Name: Fatima Hasan Reg #: 17L-4020 Section: B

National University of Computer and Emerging Sciences, Lahore Campus

Course:	Data Science	Course Code:	CS481
Program:	BS(Computer Science)	Semester:	Fall 2020
Duration:	90 Minutes	Total Marks:	41
Paper Date:	16-Oct-20	Weight	15 %
Section:	A, B	Page(s):	8
Exam:	Mid-I		

Instruction/Notes: Attempt the examination on the question paper and write concise answers. You can use extra sheet for rough work. Do not attach extra sheets used for rough with the question paper. Don't fill the table titled Questions/Marks. Soln & Best

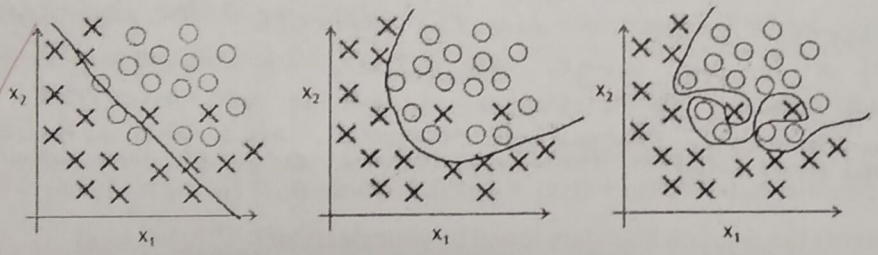
Question	Objective	1	2	3	4	Total
Marks	15 / 16	6 / 6	6 / 6	7 / 7	5.75 /6	40 / 41

Excellent

Section 1 (Objective part) [points 16]

Clearly circle the correct options and explain your choice with reasoning.

Q1. Below are the three scatter plot (A,B, C left to right) and hand drawn decision boundaries for logistic regression.



2

Suppose, above decision boundaries were generated for the different value of regularization. Which of the above decision boundary shows the maximum regularization?

(A) A B) B C) C D) All have equal regularization

Reason: Regularization is used to decrease the values of thetas to make the model simpler. In the above graphs, A is the simplest model, hence maximum regularization.

Q2. Suppose you have a dataset with n = 20 features, and m = 5000000 examples. You want to train the parameters of this regression problem using linear regression. Should you prefer gradient descent or normal equations?

(A) Gradient descent, since $(X^T X)^{-1}$ will be very slow to compute in the normal equation.
B. Gradient descent, since it will always converge to the optimal θ.

$(X^T X)^{-1}$

$(n+1 \times n+1)$

$O(n^3)$

Name: Fatima Haran Reg #: 17L-4020 Section: B

C. The normal equation, since gradient descent might be unable to find the optimal θ.

(D.) The normal equation, since it provides an efficient way to directly find the solution. (correct)

Reason: The training data is very large i.e m = 5000000 hence normal equation would be very slow. Gradient descent would be more efficient in this case

Q3. 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 Θ_0 and Θ_1. Which of the following statements are true? (select all that apply.)

(A) If the first few iterations of gradient descent cause $f(\theta_0, \theta_1)$ to increase rather than decrease, then the most likely cause is that we have set the learning rate α to too large a value.

(B) If the learning rate α is too small, then gradient descent may take a very long time to converge.

(C) Even if the learning rate α is very large; every iteration of gradient descent will decrease the value of $f(\Theta_0, \Theta_1)$.

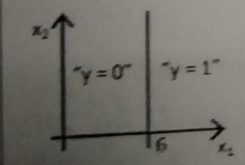
(D) No matter how Θ_0 and Θ_1 are initialized, so long as α is sufficiently small, we can safely expect gradient descent to converge to the same solution.

Reason: A and B are correct since α directly affects the working of gradient descent. If α is too large, the values of Θs may diverge and if it is too small, then GD is slow. C and D are incorrect. If α is too large, cost may increase instead of decrease. Also, the initialization of Θs determines the final solution. For a non convex ~~so~~ graph, the solutions may be different

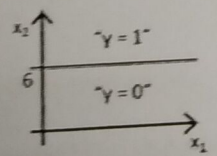
Q4. Suppose you train a logistic classifier $h_\Theta(x) = g(\Theta_0 + \Theta_1 X_1 + \Theta_2 X_2)$. Suppose $\Theta_0 = 6$, $\Theta_1 = -1$, $\Theta_2 = 0$.

Which of following figures represents the decision boundary found by your classifier?

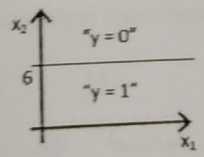
(A)



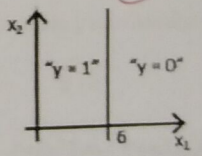
(B)



(C)



(D)



Reason (Explanation): $h_\theta(x) \geq 0.5$
$g(z) \geq 0$
$6 - x_1 \geq 0 \Rightarrow x_1 \leq 6$ → this means that y = 1 when $x \leq 6$

Q5. Which of the following statements are true? Select all that apply.

(A) The one-vs-all technique allows you to use logistic regression for problems in which each $y^{(i)}$ comes from a fixed, discrete set of values.

Name: Fatima Hasan Reg #: 17L-4020 Section: B

(B) The cost function $j(\theta)$ for logistic regression trained with m>=1 examples is always greater than or equal to zero.

(C) Since we train one classifier when there are two classes, we train two classifiers when there are three classes (and we do one-vs-all classification).

(D) Linear regression always works well for classification if you classify by using a threshold on the prediction made by the linear regression.

Reason (Explanation): A: One vs all is used for multiple class classification and the classes are fixed because it is a supervised algorithm. B: The cost function always has a positive value It is a property of the cost function. C: For n classes, we need n classifiers D: Linear regression does not work well for classification

Q6. Imagine, you are solving a classification problems with highly imbalanced class. The majority class is observed 99% of times in the training data. Your model has 99% accuracy after taking the predictions on test data. Which of the following is true in such a case?

A) Accuracy metric is not a good idea for imbalanced class problems.
B) Accuracy metric is a good idea for imbalanced class problems.
C) Precision and recall metrics are good for imbalanced class problems.
D) Precision and recall metrics aren't good for imbalanced class problems.

Reason: F score should be used instead because For skewed data, accuracy is not a good metric. It introduces bias in the model. The model cannot classify the minority class accurately and since majority of the data is skewed, accuracy does not show the complete picture

Q7. Suppose you execute logistic regression twice, once with $\lambda = 1$, and once with $\lambda = 0$. One of the times, you got parameters $\Theta = \begin{vmatrix} 71.25 \\ 11.45 \end{vmatrix}$, and the other time you got $\Theta = \begin{vmatrix} 13.65 \\ 0.85 \end{vmatrix}$. However, you forgot for which values λ corresponds to which value of Θ. Which one do you think to $\lambda = 1$?

a) $\Theta = \begin{vmatrix} 71.25 \\ 11.45 \end{vmatrix}$ b) $\Theta = \begin{vmatrix} 13.65 \\ 0.85 \end{vmatrix}$

Reason: when lambda is greater, it reduces the values of Θs

Q8. Suppose you have m=50 training examples with n=10 features (excluding the additional all-ones feature for the intercept term, which you should add). The normal equation is $\theta=(X^TX)^{-1}X^Ty$. For the given values of m and n, what are the dimensions of θ, X, and y in this equation?

Answer: θ: 11 x 1
X: 50 x 11
y: 50 x 1

Section 2 (Subjective part) (points 25)

Q1. [4+2 Marks]:

A) [4 marks] Suppose you train a logistic regression classifier in order to predict if the patient has cancer or not. Given the test data ($m_{test} = 500$), we already know that 100 patients have actually cancer. On testing, our hypothesis predicted that 40 patients have cancer. Among the predicted ones, only 16 patients are those which actually have cancer.
Draw the confusion matrix, and calculate the precision and recall for the case mentioned above.

Actual

Predicted	$y=1$	$y=0$
$\sim$ 1	16 True +ve	24 False +ve
$\sim$ 0	84 False -ve	376 True -ve

Precision = $\frac{\text{True +ve}}{\text{True +ve + False +ve}}$ = 16/40 = 0.4

Recall = $\frac{\text{True +ve}}{\text{True +ve + False -ve}}$ = $\frac{16}{100}$ = 0.16

4

B) [2 marks] A Data Scientist is evaluating different binary classification models. A false positive result is 5 times more expensive (from a business perspective) than a false negative result. The models should be evaluated based on the following criteria:
1) Must have a recall rate of at least 80%
2) Must have a false positive rate of 10% or less
3) Must minimize business costs

After creating each binary classification model, the Data Scientist generates the corresponding confusion matrix. Which confusion matrix represents the model that satisfies the requirements?

1) TN = 91, FP = 9
FN = 22, TP = 78

2) TN = 99, FP = 1
FN = 21, TP = 79

3) TN = 96, FP = 4
FN = 10, TP = 90

(4) TN = 98, FP = 2
FN = 18, TP = 82

2

1) R = 0.78
2) R = 0.79

3) R = 0.9
FP = 0.02
FN = 0.05
Total cost = 0.15

4) R = 0.82
FP = 0.01
FN = 0.09
Total cost = 0.14

[5 marks] Suppose m = 4 students have taken some machine learning course, and the course had a midterm exam and a final exam. You have collected a dataset of their scores on the two exams, which is as follows:

midterm exam	final exam
10	12
8	8
12	11
7	9

You would like to use linear regression to predict a student's final exam score from their midterm exam score. Concretely, suppose you want to fit a model of the form $h_\theta(x) = \theta_0 + \theta_1 X_1$, where X_1 is midterm score. Assume $\alpha = 0.1$, and initial $\theta_0 = 2$ and initial $\theta_1 = 3$.

The definition of the cost function is $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$.

Your task is to execute gradient descent algorithm and compute updated values of thetas (θ_0 and θ_1) and associated cost (J) for first iteration.

$$temp0 = \theta_0 - \frac{\alpha}{m} \sum_{i=1}^{m} (h_0(x^{(i)}) - y^{(i)})$$

$$= 2 - \frac{0.1}{4} \sum_{i=1}^{m} (\theta_0 + \theta_1 x_1 - y^{(i)})$$

$$= 2 - \frac{0.1}{4} (20 + 18 + 27 + 14) = 0.025$$

$$temp1 = \theta_1 - \frac{\alpha}{m} \sum_{i=1}^{m} (h_0(x^{(i)}) - y^{(i)}) \cdot x^{(i)}$$

$$= 3 - \frac{0.1}{4} (200 + 144 + 324 + 98)$$

$$= -16.15$$

6

$\theta_0 = 0.025$

$\theta_1 = -16.15$

$J(\theta_0, \theta_1) = \frac{1}{2 \times 4} (30094 + 18817 + 41933 + 14890)$

$= 13216.8$?

Q3. Multiclass Classification: (1+2+3+1 Marks)

In the table given below, we have labeled data for patients. The output is classified into classes as given below:

y = 1 if patient is "**not- ill**", y = 2 if patient has "**cold**", and y = 3 if patient has "**Flu**"

X_1	X_2	y
2	2	1
4	2	1
2	8	2
3	7	2
7	6	3
8	8	3

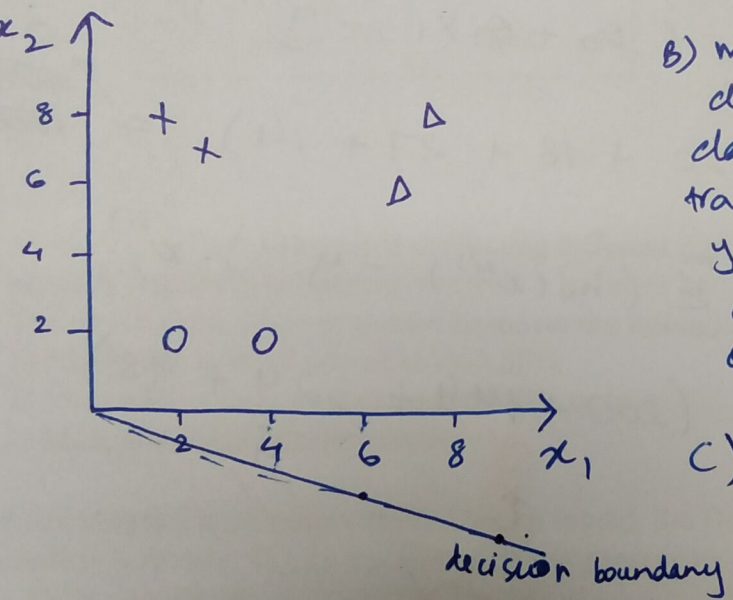
(A) Draw the plot for the training data, where each class should be represented by a different symbol. (keep X_1 on x-axis and X_2 on y-axis)

(B) How we will train logistic regression classifiers for this data?

(C) Suppose trained parameters have same values for all models ($\Theta_0 = 0$, $\Theta_1 = 1$, $\Theta_2 = 3$), draw all the decision boundaries. (Assume simple linear Hypothesis $h(x) = \Theta_0 + \Theta_1 X_1 + \Theta_2 X_2$. Predict y=1 if $h(x) >= 0.5$)

(D) On new input x (new patient), how we will predict if the patient has "flu", "cold" or is "not-ill".

A)



B) We will train 3 different classifiers, one for each class. For y=1, we will train a classifier which uses y=0 for the other two classes. Same for the other classes as well

C) $X_1 + 3X_2 \geq 0$

The decision boundary would be same for all since Os are same. Above the boundary class would be predicted +ve

D) We will pass the input to all 3 classifiers. Whichever has greatest value of $h_\theta(x)$ will be the prediction

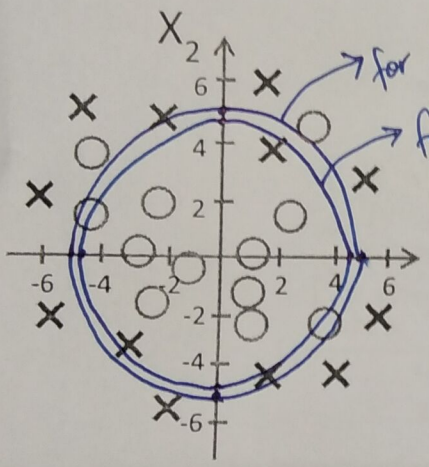
7

Name: Fatima Hasan Reg #: 17L-4020 Section: B

Q4. [6 marks] We consider the following model of logistic regression for binary classification with a sigmoid function

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

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



for $h_\theta(x) \geq 0.5$

for $h_\theta(x) \geq 0.75$

Suppose the trained parameter values are $\Theta_0 = -50$, $\Theta_1 = 2$, $\Theta_2 = 2$, $\Theta_3 = 0$, and $\Theta_4 = 0$.

Draw the decision boundary. $-50 + 2x_1^2 + 2x_2^2 \geq 0$

$x_1^2 + x_2^2 \geq 25$

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

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

on next page

$h(x) \geq 0.75$

$g(z) \geq 0.75$ where $z = \theta^T x$

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

$$\frac{4}{3} \geq 1 + e^{-z}$$

$$e^{-z} \leq \frac{1}{3}$$

$$-z \ln e \leq \ln(1/3)$$

$$z \geq -1.0986$$

$$\theta^T x \geq +1.0986$$

$$-50 + 2x_1^2 + 2x_2^2 \geq -1.0986$$

$$x_1^2 + x_2^2 \geq 24.45$$

25...

radius = 4.94