Name: Akshat Raj Saxena
Roll No.: 2022054
Page No.: 1
Q3 is at the last page.

Q1 $X = \begin{bmatrix} 0.25 & 10 & 83 \\ 0.80 & 15 & 57 \\ 0.72 & 53 & 90 \\ 0.95 & 25 & 64 \end{bmatrix}$

$m=4$ & $n=3$

$\mu_1 = \frac{1}{4}(0.25 + 0.8 + 0.72 + 0.95)$

$\mu_1 = 0.68$ — (1)

& $\mu_2 = \frac{1}{4}(10+15+53+25)$

$\mu_2 = 25.75$ — (2)

& $\mu_3 = 73.5$ — (3)

$\sigma_1 = \frac{1}{2}\sqrt{(0.25-0.68)^2 + (0.8-0.68)^2 + (0.72-0.68)^2 + (0.95-0.68)^2}$

$\Rightarrow \sigma_1 = \frac{1}{2}\sqrt{0.184 + 0.94 + 0.160 + 0.0729}$

$\sigma_1 = 0.3282$ — (4)

$\sigma_2 = \frac{1}{2}\sqrt{(10-25.75)^2 + (\cancel{0.8}\,15 - 25.75)^2 + (25.75-53)^2 + (25-25.75)^2}$

$\sigma_2 = \frac{1}{2}\sqrt{248.0625 + 115.5625 + 742.5625 + 0.5625}$

$\sigma_2 = 16.6339$ — (5)

$\sigma_3 = \frac{1}{2}\sqrt{(83-73.5)^2 + (57-73.5)^2 + (90-73.5)^2 + (64-73.5)^2}$

$= \frac{1}{2}\sqrt{90.25 + 272.25 + 272.25 + 90.25}$

$\sigma_3 = 13.4629$ — (6)

$$\rho_{jk} = \frac{\sum_{i=1}^{m} (x_{ij} - \mu_j)(x_{ik} - \mu_k)}{(m-1)\,\sigma_j \sigma_k}$$

$$R = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{21} & 1 & \rho_{23} \\ \rho_{31} & \rho_{32} & 1 \end{bmatrix}$$

$$\rho_{12} = \frac{(0.25-0.68)(10-25.75) + (0.8-0.68)(15-25.75) + (0.72-0.68)(53-25.75) + (0.95-0.68)(25-25.75)}{3 \times 0.3282 \times 16.6339}$$

$$\rho_{12} = \frac{6.777 - 1.29 + 1.09 - 0.2025}{16.372}$$

$$\rho_{12} = 0.385$$

$$\rho_{12} = \rho_{21} = 0.385$$

$$\rho_{31} = \frac{(-0.43)(9.5) + 0.12(16.5) + 0.04(16.5) + 0.27(9.5)}{3 \times 13.42 \times 0.3282}$$

$$\rho_{31} = -0.565$$

$$\rho_{13} = \rho_{31} = -0.565$$

$$\rho_{23} = \frac{(9.5)(-15.75) + (-10.75)(16.5) + 16.5(27.25) + 9.5(-0.75)}{3 \times 16.6339 \times 13.4629}$$

$$\rho_{23} = 0.5408 = \rho_{32}$$

Hence the final Correlation Matrix is

$$\begin{bmatrix} 1 & 0.365 & -0.565 \\ 0.365 & 1 & 0.540 \\ -0.565 & 0.540 & 1 \end{bmatrix} = \text{Corr}A \quad (\text{say})$$

$$\text{Corr}A \cdot \underbrace{V}_{\text{Eigenvector}} = \overbrace{\lambda}^{\text{Eigenvalue}} V$$

$$|\text{Corr}A - \lambda I| = 0$$

$$\begin{vmatrix} 1-\lambda & 0.365 & -0.565 \\ 0.365 & 1-\lambda & 0.540 \\ -0.565 & 0.540 & 1-\lambda \end{vmatrix} = 0$$

$$\Rightarrow (1-\lambda)^3 - 0.29(1-\lambda) - 0.0144(1-\lambda) - 0.114 - 0.112 - 0.3136(1-\lambda)$$

$$\Rightarrow -\lambda^3 + 3.001 \cdot \lambda^2 - \frac{45229}{20000}\lambda + \frac{36657}{1000000} = 0$$

$$\lambda = \begin{Bmatrix} 0.009 \\ 1.380 \\ 1.61 \end{Bmatrix}$$

Now for Eigenvector

$\lambda_1 = 0.009 \; ; \quad V_1 = \begin{pmatrix} 0.908 \\ -0.893 \\ 1 \end{pmatrix}$

$\lambda_2 = 1.380 \; ; \quad V_2 = \begin{pmatrix} 20.128 \\ 21.576 \\ 1 \end{pmatrix}$

$\lambda_3 = 1.61 \; ; \quad V_3 = \begin{pmatrix} -0.598 \\ 0.512 \\ 1 \end{pmatrix}$

$\boxed{\text{Now for proof}}$

We just want to show that

$a^T \Sigma a \geq 0$ where Σ is covariance Matrix &

a is any vector

We will approach this by looking at how x feature deviate from its mean.

$$\Sigma_{xx} = E\left[(x - E[x])(x - E[x])^T\right] \quad -(1)$$

for any vector a, $a^T \Sigma_{xx} a$ is its Quadratic form

$$a^T \Sigma_{xx} a = E\left[a^T (x - E[x])(x - E[x])^T a\right]$$

Let's say $J^2 = E\left[a^T (x - E[x]))^2\right]$

Hence $a^T \Sigma_{xx} a = E[J^2]$

As square can never attain -ve value in their domain

Hence covariance Matrix is always +ve definite.

Q2 $y_0 = a_0 + a_1 x_1 + a_2 x_2$

$$X = \begin{bmatrix} 1 & 2 & 6 \\ 1 & 4 & 12 \\ 1 & 9 & 28 \\ 1 & 12 & 31 \\ 1 & 18 & 34 \\ 1 & 23 & 30 \end{bmatrix} ; \quad y = \begin{bmatrix} 23 \\ 34 \\ 25 \\ 50 \\ 60 \\ 100 \end{bmatrix}$$

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

$$(X^T X) = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 4 & 9 & 12 & 18 & 23 \\ 6 & 12 & 28 & 31 & 34 & 30 \end{bmatrix} \begin{bmatrix} 1 & 2 & 6 \\ 1 & 4 & 12 \\ 1 & 9 & 28 \\ 1 & 12 & 31 \\ 1 & 18 & 34 \\ 1 & 23 & 30 \end{bmatrix}$$

$$= \begin{bmatrix} 6 & 2+4+9+12+18+23 & 6+12+28+31+34+30 \\ 2\times1+4+9+12+18+23 & 4+16+81+144+324+(23\times23) & 2\times6+4\times12+9\times28+12\times31+18\times34+23\times30 \\ 6+12+28+31+34+30 & 12+48+28\times9+31\times12+34\times18+30\times23 & 36+144+28\times28+31\times31+34\times34+30\times30 \end{bmatrix}$$

$$= \begin{bmatrix} 6 & 68 & 141 \\ 68 & 1098 & 1986 \\ 141 & 1986 & 3981 \end{bmatrix}$$

$$(X^T X)^{-1} = \quad \det(t) = 6(-1)^{1+1} \begin{vmatrix} 982/3 & 388 \\ 388 & 1335/2 \end{vmatrix} + 0 + 0$$

$$\Rightarrow 6\left(\frac{982 \times 1335}{6} - 388(388)\right)$$

$$\Rightarrow 6(67951) = 407706$$

$$\text{adj } A = \begin{matrix} 426942 & 9318 & -19770 \\ 9318 & 4005 & -2328 \\ -19770 & -2328 & 1964 \end{matrix}$$

$$(X^T X)^{-1} = \frac{1}{67951} \begin{bmatrix} 71157 & 1553 & -3295 \\ 1553 & 1335/2 & -388 \\ -3295 & -338 & 982/3 \end{bmatrix}$$

$$\left((X^T X)^{-1} \cdot X^T\right) = \quad 1771$$

$$\frac{1}{67951} \begin{bmatrix} 54493 & 40935 & -7126 & -12352 & -12919 & -57874 \\ 560 & 902 & -6607/2 & -2465 & 376 & -4989/2 \\ -2107 & -1695 & 7135/3 & 6589/3 & 2551/3 & 12449/3 \end{bmatrix}$$

$$\left((X^T X)^{-1} \cdot X^T\right) \cdot Y \Rightarrow \begin{bmatrix} 26.066 \\ 5.024 \\ -1.461 \end{bmatrix}$$

a) $y = 26.066 + 5.024(26) - 1.461(50)$

$$\boxed{y \Rightarrow 83.6336}$$

b) (i) For linear regression, we have to maximise the likelihood of residuals

Likelihood $L(\Theta) = \prod_{i=1}^{n} \frac{1}{2b} e^{-\frac{|\epsilon - a|}{b}}$

$\log L(\Theta) \Rightarrow L(\theta_0, \theta_1, \theta_2 \mid x_1, x_2, x_3 y)$

$$= \prod_{i=1}^{n} \frac{1}{2b} e^{-\frac{|y_i - (\theta_0 + \theta_1 x_1 + \theta_2 x_2)|}{b}}$$

$$\log L = -N \log(2b) - \frac{1}{b} \sum_{i=1}^{N} |y_i - (\theta_0 + \theta_1 x_1 + \theta_2 x_2)|$$

To Maximize the above, we need to minimize the sum of the above.

$$|y_i - (\theta_0 + \theta_1 x_1 + \theta_2 x_2)|$$

This is Mean Absolute fn.

(ii) When, $\epsilon \sim$ Exponential (λ)

$P(\epsilon) = \lambda e^{-\lambda t}$ for $t \geq 0$ & $\lambda > 0$

$$L(\theta_0, \theta_1, \theta_2 \mid X_1, X_2, Y) = \prod \lambda e^{-\lambda(y_i - \theta_0 - \theta_1 x_1 - \theta_2 x_2)}$$

$$\log L = N \log \lambda - \lambda \sum_{i=1}^{N} (y_i - \theta_0 - \theta_1 x_1 - \theta_2 x_2)$$

the term $\sum (y_i - \hat{y})$ is minimize when $y_i - \hat{y} = 0$

MLE estimate for expontial noise minimizes the sum of residuals (L1 loss)

Q2(d) PDF of y

$$P(y) = Ce^{-(y - \mu_y)^2}$$

For N i.i.d samples, the joint likelihood fn is:

$$L(w) = \prod_{1}^{N} P(y_i) = \prod_{i=1}^{N} Ce^{-(y_i - \mu_y)^2}$$

$$\ln L(w) = \sum_{i=1}^{N} \overset{\text{Constant}}{\cancel{\ln C}} - \sum_{i=1}^{N} (y_i - \mu_y)^2$$

$$\ln L(w) - \sum_{i=1}^{N} (y_i - \mu_y)^2 \quad \text{—(1)}$$

Maximizing $\ln L(w)$ is equivalent to minimizing:

$$J(w) = \sum_{i=1}^{N} (y_i - \mu_y)^2 \quad \text{—(2)}$$

Mean of y: $\mu_y = E[y] = E[w^T x + b) = w^T \mu + b$ —(3)

$\sigma_y^2 = var(y) = w^T S w$ (S is covariance Matrix) —(4)

$\Rightarrow \mu_y^2 = \sigma_y^2 \Rightarrow (w^T \mu + b)^2 = w^T S w$

Now, To find MLE estimate of w,

$$\left[\min_{w} \sum_{i=1}^{N} (y_i - \mu_y)^2 \text{ subject to } (w^T \mu + b)^2 = w^T S w \right]$$

This is a constrained optimization, $\Rightarrow$

Q2(c) The MAE Loss fn:

$$L(\theta_0, \theta_1, \theta_2) = \frac{1}{N} \sum_{i=1} |y_i - \hat{y}_i|$$

Computing gradient of MAE loss fn wrt $(\theta_0, \theta_1, \theta_2)$

$$\Rightarrow \frac{\partial L}{\partial \theta_j} = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial}{\partial \theta_j} |y_i - \hat{y}_i|$$

$$\Rightarrow \frac{\partial |y_i - \hat{y}_i|}{\partial \theta_j} = \begin{cases} -1 & ; \cancel{x>0}\ y_i - \hat{y}_i > 0 \\ 1 & ; y_i - \hat{y}_i < 0 \\ \text{undefined} & ; y_i - \hat{y}_i = 0 \times \end{cases}$$

$$\Rightarrow \frac{\partial |y_i - \hat{y}_i|}{\partial \theta_j} = -\operatorname{sgn}(y_i - \hat{y}_i) \cdot x_{ji}$$

$$\Rightarrow \frac{\partial L}{\partial \theta_0} = -\frac{1}{N} \sum_{i=1}^{N} \operatorname{sgn}(y_i - \hat{y}_i)$$

$$\frac{\partial L}{\partial \theta_1} = -\frac{1}{N} \sum_{i=1}^{N} \operatorname{sgn}(y_i - \hat{y}_i) \cdot x_{1i}$$

$$\frac{\partial L}{\partial \theta_2} = -\frac{1}{N} \sum_{i=1}^{N} \operatorname{sgn}(y_i - \hat{y}_i) \cdot x_{2i}$$

As we know $\theta_j = \theta_j - \eta \frac{dL}{\partial \theta_j}$

$$\Rightarrow \boxed{\begin{aligned} \theta_0 &= \theta_0 + \eta \cdot \frac{1}{N} \sum \operatorname{sgn}(y_i - \hat{y}_i) \\ \theta_1 &= \theta_1 + \eta \cdot \frac{1}{N} \sum \operatorname{sgn}(y_i - \hat{y}_i) \cdot x_{1i} \\ \theta_2 &= \theta_2 + \eta \cdot \frac{1}{N} \sum \operatorname{sgn}(y_i - \hat{y}_i) \cdot x_{2i} \end{aligned}}$$

Q4 $\hat{\theta}_\lambda = \arg\min_\theta \|X\theta - y\|^2 + \lambda\|\theta\|^2$ — (1) 10

$X = U\Sigma V^T = \sum_{i=1}^{N} \sigma_i u_i v_i^T$ — (2)

(i) To prove: $\hat{\theta}_\lambda = \sum_{i=1}^{d} \frac{\sigma_i}{\sigma_i^2 + \lambda} \langle u_i, y\rangle v_i$

Since U & V both are orthogonal Matrix — (3)

$X\theta = U\Sigma V^T\theta \Rightarrow \|X\theta - y\|^2 = \|U\Sigma V^T\theta - y\|^2$

from (3) $\|Ux\| = \|x\|$

$\|X\theta - y\|^2 = \|\Sigma V^T\theta - U^T y\|^2$

Let's say $\beta = V^T\theta$; then $\|\beta\| = \|\theta\|$

$\Rightarrow \beta_\lambda = \arg\min_\beta \|\Sigma\beta - U^T y\|^2 + \lambda\|\beta\|^2$

The Problem can now be seperable in terms of β_i

$\beta_i = \frac{\sigma_i \langle u_i, y\rangle}{\sigma_i^2 + y}$ from 4

$$\Rightarrow \boxed{\hat{\theta}_\lambda = \sum_{i=1}^{d} \frac{\sigma_i}{\sigma_i^2 + \lambda} \cdot v_i \cdot \langle u_i, y\rangle}$$

(ii) To prove: $\|\hat{\theta}_\lambda\|^2 = \sum_{i:\sigma_i > 0} \left(\frac{\sigma_i}{\sigma_i^2 + \lambda}\right)^2 \langle u_i, y\rangle^2$

Norm of ridge estimator, calculating squared Norm

$\|\hat{\theta}_\lambda\|^2 = \|\sum_{i=1}^{d} \frac{\sigma_i}{\sigma_i^2 + \lambda} v_i \langle u_i, y\rangle\|^2$

Since v_i is orthonormal

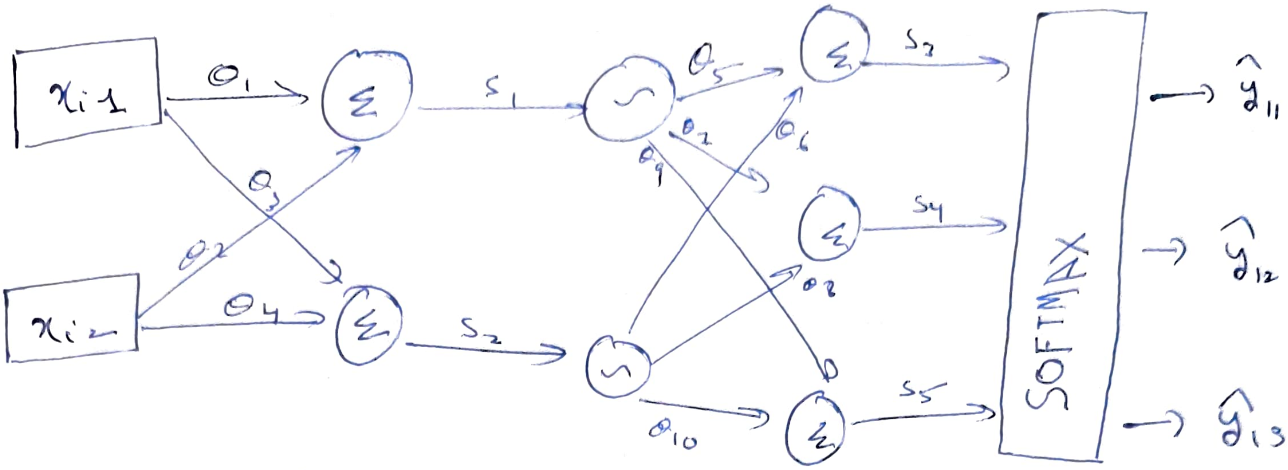
$\Rightarrow \langle v_i, v_j\rangle = 0$ if $i = 0$ & 1 if $i = j$

$\Rightarrow \|\hat{\theta}_\lambda\|^2 = \sum_{i=1}^{d} \left(\frac{\sigma_i}{\sigma_i^2 + \lambda} \langle u_i, y\rangle\right)^2$

$\forall\ \sigma_i > 0$;

$$\boxed{\|\theta_\lambda\|^2 = \sum_{i:\sigma_i > 0} \left(\frac{\sigma_i}{\sigma_i^2 + \lambda}\right)^2 \langle u_i, y\rangle^2}$$

Q5 a)



Cross Entropy loss fn: $-\sum y_i \log(\hat{y}_i)$

Sigmoid Activation: $\dfrac{1}{1+e^{-x}}$

$S_1 = \theta_1 x_{i1} + \theta_2 x_{i2}$; $S_2 = \theta_3 x_{i1} + \theta_4 x_{i2}$

$h_1 = \sigma(s_1)$ & $h_2 \ \sigma(S_2)$

$S_3 = \theta_5 h_1 + \theta_6 h_2$; $S_4 = \theta_7 h_1 + \theta_8 h_2$

$S_5 = \theta_9 h_1 + \theta_{10} h_2$

Backpropagation

$$\frac{\partial h}{\partial s_j} = \hat{y}_{ij} - y_{ij}$$

$$\frac{\partial L}{\partial \theta_5} = (\hat{y}_{i1} - y_{i1}) \cdot h_1$$

$$\frac{\partial L}{\partial h_1} = \sum_{j=3} \frac{\partial L}{\partial s_j} \frac{\partial s_j}{\partial h_1} = \sum (\hat{y}_{ij} - y_{ij})\theta_{j1}$$

$$\frac{\partial L}{\partial s_1} = \frac{\partial L}{\partial h_1} \cdot \sigma'(s_1) \Rightarrow \left[\sigma'(s) = \frac{\sigma(s)}{1-\sigma(s)} \right]$$

$$\frac{\partial L}{\partial \theta_1} = \frac{\partial L}{\partial s_1} \cdot \frac{\partial s_1}{\partial \theta_1}$$

$$\theta_K \leftarrow \theta_K - \eta \frac{\partial L}{\partial \theta_K}$$

Final Weights

$$\theta_1 \leftarrow \theta_1 - \eta \left[\sum_{j=3}^{5} (\hat{y}_{ij} - y_{ij}) \cdot \theta_{j1} \right] \cdot \sigma'(s_1) \cdot x_{i1}$$

$$\theta_2 \leftarrow \theta_2 - \eta \left[\sum_{j=3}^{5} (\hat{y}_{ij} - y_{ij}) \cdot \theta_{j2} \right] \cdot \sigma'(s_2) \cdot x_{i2}$$

$$\theta_5 \leftarrow \theta_5 - \eta \left[(\hat{y}_{i2} - y_{i2}) \cdot h_i \right]$$

b) (i) $\alpha = 0.01$

$x_1 = [0.6, 0.7, 1.3]^T$

$1(0.6) + 0.5(0.7) + 3(1.3) + 0.3$

$h_1 = \sigma(4.85)$

$h_1 = 0.992$

$\Rightarrow 2(0.6) + 2.5(0.7) - 0.5(1.3) + 0.8$

$\Rightarrow h_2 = \sigma(1.22)$

$h_2 = 0.772$

o/p $2h_1 - 1.5h_2 \Rightarrow$

$\sigma(0.826)$

o/p$_1$ = 0.6955

$x_2 = [-1.8, 0.1, 0.2]^T$

$1(1.8) + 0.5(0.1) + 3(0.2) + 0.3$

$\sigma(-0.85) \Rightarrow h_1 = 0.2994$

$h_2 = 2(-1.8) + 2.5(0.1) - 0.5(0.2) + 0.8$

$\sigma(2.65)$

$h_2 = 0.0659$

o/p

$2(0.299) - 1.5(0.065)$

$\sigma(0.5005)$

o/p$_2$ = 0.6225

b) (ii) MSE $\Rightarrow \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

$\hat{y}_1 = 0.6955$ & $\hat{y}_2 = 0.6225$

$$MSE = \frac{1}{2}\left[(1-0.6955)^2 + (0.6225)^2\right]$$

$$= \frac{1}{2}\left[0.0927 + 0.3912\right]$$

$$MSE = 0.24195$$

$$BCE = -\frac{1}{2} \sum_{i=1}^{2} \left(y_i \log \hat{y}_i + (1-y_i) \log (1-\hat{y}_i) \right]$$

$$= -\frac{1}{2}\left[\log(0.6955) + 0 + 0 + (1)(\log(0.3775))\right]$$

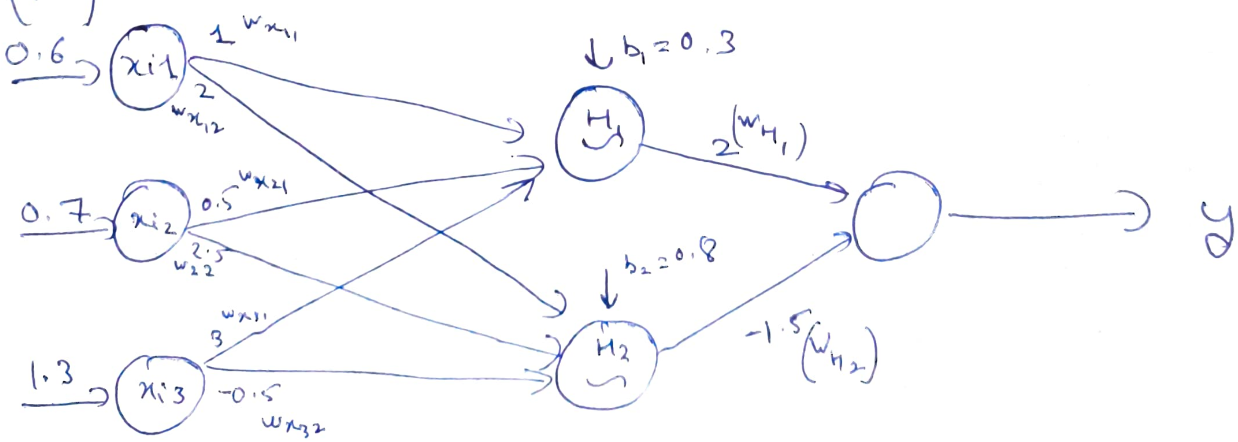
$$\Rightarrow \frac{1}{2}\left[0.363 + 0.974\right]$$

$$\Rightarrow 0.6685$$

Since the predicted Probability was closer to their actual hence MSE comes out to be lower than BCE

The higher BCE loss indicates that the weight can be more updated to make these loss more lower

(iii)



$\alpha = 0.01$

$BCE = 0.6685$ for x_1 input class 1

for $O_1 = 0.6955$

$H_1 = 0.992$

$H_2 = 0.772$

$$\Delta w_{ji} = \eta \, \delta_j \, O_i$$

$$\delta_j = O_j(1 - O_j)(t_j - O_j)$$

$$\delta_j = O_j(1 - O_j) \sum_{k=1}^{m} \delta_k w_{kj}$$

for $\delta_1 = O_1(1 - O_1)(\text{Actual} - O_1)$

$$= 0.69(0.31)(0.31)$$

$$\delta_1 = 0.066 - 1$$

$$\delta_{H_1} = H_1(1 - H_1)(2)(\delta_1) =$$

$$\delta_{H_1} = 0.00130 - 2$$

$$\delta_{H_2} = -0.77(0.23)(1.5)(0.66)$$

$$\delta_{H_2} = -0.0175$$

$$\Delta w_{H_1} = \eta \, \delta_1 H_1 = (0.01)(0.066)(0.992)$$

$$\Delta w_{H_1} = 0.00065472$$

$$\Delta w_{H_2} = (0.01)(0.066)(0.772)$$

$$\Delta w_{H_2} = 0.00050952$$

$$W_{H_1(new)} = 2.00065472 \quad \text{— (4)}$$

$$W_{H_2(new)} = -1.49949048 \quad \text{— (5)}$$

$$b_{(new)} = 0.3 + 0.01(0.0013)$$
$$= 0.3113$$

$$b_{2(new)} = 0.8 + (0.01)(-0.0175)$$
$$= 0.799825$$

$$\Delta w_{x_{11}} = (0.01)(0.0013)(0.6) = 0.0000078$$

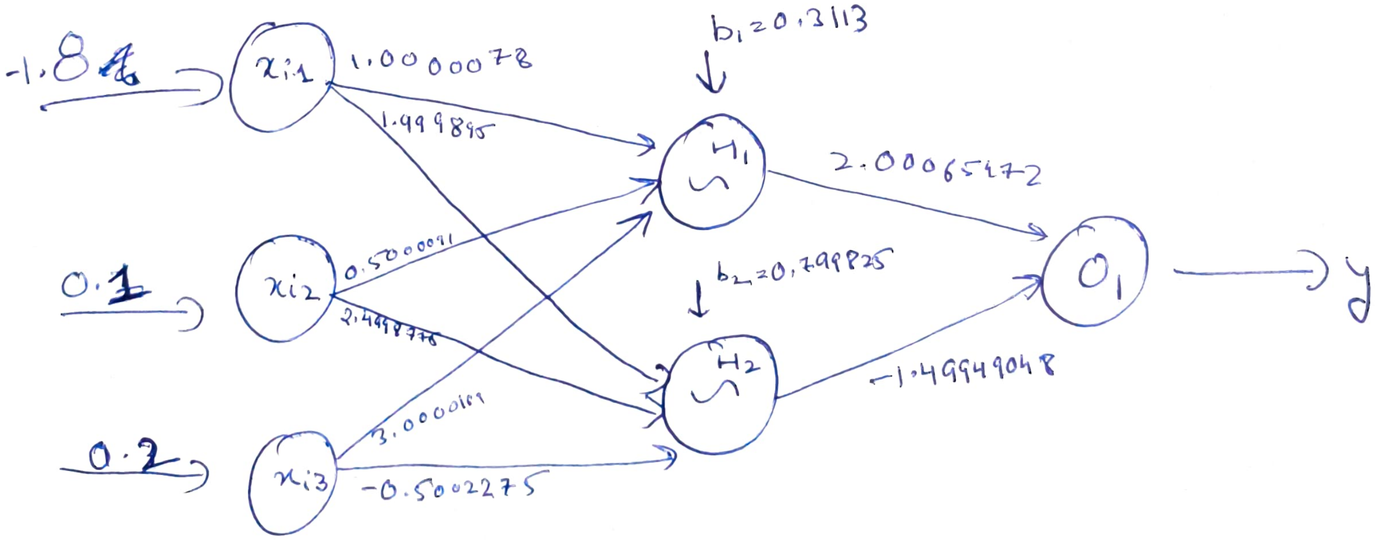
$$\Delta w_{x_{12}} = (0.01)(-0.0175)(0.6) = -0.000105$$

$$\Delta w_{x_{21}} = (0.01)(0.0013)(0.7) = 0.0000091$$

$$\Delta w_{x_{22}} = (0.01)(-0.0175)(0.7) = -0.0001225$$

$$\Delta w_{31} = (0.01)(0.0013)(1.3) = 0.0000169$$

$$\Delta w_{x_{32}} = (0.01)(-0.0175)(1.3) = -0.0002275$$



$$H_1 = \sigma\left(-1.8(1.0000078) + 0.2(3.0000169) + 0.1(0.5000091)\right)$$

$$H_1 = \sigma(-1.15000975)$$

$$H_1 = 0.2404$$

$$\cancel{H_2 = \sigma(2.00065472(0.2404) + 0.03077(-1.49949048))}$$

$$H_2 = \sigma(-1.8(1.999895) + 0.1(2.49987775) - 0.2(0.50022275))$$

$$\sigma(-3.44986875)$$

$$H_2 = 0.03077$$

$$\sigma(2.00065472(0.2404) + 0.03077(-1.49949048)) = O_1$$

$$\Rightarrow O_1 = 0.6070236$$

Hence this result is more ~~to~~ closer to class ①, the Previous

Q3 The hypothesis fn using tanh activation

$$h_w(x) = \tanh(w^T x + b)$$

$w \to$ weights , $b \to$ bias &

$$\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

For binary classification with labels $y \in \{0, 1\}$, using Binary - cross Entropy Loss:

$$L = -y \log(p) - (1-y) \log(1-p)$$

$$p = \frac{\tanh(W^T x + b) + 1}{2}$$

$\Rightarrow$ Derivative of tanh

$$\frac{\partial}{\partial z} \tanh(z) = 1 - \tanh^2(z)$$

$\Rightarrow$ Gradient of loss wrt weights

$$\frac{\partial L}{\partial w} = (\tanh(z) - y) \cdot (1 - \tanh^2(z)) \cdot x$$

$z = w^T x + b$

The weights & bias are updated as:

$$w := w - \eta (\tanh(z) - y) \cdot (1 - \tanh^2(z)) \cdot x$$

$$b := b - \eta (\tanh(z) - y) \cdot (1 - \tanh^2(z))$$