UNIVERSITY OF RUHUNA

Faculty of Engineering

End-Semester 7 Examination in Engineering: March 2022

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

[Three Hours]

[Answer **all questions**, each question carries **10** marks]
Please attach page 6 and page 7 to the answer script.

Q1 a) (i) Briefly explain the difference between testing and validation data set.
(ii) What is overfitting? List two (2) methods to overcome it.

[2.0 Marks]

b) (i) How does Machine Learning differ from Deep Learning?
(ii) A machine learning problem is given as "salary prediction based on years of experience." What type of machine learning problem is this?

[2.0 Marks]

c) (i) Briefly explain what you understand by a confusion matrix.
(ii) Which is worse to have; false negatives or false positives? Briefly explain your answer.

[2.0 Marks]

d) Are each of the following a Regression (R) problem or Classification (C) problem?
(i) Logistic Regression
(ii) Random Forest
(iii) Artificial Neural Network (ANN)
(iv) Support Vector Machine (SVM)

[2.0 Marks]

e) You have been hired as a Data Scientist to a company that has a project with Teaching Hospital, Karapitiya to decide whether a new visiting patient has any of heart problems, asthma, or COVID-19 (a patient can have one or more of these diseases). You have been given data from the past two years (2020 and 2021) regarding patient vitals (e.g., temperature, height, weight, diseases etc) in regard to the above three (3) conditions.
(i) Briefly explain how you would set the machine learning problem and list the steps for solving it.
(ii) If you were to use Artificial Neural Network (ANN) for this, would you use a separate neural network for each of the diseases or a single neural network with one output neuron for each disease? Justify your answer.

[2.0 Marks]

Q2 a) (i) Briefly explain giving a graphical example how outliers are removed using principal component analysis (PCA).
(ii) State one (1) similarity and one (1) difference between linear discriminant analysis (LDA) and PCA.

[2.0 Marks]

b) (i) What is variance accounted for? How would you use this in PCA?
(ii) When should you NOT use PCA?

[2.0 Marks]

c) (i) What is the first step of PCA?

(ii) Implement the first step of PCA for the data given in Table Q2c.

[2.0 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) Machine learning is used to predict if a client will subscribe to deposit in a bank. Two of the features are age of the client and gender of the client. How would you handle missing values in each of these features?

[2.0 Marks]

e) (i) The ABC water board company uses machine learning to check potability of water (i.e. how safe it is for drinking). There are 3726 labeled data samples which are all numerical data. 10% of data is from class I and remaining from class II. Briefly describe how you would handle the misbalanced classes.

(ii) Combined Cycle Power Plant Output prediction is done using decision tree algorithm. There are some outliers in the data. How should these outliers be handled before executing the machine learning algorithm?

[2.0 Marks]

Q3 a) (i) Briefly explain the 'explore' vs 'exploit' concepts in Reinforcement Learning (RL) and how this affects the final outcome.

(ii) Give two (2) examples of when you should not use RL in an application.

[2.0 Marks]

b) (i) Briefly explain two (2) roles of the activation functions in Neural Networks?

(ii) Give a difference in application for sigmoid vs tanh activation functions.

[2.0 Marks]

c) We want to classify some data using the network shown in Figure Q3c. There are three input parameters and two output classes. The network has one hidden layer and one output layer. There is no bias or activation function in the network (i.e. $\sigma(x) = x$). Derive the general expressions for the output y in terms of x, u and w.

[2.0 Marks]

d) Can you represent the boolean function in Table Q3d with a single unit from a neural network? If yes, show the architecture and calculation. If not, briefly explain why not.

[2.0 Marks]

e) Table Q3e gives symptoms and diagnosis for ten (10) patients.

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

(ii) Show with justification how a patient with following symptoms will be classified from the above algorithm used in Q3e (i)

- Runny Nose = Yes
- Fever = Yes

[2.0 Marks]

Q4 a) Are the following statements TRUE or FALSE? Justify your answer.

(i) When the training data set is small, a model is more likely to overfit.

(ii) MODEL 1: $ax+b+\xi$

MODEL 2: $ax^7+bx^6+cx^5+dx^4+ex^3+fx^2+gx+h+\xi$

If there are only 8 data points, MODEL 2 is likely to overfit the data.

[2.0 Marks]

b) Figure Q4b shows how the error of an algorithm varies with the number of epochs for training and validation datasets.

(i) What are the consequences of stopping the training at each epoch given by A, B and C?

(ii) What is the best place to stop the training A, B, C or a different iteration? If it is a different location than A, B or C, you may mark your solution on Figure Q4b and attach the page with the answer script.

[2.0 Marks]

c) You are given the 2 identical plots shown in Figure Q4c, which illustrates a dataset with two classes as 'o' in a larger circle and '▪' in a smaller circle. Classes have equal number of instances. Draw the decision boundary when you train an SVM classifier with linear and polynomial (order 2) respectively. Mark your solution on Figure Qc and attach the page with the answer script.

[2.0 Marks]

d) Give two (2) advantages and two (2) disadvantages of using decision trees for classification of real world problems?

[2.0 Marks]

e) Table Q4e 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 Q4d 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

[2.0 Marks]

Q5 a) (i) In decision trees do you prefer a shorter tree or a longer tree? Justify your answer.

(ii) Figure Q5a shows the decision boundaries obtained from three learning algorithms: decision trees, logistic regression, and nearest neighbor classification. '+' and 'o' depict two different classes. Beside each of the three plots, write the name of the learning algorithm and the number of misclassifications from each method.

[2.0 Marks]

b) There are six (6) data points on a one-dimensional (1D) number line: x=1, x=2, x=3, x=7, x=8, x=9. Two cluster centres A and B are initiated as x=0.5 and x=3. Starting with this initialization, graphically show how the k-means algorithm would estimate the points-to-cluster partitioning and the positions of the updated cluster centres (approximately). Repeat the process until convergence. Mark each iteration with t. Example t=1, t=2 etc. Each iteration should be on a new graph.

[2.0 Marks]

c) You are given a scatter plot of a 2D dataset in Figure Q5c. Draw the first and second principal components on the plot.

[2.0 Marks]

d) Table Q5d 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 Q5d.

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

(ii) What is the entropy of "Practice"?
Hint: $entropy = \boldsymbol{H(S)} = -\sum_{i=1}^{n} \boldsymbol{P(x_i) \times \log_2 P(x_i)}$

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

[4.0 Marks]

Table Q2c

Student	Telecommunication	Renewable Power	Computer Engineering
1	90	60	90
2	90	90	30
3	60	60	60
4	60	60	90
5	30	30	30

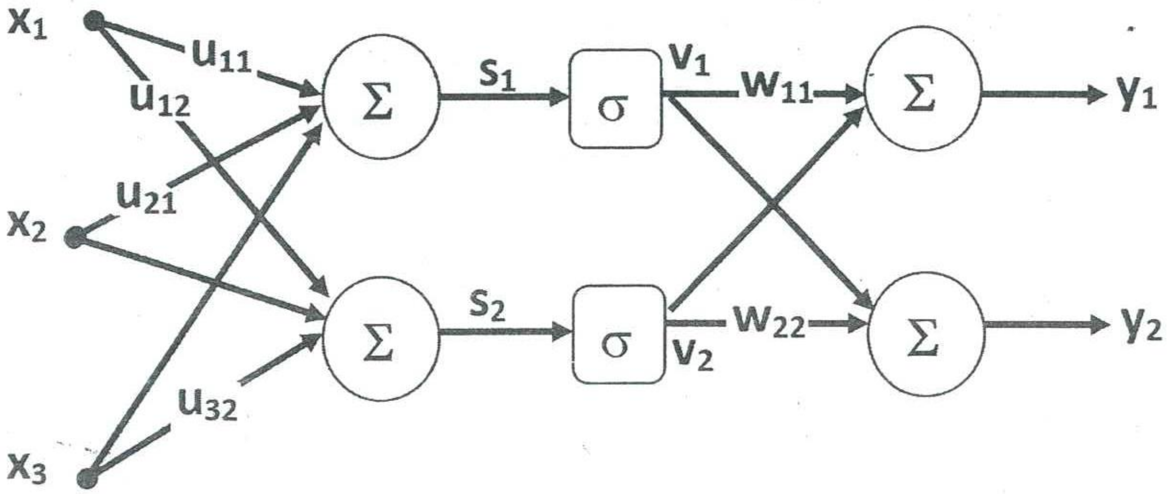


Figure Q3c

Table Q3d

A	B	output
1	1	0
0	0	0
1	0	1
0	1	0

Table Q3e

Patient ID	Cold	Runny Nose	Fever	Diagnosis
1	Yes	Yes	No	Flu
2	Yes	Yes	No	Flu
3	No	No	Yes	Covid-19
4	No	No	Yes	Covid-19
5	No	No	Yes	Covid-19
6	Yes	Yes	No	Flu
7	Yes	Yes	No	Flu
8	No	No	Yes	Covid-19
9	Yes	Yes	No	Flu
10	No	No	Yes	Covid-19

Please attach page 6 and page 7 to the answer script. **Index No EG/______/______:**

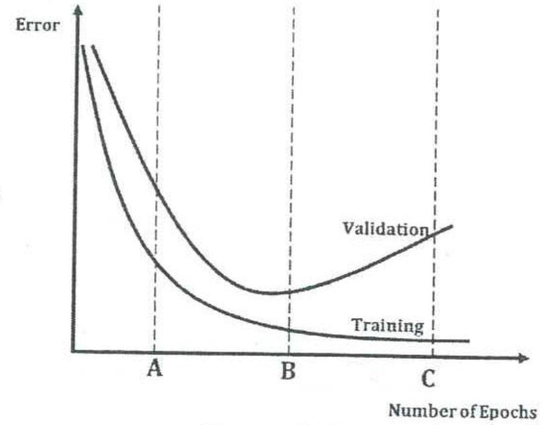
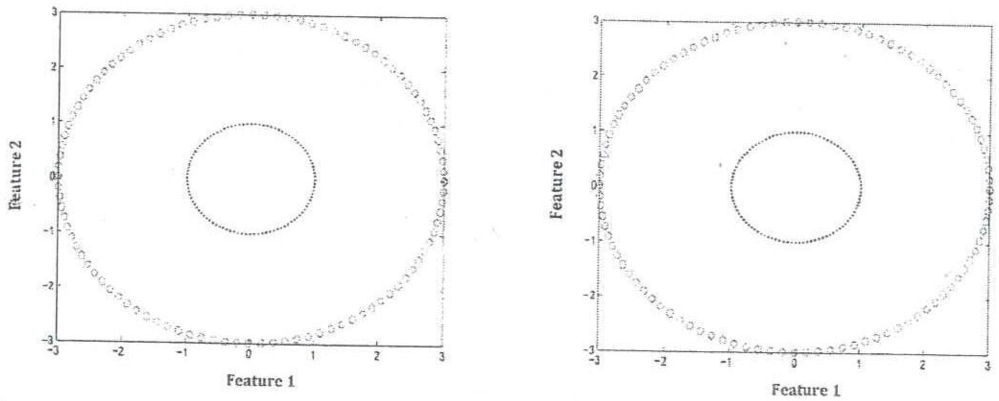


Figure Q4b



linear

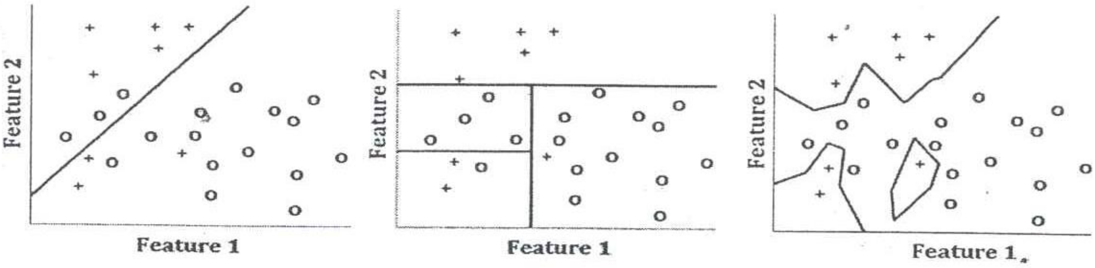
polynomial (order 2)

Figure Q4c
(Mark your solution on the given figures.)

Table Q4e

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

Please attach page 6 and page 7 to the answer script. **Index No EG/______/______.**



Algorithm ____________________ ____________________ ____________________

Misclassification # ________________ ________________ ________________

Figure Q5a
(Mark your solution on the given figure)

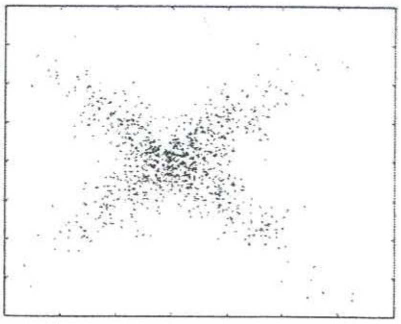


Figure Q5c
(Mark your solution on the given figure)

Table Q5d

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