1 Linear Regression

Model

A simple linear regression predicts the output as a linear function of the input feature x:

$$f(x; w_0; w_1) = w_0 + w_1 x$$

We refer to w_0 and w_1 as the parameters of the model. To choose w_0 and w_1, we are given a data set of previous input-output measurements:

$$\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(N)}, y^{(N)})\}$$

or

$$\{(x^{(N)}, y^{(N)})\}_{n=1}^{N}$$

Loss Function

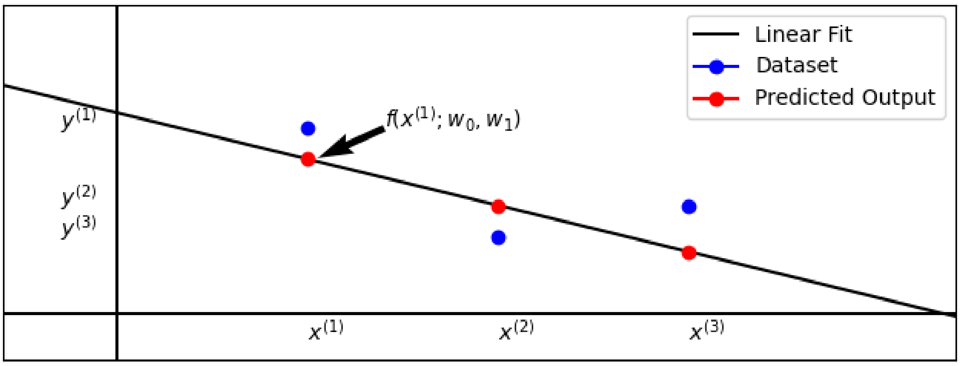

Figura 1: Loss function and its Prediction

the fit is adjusted using the "squared loss"or the "residual squares"(RSS)

$$J(w_0, w_1) = \sum_{n=1}^{N} \left(y^{(n)} - f(x^{(n)}; w_0; w_1)\right)^2$$

Optimization Want to find w_0, w_1 to minimise $J(w_0, w_1)$, $\hat{w_0}, \hat{w_1} = argmin_{w_0,w_1} J(w_0, w_1)$
strategy: Set $\frac{\partial J}{\partial w_0} = 0$ and $\frac{\partial J}{\partial w_1} = 0$

$$J(w_0, w_1) = \sum_{n=1}^{N} \left(y^{(n)} - f(x^{(n)}; w_0, w_1)\right)^2 = \sum_{n=1}^{N} \left(y^{(n)} - (w_0 + w_1 x^{(n)})\right)^2$$

$$\frac{\partial J}{\partial w_0} = \sum_{n=1}^{N} \frac{\partial}{\partial w_0}\left(y^{(n)} - f(x^{(n)}; w_0, w_1)\right)^2 = \sum_{n=1}^{N} 2\left(y^{(n)} - w_0 - w_1 x^{(n)}\right)(-1) = 0$$

$$\iff \sum_{n=1}^{N} w_0 = \sum_{n=1}^{N} y^{(n)} - w_1 \sum_{n=1}^{N} x^{(n)} \iff N w_0 = \sum_{n=1}^{N} y^{(n)} - w_1 \sum_{n=1}^{N} x^{(n)}$$

$$\iff w_0 = \frac{1}{N}\sum_{n=1}^{N} y^{(n)} - w_1 \frac{1}{N}\sum_{n=1}^{N} x^{(n)} \implies \hat{w}_0 = \overline{y} - w_1 \cdot \overline{x}$$

Hat used to indicate particular value, bar used to indicate estimated mean, the same accours to w_1

$$\frac{\partial J}{\partial w_1} = \sum_{n=1}^{N} \frac{\partial}{\partial w_0}\left(y^{(n)} - f(x^{(n)}; w_0, w_1)\right)^2 = \sum_{n=1}^{N} 2\left(y^{(n)} - \overline{y} - w_1(x^{(n)} - \overline{x})\right)(-1)(x^{(n)} - \overline{x}) = 0$$

$$\implies \hat{w_1} = \frac{\sum\limits_{n=1}^{N} (y^{(n)} - \overline{y})(x^{(n)} - \overline{x})}{\sum\limits_{n=1}^{N} \left(x^{(n)} - \overline{x}\right)^2}$$

These parameters settings are called the "least squared estimates".

For multiple linear regression,

$$f(x_1, x_2, \ldots, x_d; w_0, w_1, \ldots, w_d) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_D x_D = f(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T \mathbf{x}$$

$$\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_D \end{bmatrix} \qquad and \qquad \mathbf{x} = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_D \end{bmatrix}$$

Loss function

Squared loss:

$$\begin{aligned} J(\mathbf{w}) &= \sum_{n=1}^{N} (y^{(n)} - f(\mathbf{x}^{(n)}; \mathbf{w}))^2 \\ &= \sum_{n=1}^{N} (y^{(n)} - (w_0 + w_1 x_1^{(n)} + \cdots + w_D x_D^{(n)}))^2 \end{aligned}$$

Derivating, $\frac{\partial J}{\partial w_0}, \frac{\partial J}{\partial w_1}, \ldots, \frac{\partial J}{\partial w_D}$ and set equal to zero, $\frac{\partial J}{\partial \mathbf{w}} = 0$, getting the followoing result

$$\hat{\mathbf{w}} = (\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T \mathbf{y}$$

This is called the normal equation.