CSE343/CSE543/ECE363/ECE563: Machine Learning Sec A (Monsoon 2023)
Quiz - 1

Date of Examination: 29.08.2023   Duration: 45 mins   Total Marks: 20 marks

Instructions –

- Attempt all questions.
- MCQs have a single correct option.
- State any assumptions you have made clearly.
- Standard institute plagiarism policy holds.
- No evaluation without suitable justification.

0 marks if the option or justification of MCQs is incorrect.

1. How can you solve overfitting? - [**1 mark**]

   (A) Gather more training data

   (B) Use a more complex model

   (C) Reduce your training data

   (D) Decrease the model complexity

   (a) B and C.

   (b) A and B.

   (c) A and D.

   (d) C and D

   Solution: c) A and D - because introducing more data and decreasing the model complexity can help the model generalize better on unseen data.

2. Why cost function (Mean Squared Error) which has been used for linear regression can't be used for logistic regression? [**1 mark**]

   1. Because it will be a non-convex function of its parameters. Gradient descent will converge into the global minimum only if the function is convex.
   2. Because it will be a convex function of its parameters. Gradient descent will not converge into a global minimum if the function is convex.
   3. Because it will be a convex function of its parameters. Gradient descent will converge into the global minimum only if the function is non-convex.
   4. None of the above.

   Solution: 1) Because it will be a non-convex function of its parameters. Gradient descent will converge into the global minimum only if the function is convex

3. In the context of Linear Regression, consider a dataset with potential outliers. You aim to estimate the regression coefficients using robust methods. Which statement accurately differentiates between Maximum Likelihood Estimation (MLE) and Maximum A Posteriori (MAP) estimation regarding robustness and outliers? [**1 mark**]

1. MLE is robust to outliers due to its reliance on prior information, while MAP is sensitive to outliers as it directly maximizes the likelihood of observed data.
2. Both MLE and MAP are robust to outliers due to their ability to handle data points deviating from the model.
3. MLE is sensitive to outliers as it directly maximizes the likelihood of observed data, while MAP can be more robust by incorporating a robust prior distribution.
4. MLE is sensitive to outliers due to its focus on prior information, while MAP is robust because it considers the posterior distribution of parameters.
   Solution: (C) MLE finds the optimal parameters that maximizes the likelihood of the sample fitting the regression line. However, if the data sample does not represent the actual population well and contains outliers, the regression line will not generalize well on the actual population and, hence is susceptible to overfitting. On the other hand, MAP overcomes this as it also considers the prior distribution of the parameters.

4. Increasing the complexity of a machine learning model is likely to result in: **[1 mark]**

   1. Higher bias and lower variance
   2. Lower bias and higher variance
   3. Both higher bias and variance
   4. Neither higher bias nor variance

   Solution: b) Lower Bias and Higher Variance. This is because of the reason that as we increase the model complexity, the model will start to fit the training data better, and doing so will reduce the bias. Also, in order to capture more data points better on the training set, the model will lose the overall general trend present, also known as overfitting and therefore, the variance will increase.

5. Let f be some function so that f$(\theta 0, \theta 1)$ outputs a number. For this problem,f is some arbitrary/unknown smooth function (not necessarily the cost function of linear regression, so f may have local optima). Suppose we use gradient descent to try to minimize f$(\theta 0, \theta 1)$ as a function of $\theta 0$ and $\theta 1$. Choose the True statement(s) and Justify it.**[1 mark]**

   1. If the learning rate is too small, then gradient descent may take a very long time to converge.
   2. If $\theta 0$ and $\theta 1$ are initialized at a local minimum, then one iteration will not change their values.
   3. If $\theta 0$ and $\theta 1$ are initialized so that $\theta 0 = \theta 1$, then by symmetry (because we do simultaneous updates to the two parameters), after one iteration of gradient descent, we will still have $\theta 0 = \theta 1$.
   4. Even if the learning rate α is very large, every iteration of gradient descent will decrease the value of f$(\theta 0, \theta 1)$.

   Solution: 1) and 2). Reasoning: If the learning rate is too low, the update will happen slowly, and it will take longer to converge to the optimum. If both parameters are initialized at a local minimum, the gradient would be 0. During their updates;

   $$W_{i+1} = W_i - \eta(\nabla(Loss))$$

   $$\nabla(Loss) = 0$$

   making no change in the updated values. Hence, one iteration won't change their values.

6. You are given two binary classification models for predicting cancer. Both are tested on unseen 2000 samples and the results are given in table. According to medical specifications, a false negative result is 3 times more dangerous than a false positive result and adds up to the medical risk. State which model is better considering the medical risk. Also calculate precision and recall for that model.[**5 marks**]

Model No	TP	FP	TN	FN
Model 1	760	110	910	220
Model 2	780	90	930	200

Solution:

Risk associated with model 1 = 110 + 3*220 = 770 (1 mark)
Risk associated with model 2 = 90 + 3*200 = 690 (1 mark)
Therefore, Model 2 is better as it has lower risk.
We calculate precision and recall for model 2. (1 mark)
Note - 50 % marks will be awarded if the student has not calculated the risk and inferred directly
Precision = (TP/TP+FP) = 780/(780+90) = 0.8966 (1 mark)
Recall = (TP/TP+FN) = 780/(780+200) = 0.7959 (1 mark)

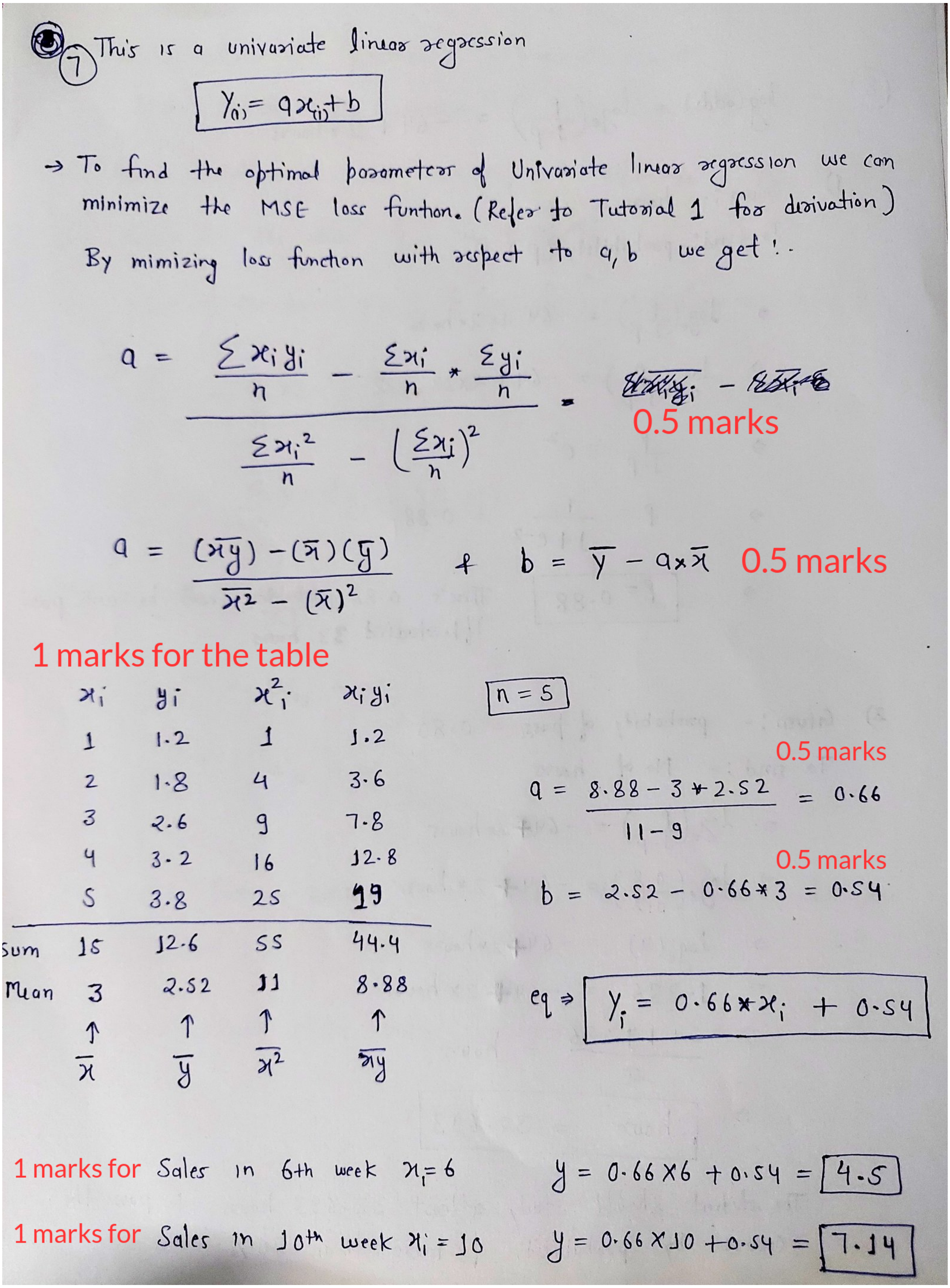
7. Apply Linear Regression techniques on the given dataset and predict the sales in the 6th and 10th week. **Refer to Week vs Sales table for this question.** [**5 marks**]

Week (x)	Sales (y)
1	1.2
2	1.8
3	2.6
4	3.2
5	3.8

Hours study	Pass (1) / Fail (0)
29	0
15	0
33	1
28	1
39	1

8. Use the logistic regression on the dataset to answer the following questions. **Refer to Hours study vs Pass/Fail table for this question**. Given log(odds) or logit = -64 + 2*hours [**5 marks**]

    1. Calculate the probability of a pass for the student who studied 33 hours.
    2. At least how many hours a student should study that will make him pass the course by the probability of more than 80%.



(7) This is a univariate linear regression

$$\boxed{Y_{(i)} = a x_{(i)} + b}$$

→ To find the optimal parameter of Univariate linear regression we can minimize the MSE loss funtion. (Refer to Tutorial 1 for derivation)

By mimizing loss function with respect to a, b we get :-

$$a = \frac{\frac{\Sigma x_i y_i}{n} - \frac{\Sigma x_i}{n} * \frac{\Sigma y_i}{n}}{\frac{\Sigma x_i^2}{n} - \left(\frac{\Sigma x_i}{n}\right)^2} = \cancel{\Sigma x_i y_i} - \cancel{\Sigma x_i}$$

0.5 marks

$$a = \frac{(\overline{xy}) - (\bar{x})(\bar{y})}{\overline{x^2} - (\bar{x})^2} \quad \& \quad b = \bar{y} - a \times \bar{x}$$

0.5 marks

1 marks for the table

	x_i	y_i	x_i^2	$x_i y_i$
	1	1.2	1	1.2
	2	1.8	4	3.6
	3	2.6	9	7.8
	4	3.2	16	12.8
	5	3.8	25	19
Sum	15	12.6	55	44.4
Mean	3	2.52	11	8.88
	↑	↑	↑	↑
	$\bar{x}$	$\bar{y}$	$\overline{x^2}$	$\overline{xy}$

$\boxed{n = 5}$

0.5 marks

$$a = \frac{8.88 - 3 * 2.52}{11 - 9} = 0.66$$

0.5 marks

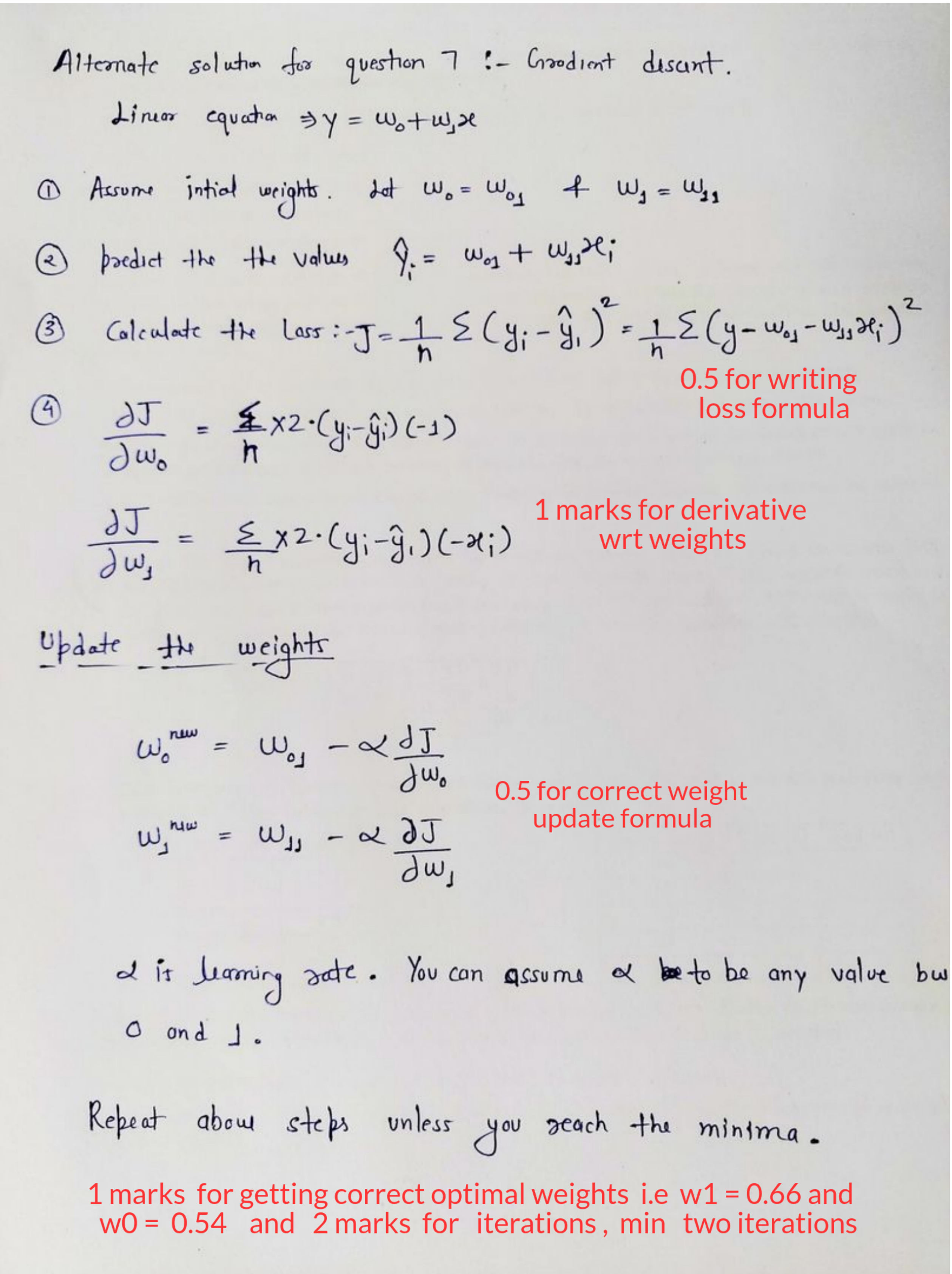
$$b = 2.52 - 0.66 * 3 = 0.54$$

$$eq \Rightarrow \boxed{Y_i = 0.66 * x_i + 0.54}$$

1 marks for Sales in 6th week $x_i = 6$ $\quad y = 0.66 \times 6 + 0.54 = \boxed{4.5}$

1 marks for Sales in 10th week $x_i = 10$ $\quad y = 0.66 \times 10 + 0.54 = \boxed{7.14}$

Figure 1: Q7



Alternate solution for question 7 :- Gradient descent.

Linear equation $\Rightarrow y = w_0 + w_1 x$

① Assume intial weights. let $w_0 = w_{01}$ & $w_1 = w_{11}$

② predict the the values $\hat{y}_i = w_{01} + w_{11} x_i$

③ Calculate the Loss :- $J = \frac{1}{n} \sum (y_i - \hat{y}_i)^2 = \frac{1}{n} \sum (y - w_{01} - w_{11} x_i)^2$

0.5 for writing loss formula

④ $$\frac{\partial J}{\partial w_0} = \frac{\sum}{n} \times 2 \cdot (y_i - \hat{y}_i)(-1)$$

$$\frac{\partial J}{\partial w_1} = \frac{\sum}{n} \times 2 \cdot (y_i - \hat{y}_i)(-x_i)$$

1 marks for derivative wrt weights

Update the weights

$$w_0^{new} = w_{01} - \alpha \frac{\partial J}{\partial w_0}$$

$$w_1^{new} = w_{11} - \alpha \frac{\partial J}{\partial w_1}$$

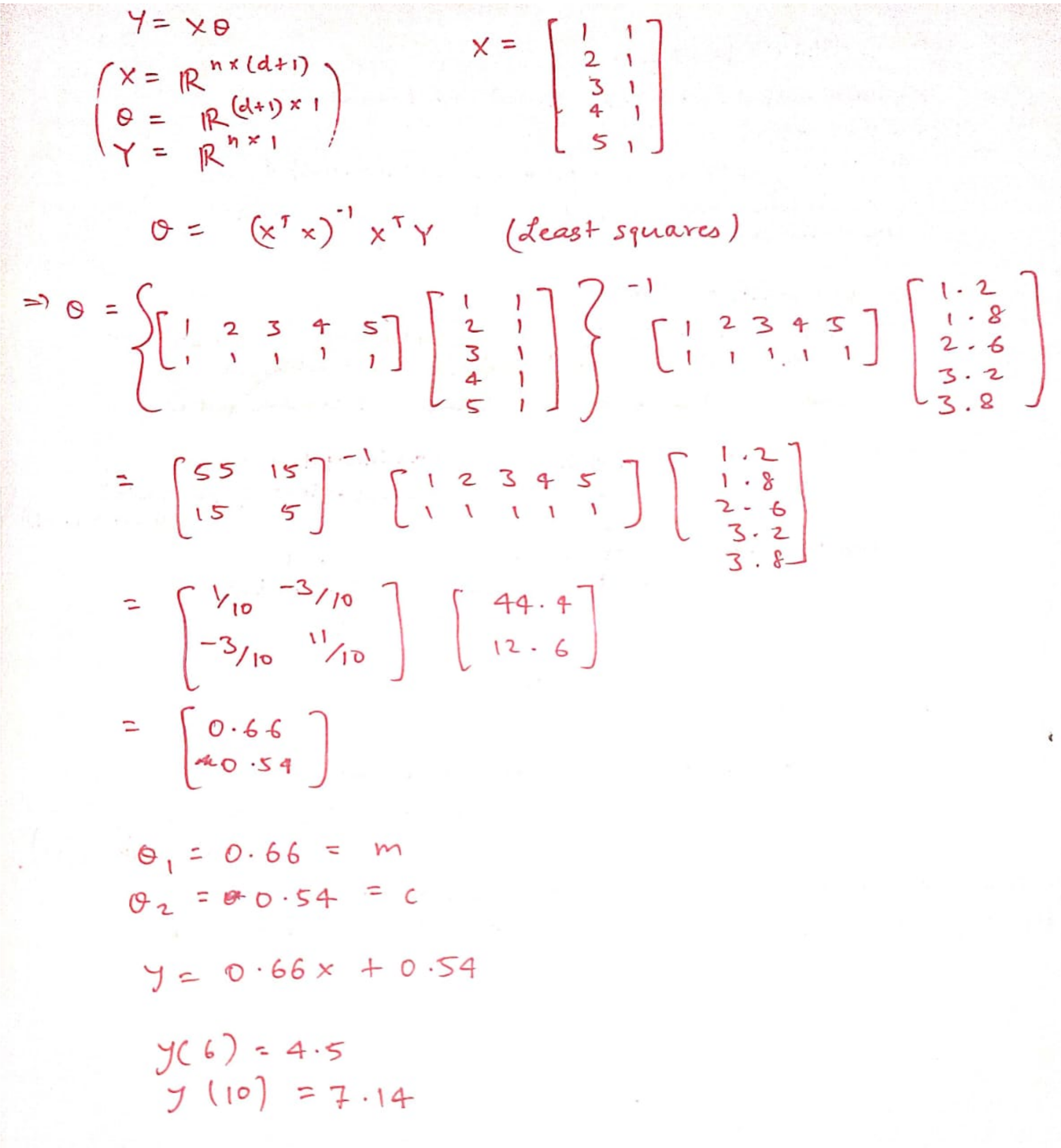
0.5 for correct weight update formula

α is learning rate. You can assume α ~~be~~ to be any value bw 0 and 1.

Repeat abou steps unless you reach the minima.

1 marks for getting correct optimal weights i.e w1 = 0.66 and w0 = 0.54 and 2 marks for iterations, min two iterations

Figure 2: Q7(gradient descent)



$$Y = X\theta$$

$$\left(\begin{array}{l} X = \mathbb{R}^{n \times (d+1)} \\ \theta = \mathbb{R}^{(d+1) \times 1} \\ Y = \mathbb{R}^{n \times 1} \end{array} \right) \qquad X = \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 1 \\ 4 & 1 \\ 5 & 1 \end{bmatrix}$$

$$\theta = (X^T X)^{-1} X^T Y \quad (\text{Least squares})$$

$$\Rightarrow \theta = \left\{ \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 1 \\ 4 & 1 \\ 5 & 1 \end{bmatrix} \right\}^{-1} \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1.2 \\ 1.8 \\ 2.6 \\ 3.2 \\ 3.8 \end{bmatrix}$$

$$= \begin{bmatrix} 55 & 15 \\ 15 & 5 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1.2 \\ 1.8 \\ 2.6 \\ 3.2 \\ 3.8 \end{bmatrix}$$

$$= \begin{bmatrix} 1/10 & -3/10 \\ -3/10 & 11/10 \end{bmatrix} \begin{bmatrix} 44.4 \\ 12.6 \end{bmatrix}$$

$$= \begin{bmatrix} 0.66 \\ 0.54 \end{bmatrix}$$

$$\theta_1 = 0.66 = m$$

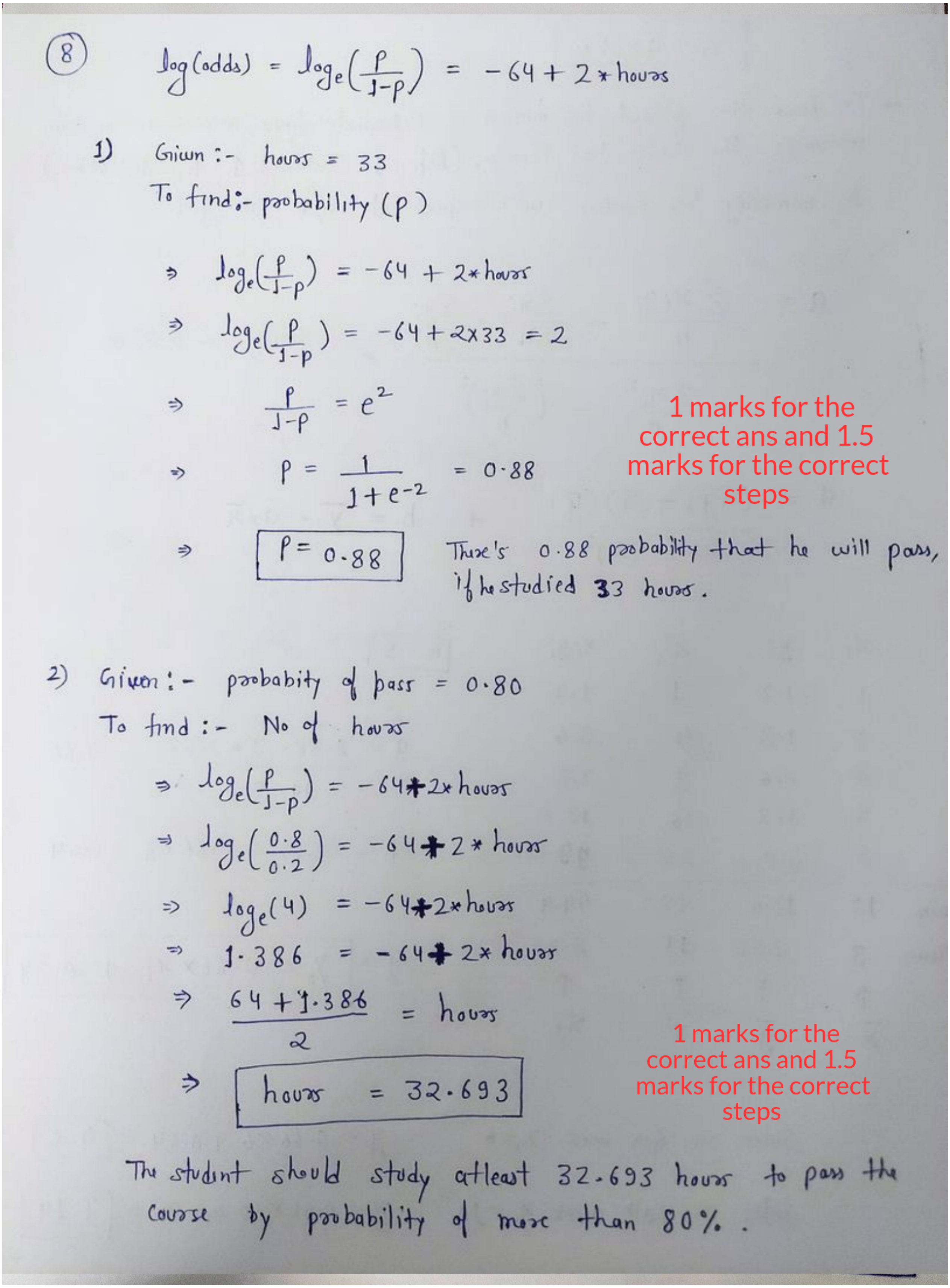
$$\theta_2 = 0.54 = c$$

$$y = 0.66x + 0.54$$

$$y(6) = 4.5$$

$$y(10) = 7.14$$

Figure 3: Q7(least squares closed form)



(8) $\log(\text{odds}) = \log_e\left(\frac{p}{1-p}\right) = -64 + 2 * \text{hours}$

1) Given :- hours = 33

To find:- probability (p)

$\Rightarrow \log_e\left(\frac{p}{1-p}\right) = -64 + 2 * \text{hours}$

$\Rightarrow \log_e\left(\frac{p}{1-p}\right) = -64 + 2 \times 33 = 2$

$\Rightarrow \frac{p}{1-p} = e^2$

$\Rightarrow p = \frac{1}{1+e^{-2}} = 0.88$

$\Rightarrow \boxed{p = 0.88}$ There's 0.88 probability that he will pass, if he studied 33 hours.

1 marks for the correct ans and 1.5 marks for the correct steps

2) Given :- probabity of pass = 0.80

To find :- No of hours

$\Rightarrow \log_e\left(\frac{p}{1-p}\right) = -64 + 2 * \text{hours}$

$\Rightarrow \log_e\left(\frac{0.8}{0.2}\right) = -64 + 2 * \text{hours}$

$\Rightarrow \log_e(4) = -64 + 2 * \text{hours}$

$\Rightarrow 1.386 = -64 + 2 * \text{hours}$

$\Rightarrow \frac{64 + 1.386}{2} = \text{hours}$

$\Rightarrow \boxed{\text{hours} = 32.693}$

1 marks for the correct ans and 1.5 marks for the correct steps

The student should study atleast 32.693 hours to pass the course by probability of more than 80%.

Figure 4: Q8