivative of the cost function with respect to b is

$$\boxed{\frac{\partial C}{\partial b} = \frac{1}{m} \boldsymbol{o}^T \left(\boldsymbol{h}(\boldsymbol{\theta}, b) - \boldsymbol{y}\right)}$$

1.2.2 Hessian matrix

The first derivative of the cost function with respect to θ_i can be written as

$$\frac{\partial C}{\partial \theta_i} = \frac{1}{m} \sum_{k=0}^{\mu} X_{ik}^T \, e_k + \frac{\lambda}{m} \, \theta_i \tag{1.65}$$

where

$$e_i(\boldsymbol{\theta}, b) = h_i(\boldsymbol{\theta}, b) - y_i \tag{1.66}$$

Thus, the Hessian matrix with respect to $\boldsymbol{\theta}$ is

$$H_{ij} = \frac{\partial^2 C}{\partial \theta_i \partial \theta_j} = \frac{1}{m} \sum_{k=0}^{\mu} X_{ik}^T \, \frac{\partial e_k}{\partial \theta_j} + \frac{\lambda}{m} \, \delta_{ij} \tag{1.67}$$

The following derivative is required

$$\frac{\partial e_k}{\partial \theta_j} = \frac{\partial h_k}{\partial \theta_j} = \frac{\partial (p_k)^{-1}}{\partial \theta_j} = -\frac{1}{p_k^2} \frac{\partial p_k}{\partial \theta_j} = \frac{\mathrm{e}^{-\ell_k}}{p_k^2} \, X_{kj} \tag{1.68}$$

We can show that

$$\boxed{d_i \equiv \frac{\mathrm{e}^{-\ell_i}}{p_i^2} = g(\ell_i) \, [1 - g(\ell_i)]}$$

thus

$$\frac{\partial e_k}{\partial \theta_j} = d_k \, X_{kj} \tag{1.69}$$

Therefore, we obtain

$$H_{ij} = \frac{\partial^2 C}{\partial \theta_i \partial \theta_j} = \frac{1}{m} \sum_{k=0}^{\mu} X_{ik}^T \, d_k \, X_{kj} + \frac{\lambda}{m} \, \delta_{ij} \tag{1.70}$$

or

$$\boxed{\boldsymbol{H} = \frac{1}{m} \boldsymbol{X}^T \boldsymbol{D} + \frac{\lambda}{m} \boldsymbol{I}}$$

where the following matrix is defined

$$\boxed{D_{ij} = d_i \, X_{ij}}$$

We require also the cross derivative vector

$$v_i = \frac{\partial^2 C}{\partial \theta_i \partial b} = \frac{1}{m} \sum_{k=0}^{\mu} X_{ik}^T \, \frac{\partial e_k}{\partial b} \tag{1.71}$$

The following derivative is required

$$\frac{\partial e_k}{\partial b} = \frac{\partial h_k}{\partial b} = \frac{\partial (p_k)^{-1}}{\partial b} = -\frac{1}{p_k^2}\,\frac{\partial p_k}{\partial b} = \frac{\mathrm{e}^{-\ell_k}}{p_k^2} = d_k \tag{1.72}$$

or

$$v_i = \frac{\partial^2 C}{\partial \theta_i \partial b} = \frac{1}{m}\sum_{k=0}^{\mu} X_{ik}^T\, d_k \tag{1.73}$$

In vector notation

$$\boxed{\boldsymbol{v} = \frac{1}{m}\,\boldsymbol{X}^T\,\boldsymbol{d}}$$

The last Hessian term is

$$\frac{\partial^2 C}{\partial b^2} = \frac{1}{m}\sum_{i=0}^{\mu} \frac{\partial h_i}{\partial b}\, o_i = \frac{1}{m}\sum_{i=0}^{\mu} d_i\, o_i \tag{1.74}$$

or

$$\boxed{\frac{\partial^2 C}{\partial b^2} = \frac{1}{m}\boldsymbol{o}^T\boldsymbol{d}}$$