Name: _____ Reg #: _____ Section: _____

National University of Computer and Emerging Sciences, Lahore Campus

Course:	Data Science	Course Code:	DS2001
Program:	BS (Data Science)	Semester:	Fall 2021
Duration:	30 Minutes	Total Marks:	10
Paper Date:	26-Nov-21	Weight	
Section:	B	Page(s):	2
Quiz:	2		

Instruction/Notes: Attempt the quiz on the question paper and write concise answers.

Marks	
Total	10

Q1. [1 points] 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.

Q2. [1 point] The model high variance (overfitting) typically tends to reduce as the number of training data points tends to infinity? Explain your choice with reasoning. a) False b) True

Reason: **An increase in data set size leads to a reduction in outliers/noise during averaging. Model makes fewer assumptions due to availability of more data, leading to a decision boundary of best fit.**

Q3. [3 points] Suppose you train a logistic regression classifier in order to predict if the aircraft engine is faulty or not. Our model predicts 1 if h(x) > 0.6. Given the test data (m_{test} = 250), we already know that 40% of the aircrafts have actually fault. On testing, our hypothesis predicted that 30% of the aircrafts have fault. Only 50% of the predicted ones (it's not 50% of the total), which actually have fault.

b. If we want to predict faculty engines only if we are very confident, what we will do (How we will change the threshold)? **We are essentially trying to increase precision i.e. TP/(TP+FP), by decreasing the number of False Positives and increasing the number of True Positives. This can be achieved by increasing the threshold e.g to h(x) > 0.9**

a. Create a confusion matrix with actual number of true positive, true negative, false positive and false negative examples. Moreover, Calculate the Precision, Recall, and F score.

	Actual y = 1	Actual y = 0
predicted ŷ = 1	True +ve 37.5	False +ve 37.5
predicted ŷ = 0	False -ve 62.5	True -ve 112.5

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{37.5}{37.5 + 37.5} = \frac{37.5}{75} = 0.5$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{37.5}{37.5 + 62.5} = \frac{37.5}{100} = 0.375$$

$$\text{F-score} = \frac{2PR}{P+R} = \frac{2 \times 0.5 \times 0.375}{0.5 + 0.375} = \frac{0.375}{0.875} = 0.429$$

Q4. [2 marks] **Diagnosing bias vs. variance: Answer the following questions:**

(1). If $J_{cv}(\theta)$ and $J_{train}(\theta)$ are high such that ($J_{cv}(\theta) \approx J_{train}(\theta)$). Is it a bias problem or variance problem?

Bias Problem: High Bias/underfitting means the model returns a high training error and a high cross validation/testing error.

(2). If $J_{train}(\theta)$ is low and $J_{cv}(\theta) >> J_{train}(\theta)$. Is it a bias problem or variance problem?

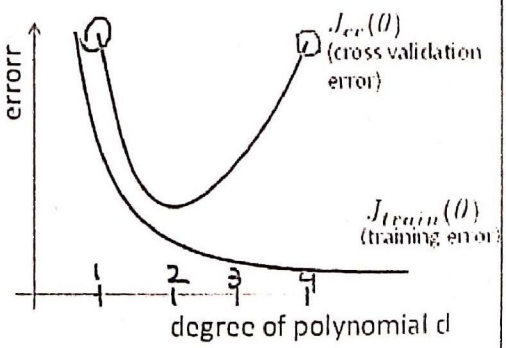
Variance Problem: High Variance/overfitting means the model returns a low training error but a high cross validation/testing error.

(3). For what value of d (degree of polynomial), the problem is underfit?

At d=1 both training and CV errors are high.

(4). For what value of d (degree of polynomial), the problem is overfit?

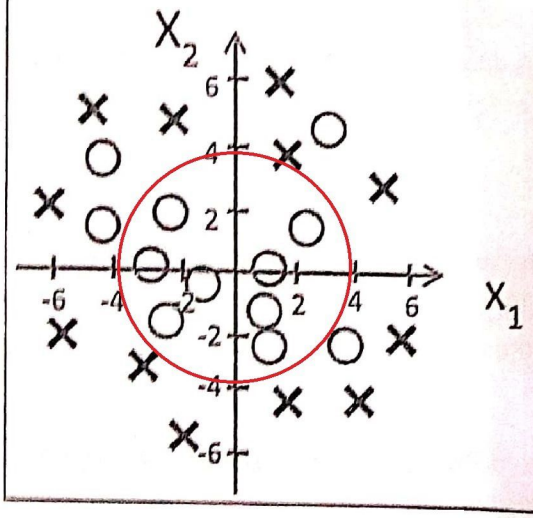
At d=4 training error is low but CV error is high.



Q5: [3 points] 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 $\Theta_0 = -30$, $\Theta_1 = 2$, $\Theta_2 = 2$, $\Theta_3 = 0$, and $\Theta_4 = 0$.

Predict "y = 1" if h(x) >= 0.5

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.

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

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

$$\theta_0 = -30, \theta_1 = 2, \theta_2 = 2, \theta_3 = 0, \theta_4 = 0$$

$$g(-30 + 2x_1^2 + 2x_2^2) = \frac{1}{1 + e^{-(-30+2x_1^2+2x_2^2)}} = 0.5$$

$$\frac{1}{0.5} = 1 + e^{-(-30+2x_1^2+2x_2^2)}$$

$$2 = 1 + e^{-(-30+2x_1^2+2x_2^2)}$$

$$1 = e^{-(-30+2x_1^2+2x_2^2)}$$

Taking ln on both sides

$$ln(1) = ln\left(e^{-(-30+2x_1^2+2x_2^2)}\right)$$

$$0 = -(-30 + 2x_1^2 + 2x_2^2)$$

Multiply both sides by -1

$$-30 + 2x_1^2 + 2x_2^2 = 0$$

$$2x_1^2 + 2x_2^2 = 30$$

$$2(x_1^2 + x_2^2) = 30$$

$$x_1^2 + x_2^2 = \frac{30}{2}$$

$$x_1^2 + x_2^2 = 15$$

Equation of Circle: $(x-c_1)^2 + (y-c_2)^2 = r^2$

Where r is the radius and (c_1,c_2) are center coordinates which are (0,0) in this case

$$x_1^2 + x_2^2 = (3.873)^2$$