LINEAR REGRESSION

Data:

$\{(x_i, y_i)\}_{i=0,\dots,N-1}$, $x_i = (x_{i0}, x_{i1}, \dots, x_{i,M-1}) \in \mathbb{R}^{1\times M}$

$y_i \in \mathbb{R}$

Model:

$$\Phi(X; w, b) = Xw + b$$

where

- $X \in \mathbb{R}^{1\times M}$ is a random feature vector
- $w \in \mathbb{R}^{1\times M}$ is a weight vector
- $b \in \mathbb{R}$ is a bias

Objective:

$$\min_{w,b} MSE(w, b)$$

where

$$MSE(w, b) = \frac{1}{N} \sum_{i=0}^{N-1} \left(\Phi(x_i; w, b) - y_i \right)^2$$

Let us impose:

$$\begin{cases} \partial_{w_j} MSE(w,b) = 0 & , j = 0, \dots, M-1 \\ \partial_b MSE(w,b) = 0 \end{cases}$$

$$MSE(w,b) = \frac{1}{N} \sum_{i=0}^{N-1} \left(\sum_{j=0}^{M-1} x_{ij} w_j + b - y_i \right)^2$$

$$\begin{cases} \partial_{w_j} MSE(w,b) = \frac{2}{N} \sum_{i=0}^{N-1} x_{ij} (\Phi(x_i) - y_i) & j = 0, \dots, M-1 \\ \partial_b MSE(w,b) = \frac{2}{N} \sum_{i=0}^{N-1} (\Phi(x_i) - y_i) \end{cases}$$

$\Rightarrow$

$$\begin{cases} \frac{1}{N} \sum_{i=0}^{N-1} \left(\sum_{k=0}^{M-1} (x_{ij} x_{ik} w_k) + x_{ij} b - x_{ij} y_i \right) = 0, \; j = 0, \dots, M-1 \\ \frac{1}{N} \sum_{i=0}^{N-1} \left(\sum_{k=0}^{M-1} (x_{ik} w_k) + b - y_i \right) = 0 \end{cases}$$

$$\Rightarrow \begin{cases} \sum_{k=0}^{M-1} \left(\frac{1}{N} \sum_{i=0}^{N-1} x_{ij} x_{ik} \right) w_k + \left(\frac{1}{N} \sum_{i=0}^{N-1} x_{ij} \right) b = \frac{1}{N} \sum_{i=0}^{N-1} x_{ij} y_i \\ \sum_{k=0}^{M-1} \left(\frac{1}{N} \sum_{i=0}^{N-1} x_{ik} \right) w_k + b = \frac{1}{N} \sum_{i=0}^{N-1} y_i \end{cases}$$

This can be written as:

$$A \begin{pmatrix} w \\ b \end{pmatrix} = c$$

where

$$A = \frac{1}{N} \tilde{x}^T \tilde{x} \quad , \quad \tilde{x} = (x \mid 1) = \begin{pmatrix} x_{00} & \cdots & x_{0,M-1} & 1 \\ x_{10} & \cdots & x_{1,M-1} & 1 \\ \vdots & \ddots & \vdots & \\ x_{N-1,0} & \cdots & x_{N-1,M-1} & 1 \end{pmatrix}$$

$$= \frac{1}{N} \begin{pmatrix} x_{00} & \cdots & x_{N-1,0} \\ x_{01} & \cdots & x_{N-1,1} \\ \vdots & \ddots & \vdots \\ x_{0,M-1} & \cdots & x_{N-1,M-1} \\ 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} x_{00} & \cdots & x_{0,M-1} & 1 \\ x_{10} & \cdots & x_{1,M-1} & 1 \\ \vdots & \ddots & \vdots & \\ x_{N-1,0} & \cdots & x_{N-1,M-1} & 1 \end{pmatrix}$$

$$= \frac{1}{N} \begin{pmatrix} \sum_i x_{i0} x_{i0} & \sum_i x_{i0} x_{i1} & \cdots & \sum_i x_{i0} x_{i,M-1} & \sum_i x_{i0} \\ \sum_i x_{i1} x_{i0} & \sum_i x_{i1} x_{i1} & \cdots & \sum_i x_{i1} x_{i,M-1} & \sum_i x_{i1} \\ \vdots & & & & \\ \sum_i x_{i,M-1} x_{i0} & \sum_i x_{i,M-1} x_{i1} & \cdots & \sum_i x_{i,M-1} x_{i,M-1} & \sum_i x_{i,M-1} \\ \sum_i x_{i0} & \sum_i x_{i1} & \cdots & \sum_i x_{i,M-1} & 1 \end{pmatrix}$$

and

$$c = \frac{1}{N} \tilde{X}^T y = \frac{1}{N} \begin{pmatrix} x_{00} & \cdots & x_{N-1,0} \\ x_{01} & \cdots & x_{N-1,1} \\ \vdots & \ddots & \vdots \\ x_{0,M-1} & & x_{N-1,M-1} \\ 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_{N-1} \end{pmatrix}$$

[See Hastie, Tibshirani, Friedman
"the elements of statistical learning"]