UNIVERSITY OF MALAYA

EXAMINATION FOR THE DEGREE OF MASTER OF DATA SCIENCE

ACADEMIC SESSION 2023/2024 : SEMESTER I

WQD7006 : Machine Learning for Data Science

Jan 2024 Time : 1.15 hour

INSTRUCTIONS TO CANDIDATES :

*Answer **ALL** questions. (50 marks)*

(Kertas soalan ini mengandungi 3 soalan dalam 4 halaman yang bercetak)
(This question paper consists of 3 questions on 4 printed pages)

Online assessment/30 marks – 1.15 hours
Submission through Spectrum. Specific instructions to be given on Spectrum as well.

Question 1: 12 marks

1. (a) You are given the dataset below:

Training data

Example	Feature X1	Feature X2	Target Y
1	3	5	18
2	7	2	28

Testing data

Example	Feature X1	Feature X2	Target Y
3	4	1	18

Parameter θ is the constant weight., θ1 and θ2 are the weights of features X1 and X2, respectively. Write the general linear regression model using these symbols.

(2 marks)

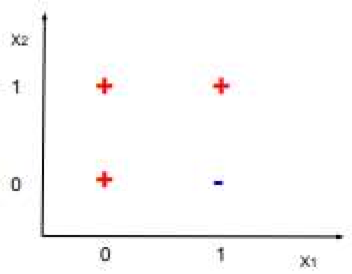
(b) Assume that you are given Weights = [θ = 1, θ1 = 3, θ2 = 2]. Calculate the predicted value for the three examples above. Show the steps.

(3 marks)

(c) The Mean Squared Error (MSE) is calculated as MSE = (1/m) Σ_j (Y_j - $\bar{Y}_j$)^2, where m is the number of examples. Calculate MSE for training and testing data separately.

(2 marks)

(d) We are interested in predicting whether a person makes over 50K a year, and we model the two features with two boolean variables X1, X2 ∈ {0,1}, and label Y ∈ {0,1} where Y = 1 indicates a person makes over 50K. Figure below shows three positive samples ("+" for Y = 1) and one negative sample ("-" for Y = 0). Answer the following questions:



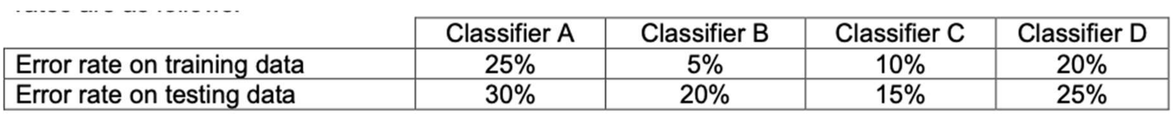
i. For the above scenario, which model would be better in predicting: linear or logistic regression? Why? (2 marks)

ii. Is there any logistic regression classifier using X1 and X2 that can perfectly classify the examples in the figure above? Explain. (2 marks)

iii. If we change the label of point (0,1) from "+" to "-", will there be a perfect logistic regression classifier? (1 mark)

Question 2: 8 marks

2. The performance of four classifiers (i.e. misclassification rates) are as follows:



	Classifier A	Classifier B	Classifier C	Classifier D
Error rate on training data	25%	5%	10%	20%
Error rate on testing data	30%	20%	15%	25%

a). Which classifier has the best generalization performance, i.e., most likely would perform the best when applied to unseen data? Why? (2 marks)

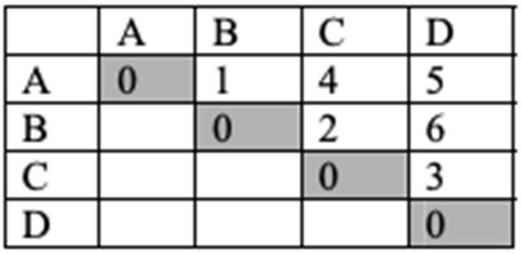
b). Which classifier is underfitting the most? Why? (2 marks)

c). Which classifier is overfitting the most? Why? (2 marks)

d). Given a dataset, assume we are using linear regression to make a prediction. The dataset is split randomly into training and testing. We increase the training set size gradually. What happens to the mean training and mean testing errors, that is, increase or decrease? (2 marks)

Question 3: 10 marks

You are given the following distance matrix:



	A	B	C	D
A	0	1	4	5
B		0	2	6
C			0	3
D				0

(a) Perform hierarchical clustering using single link technique. Show the distance matrix at each step and draw the final dendogram. (5 marks)

(b) Perform the same as (a) using complete link. (5 marks)

END