


National University of Computer and Emerging Sciences, Lahore Campus

	Course Name:	Introduction to Data Science	Course Code:	DS2001
	Degree Program:	BSDS	Semester:	Fall 2021
	Exam Duration:	60 Minutes	Total Marks:	21
	Paper Date:	1-12-2021	Weight	15 %
	Section:	ALL	Page(s):	8
	Exam Type:	Midterm-II		

Student : Name: _____ **Roll No.** _____ **Section:** _____

Instruction/Notes: Attempt the examination on the question paper and write concise answers. Extra pages are provided for rough work at the end. Do not attach extra sheets with the question paper. Don't fill the table titled Questions/Marks.

Question	Objective	1	2	3	Total
Marks	/ 6	/ 8	/ 4	/ 3	/ 21

Section 1 (Objective part) [points 6]

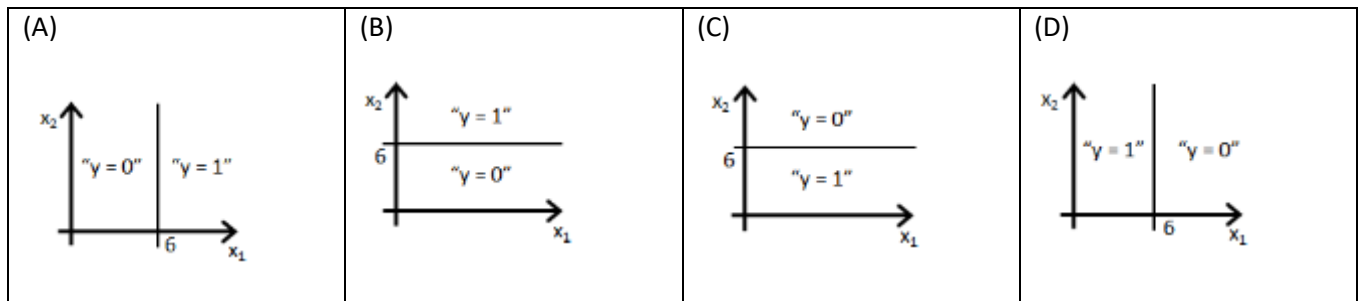
Clearly circle the correct options and explain your choice with reasoning.

Q1. Regularization usually decreases the training error?

- a) False b) True

Reason:

Q2. Suppose you train a logistic classifier $h_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$. Suppose $\theta_0 = -6$, $\theta_1 = 0$, $\theta_2 = 1$. Which of following figures represents the decision boundary found by your classifier?



Compute the Decision Boundary:

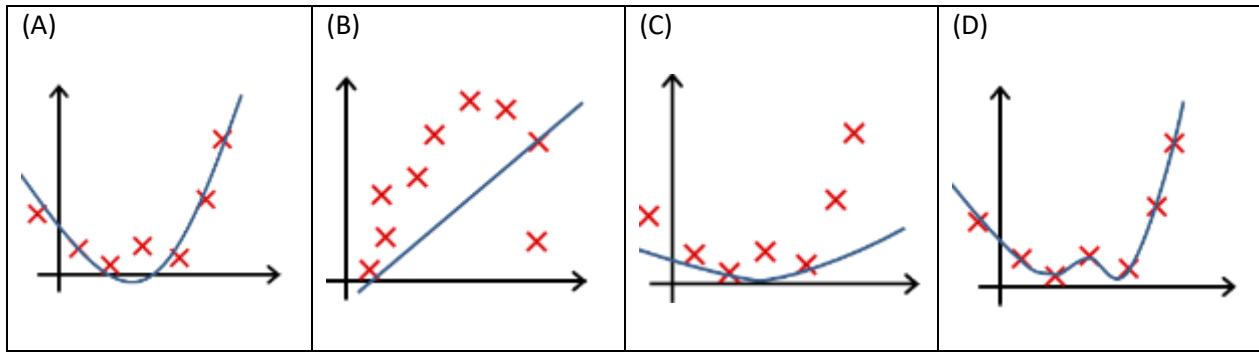
Q3: You are training a classification model with logistic regression, which of the following statement are true. Select all that apply.

- (A) Introducing regularization to the model always results in equal or better performance on examples not in the training set.
- (B) Adding many new features to the model helps prevent overfitting on the training set.
- (C) Adding many new features to the model makes it more likely to overfit the training set.
- (D) Adding a new feature to the model always results in equal or better performance on examples not in the training set.

Q4. Which of the following statement about regularization are true. Select all that apply.

- (A) Using too large value of λ can cause your hypothesis to overfit the data; this can be avoided by reducing λ .
- (B) Using too large value of λ can cause your hypothesis to underfit the data.
- (C) Using too small value of λ can cause your hypothesis to overfit the data.
- (D) Using very large value of λ cannot hurt the performance of your hypothesis; the only reason we do not set λ to be too large is to avoid numerical problems.

Q5. In which of the following figure do you think the hypothesis is over-fitting the training set?



Q6. Suppose that you have trained a logistic regression classifier, and it outputs a new example \mathbf{x} a prediction $\mathbf{h}_{\theta}(\mathbf{x}) = 0.6$. This means (select all that apply):

- (A) our estimate for $P(y = 0 \mid \mathbf{x}; \theta)$ is 0.4
- (B) our estimate for $P(y = 0 \mid \mathbf{x}; \theta)$ is 0.6
- (C) our estimate for $P(y = 1 \mid \mathbf{x}; \theta)$ is 0.6
- (D) our estimate for $P(y = 1 \mid \mathbf{x}; \theta)$ is 0.4

Section 2 (Subjective part) (marks 12)

Q1. [8 Marks] Short Questions:

A) [1.5 points] **Bias and Variance:** A set of data points is generated by the following process: $Y = w_0 + w_1X_1 + w_2X_2 + w_3X_3 + w_4X_4$, where X is a real-valued random variable. You use two models to fit the data:

Model 1: $Y = aX + b$

Model 2: $Y = w_0 + w_1X_1 + \dots + w_9X_9$

i. Model 1, when compared to Model 2 using a fixed number of training examples, has a bias which is:

- (a) Lower
- (b) Higher
- (c) The Same

ii. Model 1, when compared to Model 2 using a fixed number of training examples, has a variance which is:

- (a) Lower
- (b) Higher
- (c) The Same

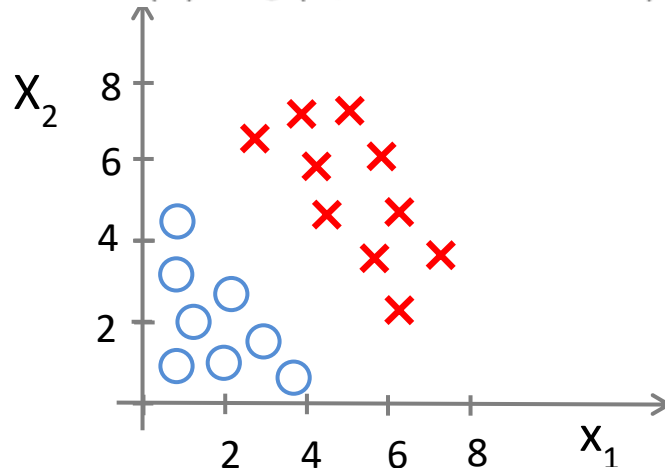
iii. Given 10 training examples, which model is more likely to overfit the data?

- (a) Model 1 (b) Model 2 (c) cannot say (d) none

B) [2 marks] Logistic Regression – Decision Boundary:

We consider the following model of logistic regression for binary classification with a sigmoid function

Model: $h_{\theta}(x) = g(\theta_0 + \theta_1 x_1^2 + \theta_2 x_2)$



Suppose the trained parameter values are $\theta_0 = -8$, $\theta_1 = 2$ AND $\theta_2 = 2$

Predict "y = 1" if $h(x) \geq 0.25$

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.

- a) [2.5 points] Suppose you have implemented regularized linear regression to predict housing prices. However, when you test your hypothesis in a new set of houses, you find that it makes unacceptably large errors (high bias or high variance) in its prediction. You can try some of the options given in the first column in order to fix the problem. Mark 'y' in the second or third column for all 6 options.

If you try	Fixes high bias	Fixes High variance
Get more training examples		
Try smaller sets of features		
Try decreasing λ		
Try increasing λ		
Try getting additional features		
Try adding polynomial features		

- c) [2 marks] Suppose you train a logistic regression classifier in order to predict if the patient has cancer or not. Given the test data ($m_{\text{test}} = 100$), we already know that 20 patients have actually cancer. On testing, our hypothesis predicted that 26 patients have cancer. Among the predicted ones, only 12 patients are those which actually have cancer.

Calculate the precision and recall for the case mentioned above.

Q2. [4 Marks] A friend of yours is faced with a regression problem with two possible inputs, X_1 and X_2 . he/she considers three linear regression models:

(Model 1) $h(x) = \theta_0 + \theta_1 X_1$

(Model 2) $h(x) = \theta_0 + \theta_1 X_1 + \theta_2 X_2$

The data set is given in the following table:

X_1	X_2	Y
2	2	18
3	3	24
1	4	15
5	3	32
4	2	20
6	6	23
5	4	12
10	8	25

Training Data: $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), (x^{(3)}, y^{(3)}), (x^{(4)}, y^{(4)})\}$

Validation Data: $\{(x^{(5)}, y^{(5)}), (x^{(6)}, y^{(6)})\}$

Test Data: $\{(x^{(7)}, y^{(7)}), (x^{(8)}, y^{(8)})\}$

All models are fitted to a training data set using mean-squared-errors, resulting in the three prediction models respectively:

(P1) $h(x) = 12 + 3X_1$,

(P2) $h(x) = 11 - X_1 + 2X_2$

Your friend is puzzled by these results and comes to you for advice.

- What do you think, which model will be the best?
- How well does the model generalize?

Q3. [3 marks] Data Wrangling: Suppose you want to train a model to predict the sale price of a house, given its size, number of stories and number of bedrooms. In order to build the prediction model, you need data for training, which you have collected from local property dealers. After looking at the dataset, you realize that some of the data is missing and you want to handle the missing data before you use this it for the training. The data set is given in the following table:

X1 (Size of house in Marla's)	X2 (number of stories)	X3 (number of bed rooms)	Y (Price in millions)
10	2	4	100
20	2	8	150
30	1	6	200
1	1		10
100	3		400

Your task is to impute the missing values of X3.

A) Impute using Mean

- B) Impute using Linear Regression. For this you can assume a linear Model: $h(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$, where $h(x)$ is predicted number of bed rooms. Suppose your model is fitted to the training data set using mean-squared-errors, resulting in the following prediction model:
 $h(x) = 1 + 0.3x_1 + 0.5x_2$. Your task is to impute the missing values.
- C) What do you think, what will be the advantage and disadvantage of imputation using Mean? Explain your answer according to the dataset given above.

