Name: _____ Reg #: _____ Section: _____

National University of Computer and Emerging Sciences, Lahore Campus

Course:	**Data Science**	**Course Code:**	**CS4048**
Program:	**BS(Computer Science)**	**Semester:**	**Spring 2023**
Duration:	**60 Minutes**	**Total Marks:**	**53**
Paper Date:	**27-Feb-23**	**Weight**	**15 %**
Section:	**B**	**Page(s):**	**3**
Exam:	**Mid-I**		

Instruction/Notes: Attempt the examination on the answer sheets and write concise answers. Clearly write the question number and your answers in the answer Booklet(s) provided. You can use extra sheet for rough work. Do not attach extra sheets used for rough work with the answer sheets. Do not use pencil or red ink to answer the questions. In case of any confusion or ambiguity, make a reasonable assumption.

Question	1	2	3	4	5	6	Total
Marks	/ 10	/ 5	/ 5	/ 8	/15	/10	/ 53

Question # 1:

Let us imagine we have a tree with 15 apples on it. Nine are ripe and six are still unripe, but we do not know which one is which. We have a model which is trained to recognize which apples are ripe for picking, and pick all the ripe apples and no unripe apples. Then model picked 8 apples out of which 6 are ripe and 2 are unripe. We would like to calculate the F-score, and we consider both precision and recall to be equally important. Show complete steps.

Question # 2:

Why we need logistic regression, when we already have linear regression? Can we use linear regression for classification? Support your answer with valid reasons and examples.

Question # 3:

Suppose you have a dataset with n=100 features and m=6000000 examples. You want to train the parameters of this regression problem using linear regression. Should you prefer gradient descent or Normal equation? Support your answer with valid reasons.

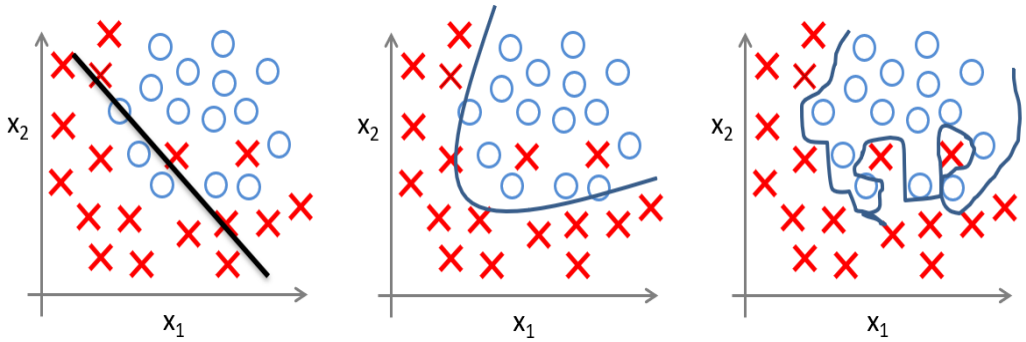
Question # 4:

Part a:

Why we need regularization parameter? What will happen if I take this parameter very large or very small?

Name: _____ Reg #: _____ Section: _____

Part b:



A B C

Suppose, above decision boundaries were generated for the different value of regularization. Which of the above decision boundary shows maximum regularization? Support your answer with valid reasons.

Question # 5: Suppose you have the dataset given below.

Part 1:

X	Y
1	1
2	2
3	3
4	4
5	5

Alpha =0.01

Θ_0= 0.22 (fixed)

Θ_1= {0.8,0.9,1,2}

Your task is to use linear regression and compute all predicted values i.e. ŷ and their corresponding errors and report best Θ_1 value i.e. where error is going to be minimum.

Hint: Θ_0 is fixed but you have to try all Θ_1 values given above.

Part 2:

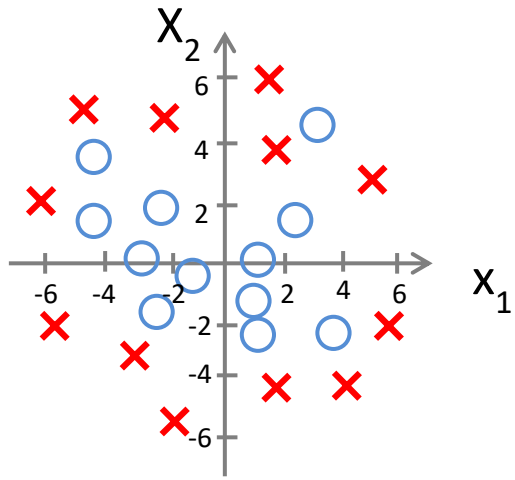
In this part you have to draw cost function for all the errors that you have computed in the part 1. You can take Θ_1 on x-axis and cost on y-axis.

Name: _____ Reg #: _____ Section: _____

Question # 6: We consider the following model of logistic regression for binary classification with a sigmoid function

$$g(z) = \frac{1}{1+e^{-z}}$$

Model: $$h_\theta(x) = g(\theta_0 + \theta_1 x_1^2 + \theta_2 x_2^2 + \theta_3 x_1^2 x_2 + \theta_4 x_1^3 x_2)$$



Suppose the trained parameter values are **Θ_0 = -50, Θ_1 = 2, Θ_2 = 2, Θ_3 = 0,** and **Θ_4 = 0**.

Draw the decision boundary.

Predict "y = 1" if h(x) >= 0.75

Calculate and Draw the decision boundary according to the threshold given above. Show your working as well. If you just draw the boundary without working, you will not get any point.

Question #1

		Act 1	Act 0
Sys	1	6 TP	2 FP
	0	3 FN	4 TN

10

$$P = \frac{TP}{TP + FP} = \frac{6}{6+2} = \frac{6}{8} = 0.75$$

$$R = \frac{TP}{TP + FN} = \frac{6}{6+3} = \frac{6}{9} = \frac{2}{3} = 0.67$$

$$R = \frac{TP}{TP + FN}$$

$$F_1 \text{ Score} = \frac{2PR}{P+R} = \frac{2 \times 0.75 \times \frac{2}{3}}{0.75 + \frac{2}{3}}$$

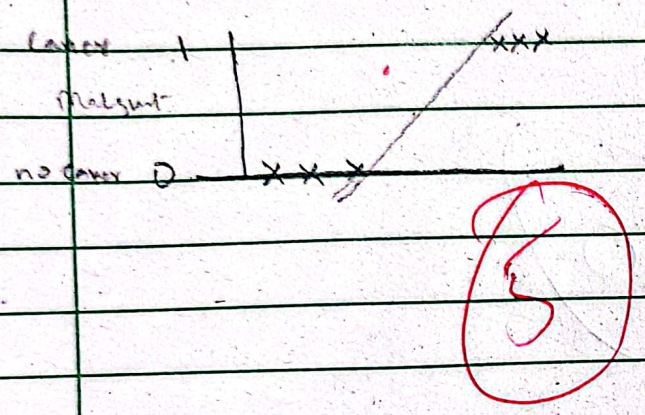
$$= \frac{12}{17} = 0.706$$

Lecture #2

- Linear Regression works for predicting real continuous values
- we need logistic regression to deal with classification problems

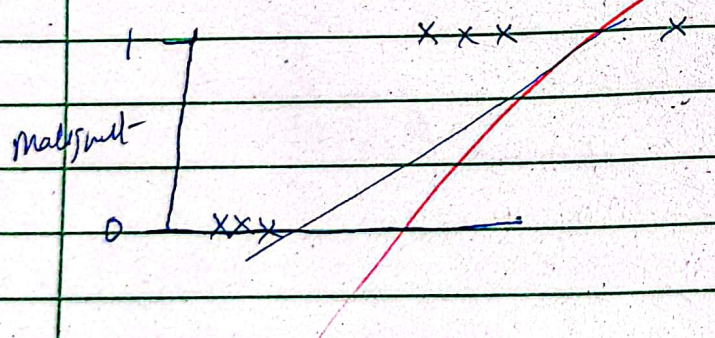
- This is because using linear regression doesnot always give right answers for classification

for eg:



take an example of cancer classification.
if we use linear regression we might get an
$h_\theta(x) = \theta_0 + \theta_1 x$
and use this to decide a threshold that
$< h_\theta(x) <$

Even though this might work here
Lets say we take another training example



our line $h_\theta(x)$ will now look like this and prev assumption of boundary ie $h(x)$ will not be correct.

So as shown by eg this is why we need Logistic Regression for classification problems.

Question #3

$n = 100$ features

$m = 6000\,000$ examples

In this case, the number of training examples are alot more than the number of features

i.e $m >>> n$

~~as~~ We should prefer Normal Equation here as it does not require ~~alot of~~ iterations of training data

as compared to gradient descent which requires alot of iterations for 6000 000 ~~examples~~ dataset.

5

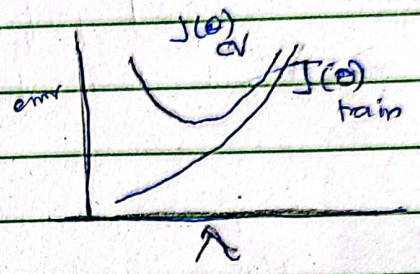
GD	NE
• req α	• does not req α
• req alot of iter	• no iteration
• large n	• large n req Larg $O(n^3)$

Question # 4

a) We need a regularization parameter to solve the problem of overfitting.

$\lambda \rightarrow$ regularization parameter affects how much influence feature x_i will have using θ_j (modifying value of θ_j).

λ very large $\rightarrow$ the model will underfit as influence of feature will decrease due to $x_i \theta_j$



- high bias
- underfit
- high training error & cross validation error.

5

λ very small $\rightarrow$ model will overfit as influence of features increases by $x \theta_j$

- high variance
- overfit
- Low training error, high cross validation error

Part B)

Figure A shows maximum Regularization. This is because regularization is done to decrease overfitting which is least in figure A.

3

This is done by making λ very large so that the influence of features decreases on hypothesis value due to $x \theta_j$ where θ_j parameters are high so error increases

Part No.

Question #5 $\alpha = 0.01$, $\theta_0 = 0.22$, $\theta_1 = \{0.8, 0.9, 1, 2\}$

when $\theta_1 = 0.8$

$h_0(x) = 0.22 + 0.8(1) = 1.02$

$= 0.22 + 0.8(2) = 1.82$

$= 0.22 + 0.8(3) = 2.62$

$= 0.22 + 0.8(4) = 3.42$

$= 0.22 + 0.8(5) = 4.22$

$$\text{Error} = \frac{1}{2m} \sum (h_0(x^i) - y^i)^2$$

$$= \frac{1}{2 \times 5} \left[(1.02-1)^2 + (1.82-2)^2 + (2.62-3)^2 + (3.42-4)^2 + (4.22-5)^2 \right]$$

$= 0.5202$

when $\theta_1 = 0.9$

$h_0(x) = 0.22 + 0.9(1) = 1.12$

$= 0.22 + 0.9(2) = 2.02$

$= 0.22 + 0.9(3) = 2.92$

$= 0.22 + 0.9(4) = 3.82$

$= 0.22 + (0.9)(5) = 4.72$

$$\text{Error} = \frac{1}{10} \left[(1.12-1)^2 + (2.02-2)^2 + (2.92-3)^2 + (3.82-4)^2 + (4.72-5)^2 \right]$$

$= 0.0132$

- when $\theta_0 = 1$

$$h_\theta(x) = 0.22 + 1 = 1.22$$
$$= 0.22 + 2 = 2.22$$
$$= 0.22 + 3 = 3.22$$
$$= 0.22 + 4 = 4.22$$
$$= 0.22 + 5 = 5.22$$

$$\text{Error} = \frac{1}{10}\left[(1.22-1)^2 + (2.22-2)^2 + (3.22-3)^2 + (4.22-4)^2 + (5.22-5)^2\right]$$

$$= 0.0242$$

- when $\theta_1 = 2$

$$h_\theta(x) = 0.22 + 2(1) = 2.22$$
$$= 0.22 + 2(2) = 4.22$$
$$= 0.22 + 2(3) = 6.22$$
$$= 0.22 + 2(4) = 8.22$$
$$= 0.22 + 2(5) = 10.22$$

$$\text{Error} = \frac{1}{10}\left[(2.22-1)^2 + (4.22-2)^2 + (6.22-3)^2 + (8.22-4)^2 + (10.22-5)^2\right]$$

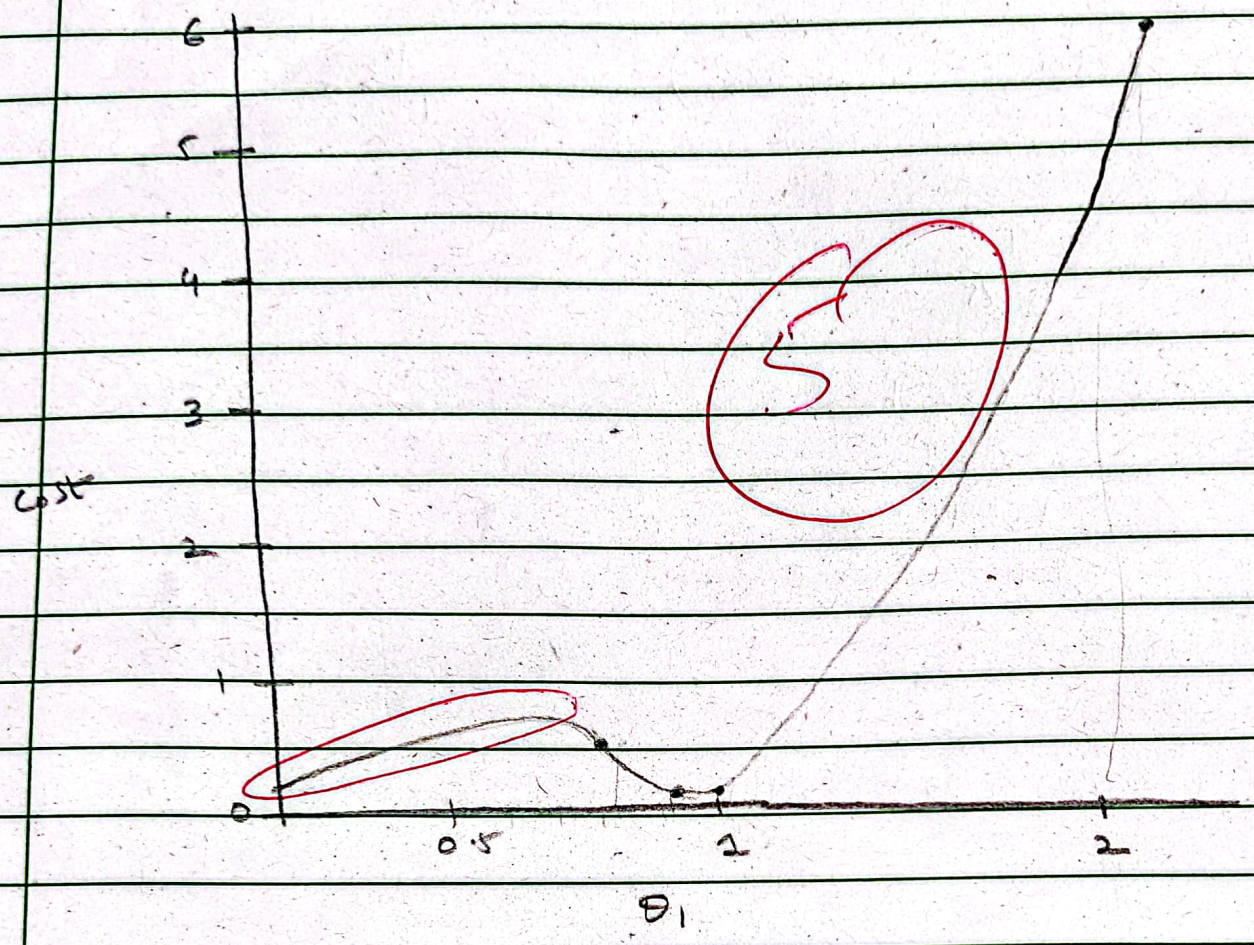
$$= 6.1842$$

Best $\theta_1 = 0.9$

Error $= 0.0132$

Part No.

Part 2



5

$h_\theta(x) = g(-50 + 2x_1^2 + 2x_2^2)$ $\qquad \frac{1}{1+e^{-z}} = h_\theta(x)$

$1+e^{-z} = \frac{1}{0.75}$

$-z = \ln(1/3)$

$z = 1.09$

≈ 1.1

$-2+1 > -4$

$1 > -4+2$

$1 > -2$

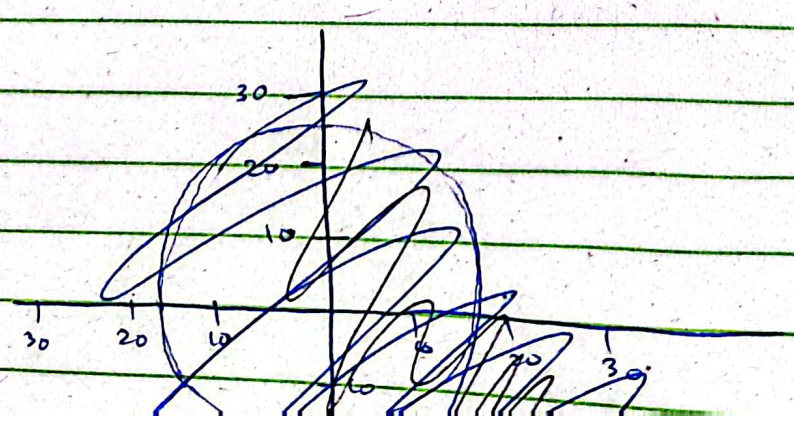
$-50 + 2x_1^2 + 2x_2^2 > 1.1$ then $y=1$

$\frac{2x_1^2 + 2x_2^2}{2} > \frac{1.1+50}{2}$

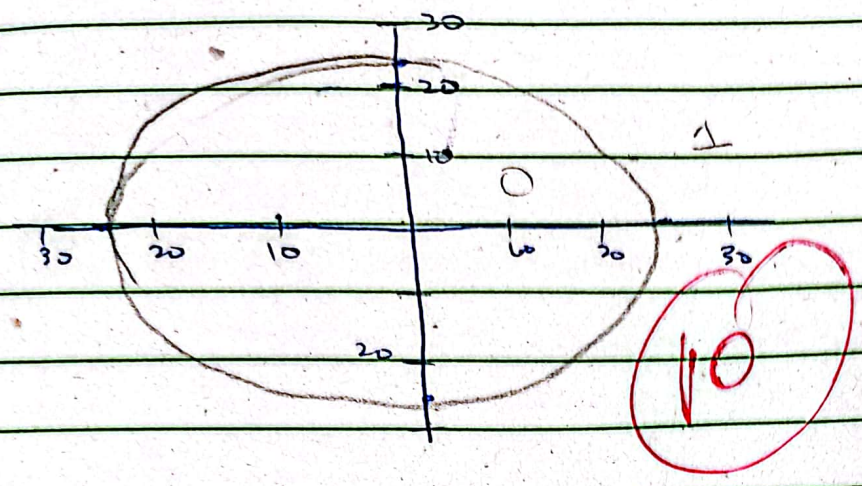
$x_1^2 + x_2^2 > 25.55$

$-50 + 2x_1^2 + 2x_2^2 \leq 1.1$ then $y=0$

$x_1^2 + x_2^2 \leq 25.55$ then $y=0$



t No.



Decision Boundary = Circle of radius 25.55