Index No EG/_____/_____.

UNIVERSITY OF RUHUNA

Faculty of Engineering

End-Semester 7 Examination in Engineering: May 2023

Module Number: EE7209 **Module Name: Machine Learning**

[Three Hours]

[Answer **all questions**, each question carries **10** marks]
Please attach the question paper to the answer script.

Q1 a) (i) Briefly define the three types of machine learning and the type of problems each can handle.

(ii) List out the differences between the three types of machine learning with respect to the input to the system, driving force and approach.

(iii) Categorize each of the following algorithm based on the types of machine learning given in part a) (i).

a) Naïve Bayes
b) K- means
c) Logistic Regression
d) Linear Regression
e) Actor- Critic

[4.0 Marks]

b) Circle the most appropriate answer in (i) to (xii) below on the question paper itself.

(i) Which of the following statements are/is true regarding I and II below?

I: When the hypothesis space is richer, overfitting is more likely.
II: When the feature space is larger, overfitting is more likely.

(A) I only (B) II only (C) Both I and II (D) Neither I nor II

(ii) Which of the following algorithms can be used for clustering?

I: K-Nearest Neighbour
II: K-means

(A) I only (B) II only (C) Both I and II (D) Neither I nor II

(iii) In neural networks, nonlinear activation functions such as sigmoid, tanh, and ReLU

(A) speed up the gradient calculation in back-propagation, as compared to linear units.
(B) help to learn nonlinear decision boundaries.
(C) are applied only to the output units.
(D) always output values between 0 and 1.

(iv) A student proposes to use the following two algorithms to learn $f : X \rightarrow Y$, where X is the feature vector $X = < X1, X2, X3>$. Which of them contains sufficient information to allow calculating P(X1, X2, X3, Y)?

I: Naïve Bayes
II: Logistic Regression

(A) I only (B) II only (C) Both I and II (D) Neither I nor II

(v) As the number of training samples goes to infinity, your model trained on that data will have:

(A) Lower variance
(B) Higher variance
(C) Same variance
(D) None of the above

(vi) Which one of the following is the main reason for pruning a Decision Tree?

(A) To save computing time during testing
(B) To save space for storing the Decision Tree
(C) To make the training set error smaller
(D) To avoid overfitting the training set

(vii) A decision tree is used for spam classification, and it is getting abnormally bad performance on both your training and test sets. What could be causing the problem?

(A) The decision trees are too shallow.
(B) Need to increase the learning rate.
(C) Decision Tree has overfitted.
(D) None of the above.

(viii) Which of these regression models is more appropriate to fit the training data better?

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

(A) Model I (B) Model II (C) Both will equally fit (D) Not enough information

(ix) Which of the following tasks can be best solved using Clustering?

I: Predicting the amount of rainfall based on various cues.
II: Detecting fraudulent credit card transactions.
III: Training a robot to solve a maze.

(A) I and II only (B) II only (C) III only (D) All of the above

(x) Decision trees can work with _____

I: Numerical Values.
II: Nominal Values.

(A) I only (B) II only (C) Both I and II (D) Neither I nor II

(xi) What can help to reduce overfitting in an SVM classifier?

(A) High-degree polynomial features.
(B) Setting a very low learning rate.
(C) Use of slack variables.
(D) Normalizing the data.

(xii) Which of the following can classify the data shown by the XOR function?

I: Decision Tree.
II: Logistic Regression.
III: Gaussian Naïve Bayes.

(A) I only (B) I and II only (C) I and III only (D) All of the above

[6.0 Marks]

Q2 a) You are hired as a data scientist to evaluate different binary classification models in a business setting. A false positive result is 5 times more expensive (from a business perspective) than a false negative result. The models should be evaluated based on the following criteria:

I: Must have a recall rate of at least 80%
II: Must have a false positive rate of 10% or less
III: Must minimize business costs

After creating each binary classification model, the data scientist generates the corresponding confusion matrix. Which of the confusion matrices below represents the best model that satisfies the requirements? Justify your answer.

(A) TN = 91, FP = 9, FN = 22, TP = 78
(B) TN = 99, FP = 1, FN = 21, TP = 79
(C) TN = 96, FP = 4, FN = 10, TP = 90
(D) TN = 98, FP = 2, FN = 18, TP = 82

[2.0 Marks]

b) Can you represent the boolean function shown in Table Q2 with a single logistic threshold unit (i.e., a single unit from a neural network)? Justify your answer.

Table Q2

A	B	f(A,B)
1	1	0
0	0	0
1	0	1
0	1	0

[2.0 Marks]

c) You are required to train a Support Vector Machine (SVM) on a tiny dataset with 4 points shown in Figure Q2. 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 Q2 provided.

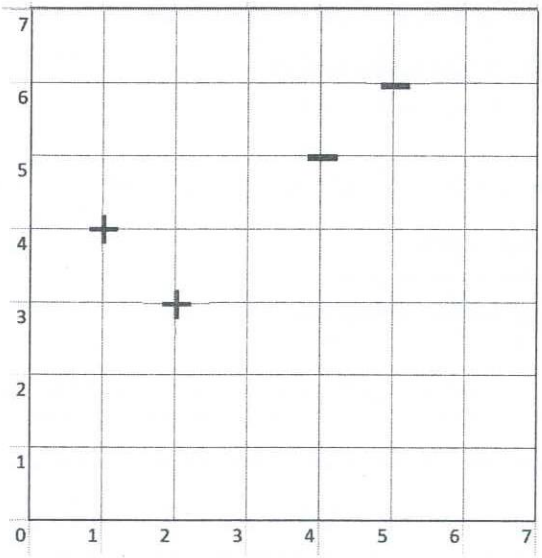

Figure Q2

[4.0 Marks]

d) (i) In a problem for "Customer segmentation using machine learning", the following features have been used. There are no missing values or outliers in this dataset. List two (2) possible pre-processing techniques to be used in this data.

- Customer's Age
- Customer's Gender
- Customer's Annual income
- Customer's Spending score

(ii) A leading telecommunication company analyses its data for customer satisfaction. There are 20,356 labeled data samples with approximately 90% of customers satisfied with the service and approximately 10% not satisfied. Briefly describe how you would handle the misbalanced classes.

[2.0 Marks]

Q3 a) Figure Q3.a shows the same 2D data set in two different spaces. Which plot contains the first and second principal components, subplot A or B?

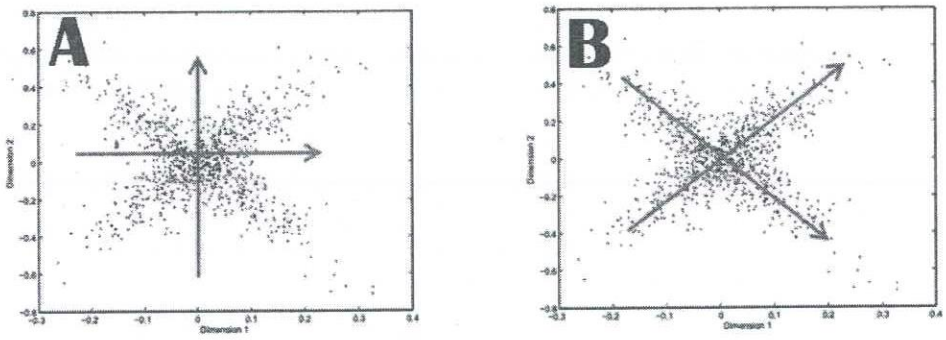

Figure Q3.a

[2.0 Marks]

b) Figure Q3.b shows two plots of 2D datasets. Draw the first and second principal components on each plot clearly marking them. Submit the question paper with the answer script.

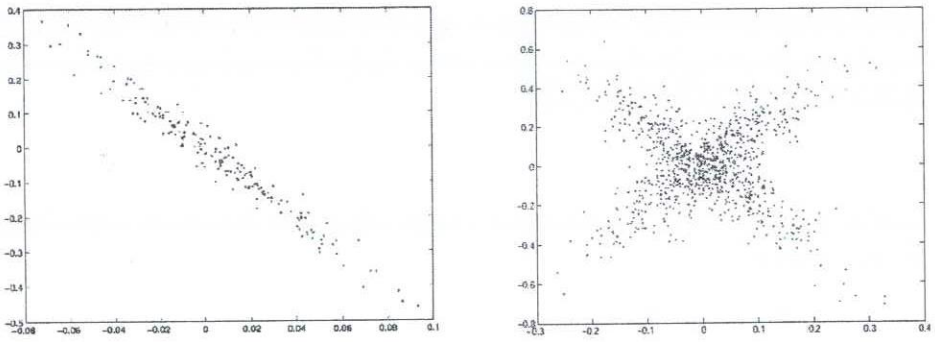

Figure Q3.b

[2.0 Marks]

c) Say the incidence of a disease D is about 5 cases per 100 people (i.e., P(D) = 0.05). Let Boolean random variable D mean a patient "has disease D" and let Boolean random variable TP stand for "tests positive." Tests for disease D are known to be very accurate in the sense that the probability of testing positive when you have the disease is 0.99, and the probability of testing negative when you do not have the disease is 0.97. What is P(TP), the prior probability of testing positive.

[2.0 Marks]

d) Figure Q3.d shows an example dataset.

(i) Mark any outliers in the data.

(ii) Your colleague proposes that support vector machine (SVM) is the best to classify this data. Do you agree with your colleague? Justify your answer.

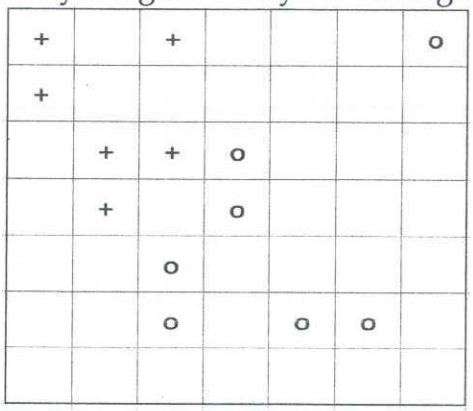

Figure Q3.d

[2.0 Marks]

e) Logistic regression is to build a fraud detection model with model accuracy 99%. However, 90% of the fraud cases are not detected by the model. Explain how you will help the model detect more than 10% of fraud cases. What is the compromise in the solution you propose?

[2.0 Marks]

Q4 a) Answer the following questions regarding the Receiver Operating Characteristic (ROC) curve.

(i) What is the ROC?

(ii) What does the ROC represent?

(iii) Sketch all possibilities of a ROC and explain how you would use the ROC to make decisions.

[4.0 Marks]

b) Figure Q4 shows results of three machine learning models. Choose the best model (A, B or C). Justify your answer.

A		Actual: positive	Actual: negative
Predicted	positive	190	9
	negative	0	1

B		Actual: positive	Actual: negative
Predicted	positive	180	5
	negative	10	5

C		Actual: positive	Actual: negative
Predicted	positive	170	1
	negative	20	9

Figure Q4

[2.0 Marks]

c) Table Q4 shows whether students will pass or fail EE7209 based on whether or not they studied, cheated, and slept well before the exam. You are given the following data for five students. There are three features, "Studied," "Slept," and "Cheated." The column "Result" shows the label we want to predict.

Table Q4

	Studied?	Slept?	Cheated?	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.
(iii) Did the tree you built implicitly perform feature subset selection? Justify your answer.

[4.0 Marks]

Q5 a) Write in point form how you would advise your junior batch on uses of chat gpt for assignments, stating the positive and negatives of it.

[2.0 Marks]

b) Table Q5.a shows a dataset used to learn a decision tree for predicting if a person is sad (S) or happy (H) based on the colour of the shirt/ blouse (Green, Blue or Red), whether they are wearing a jacket and the number of toes they have. Answer the following questions based on Table Q4.a and assume no pruning.

(i) What is H(emotion | Jacket=Yes)?
(ii) What is H(emotion | toes=11)?
(iii) Which attribute would the decision tree building algorithm choose for the root of the tree?
(iv) Draw the full decision tree that would be learnt for this data.

Table Q5.a

Colour of Shirt/ Blouse	Wearing Jacket	Number of Toes	Emotion (Output)
G	Yes	10	S
G	Yes	10	S
G	No	10	S
B	No	10	S
B	No	10	H
R	Yes	10	H
R	Yes	10	H
R	No	10	H
R	Yes	11	H

[2.0 Marks]

c) A psychology student wanted to see if three things can be used to predict if a child was overprotected or neglected in school. Table Q5.b shows the data from 10 children.

(i) Would you use Naïve Bayes algorithm to classify this data? Justify your answer.

(ii) Show with justification how a child with following will be classified from the above algorithm.

- Does well in school = Yes
- Plays sports = Yes

Table Q5.b

ID	Has a sibling?	Does well in school?	Plays Sports?	Overprotected or Neglected?
1	Yes	Yes	No	Neglected
2	Yes	Yes	No	Neglected
3	No	No	Yes	Overprotected
4	No	No	Yes	Overprotected
5	No	No	Yes	Overprotected
6	Yes	Yes	No	Neglected
7	Yes	Yes	No	Neglected
8	No	No	Yes	Overprotected
9	Yes	Yes	No	Neglected
10	No	No	Yes	Overprotected

[2.0 Marks]

d) Table Q5.c shows data from 8 different days that the University of Ruhuna Cricket team decided to practice or not based on four (4) different conditions. Answer the following questions using information in Table Q5.c.

(i) Calculate the eight (8) conditional probabilities of the attributes.
Eg: $\boldsymbol{P(Outlook|Practice = Yes)}$

(ii) What is the entropy of "Practice"?

(iii) Which attribute should you choose as the root of a decision tree?

Table Q5.c

Day	*Outlook*	*Humidity*	*Wind*	*Captain Present?*	*Practice?*
1	Sunny	Normal	Weak	No	No
2	Sunny	Normal	Strong	No	No
3	Overcast	High	Weak	Yes	No
4	Overcast	Normal	Weak	Yes	Yes
5	Sunny	High	Strong	No	Yes
6	Sunny	Normal	Strong	Yes	Yes
7	Sunny	Normal	Weak	Yes	Yes
8	Overcast	High	Weak	No	Yes

[4.0 Marks]