-: Assignment -2 :-

Q1) Construct a linear regression model for the following data set. Predict the price for a house with area 2200 sq ft and 3 bedrooms.

Size (sq ft)	Number of Bedrooms	Price (in \$)
1000	2	300000
1500	3	450000
2000	3	500000
2500	4	600 000
3000	4	700000

Ans → Simplified →

Size (x_1)	No. of Bedrooms (x_2)	Price (y)
1	2	3
1.5	3	4.5
2	3	5
2.5	4	6
3	4	7

$$\vec{y} = w_0 + w_1 x_1 + w_2 x_2 + \dots$$

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

$$X = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 1.5 & 3 \\ 1 & 2 & 3 \\ 1 & 2.5 & 4 \\ 1 & 3 & 4 \end{bmatrix} \quad X^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1.5 & 2 & 2.5 & 3 \\ 2 & 3 & 3 & 4 & 4 \end{bmatrix} \quad X^T X = \begin{bmatrix} 5 & 10 & 16 \\ 10 & 22.5 & 34.5 \\ 16 & 34.5 & 54 \end{bmatrix}$$

$$[X^T X]^{-1} = \begin{bmatrix} 6.6 & 3.2 & -4 \\ 3.2 & 3.73 & -3.33 \\ -4 & -3.33 & 3.33 \end{bmatrix} \quad X^T Y = \begin{bmatrix} 25.5 \\ 55.75 \\ 86.75 \end{bmatrix} \quad w = \begin{bmatrix} 0.7 \\ 1.5025 \\ 0.3975 \end{bmatrix}$$

For, $x_1 = 2200$ $\quad x_2 = 3$

$= 2.2 \times 10^3$

$$y\ \text{predict} = (0.7) + (1.5 \times 2.2) + (0.4 \times 3)$$
$$= 5.2$$

Thus, the predicted value/price of a house with area 2200 sq ft & 3 bedrooms is \$520000.

2) Construct the a Linear Regression Model for the following data set. Predict the price for a car with age 3 years and mileage 50,000 miles.

Age (years)	Mileage (miles)	Price (in $)
1	10,000	25,000
3	30,000	19,000
5	50,000	15,000
7	70,000	12,000
10	100,000	8000

Ans → Simplified Table →

Age (years)	Mileage (miles)	Price (in $)
1	10	2.5
3	30	1.9
5	5	1.5
7	7	1.2
10	10	0.8

$$X = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 3 & 3 \\ 1 & 5 & 5 \\ 1 & 7 & 7 \\ 1 & 10 & 10 \end{bmatrix} \quad X^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 5 & 7 & 10 \\ 1 & 3 & 5 & 7 & 10 \end{bmatrix} \quad (X^T X) = \begin{bmatrix} 5 & 26 & 25 \\ 26 & 184 & 177 \\ 26 & 184 & 177 \end{bmatrix}$$

$$(X^T Y) = \begin{bmatrix} 7.9 \\ 32.1 \\ 32.1 \end{bmatrix}$$

$(X^T x)^{-1}$ not possible because $\det(X^T x) = 0$.

Q3) Actual and predicted values are given for 10 students in the following data set.

Students	Actual Data	Predicted Data
1	Pass	Pass
2	Pass	Fail
3	Fail	Pass
4	Fail	Fail
5	Pass	Pass
6	Fail	Fail
7	Pass	Pass
8	Pass	Fail
9	Fail	Pass
10	Fail	Fail

Ans→ Data for Pass

$TP(P) = 3$

$TN(P) = 3$

$FP(P) = 2$

$FN(P) = 2$

Data for Fail

$TP = 3$

$TN = 3$

$FP(F) = 2$

$FP(F) = 2$

a) Precision ~~Recall~~ for Pass $= \dfrac{TP}{TP + FP}$

$= \dfrac{3}{3+2}$

$= 3/5$

Precision ~~Recall~~ for Fail $= 3/5$

Recall for Pass $= \dfrac{TP}{TP + FN}$

$= \dfrac{3}{3+2}$

$= 3/5$

Recall for fail $= 3/5$

b) F-1 score of Pass $= \dfrac{2}{\dfrac{1}{Precison} + \dfrac{1}{Recall}}$

$= \dfrac{2}{\dfrac{1}{3/5} + 1/3/5}$

$= 0.6$

b) F-1 score of fail $= 0.6$

Confusion matrix →

Actual Value →

Predicted Value	Pass	Fail
Pass	3	2
Fail	2	3

Q) Use logistic regression to predict the result (pass or fail) of a student based on number of study hour. Use the model to predict whether a student will pass if the number of study hours is a) 4 b) 6

Hours Studied	Passed (1 = Yes, 0 = No)
1	0
2	0
3	0
4	1
5	1
6	1
7	1

Use the threshold function as $\text{Thresh}(x) = \begin{cases} 0, & x < 0.5 \\ 1, & x \geq 0.5 \end{cases}$ and $w = (w_0, w_1) = -6, 1.5$

Ans → a) Sigmoid Function $= \dfrac{1}{1+e^{-x}}$

$$= \frac{1}{1+e^{-(-6+1.5(4))}}$$

$$= \frac{1}{1+e^{-0}} = \frac{1}{1+1} = 1/2 = 0.5$$

Thresh $(0.5) = 1$ (Pass)

b) $\cancel{x = -6 +}$ $x = -6 + 1.5 \times 6$

$= 3$

Sigmoid Function $= \dfrac{1}{1+e^{-x}}$

$$= \frac{1}{1+e^{-3}} \approx 0.9526$$

Thresh $(0.9526) = 1$ (Pass)

Q5) Construct the decision tree for predicting whether a student will pass an exam based on features: Study Hours and number of homework completed.

Study Hours	~~Number of~~ Home work completed	Passed Exam (Target)
A1	B2	0
A2	B1	0
A2	B3	1
A3	B1	0
A3	B2	1
A4	B4	1

Ans →

$$\text{Entropy (Whole data)} = \frac{-p}{n+p}\log\left(\frac{p}{n+p}\right) - \frac{n}{n+p}\log\left(\frac{n}{n+p}\right)$$

$$= \frac{-3}{6}\log\left(\frac{3}{6}\right) - \frac{3}{6}\log\left(\frac{3}{6}\right)$$

$$\text{Entropy (A1)} = -\frac{0}{1}\log\left(\frac{0}{1}\right) - \frac{1}{0}\log\left(\frac{1}{0}\right) = 0$$

$$\text{Entropy (A2)} = \frac{-1}{2}\log(1/2) - \frac{1}{2}\log\frac{1}{2} = 1$$

$$\text{Entropy (A3)} = \frac{-1}{2}\log(1/2) - \frac{1}{2}\log\frac{1}{2} = 1$$

$$\text{Entropy (A4)} = \frac{1}{0}\log\frac{1}{0} - \frac{0}{1}\log\frac{0}{1} = 0$$

$$\text{Information gain of Study hours} = E(\text{Whole data}) - \frac{1}{6}E(A1) - \frac{2}{6}E(A2) - \frac{2}{6}E(A3) - \frac{1}{6}E(A4)$$

$$= 1 - \frac{1}{6}(0) - \frac{2}{6}(1) - \frac{2}{6}(1) - \frac{1}{6}(0)$$

$$= 1 - \frac{2}{6} - \frac{2}{6} = 0.33$$

$E(B1) = \cancel{\frac{0}{2}\log\frac{0}{2}}\ 0\ [\because p=0, n=2]$

$E(B3) = 0\ [\because p=1, n=0]$

$E(B4) = 0\ [\because p=1, n=0]$

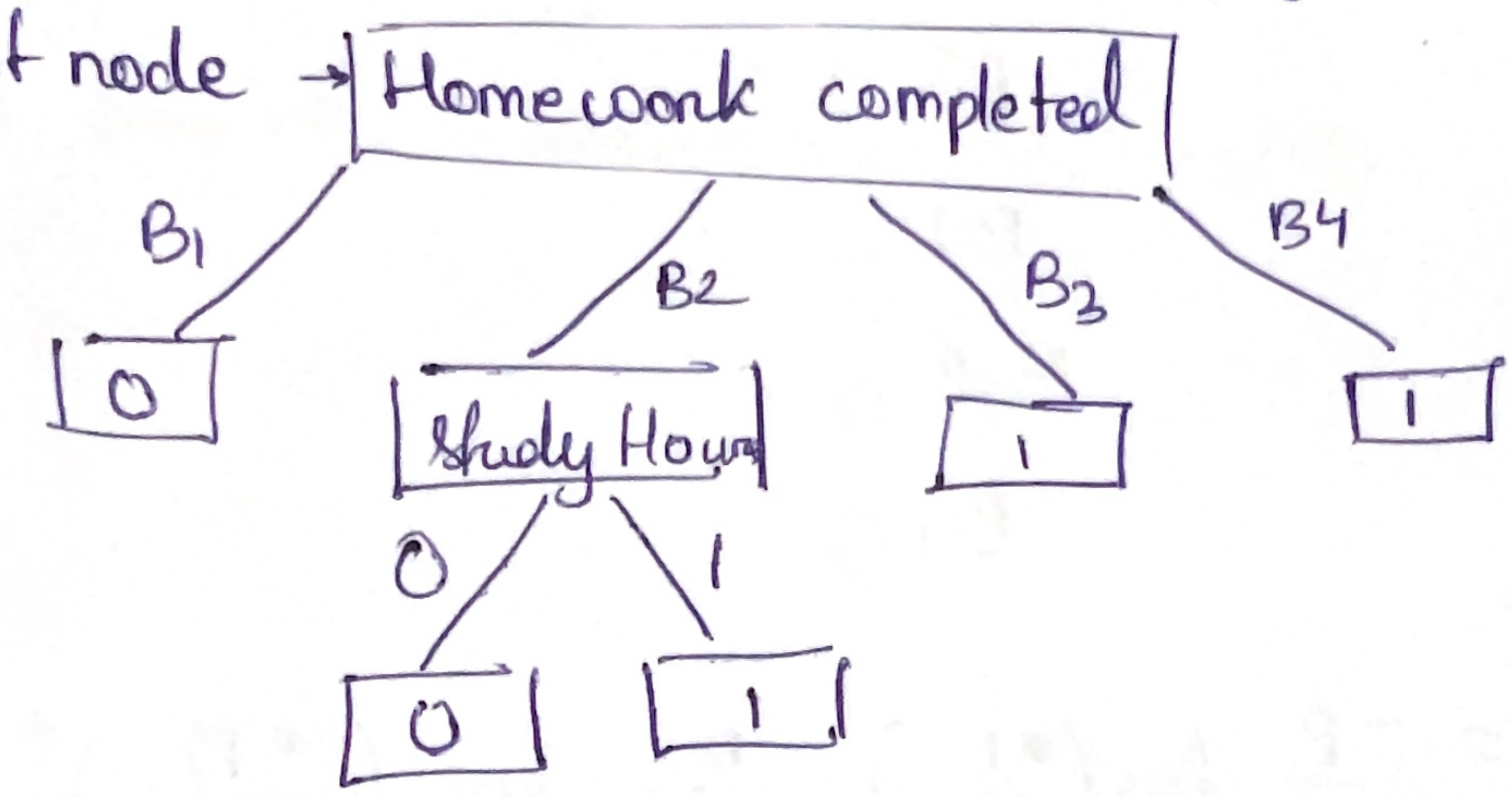
$$E(B2) = \frac{-1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1$$

Information gain of ~~Q~~ Homeworks completed $= 1 - \frac{2}{6}E(B1) - \frac{2}{6}E(B2) - \frac{1}{6}E(B3) - \frac{1}{6}E(B4)$

$= 1 - \frac{2}{6}(0) - \frac{2}{6}(1) - \frac{1}{6}(0) - \frac{1}{6}(0)$

$= 1 - \frac{2}{6} = 0.66$

∴ Root node →



Q6) Suppose we have the following dataset. In this dataset, there are four attributes. And on the basis of these attributes, make a Decision Tree.

Age	Competition	Type	Profit
Old	Yes	software	Down
Old	No	software	Down
Old	No	hardware	Down
Mid	Yes	software	Down
mid	Yes	hardware	Down
mid	No	hardware	Up
Mid	~~Y~~ No	software	Up
new	Yes	software	Up
new	No	~~softw~~ hardware	Up
new	Yes	software	Up

Ans → Entropy (Whole data) $= -\frac{5}{10}\log(5/10) - 5/10 \log 5/10 = 1$

Entropy (Age-old) $= 0$ [∵ $P=0$, $N=3$]

Entropy (Age-new) $= 0$ [∵ $p=3$, $N=0$]

Entropy (Age-mid) $= -\frac{2}{4}\log\left(\frac{2}{4}\right) - \frac{2}{4}\log\frac{2}{4} = 1$

$$\text{Information Gain}(Age) = E(WD) - \tfrac{3}{10}E(Old) - \tfrac{4}{10}E(mid) - \tfrac{3}{10}E(new)$$

$$= 1 - \tfrac{3}{10}(0) - \tfrac{4}{10}(1) - \tfrac{3}{10}(0)$$

$$= 1 - \tfrac{4}{10} = \tfrac{6}{10} = 0.6$$

For Competition →

$$E(Yes) = -\tfrac{1}{4}\log\tfrac{1}{4} - \tfrac{3}{4}\log\tfrac{3}{4} = 0.811$$

$$E(No) = -\tfrac{4}{6}\log\tfrac{4}{6} - \tfrac{2}{6}\log\tfrac{2}{6} \approx 0.585$$

$$IG(Competition) = E(WD) - \tfrac{4}{10}E(Yes) - \tfrac{6}{10}E(No)$$

$$= 1 - \tfrac{4}{10}(0.811) - \tfrac{6}{10}(0.585)$$

$$= 0.3246$$

For Type →

$$E(Hardware) = -\tfrac{2}{4}\log\tfrac{2}{4} - \tfrac{2}{4}\log\tfrac{2}{4} = 1$$

$$E(Software) = -\tfrac{3}{6}\log\tfrac{3}{6} - \tfrac{3}{6}\log\tfrac{3}{6} = 1$$

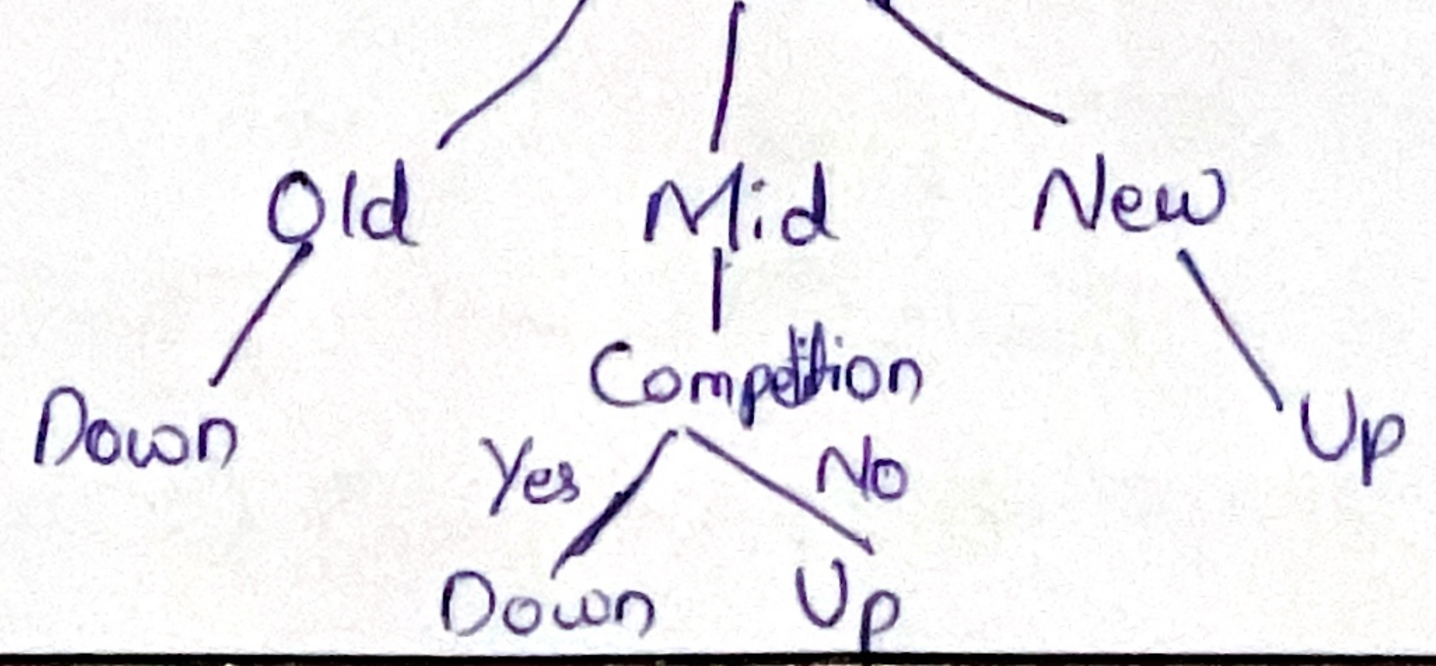
$$IG(Type) = E(WD) - \tfrac{4}{10}E(Hardware) - \tfrac{6}{10}E(Software)$$

$$= 1 - \tfrac{4}{10}(1) - \tfrac{6}{10}(1)$$

$$= 0$$

$$IG(AGE) > IG(Competition) > IG\ Age$$

Root node →



7) Derive the Mean Squared Error function in matrix form for the Linear Regression Model, i.e. Derive $L(\omega) = \frac{1}{m}(X\omega - Y_{true})^T(X\omega - y_{true})$

Ans → $MSE = \frac{1}{m}(|y^1_{predict} - y^1_{true}| + |y^2_{predict} - y^2_{true}| + |y^3_{pred} - y^3_{true}| + \dots + |y^m_{pred} - y^m_{true}|)$

$= \frac{1}{m}\sum_{i=1}^{m}(y^i_{pred} - y^i_{true})$

$= \frac{1}{m}(\vec{y}_{pred} - \vec{y}_{true})^T(\vec{y}_{pred} - \vec{y}_{true})$

$= \frac{1}{m}\|\vec{y}_{pred} - \vec{y}_{true}\|^2_1$

As $\|y\| = \sqrt{y_1^2 + y_2^2 + y_3^2 + \dots + y^2_m}$

$y^t = (y_1 \dots y_m)$

$y^1_{pred} = w_0 + w_1 x_1^1 + w_2 x_2^1 + \dots + w_m x^1_m$

$y^m_{pred} = w_0 + w_1 x_1^m + w_2 x_2^m + \dots + w_m x^m_m$

$$\therefore \begin{bmatrix} y^1_{pred} \\ y^2_{pred} \\ \vdots \\ y^m_{pred} \end{bmatrix} = \begin{bmatrix} 1 & x_1^1 & x_2^1 & \dots & x^1_m \\ 1 & x_1^2 & x_2^2 & \dots & x^2_m \\ 1 & x_1^3 & x_2^3 & \dots & x^3_m \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_1^m & x_2^m & \dots & x^m_m \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_3 \end{bmatrix}$$

$\vec{y}_{pred} = X\vec{w}$ (Thus proved)

8) Derive the minimum value for the MSE $L(\omega)$ using gradient i.e. Derive $\omega = (X^TX)^{-1}X^Ty_{true}$.

Ans → $L(\vec{w}) = \frac{1}{m}(X\vec{w} - \vec{y}_{true})^T(X\vec{w} - \vec{y}_{true})$

$= \frac{1}{m}(\vec{w}^TX^T - \vec{y}_{true}^T)(X\vec{w} - \vec{y}_{true})$

$= \frac{1}{m}(\vec{w}X^TX\vec{w} - \vec{w}^TX^T\vec{y}_{true} - \vec{y}^T_{true}X\vec{w} + \vec{y}^T_{true}\vec{y}_{true})$

$= \frac{1}{m}(\vec{w}^TS\vec{w} - \vec{w}^T\vec{a} - \vec{a}^T\vec{w} + \vec{y}^T_{true}\vec{y}_{true})$ $(\because X^TX = S,\ y_{true}X^T = \vec{a})$

$\Rightarrow \nabla L(\vec{w}) = \frac{1}{m}(2S\vec{w} - 2\vec{a} + 0)$ & $\nabla L(\vec{w}) = 0$

$\therefore \frac{1}{m}(2S\vec{w} - 2\vec{a}) = 0$

$\Rightarrow 2S\vec{w} = 2\vec{a}$

$\Rightarrow \vec{w} = S^{-1}\vec{a}$

$= (X^T X^{-1})(X^T \vec{y}_{true})$ Thus proved

9) Use K-means Algorithm (up to the second iteration) to cluster the following eight points into three clusters. A(2,10), B(2,5), C(8,4), D(5,8), E(7,5), F(6,4), G(1,2), H(4,9) with A(2,10), D(5,8) and G(1,2) as the initial cluster centroids. Consider the following two distance functions in two different cases →

a) The distance function between two points $a = (x_1, y_1)$ & $b = (x_2, y_2)$ is defined as $d(a,b) = |x_1 - x_2| + |y_1 - y_2|$.

b) The distance function between two points $a = (x_1, y_1)$ & $b = (x_2, y_2)$ is defined as $d(a,b) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$

Ans → a)

	Iteration 1			Iteration 2		
Points	A(2,10)	D(5,8)	G(1,2)	X_1(2,10)	X_2(6,6)	X_3(1.5,3.5)
A(2,10)	0	5	9	0	8	7
B(2,5)	5	6	4	5	8	2
C(8,4)	12	7	9	12	4	7
D(5,8)	5	0	9	5	3	8
E(7,5)	10	5	9	10	2	7
F(6,4)	10	5	7	10	2	5
G(1,2)	9	10	0	9	9	2
H(4,9)	3	2	10	3	8	8

Centroid X_1(A) = 2,10

X_2(C, D, E, F, H) = $\left(\frac{8+5+7+6+4}{5}, \frac{4+8+5+4+9}{5}\right) = (6,6)$

X_3(B, G) = $\left(\frac{2+1}{2}, \frac{5+2}{2}\right) = (1.5, 3.5)$

Centroid 1 → X_1(A, H) : $\left(\frac{2+4}{2}, \frac{10+9}{2}\right) = (3, 9.5)$

Centroid 2 $(C,D,E,F) = \left(\frac{8+5+7+6}{4}, \frac{4+8+5+4}{4}\right) = (6.5, 5.25)$

Centroid 3 $(B,G) = \left(\frac{2+1}{2}, \frac{5+2}{2}\right) = (1.5, 3.5)$

b)

	Iteration 1			Iteration 2		
Points	A(2,10)	~~B~~ D(5,8)	G(1,2)	x_1(2,10)	x_2(6,6)	x_3(1.5,3.5)
A(2,10)	0	3.60	8.06	0	5.66	6.52
B(2,5)	5	4.24	3.16	5	4.12	1.58
C(8,4)	8.48	5	7.28	8.48	2.83	6.52
D(5,8)	3.60	0	7.21	3.6	2.24	5.7
E(7,5)	7.07	3.6	6.71	7.07	1.41	5.7
F(6,4)	7.21	4.1	5.38	7.21	2	4.53
G(1,2)	8.06	7.21	0	8.06	6.4	1.58
H(4,9)	2.23	1.41	7.61	2.24	3.6	6.64

x_1 (A) = (2,10)

x_2 $(C, D, E, F, H) = \left(\frac{8+5+7+6+4}{5}, \frac{4+8+5+4+9}{5}\right) = (6,6)$

x_3 $(B, G) = \left(\frac{2+1}{2}, \frac{5+2}{2}\right) = (1.5, 3.5)$

Centroid 1 $(x_1) = (3, 9.5)$

Centroid 2 $(x_2) = (6.5, 5.25)$

Centroid 3 $(x_3) = (1.5, 3.5)$

10) The dataset of pass or fail in an exam of 6 students is given in the ~~middle~~ table.

Study Hour (x)	29	16	33	28	29	30
Pass (1), Fail (0)	0	0	1	0	1	1

Use logistic regression as classifier, calculate the probability of pass for the student who studied 30 hours, where $w = (w_0, w_1) = (-64, 2)$.

Ans→ Sigmoid function $= \frac{1}{1+e^{-x}}$ Here, $x = w_0 + w_1 x_1$

$= -64 + 2(30)$

$= -4$

$= \frac{1}{1+e^{4}}$

≈ 0.018 ∴ Thres $(0.018) = 0$ (Student fails)