4.1) Loss for whole dataset -

$$\text{Loss} = \frac{1}{N} \sum_{i=1}^{N} y_i \log(p_i) - (1-y_i)\log(1-p_i)$$

Here $$p_i = \frac{1}{1+e^{-w^T x_i - b}}$$

2) The new labels $y'_i \in \{-1, +1\}$

we try changing the labels in a way that for $y'_i \rightarrow$

-1 becomes 0 & $+1$ stays $+1$

If we apply the below function on y'_i the labels change to the required label form

$$f(y) = \frac{1}{2}(y+1)$$

$\therefore$ for $y = 1$; $f(y) = 1$

$y = -1$; $f(y) = 0$

Thus we can now use the same cost function we derived in the previous part; but with $f(y'_i)$ instead of y_i

$$\text{Loss} = \frac{1}{N}\sum_{i=1}^{N} f(y_i')\log(p_i) - (1-f(y_i'))\log(1-p_i)$$

$$= \frac{1}{N}\sum_{i=1}^{N} \frac{1}{2}(y_i'+1)\log(p_i) - \left(1-\frac{1}{2}(y_i'+1)\right)\log(1-p_i)$$

$$= \frac{1}{2N}\sum_{i=1}^{N} (1+y_i')\log(p_i) - (1-y_i')\log(1-p_i)$$

where $p_i = \dfrac{1}{1+e^{-(w^T x+b)}}$

Ex 4.2

$$f(x) = \frac{1}{1 + \exp(-(x_1^2 + x_2^2))}$$

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix}$$

$$\frac{\partial f}{\partial x_1} = \frac{\partial}{\partial x_1} \left(1 + \exp(-(x_1^2 + x_2^2))\right)^{-1}$$

$$= -1\left(1 + \exp(-x_1^2 - x_2^2)\right)^{-2} \cdot \exp(-(x_1^2 + x_2^2)) \cdot \frac{\partial}{\partial x_1}\left(-(x_1^2 + x_2^2)\right)$$

$$= \left(1 + \exp(-x_1^2 - x_2^2)\right)^{-2} \cdot \exp(-x_1^2 - x_2^2) \cdot 2x_1$$

$$\frac{\partial f}{\partial x_2} = \left(1 + \exp(-x_1^2 - x_2^2)\right)^{-2} \cdot \exp(-x_1^2 - x_2^2) \cdot 2x_2$$

$$\boxed{\nabla f(x) = \begin{bmatrix} \dfrac{2x_1 \cdot \exp(-(x_1^2 + x_2^2))}{(1 + \exp(-(x_1^2 + x_2^2)))^2} \\[2ex] \dfrac{2x_2 \cdot \exp(-(x_1^2 + x_2^2))}{(1 + \exp(-(x_1^2 + x_2^2)))^2} \end{bmatrix}}$$

iteration 1

$$x_0 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

So, $\exp(-(1+1))$
$= \exp(-2) = 0.135$

$$\nabla f(x) = \begin{bmatrix} \dfrac{2.1.(0.135)}{(1+0.135)^2} \\ \\ \dfrac{2(-1)(0.135)}{(1+0.135)^2} \end{bmatrix}$$

$$\boxed{\nabla f(x)_{i1} = \begin{bmatrix} 0.21 \\ \\ -0.21 \end{bmatrix}}$$

$$x_{i1} = x_0 - \epsilon * \nabla f(x)$$

$$= \begin{bmatrix} 1 \\ -1 \end{bmatrix} - 0.1 * \begin{bmatrix} 0.21 \\ -0.21 \end{bmatrix}$$

$$= \begin{bmatrix} 1 - 0.021 \\ -1 + 0.021 \end{bmatrix}$$

$$\boxed{x_{i1} = \begin{bmatrix} 0.98 \\ -0.98 \end{bmatrix} \qquad f(x_{i1}) = 0.87}$$

iteration 2

$$\nabla f(x) = \begin{bmatrix} \dfrac{2(0.98)(0.146)}{(1+0.146)^2} \\ \dfrac{2(-0.98)(0.146)}{(1+0.146)^2} \end{bmatrix}$$

$$\Rightarrow \exp(-(x_1^2 + x_2^2)) = 0.146$$

$$\nabla f(x) = \begin{bmatrix} 0.22 \end{bmatrix}$$

$$\nabla f(x) = \begin{bmatrix} -0.22 \end{bmatrix}$$

$$x_{i_2} = x_{i_1} - \epsilon * \nabla f(x)$$

$$x_{i_2} = \begin{bmatrix} 0.96 \\ -0.96 \end{bmatrix}$$

$$f(x_{i_2}) = 0.86$$

iteration 3

$$\nabla f(x) = \begin{bmatrix} \frac{2(0.96)(0.158)}{(1+0.158)^2} \\ \frac{2(-0.96)(0.158)}{(1+0.158)^2} \end{bmatrix}$$

$$\nabla f(x) = \begin{bmatrix} 0.23 \\ -0.23 \end{bmatrix}$$

$$x_{i_3} = \begin{bmatrix} 0.94 \\ -0.94 \end{bmatrix}$$

$$f(x_{i_3}) = 0.8633$$

Ex 4.3.2

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

$$\frac{df(x, y : w)}{dw} \qquad \left[\frac{d(x^T a)}{dx} = a^T \right]$$

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

$$\boxed{= X^T (Xw - y) + \lambda w}$$

Ex. 4.4

X = samples
Y = observation / ground truth
$X_{[i]}$ = X with i^{th} sample (row) removed
$Y_{[i]}$ = Y with i^{th} outcome (row) removed
$\hat{Y}$ = predicted value when model is estimated with all samples included
$\hat{Y}_{[i]}$ = predicted value when model is estimated with all except the i^{th} sample
$\hat{w}_{[i]}$ = estimated weights without the i^{th} sample

$$\hat{Y} = X\hat{w}$$

We aim to select $\hat{\beta}$ in a way which minimises the mean square error between Y & $\hat{Y}$

$$MSE = \frac{1}{n} \|Y - Xw\|^2$$

$$= \frac{1}{n} (Y - Xw)^T (Y - Xw)$$

For best value of w we do:

$$\frac{d(MSE)}{dw} = 0$$

$$\Rightarrow -\frac{2X^T}{n}(Y - Xw) = 0$$

$$\Rightarrow X^T Y - X^T X w = 0$$

$$\Rightarrow \boxed{\hat{w} = (X^T X)^{-1} X^T Y} \quad \text{— (1)}$$

Also $\rightarrow$ $\hat{Y} = HY$

$$\Rightarrow X\hat{w} = HY$$

$$\Rightarrow X(X^T X)^{-1} X^T Y = HY$$

$$\Rightarrow \boxed{H = X(X^T X)^{-1} X^T} \quad \text{— (2)}$$

To compute $\hat{w}_{[i]}$ we use (1)

$$\hat{w}_{[i]} = (X_{[i]}^T X_{[i]})^{-1} X_{[i]}^T Y_{[i]}$$

from (2) $\rightarrow$ $\boxed{h_i = x_i (X^T X)^{-1} x_i^T}$ — (3)

$$\therefore\; X^T X = \sum_{i=1}^{n} x_i^T x_i$$

$$\Rightarrow \boxed{X_{[i]}^T X_{[i]} = X^T X - x_i^T x_i} \quad \text{—} (4)$$

According to Sherman-Morrison formula -

$$\boxed{(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1} u v^T A^{-1}}{1 + v^T A^{-1} u}} \quad \text{—} (5)$$

Putting the following values for variables in eq (5)

$$A = X^T X$$
$$u = -x_i^T$$
$$v = x_i^T$$

$$\Rightarrow \left(X^T X + (-x_i^T x_i)\right)^{-1}$$
$$= (X^T X)^{-1} - \frac{(X^T X)^{-1}(-x_i^T x_i)(X^T X)^{-1}}{1 + x_i (X^T X)^{-1} (-x_i^T)}$$

$$\Rightarrow \left(X_{[i]}^T X_{[i]}\right)^{-1} = (X^TX)^{-1} + \frac{(X^TX)^{-1}(x_i^T x_i)(X^TX)^{-1}}{1 - \underbrace{x_i (X^TX)^{-1}(x_i^T)}_{\hookrightarrow h_i \{\text{from } 3\}}}$$

$$\Rightarrow \left(X_{[i]}^T X_{[i]}\right)^{-1} = (X^TX)^{-1} + \frac{(X^TX)^{-1}(x_i^T x_i)(X^TX)^{-1}}{1-h_i}$$

Multiplying x_i^T on both sides

$$\Rightarrow \left(X_{[i]}^T X_{[i]}\right)^{-1} x_i^T$$

$$= (X^TX)^{-1} x_i^T + \frac{(X^TX)^{-1} x_i^T \overbrace{x_i (X^TX)^{-1} x_i^T}^{h_i}}{1-h_i}$$

$$\Rightarrow \left(X_{[i]}^T X_{[i]}\right)^{-1} x_i^T = (X^TX)^{-1} x_i^T \left[1 + \frac{h_i}{1-h_i}\right]$$

$$\Rightarrow \boxed{\left(X_{[i]}^T X_{[i]}\right)^{-1} x_i^T = (X^TX)^{-1} x_i^T \left(\frac{1}{1-h_i}\right)} \quad - 6$$

From 1 we have $\rightarrow$

$$X^T X \hat{\omega} = X^T Y$$

From (4) $\rightarrow$

$$X^T X = X_{[i]}^T X_{[i]} + x_i^T x_i$$

Similarly $\rightarrow$

$$X^T Y = X_{[i]}^T Y_{[i]} + x_i^T y_i$$

$$\Rightarrow \left[X_{[i]}^T X_{[i]} + x_i^T x_i \right] \hat{\omega} = X_{[i]}^T Y_{[i]} + x_i^T y_i$$

Multiplying $(X_{[i]}^T X_{[i]})^{-1}$ on both sides:

$$\left[I + (X_{[i]}^T X_{[i]})^{-1} x_i^T x_i \right] \hat{\omega}$$

$$= \underbrace{(X_{[i]}^T X_{[i]})^{-1} X_{[i]}^T Y_{[i]}}_{\hat{\omega}_{[i]}} + (X_{[i]}^T X_{[i]})^{-1} x_i^T y_i$$

As we know $\rightarrow$

$$e_i = y_i - \hat{y}_i$$

$$\Rightarrow e_i = y_i - x_i \hat{\omega}$$

$$\Rightarrow \boxed{y_i = e_i + x_i \hat{w}} \quad (7)$$

Using this relation in the above equation, we get →

$$\hat{w} + \cancel{(X_{[i]}^T X_{[i]})^{-1}} x_i^T x_i \hat{w}$$
$$= \hat{w}_{[i]} + (X_{[i]}^T X_{[i]})^{-1} x_i^T (e_i + \cancel{x_i \hat{w}})$$

$$\Rightarrow \hat{w} = \hat{w}_{[i]} + (X_{[i]}^T X_{[i]})^{-1} x_i^T e_i$$

Replacing with equation (6) we get -

$$\Rightarrow \hat{w} = \hat{w}_{[i]} + \frac{(X^T X)^{-1} x_i^T e_i}{1 - h_i}$$

Multiplying with x_i on both sides -

$$\Rightarrow x_i \hat{w} = x_i \hat{w}_{[i]} + \frac{x_i (X^T X)^{-1} x_i^T e_i}{1 - h_i}$$

Subtracting y_i on both sides -

$$\Rightarrow \boxed{x_i \hat{w} - y_i = x_i \hat{w}_{[i]} - y_i + \frac{h_i e_i}{1 - h_i}} \quad (8)$$

$\Rightarrow$ <u>We know</u> $\rightarrow$

$$e_i = y_i - x_i \hat{\omega}$$

$$e_{[i]} = y_i - x_i \hat{\omega}_{[i]}$$

$$h_i = x_i (X^T X)^{-1} x_i^T e_i$$

Placing these in equation 8, we get

$$-e_i = -e_{[i]} + \frac{h_i e_i}{1-h_i}$$

$$\Rightarrow \boxed{e_{[i]} = \frac{e_i}{1-h_i}} \quad \text{— } 9$$

As we know LOOCV is given as -

$$CV = \frac{1}{n} \sum_{i=1}^{n} e_{[i]}^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left(\frac{e_i}{1-h_i} \right)^2$$

$$\Rightarrow \boxed{CV = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1-h_i}\right)^2}$$