Exercise 3.1: Linear Regression

(a)

Derivation of the normal equation

Given,

$$f(w) = \frac{1}{N} \sum_{i=1}^{N} \left(w^T x - y\right)^2$$

Writing in matrix form,

$$f(w) = \frac{1}{N} \sum_{i=1}^{N} \|Xw - y\|^2$$

Taking the gradient of $f(w)$ with respect to w,

$$\nabla_w f(w) = \frac{2}{N} X^T (Xw - y)$$

Setting $\nabla_w f(w) = 0$,

$$0 = \frac{2}{N} X^T Xw - \frac{2}{N} X^T y$$

$$\Rightarrow X^T Xw = X^T y$$

$$\Rightarrow w = (X^T X)^{-1} X^T y$$

(i) The normal equation is prone to overfit the data, which causes problems with generalization.

Moreover, in the case of a large dataset, it can be very expensive in terms of both computation and memory.

(ii) When X is not invertible, it can still work by using the Moore-Penrose Pseudoinverse,

$$w = X^+ y$$

where

$$X^+ = \lim_{\alpha \to 0} \left(X^T X + \alpha I\right)^{-1} X^T$$

(b)

Given,

$$x = \begin{bmatrix} 9.0 & 2.0 & 6.0 & 1.0 & 8.0 \end{bmatrix}$$

$$y = \begin{bmatrix} 1.0 & 0.0 & 3.0 & 0.0 & 1.0 \end{bmatrix}$$

The normal equation is

$$w = \left(X^T X\right)^{-1} X^T y$$

Calculating $X^T X$,

$$\begin{aligned} X^T X &= 9^2 + 2^2 + 6^2 + 1^2 + 8^2 \\ &= 186 \end{aligned}$$

Calculating $X^T y$,

$$\begin{aligned} X^T y &= (9 \times 1) + (2 \times 0) + (6 \times 3) + (1 \times 0) + (8 \times 1) \\ &= 35 \end{aligned}$$

Finally,

$$\begin{aligned} w &= \frac{X^T y}{X^T X} \\ &= \frac{35}{186} \\ &= 0.188 \end{aligned}$$

(c)

In the mentioned case, $w_1 \neq w_2$ and $b_1 \neq b_2$ holds.

It is given that $w_1 y \neq \eta$, which suggests that the shift in input and output is not proportional; therefore, w_1 and w_2 will be different.

Moreover, if y in D_2 is shifted by η, that means b_2 will adjust accordingly, making $b_1 \neq b_2$.

Exercise 3.2: PCA I

(a) Normalization in PCA ensures equal contribution of all the features regardless of their original value. It makes sure no specific feature is unfairly dominating over others.

(b) PCA performs badly when the data has nonlinearity and complexity because PCA assumes linearity in variance maximization.

Exercise 3.3: PCA II

(a) The dataset has 4 features x_1, x_2, x_3, and x_4, which is why there are 4 principal components, because PCA reduces dimension based on the number of features. For 4 features, there can exist 4 principal components maximum.

(b) Variance preserved by the first two components,

$$= \frac{\mathrm{PC}_1 + \mathrm{PC}_2}{\text{total variance}}$$

$$= \frac{0.739 + 0.685}{0.739 + 0.685 + 0.239 + 0.084}$$

$$= 0.85 \text{ or } 85\%$$

(c) Variance preserved by the first and third principal components,

$$= \frac{\mathrm{PC}_1 + \mathrm{PC}_3}{\text{total variance}}$$

$$= \frac{0.739 + 0.239}{0.739 + 0.685 + 0.239 + 0.004}$$

$$= 0.58 \text{ or } 58\%$$