Name: _____ Reg #: _____ Section: _____

National University of Computer and Emerging Sciences, Lahore Campus

Course:	Data Science	Course Code:	CS481
Program:	BS(Computer Science)	Semester:	Spring 2018
Duration:	60 Minutes	Total Marks:	20
Paper Date:	28-Feb-18	Weight	13 - 15 %
Section:	ALL	Page(s):	6
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.

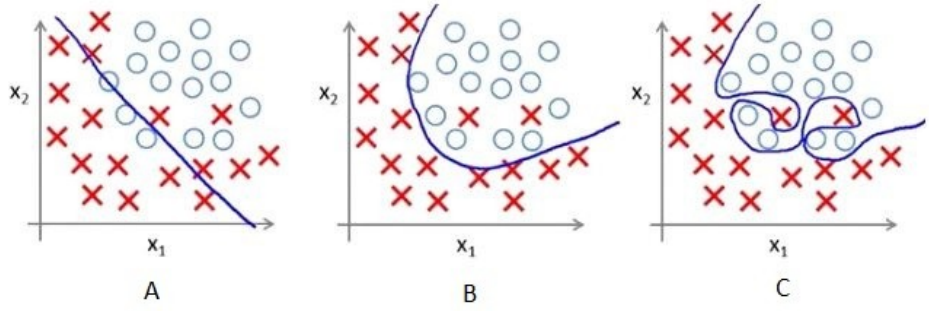
Question	Objective	1	2	3	Total
Marks	/ 8	/ 4	/ 4	/ 4	/ 20

Section 1

(Objective part) [points 8]

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.



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:

Name: _____ Reg #: _____ Section: _____

Q2. Consider a following model for logistic regression: P (y =1|x, w)= g(w0 + w1x)

where g(z) is the logistic function.

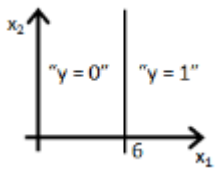
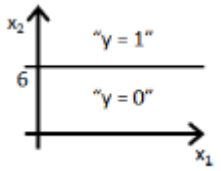
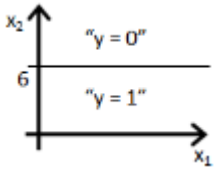
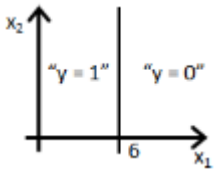
In the above equation the P (y =1|x; w) , viewed as a function of x, that we can get by changing the parameters w.

What would be the range of p in such case?

A) (0, inf) B) (-inf, 0) C) (0, 1) D) (-inf, inf)

Reason:

Q3. Suppose you train a logistic classifier $\mathbf{h_\theta(x) = g(\theta_0 + \theta_1 X_1 + \theta_2 X_2)}$. 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):

Q4. Which of the following statements are false? Select all that apply.

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

(B) The cost function ***j(θ)*** 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.

Name: _____ Reg #: _____ Section: _____

Reason:

Q5. Focusing on describing or explaining data versus going beyond immediate data and making inferences is the difference between _____.

a. Central tendency and common tendency b. Mutually exclusive and mutually exhaustive properties

c. Descriptive and inferential d. Positive skew and negative skew

Reason:

Q6. The _____ is often the preferred measure of central tendency if the data are severely skewed.

a. Mean b. Median c. Mode d. Range

Reason:

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

a) Θ ¿ $\begin{vmatrix} 71.25 \\ 11.45 \end{vmatrix}$ b) Θ ¿ $\begin{vmatrix} 13.65 \\ 0.85 \end{vmatrix}$

Reason:

Q8. Suppose you have m=20 training examples with n=4 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?

Name: _____ Reg #: _____ Section: _____

Answer:

Section 2 (Subjective part) (marks 12)

Q1. Multiclass Classification: (1+2+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**"

$\mathbf{x_1}$	$\mathbf{x_2}$	**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 $\mathbf{x_1}$ on x-axis and $\mathbf{x_2}$ on y-axis)

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

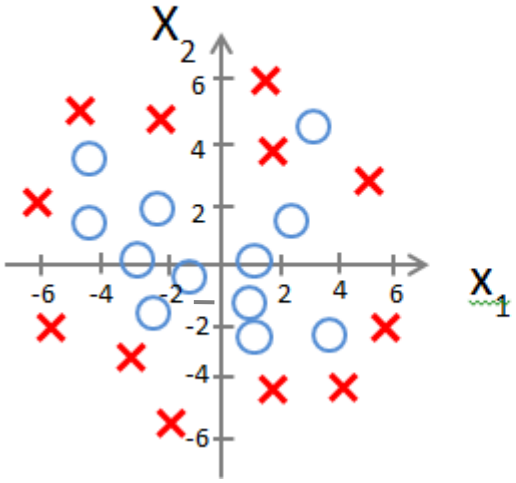
Name: _____ Reg #: _____ Section: _____

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

Q2: [4 points] Logistic Regression – Decision Boundary:

We consider the following model of logistic regression for binary classification with a sigmoid function as discussed in the course.

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



Suppose the trained parameter values are **Θ_0 = -1, Θ_1 = 1, Θ_2 = 1, Θ_3 = 0,** and **Θ_4 = 0**.

Predict ‘’y = 1’’ if h(x) >= 0.25

Name: _____ Reg #: _____ Section: _____

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.

Q3. [4 Marks] **Short Questions**:

A) **[**2 points**]** In classification problems, why do we use equation (ii) for cost instead of equation (i) given below. Here in both equations h(x) is a sigmoid logistic function.

$\text{Cost}(h_\theta(x^{(i)}), y^{(i)}) = \frac{1}{2}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$	Equation (i)
$\text{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$	Equation (ii)

Name: _____ Reg #: _____ Section: _____

B) **[**2 points**]** What is the **overfitting problem** and what can be the possible cause for this problem? Write down all possible options for addressing the overfitting problem.