Machine Learning

Bhaswar Chakma

2021-07-06

Contents

5

1 OLS **7**

1.1 SLR 7

Source: Google ML Comic

Chapter 1

OLS

1.1 SLR

We define our column vector of parameters $\mathbf{w}$ (β in econometrics):

$$\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

A single observation $\mathbf{x_n}$ is:

$$\mathbf{x_n} = \begin{bmatrix} 1 \\ x_n \end{bmatrix}$$

And thus the value predicted by the model can be written as:

$$f(x_n; w_0, w_1) = w_0 + w_1 x_n = \begin{bmatrix} w_0 & w_1 \end{bmatrix} \begin{bmatrix} 1 \\ x_1 \end{bmatrix} = \mathbf{w}^T \mathbf{x_n}$$

The matrix representing all the input data $\mathbf{X}$ and the column vector of outcomes $\mathbf{t}$ ($\mathbf{y}$ in econometrics) are:

$$\mathbf{X} = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix}$$

$$\mathbf{t} = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{bmatrix}$$

The loss function $\mathcal{L}$ becomes

$$\mathcal{L} = \frac{1}{N}(\mathbf{t} - \mathbf{X})^T(\mathbf{t} - \mathbf{X})$$

The column vector of values predicted by the function $f(x_n; w_0, w_1)$ are:

$$\mathbf{Xw} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix} \times \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} w_0 & w_1 x_1 \\ w_0 & w_1 x_2 \\ \vdots & \vdots \\ w_0 & w_1 x_N \end{bmatrix}$$

$$\underset{(N\times 2)}{\mathbf{X}} \times \underset{(2\times 1)}{\mathbf{w}} = \underset{(N\times 1)}{\mathbf{Xw}}$$

The column vector of residual is

$$\mathbf{t} - \mathbf{Xw} = \begin{bmatrix} t_1 - w_0 + w_1 x_1 \\ t_2 - w_0 + w_1 x_2 \\ \vdots \\ t_N - w_0 + w_1 x_N \end{bmatrix}$$

The sum of squared errors can be expressed as:

$$\sum_{n=1}^{N}(t_n - (w_0 + w_1 x_1))^2$$

$$= \begin{bmatrix} t_1 - (w_0 + w_1 x_1) & t_2 - (w_0 + w_1 x_2) & \dots & t_N - (w_0 + w_1 x_N) \end{bmatrix} \times \begin{bmatrix} t_1 - (w_0 + w_1 x_1) \\ t_2 - (w_0 + w_1 x_2) \\ \vdots \\ t_N - (w_0 + w_1 x_N) \end{bmatrix} = (\mathbf{t} - \mathbf{Xw})^T$$

We can now write our loss function compactly with matrix notation, which produces a scalar (1×1) value:

$$\underset{(1\times 1)}{\frac{1}{N}} \underset{(1\times N)}{(\mathbf{t} - \mathbf{Xw})^T} \underset{(N\times 1)}{(\mathbf{t} - \mathbf{Xw})} = \underset{(1\times 1)}{\mathcal{L}}$$

We can expand the sum of squared errors as follows:

$$(\mathbf{t} - \mathbf{Xw})^T(\mathbf{t} - \mathbf{Xw})$$

$$= (\mathbf{t}^T - (\mathbf{Xw})^T)(\mathbf{t} - \mathbf{Xw})$$

$$= (\mathbf{t}^T - (\mathbf{w}^T \mathbf{X}^T))(\mathbf{t} - \mathbf{X}\mathbf{w})$$

$$= \underset{(1\times1)}{\mathbf{t}^T\mathbf{t}} - \underset{(1\times N)}{\mathbf{t}^T} \underset{(N\times2)}{\mathbf{X}} \underset{(2\times1)}{\mathbf{w}} - \underset{(1\times2)}{\mathbf{w}^T} \underset{(2\times N)}{\mathbf{X}^T} \underset{(N\times1)}{\mathbf{t}} + \underset{(1\times2)}{\mathbf{w}^T} \underset{(2\times N)}{\mathbf{X}^T} \underset{(N\times2)}{\mathbf{X}} \underset{(2\times1)}{\mathbf{w}} \tag{1.1}$$

Note the following:

- each of the term becomes 1×1
- the terms $\mathbf{t}^T\mathbf{X}\mathbf{w}$ and $\mathbf{w}^T\mathbf{X}^T\mathbf{t}$ are transposes of each other and are scalars. So they are the same value and can be combined.

So, sum of squared residuals can be written as

$$\mathbf{t}^T\mathbf{t} - 2\mathbf{w}^T\mathbf{X}^T\mathbf{t} + \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}$$

Hence, the loss function becomes:

$$\mathcal{L} = \frac{1}{N}\mathbf{t}^T\mathbf{t} - 2\mathbf{w}^T\mathbf{X}^T\mathbf{t} + \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}$$