Ex 3.1 (a)

(i) Correlation $[\rho_{xy}] = \dfrac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum (x_i-\bar{x})^2 (y_i-\bar{y})^2}}$

$$= \frac{\dfrac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{n^2}}}{\sqrt{\dfrac{\sum (x_i-\bar{x})^2}{n} \dfrac{\sum (y_i-\bar{y})^2}{n}}}$$

$$\sigma_x^2 = \frac{\sum (x_i-\bar{x})^2}{n} \Rightarrow \boxed{\sigma_x^2 = \frac{\sum x_i^2}{n}}$$

$$\rho_{xy} = \frac{\dfrac{\sum x_i y_i}{n}}{\sigma_x \sigma_y}$$

$$\Rightarrow \sum x_i y_i = n \sigma_x \sigma_y \rho_{xy}$$

$$\boxed{a = (X^T X)^{-1} X^T y}$$

$$X^T X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}$$

$$= \sum x_i^2 = n \sigma_x^2$$

$$(X^T X)^{-1} = \frac{1}{n \sigma_x^2}$$

$$(X^T Y) = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix}$$

$$= \sum x_i y_i$$

$$= n \sigma_x \sigma_y \rho_{xy}$$

$$\Rightarrow a = \frac{n \sigma_x \sigma_y \rho_{xy}}{n \sigma_x^2}$$

$$\boxed{a = \frac{\sigma_y \rho_{xy}}{\sigma_x}}$$

(ii) Given $X' = sX$ and $Y' = tY$

$\Rightarrow$ if $\bar{X} = \bar{Y} = 0$

$\Rightarrow \overline{X'} = \overline{Y'} = 0$

$$\sigma_{x'} = \sqrt{\frac{\sum x_i'^2}{n}} = s\sqrt{\frac{\sum x_i^2}{n}} = s\sigma_x$$

Similarly $\rightarrow \sigma_{y'} = t\sigma_y$

$$\rho_{x'y'} = \frac{\sum x_i' y_i'}{n\sigma_{x'}\sigma_{y'}}$$

$$= \frac{\cancel{s}\cancel{t}\sum x_i y_i}{n\cancel{s}\sigma_x \cancel{t}\sigma_y}$$

$$= \frac{\sum x_i y_i}{n\sigma_x\sigma_y}$$

$$= \rho_{xy}$$

$$\Rightarrow \boxed{\rho_{x'y'} = \rho_{xy}}$$

Here $a' = \dfrac{\sigma_{y'} \rho_{x'y'}}{\sigma_{x'}}$

$$= \frac{t\sigma_y \rho_{xy}}{s\sigma_x}$$

$$= \frac{t}{s} \frac{\sigma_y \rho_{xy}}{\sigma_x}$$

$$\Rightarrow \boxed{a' = \frac{t}{s} a}$$

b) MSE $f(x, y : w) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle w, x_i \rangle)^2$

Adding the noise term:

$$f(x, y : w) = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} (y_i - \langle w, x_i \rangle)^2 - 2\epsilon \cdot w_i (y_i - \langle w, x_i \rangle) + \sum_{i=1}^{n} w_i \epsilon_i \epsilon_j \sum_{i=1}^{d} w_i \right]$$

Applying linearity of expection:

$$= \frac{1}{n} \sum_{i=1}^{n} \left((y_i - \langle w, x_i \rangle)^2 - 2\mathbb{E}[\epsilon] w_i (y_i - \langle w, x_i \rangle) + \mathbb{E}[\epsilon_i \epsilon_j] \sum_{i=1}^{d} w_i^2 \right)$$

Given $\mathbb{E}[\epsilon] = 0$ and $\mathbb{E}[\epsilon_i \epsilon_j] = \sigma^2$

$$f(w) = \frac{1}{n} \sum_{i=1}^{n} \left((y_i - \langle w, x_i \rangle)_{i=1}^{2} + \sigma^2 \sum_{i=1}^{d} w_i^2 \right)$$

Therefore we can write :

$$\mathbb{E}\left[\left(y_i - \langle w, x_i + \epsilon_i \rangle\right)^2\right] =$$

$$\mathbb{E}\left[\left(y_i - \langle w, x_i \rangle\right)^2\right] + \sigma^2 \sum_{i=1}^{d} w_i^2$$

Ex 3.3

a) The softmax function is susceptible to two vulnerabilities.

Case I : When very small numbers are passed through softmax, it is rounded up to zero ; this is known as underflo- -wing.

Case II : When the numbers are very large, Softmax approximations are interpreted as infinity. This is called overflowing.

b) The overflowing problem of Softmax can be solved by subtracting the largest numbers and shifting all the inputs. So the equation becomes:

$$\exp(x_i) - \max(x)$$

$$\text{Softmax}(x)_i = \frac{\exp(x_i) - \max(x)}{\sum_{j=1}^{n} \exp(x_j) - \max(x)}$$

It can be shown by exponential division rule that both equations are same while solving the numerical issues of Softmax.

c) For a function to be a valid probability distribution it needs to satisfy 3 attributes:

(i) The random variable should be associated with numerical.

(ii) Sum of probabilities should be 1

(iii) Each probability should be in range of 0 ~ 1

As we know exp function returns values from $0 \sim +\infty$, the result of $\exp(x_i)$ will always return a positive value. The same applies for the denominator. The denominator normalizes the value of

$\exp(x_i)$ dividing by the sum of all $\exp(x_j)$ for all possible n. Since the probabilities are a ration of these two values, they'll always sum up to 1 as the largest value will always return the largest probability. No matter how large the upper value is the denominator will always normalize it and turn into a value between $0 \sim 1$. Thus it will always return a valid probabity distribution for all values of $n \in \mathbb{R}$.

d) Jacobian matrix of Softmax:

$$S_i = \frac{e^{x_i}}{\sum_j e^{x_j}}$$

$$J_{S_i} = \begin{bmatrix} \frac{\partial s_1}{\partial x_1} & \frac{\partial s_1}{\partial x_2} & \cdots & \frac{\partial s_1}{\partial x_n} \\ \frac{\partial s_2}{\partial x_1} & \frac{\partial s_2}{\partial x_2} & \ddots & \vdots \end{bmatrix}$$

$$\begin{bmatrix} \vdots & \vdots & \ddots & \vdots \\ \frac{\partial s_n}{\partial x_1} & \frac{\partial s_n}{\partial x_2} & \cdots & \frac{\partial s_n}{\partial x_n} \end{bmatrix}$$

For easier computation we are considering the log of softmax,

$$\log s_i = \log\left(\frac{e^{x_i}}{\sum_j e^{x_j}}\right)$$

$$= x_i - \log\left(\sum_j^n e^{x_j}\right)$$

For any k th value;

$$\frac{\partial}{\partial x_k}(\log s_i) = \frac{\partial x_i}{\partial x_k} - \frac{\partial}{\partial x_k}\log\left(\sum_j e^{x_j}\right)$$

here,

$$\frac{\partial x_i}{\partial x_k} = \begin{cases} 1 & ; \quad i = k \\ 0 & ; \quad i \neq k \end{cases}$$

$$\frac{\partial}{\partial x_k}(\log s_i) = \mathbb{1}\{i=k\} - \frac{1}{\sum_j e^{x_j}} \cdot \left(\frac{\partial}{\partial x_j} e^{x_j}\right)$$

$$= 1\{i=j\} - \frac{e^{x_j}}{\sum_k e^{x_k}} \qquad \left[\frac{d}{dx}\log(x) = \frac{1}{x}\right]$$

Therefore, we can write

$$\frac{\partial}{\partial x_k} \log(s_i) = 1\{i=k\} - s_k$$

$$\frac{1}{s_i} \cdot \frac{\partial s_i}{\partial x_k} = 1\{i=k\} - s_k$$

$$\frac{\partial s_i}{\partial x_k} = s_i\,(1\{i=k\} - s_k)$$

Plugging this formula into Jacobian format we get:

$$J_s = \begin{bmatrix} s_1 \cdot (1-s_1) & -s_1 \cdot s_2 & -s_1 \cdot s_2 \cdots & -s_1 \cdot s_n \\ -s_2 \cdot s_1 & s_2 \cdot (1-s_2) & \cdots\cdots & -s_2 \cdot s_4 \\ \vdots & \vdots & \ddots & \vdots \\ -s_n \cdot s_1 & -s_n \cdot s_2 & \cdots\cdots & s_n \cdot (1-s_n) \end{bmatrix}$$

Ex 3.2 (a)

$$f(x,y;w) = \frac{1}{n} ||y - Xw||^2 + \lambda ||w||^2$$

$$f(x,y;w) = \frac{1}{n}(y - Xw)^T(y - Xw) + \lambda w^T w$$

for best w^*, we use:

$$\frac{d f(x,y;w)}{dw} = 0$$

$$\Rightarrow -\frac{\cancel{2}X^T}{n}(y - Xw) + \cancel{2}\lambda w = 0$$

$$\Rightarrow X^T(y - Xw) = n\lambda w$$

[We write $n\lambda$ as λ]

$$\Rightarrow X^T y - X^T X w = \lambda w$$

$$\Rightarrow (X^T X + \lambda I) w = X^T y$$

$$\Rightarrow \boxed{w^* = (X^T X + \lambda I)^{-1} X^T y}$$