UNIVERSITY OF RUHUNA

Faculty of Engineering

End-Semester 5 Examination in Engineering: January 2024

Module Number: EE5253 **Module Name: Machine Learning**

[1 Hour and 15 minutes]

[Answer **all questions**, each question carries **10** marks]

Attach Question Paper to the Answer Script

Q1 a) A machine learning problem is set up to predict the price of laptops. It utilizes 1-5 variables given below to predict the price (given by variable 6). A student proposes to use logistic regression to solve the above problem. Do you think it is a good choice? Justify your answer.

1) Inches – Numeric - Screen Size
2) ScreenResolution – String - Screen Resolution
3) Cpu - String -Central Processing Unit (CPU)
4) Ram – String - Laptop RAM
5) Memory – String - Hard Disk / SSD Memory
6) Price_euros – Numeric - Price (Euro)

[2.0 Marks]

b) (i) Pruning is often used in Decision Trees to avoid overfitting the training set. Briefly describe what pruning is.

(ii) Which regression model given below (Model I or Model II) is more appropriate to fit the training data better? Justify your answer.

Model I: $y = ax + e$

Model II: $y = ax + bx^2 + e$

[2.0 Marks]

c) Table Q1c shows whether students will pass or fail EE5253 based on whether or not they attended class, studied, and slept well before the exam. You are given the following data for five students. The column "Result" shows the label we want to predict.

Table Q1c

	Attended Class?	Studied?	Slept?	Result
Student 1	Yes	No	No	Passed
Student 2	Yes	No	Yes	Failed
Student 3	No	Yes	No	Failed
Student 4	Yes	Yes	Yes	Failed
Student 5	Yes	Yes	No	Passed

(i) What is the entropy H(Result) at the root node? Show your workings.

(ii) Draw the decision tree where every split maximizes the information gain. Show your workings.

[2.0 Marks]

d) Consider the data points in 2-D Euclidean space shown in Table Q1d.

Table Q1d

x	y	Class
-1	1	1
0	1	2
0	2	1
1	-1	1
1	0	2
1	2	2
2	2	1
2	3	2

(i) What is the prediction of the 3-nearest neighbour classifier at point (2,4)?
(ii) What is the prediction of the 5-nearest neighbour classifier at point (1,1)?
(iii) What is the prediction of the 7-nearest neighbour classifier at point (1,1)?
(iv) What is the prediction of the 1-nearest neighbour classifier at point (2,-1)?

[2.0 Marks]

e) You are required to train a Support Vector Machine (SVM) on a tiny dataset with 4 points shown in Figure Q1e. This dataset consists of two examples with class label -1 (-), and two examples with class label +1 (+).

(i) Find the weight vector w and bias b. What is the equation corresponding to the decision boundary?
(ii) Circle the support vectors and draw the decision boundary on Figure Q1e provided.

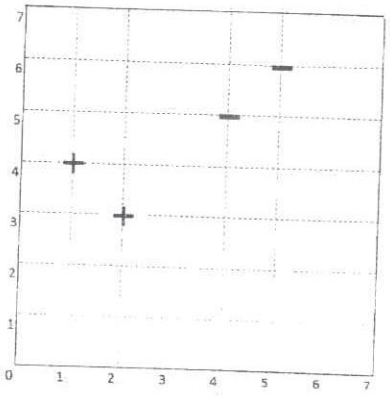

Figure Q1e

[2.0 Marks]

Q2 a) (i) List the three (3) main types of gradient descent based on the amount of data that is used.

(ii) Show the derivation for the update equations for $J(\theta_1, \theta_2) = \theta_1^2 + \theta_2^2$ using multivariate gradient descent.

[2.0 Marks]

b) Figure Q2b shows four plots with data showing low and high variance and low and high bias. Answer questions (i) to (iv) based on Figure Q2b by choosing from A, B, C or D as necessary.

(i) Which plot or plots have high variance?
(ii) Which plot or plots have high bias?

(iii) Which plot or plots have low variance?
(iv) Which plot or plots have low bias?

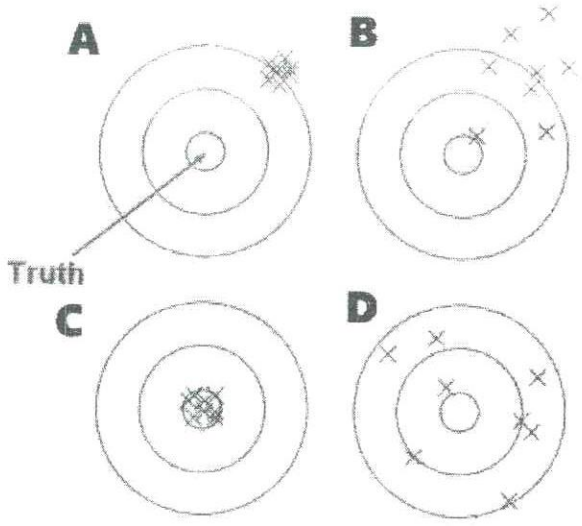

Figure Q2b

[2.0 Marks]

c) Explain stratified sampling. What is the purpose of stratified sampling in machine learning?

[2.0 Marks]

d) Briefly explain the importance of four (4) methods used for data pre-processing.

[2.0 Marks]

e) Answer the following questions regarding principal component analysis (PCA).
   (i) Briefly explain giving a graphical example how outliers are removed using PCA.
   (ii) Give two (2) instances when NOT to use PCA in a dataset?

[2.0 Marks]