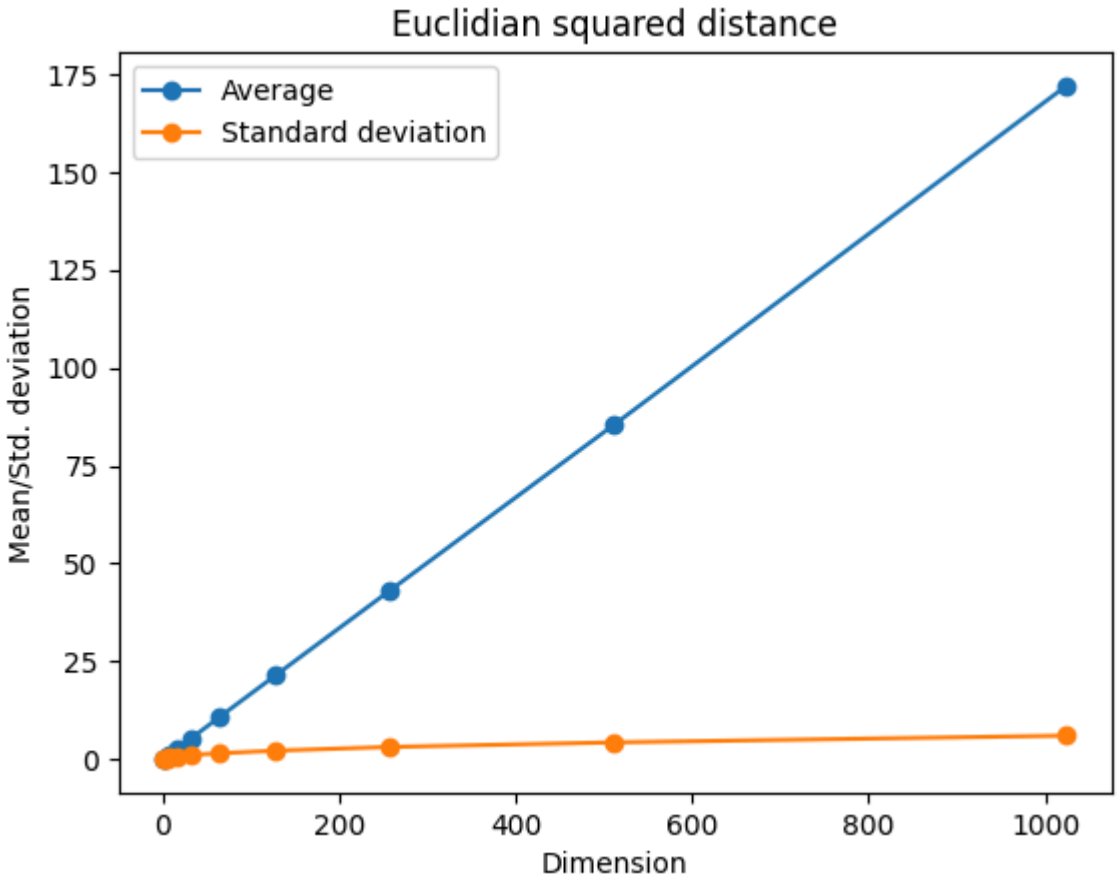
Q 1. (a)


Euclidian squared distance
Average
Standard deviation
Mean/Std. deviation
Dimension
0
25
50
75
100
125
150
175
0
200
400
600
800
1000

Aditya Agrawal
2021AM10198

Date
Page No.

APL405
HOMEWORK - 1

Q.1. (b) For a univariate random variable X with uniform distribution in the interval [0, 1], the probability density function $f_X(x)$ is given by:

$$f_X(x) = \begin{cases} 1 & , 0 \le x \le 1 \\ 0 & , \text{O.W.} \end{cases}$$

$$\therefore E[X] = \int_{-\infty}^{\infty} x f(x)\,dx = \int_0^1 x\,dx = \frac{x^2}{2}\Big|_0^1 = \frac{1}{2}$$

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x)\,dx = \int_0^1 x^2\,dx = \frac{x^3}{3}\Big|_0^1 = \frac{1}{3}$$

$\therefore$ For $Z = (X-Y)^2 = X^2 + Y^2 - 2XY$,

$$\begin{aligned} E[Z] &= E[X^2 + Y^2 - 2XY] \\ &= E[X^2] + E[Y^2] - 2E[XY] \\ &= E[X^2] + E[Y^2] - 2E[X]E[Y] \quad \{\because X \text{ \& } Y \text{ are indep.}\} \\ &= \frac{1}{3} + \frac{1}{3} - \cancel{2} \times \frac{1}{\cancel{2}} \times \frac{1}{2} \\ &= \frac{2}{3} - \frac{1}{2} \end{aligned}$$

$$\therefore E[Z] = \frac{1}{6}$$

$$Var[Z] = E[Z^2] - (E[Z])^2$$

$$= E[$$

$$Z^2 = (X-Y)^4 = X^4 + Y^4 - 4X^3Y + 6X^2Y^2 - 4XY^3$$

$$\therefore E[X^4] = \int_0^1 x^4\,dx = \frac{1}{5}$$

$$E[X^3Y] = E[X^3]E[Y] = \left(\int_0^1 x^3\,dx\right)\left(\frac{1}{2}\right) = \frac{1}{8} \ldots$$

$E[X^2Y^2] = E[X^2]E[Y^2] = \int x \frac{1}{9}$

$\therefore E[Z^2] = E[X^4] + E[Y^4] - 4E[X^3]E[1] - 4E[Y^3]E[X] + 6E[X^2]E[$

$= \frac{1}{5} + \frac{1}{5} - 4 \times \frac{1}{82} - 4 \times \frac{1}{82} + {}^2 6 \times \frac{1}{83}$

$= \frac{2}{5} - 1 + \frac{2}{3}$

$= \frac{1}{15}$

$\therefore Var[Z] = E[Z^2] - (E[Z])^2$

$= \frac{1}{15} - \frac{1}{36}$

$= \frac{7}{180}$

Q.1.(c) Given, $S = Z_1 + Z_2 + \ldots + Z_d$

$\therefore E[S] = E[Z_1] + E[Z_2] + \ldots + E[Z_d]$

$= \frac{1}{6} + \frac{1}{6} + \ldots + \frac{1}{6}$

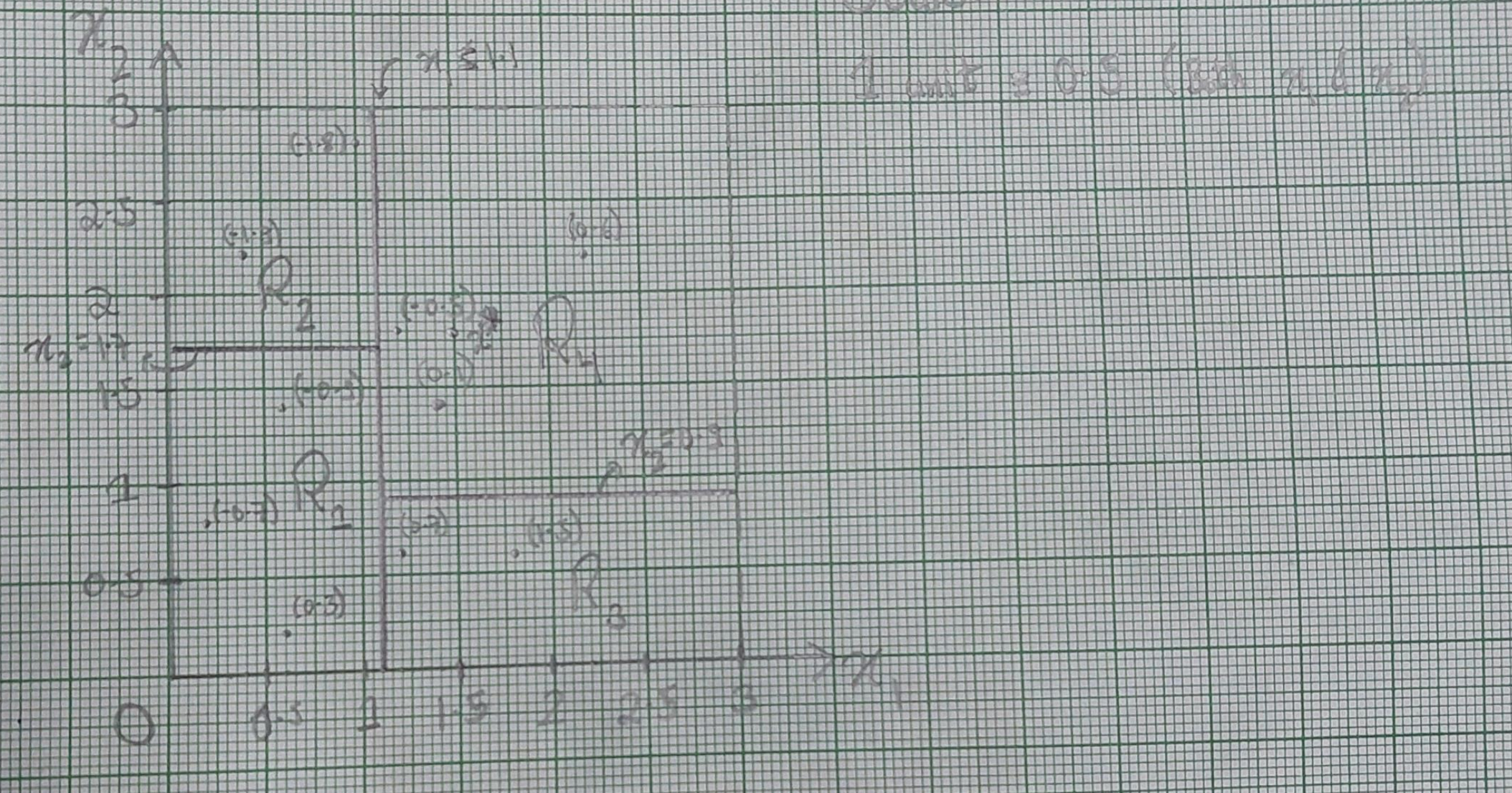
$= \frac{d}{6}$

$\therefore Var[S] = Var[Z_1 + Z_2 + \ldots + Z_d]$

$= Var[Z_1] + Var[Z_2] + \ldots \rightarrow Var[Z_d]$ {Z_is are indep.}

$= \frac{7d}{180}$

Q. 3(a)


Scale:
1 unit = 0.5 (Both x_1 & x_2)
x_2
x_1
$x_1 \geq 1$
$x_2 \geq 1$
$x_2 \geq x_1 + 1$
R_1
R_2
R_3
R_4
(1,2)
(1,3)
(2,3)
(3,3)
(3,2)
(2,2)
(2,1)
(3,1)
3
2.5
2
1.5
1
0.5
0

Q3. (b) From the graph, clearly $x^* = [1.5 \;\; 1.8]^T$ lies in R_4

$\therefore$ y for $x^* = \frac{0.6 + 0.1 - 0.5}{3} = 0.067$

(c) For R_1: $\hat{y} = \frac{0.3 - 0.7 - 0.9}{3} = -0.433$

$\therefore$

Split at	$\hat{y}_1$	$\Sigma(y_i - \hat{y}_1(j,s))^2$	$\hat{y}_2$	$\Sigma(y_i - \hat{y}_2(j,s))^2$	Loss
$x_1 = 0.5$	-0.7	0	-0.3	0.72	0.72
$x_2 = 0.5$	0.3	0	-0.8	0.02	0.02
$x_2 = 1$	-0.2	0.5	-0.9	0	0.5

$\therefore$ $x_2 = 0.5$ is the best split line for R_1.

For R_4:

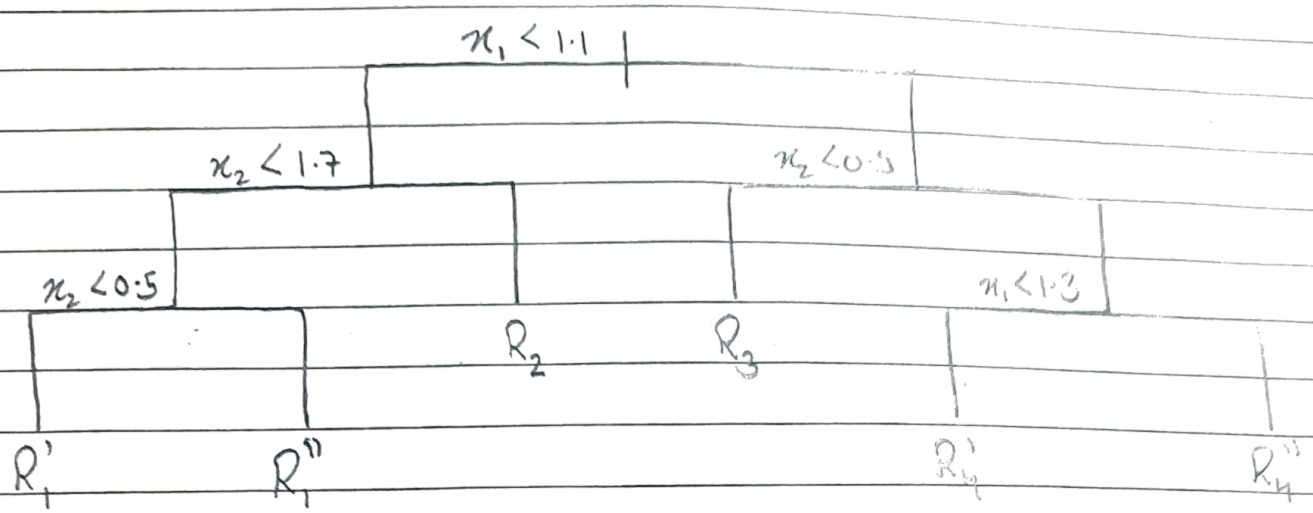
Split at	$\hat{y}_1$	$\Sigma(y_i - \hat{y}_1(j,s))^2$	$\hat{y}_2$	$\Sigma(y_i - \hat{y}_2(j,s))^2$	Loss
$x_1 = 1.3$	-0.5	0	0.35	0.125	0.125
$x_1 = 2$	-0.2	0.18	0.6	0	0.18
$x_2 = 1.5$	0.1	0	0.05	0.605	0.605
$x_2 = 2$	-0.2	0.18	0.6	0	0.18

$\therefore$ $x_1 = 1.3$ is the best split line for R_4.

Note that R_2 & R_3 already have only 2 points, thus there is no requirement of splitting them further.

Theoretically, the regions could have been split further such that each region contains a single point to minimize the error, but this would cause overfitting of the model.

The final regression tree looks like:



(d) For $x^* = [1.5 \quad 1.8]^T$, the new tree predicts $y = \frac{0.1+0.6}{2} = 0.35$

Q4. The training accuracy of the decision tree = 94.16 %
The testing accuracy of the decision tree = 92.72 %

Q5. (a) The training model for a Linear Regression is given by

$$y = X^T\theta + \epsilon$$

where, $X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{bmatrix}$; $\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{bmatrix}$; $\epsilon \rightarrow$ Random error with 0 mean

Now,
We want to find a θ such that the probability of our prediction ($\hat{y}$) matching the output label (y) is maximum, i.e.

$$\hat{\theta} = \arg\max_{\theta} p(y \mid X; \theta)$$

We assume that the error ϵ is a random variable and follows a Gaussian distribution with mean (μ) = 0 and standard deviation σ_ϵ. i.e. $\epsilon \sim \mathcal{N}(\epsilon; 0, \sigma_\epsilon^2)$

Since all data points (x_i^T, y_i) are independent, we may write:

$$p(y) = p(y_1 y_2 y_3 \dots y_N)^T) = p(y_1)\, p(y_2) \dots p(y_N)$$

$$\therefore p(y|X;\theta) = \prod_{i=1}^{N} p(y_i | x_i^T; \theta)$$

$$\text{or, } \ln p(y|X;\theta) = \sum_{i=1}^{N} \ln p(y_i | x_i^T; \theta)$$

Also, $y_i = x_i^T\theta + \epsilon$ ——— ①

$$\therefore p(y_i | x_i^T \theta) = \frac{1}{\sqrt{2\pi}\sigma_\epsilon} \exp\left(-\frac{(\epsilon-\mu)^2}{2\sigma_\epsilon^2}\right)$$

Using ① & $\mu = 0$:

$$p(y_i | x_i^T; \theta) = \frac{1}{\sqrt{2\pi}\sigma_\epsilon} \exp\left(-\frac{(y_i - x_i^T\theta)^2}{2\sigma_\epsilon^2}\right)$$

$$\therefore \ln p(y_i | x_i^T; \theta) = -\ln\sqrt{2\pi}\sigma_\epsilon - \frac{(y_i - x_i^T\theta)^2}{2\sigma_\epsilon^2}$$

$$\therefore \ln p(y|X;\theta) = \underbrace{-\sum_{i=1}^{N} \ln\sqrt{2\pi}\sigma_\epsilon}_{\text{Constant}} - \sum_{i=1}^{N} \frac{(y_i - x_i^T\theta)^2}{2\sigma_\epsilon^2}$$

Since log is an increasing function, maximising $\ln p(y|X;\theta)$ $\Rightarrow$ $p(y|X;\theta)$ maximising

$$\therefore \hat{\theta} = \arg\max_\theta \ln p(y|X;\theta) = \arg\max_\theta \left(-\frac{\sum (y_i - x_i^T\theta)^2}{2\sigma_\epsilon^2}\right)$$

$$= \arg\min_\theta \left(\sum (y_i - x_i^T\theta)^2\right)$$

Now, Define $J(\theta)$ as:

$$J(\theta) = \sum_{i=1}^{N} (y_i - x_i^T\theta)^2 = \|y - X\theta\|_2^2 = (y - X\theta)^T (y - X\theta)$$

$$= (y^T - (X\theta)^T)(y - X\theta)$$

$$= y^T y - (X\theta)^T y - y^T X\theta + (X\theta)^T (X\theta)$$

$$= y^T y - 2(\theta^T X^T) y + (X\theta)^T (X\theta)$$

To minimize $J(\theta)$ w.r.t. θ,

$$\frac{\partial J}{\partial \theta} = 0$$

or, $\frac{\partial}{\partial \theta}(y^T y) - 2\frac{\partial}{\partial \theta}((X\theta)^T y) + \frac{\partial}{\partial \theta}((X\theta)^T (X\theta)) = 0$

(y is const.)

or, $-2\frac{\partial}{\partial \theta}(\theta^T X^T y) + \frac{\partial}{\partial \theta}(\theta^T X^T X \theta) = 0$

or, $-2(X^T y)^T + \theta^T(X^T X + X X^T) = 0$

or, $-2y^T X + 2\theta^T X^T X = 0$

$\therefore (\theta^T X^T X)^T = (y^T X)^T$

$\therefore X^T X \hat{\theta} = X^T y$

(b) When $X^T X$ is not invertible, the closed form equation derived in part (a) does not have a unique solution. Rather it possesses infinite number of solutions where each solution $\hat{\theta}$ achieves the same minimal square error. This case typically arises when the number of samples of the training data (N) is less than the number of features (or the dimension) of the input data (p), i.e. $N < p$. This can be fixed by reducing dimensions by eliminating the redundant features or transforming the features in a new way

Q7. (a) For a logistic regression model, the logistic function $h(x)$ is given by:

$$h(x) = \frac{e^x}{1+e^x}$$

$$\therefore \frac{dh(x)}{dx} = \frac{e^x(1+e^x) - (e^x)^2}{(1+e^x)^2} = \frac{e^x + (e^x)^2 - (e^x)^2}{(1+e^x)^2} = \frac{e^x}{(1+e^x)^2}$$

$$= \underbrace{\frac{e^x}{1+e^x}}_{h(x)} \times \underbrace{\left(\frac{1-e^x}{1+e^x}\right)}_{1-h(x)}$$

$$\therefore \frac{dh(x)}{dx} = h(x)(1-h(x))$$

(b) For a logistic regression model, we define

$$g(\underline{x};\underline{\theta}) = h(\underline{x}^T\underline{\theta}) = \frac{e^{\underline{x}^T\underline{\theta}}}{1-e^{\underline{x}^T\underline{\theta}}}$$

Now,

$$p(y=1 \mid \underline{x};\underline{\theta}) = g(\underline{x};\underline{\theta}) = \frac{e^{\underline{x}^T\underline{\theta}}}{1+e^{\underline{x}^T\underline{\theta}}}$$

$$p(y=0 \mid \underline{x};\underline{\theta}) = 1 - g(\underline{x};\underline{\theta}) = \frac{e^{-\underline{x}^T\underline{\theta}}}{1+e^{-\underline{x}^T\underline{\theta}}}$$

$$\hat{\theta} = \arg\max_{\theta} p(\underline{y} \mid \underline{X};\underline{\theta})$$

Maximising $p(\underline{y} \mid \underline{X};\underline{\theta})$ is as good as maximising $\log(p(\underline{y} \mid \underline{X}$

$$\therefore \hat{\theta} = \arg\max_{\theta} \ln(p(\underline{y} \mid \underline{X};\underline{\theta}))$$

$$p(\underline{y} \mid \underline{X};\underline{\theta}) = \prod_{i=1}^{N} p(y^{(i)} \mid \underline{x}^{(i)};\underline{\theta})$$

$$\ln(p(\underline{y} \mid \underline{X};\underline{\theta})) = \sum_{i=1}^{N} \ln p(y^{(i)} \mid \underline{x}^{(i)};\underline{\theta})$$

$$\therefore \hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{N} \ln p(y^{(i)} \mid \underline{x}^{(i)};\underline{\theta})$$

$$= \arg\min_{\theta} -\sum_{i=1}^{N} \ln(p(y^{(i)} \mid \underline{x}^{(i)};\underline{\theta}))$$

$$\therefore \hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{N} \left[-y^{(i)} \ln\left(g(\underline{x}^{(i)}; \underline{\theta})\right) - (1-y^{(i)}) \ln\left(1 - g(\underline{x}^{(i)}; \underline{\theta})\right) \right]$$

$$\therefore J(\underline{\theta}) = -\sum_{i=1}^{N} \left[y^{(i)} \ln\left(g(\underline{x}^{(i)}; \underline{\theta})\right) + (1-y^{(i)}) \ln\left(1 - g(\underline{x}^{(i)}; \underline{\theta})\right) \right]$$

~~(e)~~

(c) $$\frac{\partial J}{\partial \theta_j} = -\sum_{i=1}^{N} \frac{\partial}{\partial \theta_j} \left(y^{(i)} \ln\left(h(\underline{x}^{(i)T}\underline{\theta})\right) + (1-y^{(i)}) \ln\left(1 - h(\underline{x}^{(i)T}\underline{\theta})\right) \right)$$

$$= -\left[\sum_{i=1}^{N} \frac{y^{(i)}}{h(\underline{x}^{(i)T}\underline{\theta})} h(\underline{x}^{(i)T}\underline{\theta})\left(1 - h(\underline{x}^{(i)T}\underline{\theta})\right) x_j^{(i)} + \frac{(1-y^{(i)})}{1 - h(\underline{x}^{(i)T}\underline{\theta})} \left(-h(\underline{x}^{(i)T}\underline{\theta})\left(1 - h(\underline{x}^{(i)T}\underline{\theta})\right) x_j^{(i)} \right) \right]$$

$$= -\left[\sum_{i=1}^{N} \left\{ \frac{y^{(i)}}{h(\underline{x}^{(i)T}\underline{\theta})} - \frac{(1-y^{(i)})}{1 - h(\underline{x}^{(i)T}\underline{\theta})} \right\} h(\underline{x}^{(i)T}\underline{\theta})\left(1 - h(\underline{x}^{(i)T}\underline{\theta})\right) x_j^{(i)} \right]$$

$$= -\left[\sum_{i=1}^{N} \left\{ y^{(i)}\left(1 - h(\underline{x}^{(i)T}\underline{\theta})\right) - (1-y^{(i)})\left(h(\underline{x}^{(i)T}\underline{\theta})\right) \right\} x_j^{(i)} \right]$$

$$= -\left[\sum_{i=1}^{N} \left\{ y^{(i)} - h(\underline{x}^{(i)T}\underline{\theta}) \right\} x_j^{(i)} \right]$$

In matrix form we can write it as:

$$\frac{\partial J}{\partial \underline{\theta}} = -\underline{\underline{X}}^T \left(\underline{y} - \underline{h}\left(\underline{\underline{X}}\underline{\theta}\right) \right)$$

where, $$\underline{\underline{X}} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_p^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdots & x_p^{(2)} \\ \vdots & \vdots & \vdots & \vdots \\ x_1^{(N)} & x_2^{(N)} & \cdots & x_p^{(N)} \end{bmatrix} ; \underline{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(N)} \end{bmatrix} ; \underline{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{bmatrix}$$

$$\therefore \frac{\partial J(\theta)}{\partial \underline{\theta}} = \underline{\underline{X}}^T \left(\underline{h}\left(\underline{\underline{X}}\underline{\theta}\right) - \underline{y} \right)$$

(d) Differentiating $\partial J / \partial \theta_j$ w.r.t. θ_k gives:

$$\frac{\partial^2 J}{\partial \theta_j \partial \theta_k} = -\sum_{i=1}^{N} \left[-h(\underline{x}^{(i)T}\underline{\theta})\left(1 - h(\underline{x}^{(i)T}\underline{\theta})\right) x_j^{(i)} x_k^{(i)} \right]$$

$$= \sum_{i=1}^{N} h(\underline{x}^{(i)T}\underline{\theta})\left(1 - h(\underline{x}^{(i)T}\underline{\theta})\right) x_j^{(i)} x_k^{(i)}$$

$\therefore$ The Hessian matrix is given by:

Date_
Page

$$H = \begin{bmatrix} \partial^2 J/\partial\theta_1\partial\theta_1 & \partial^2 J/\partial\theta_1\partial\theta_2 & \cdots & \partial^2 J/\partial\theta_1\partial\theta_n \\ \vdots & & & \\ \partial^2 J/\partial\theta_n\partial\theta_1 & \cdots & & \partial^2 J/\partial\theta_n\partial\theta_n \end{bmatrix}$$

where, $\dfrac{\partial^2 J}{\partial\theta_j \partial\theta_k} = \sum_{i=1}^{N} h(x^{(i)T}\theta)\left(1 - h(x^{(i)T}\theta)\right) x_j^{(i)} x_k^{(i)}$