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 data points and n features, the matrix $\boldsymbol{X}$ organises the data along rows such that

$$\boldsymbol{X} = \begin{bmatrix} 1 & X_{00} & X_{01} & \cdots & X_{0n} \\ 1 & X_{10} & X_{11} & \cdots & X_{1n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{\mu 0} & X_{\mu 1} & \cdots & X_{\mu n} \end{bmatrix} \tag{1.1}$$

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,

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

1.2 Logistic Regression

The Logistic function is given by

$$g(z) = \frac{1}{1 + \mathrm{e}^{-z}} \tag{1.11}$$

and the Logistic regression applied to each data point i is

$$h_i(\boldsymbol{\theta}) = g(\ell_i(\boldsymbol{\theta})) \qquad \text{or} \qquad h_i(\boldsymbol{\theta}) = \frac{1}{1 + \mathrm{e}^{-\sum_j X_{ij}\, \theta_j}} \tag{1.12}$$

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

Let's define p_i as

$$p_i(\boldsymbol{\theta}) = 1 + \mathrm{e}^{-\ell_i(\boldsymbol{\theta})} \qquad \text{hence} \qquad h_i = (p_i)^{-1} \tag{1.13}$$

Thus (considering Eq. 1.4)

$$\frac{\partial p_i}{\partial \theta_j} = -\mathrm{e}^{-\ell_i}\, \frac{\partial \ell_i}{\partial \theta_j} = -\mathrm{e}^{-\ell_i}\, X_{ij} \tag{1.14}$$

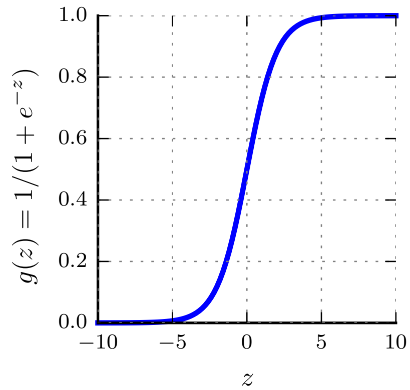


Fig. 1.1: Logistic function

Let's define q_i as

$$q_i(\boldsymbol{\theta}) = \log\left[p_i(\boldsymbol{\theta})\right] \tag{1.15}$$

Thus

$$\frac{\partial q_i}{\partial \theta_j} = \frac{1}{p_i}\frac{\partial p_i}{\partial \theta_j} = \frac{-\mathrm{e}^{-\ell_i}}{1+\mathrm{e}^{-\ell_i}}\, X_{ij} \tag{1.16}$$

Note that

$$\log h_i = \log\left(\frac{1}{p_i}\right) = -\log p_i = -q_i \tag{1.17}$$

Note also that

$$\log\left(1-h_i\right) = \log\left(\frac{p_i}{p_i} - \frac{1}{p_i}\right) = \underbrace{\log\left(p_i - 1\right)}_{-\ell_i} - \log p_i = -\ell_i - q_i \tag{1.18}$$

The cost function is defined as

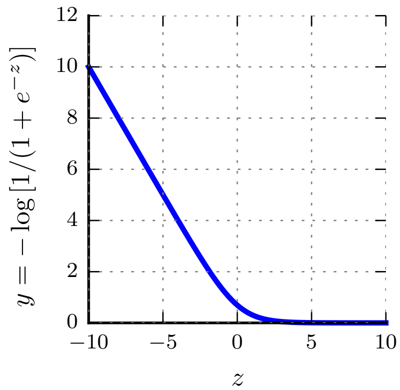
$$C(\boldsymbol{\theta}) = \frac{1}{m}\sum_{i=0}^{\mu} c_i(\boldsymbol{\theta}) \tag{1.19}$$

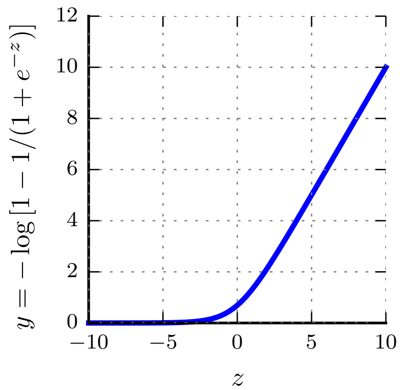
where

$$\begin{aligned} c_i(\boldsymbol{\theta}) &= -y_i\, \log\left[h_i(\boldsymbol{\theta})\right] - (1-y_i)\, \log\left[1-h_i(\boldsymbol{\theta})\right] \\ &= y_i\, q_i(\boldsymbol{\theta}) + (1-y_i)\left[\ell_i(\boldsymbol{\theta}) + q_i(\boldsymbol{\theta})\right] \\ &= \cancel{y_i\, q_i} + \ell_i + q_i - y_i\,\ell_i - \cancel{y_i\, q_i} \end{aligned} \tag{1.20}$$

or

$$c_i(\boldsymbol{\theta}) = q_i(\boldsymbol{\theta}) + (1-y_i)\,\ell_i(\boldsymbol{\theta}) \tag{1.21}$$





The cost function can hence be written as

$$C(\boldsymbol{\theta}) = \underbrace{\frac{1}{m}\sum_{i=0}^{\mu} q_i(\boldsymbol{\theta})}_{s_q} + \sum_{i=0}^{\mu} \underbrace{\frac{1-y_i}{m}}_{\bar{y}_i} \ell_i(\boldsymbol{\theta}) \tag{1.22}$$

or

$$C(\boldsymbol{\theta}) = s_q + \bar{\boldsymbol{y}}^T \boldsymbol{\ell} \tag{1.23}$$

The derivative of c_i is

$$
\begin{aligned}
\frac{\partial c_i}{\partial \theta_j} &= \frac{\partial q_i}{\partial \theta_j} + (1 - y_i)\, \frac{\partial \ell_i}{\partial \theta_j} \\
&= \left(\frac{-\mathrm{e}^{-\ell_i}}{1 + \mathrm{e}^{-\ell_i}} + 1 - y_i \right) X_{ij} \\
&= \left(\frac{-\mathrm{e}^{-\ell_i}}{1 + \mathrm{e}^{-\ell_i}} + \frac{1 + \mathrm{e}^{-\ell_i}}{1 + \mathrm{e}^{-\ell_i}} - y_i \right) X_{ij} \\
&= \left(\frac{1}{1 + \mathrm{e}^{-\ell_i}} - y_i \right) X_{ij} \qquad (1.24)
\end{aligned}
$$

or

$$\frac{\partial c_i}{\partial \theta_j} = (h_i(\boldsymbol{\theta}) - y_i)\, X_{ij} \qquad (1.25)$$

The derivative of the cost function is

$$\frac{\partial C}{\partial \theta_j} = \frac{1}{m} \sum_{i=0}^{\mu} (h_i(\boldsymbol{\theta}) - y_i)\, X_{ij} \qquad (1.26)$$

Therefore

$$\frac{\mathrm{d}C}{\mathrm{d}\boldsymbol{\theta}} = \frac{1}{m}\, (\boldsymbol{h} - \boldsymbol{y})\, \boldsymbol{X} \quad \text{or} \quad \frac{\mathrm{d}C}{\mathrm{d}\boldsymbol{\theta}} = \frac{1}{m}\, \boldsymbol{X}^T\, (\boldsymbol{h} - \boldsymbol{y}) \qquad (1.27)$$

1.3 Gradient descent

$$\boldsymbol{\theta} := \boldsymbol{\theta} - \alpha\, \frac{\mathrm{d}C}{\mathrm{d}\boldsymbol{\theta}} \qquad (1.28)$$