UNIVERSITY OF RUHUNA

Faculty of Engineering

End-Semester 7 Examination in Engineering: October 2019

Module Number: EE7206 **Module Name: Machine Learning**

[Three Hours]

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

Q1 a) Briefly explain three main uses of Machine Learning (ML). Name a real world example for each of those uses.

[3.0 Marks]

b) i) Briefly explain the main ML types.

ii) Identify the most suitable ML type for the following applications. Justify your selection.

1. Predicting the rainfall on a particular day, given historical data.
2. Categorize an unlabeled news article dataset according to the type of news article.
3. An obstacle avoiding robot.
4. A profit maximizing trading agent.

[4.0 Marks]

c) State two differences between the feature selection and the feature reduction.

[1.0 Marks]

d) Show how a derived feature can reduce the complexity of a classification problem using a graphical example.

[2.0 Marks]

Q2 a) List the four main factors one should consider when selecting a suitable ML algorithm for a given application. Briefly describe the importance of each factor.

[2.0 Marks]

b) Suppose you are given 1000 data samples to categorize into 10 classes. Would you select "k-fold cross validation" or "validation based on train-test set division"? Briefly explain the reasons for your selection.

[2.0 Marks]

c) An accuracy result obtained from a ML algorithm is shown in Figure Q2 c).

i) What is the problem with the result?

ii) Briefly describe three techniques that can be tried out to overcome this problem.

[3.0 Marks]

d) In order to solve a classification problem, suppose you are given a training data set of 1,000 columns and 1,000,000 rows. You are asked to reduce the dimension of this data set so that the model computation time can be reduced. Your machine has memory constraints. Stating any practical assumptions you make, explain how you reduce the dimensionality of the given data set.

[3.0 Marks]

Q3 a) Consider the two-dimensional data set given in Table Q3 a).

i) Plot Antenna Length Vs. Abdomen Length for the data set and draw the decision boundary for the two classes; Grasshopper and Katydid.

ii) Compute and mark the means for the two classes in part a) i).

iii) Classify the new instance (5.1, 7) into Grasshopper or Katydid using the k-Nearest Neighbors (kNN) classification method where, k = 2.

Hint: Use Euclidean distance measurement. Clearly show your workings.

iv) What is the practical limitation of selecting k = 2 to this problem? What k would you select to overcome that limitation? Justify your answer.

v) How do you mitigate the sensitivity of nearest neighbor algorithms to irrelevant features?

[5.0 Marks]

b) Using naïve Bayes theorem, deduce whether the person called Drew is a Male or a Female based on the previously seen data given in Table Q3 b).

[2.0 Marks]

c) i) Explain the two main types of clustering.

ii) What is the use of a dendriogram in clustering?

iii) Write a psudo code for k means clustering.

[2.0 Marks]

d) When should you use classification over regression?

[1.0 Mark]

Q4 a) Consider a perceptron, the simplest possible Neural Network (NN). Its trainable parameters are the weights and the bias.

i) What is defined by the weights in the NN?

ii) What is the purpose of using the bias parameter?

[2.0 Marks]

b) Consider a perceptron that takes in a 4 dimensional feature vector. Its weight vector is randomly initialized to [0.2, 0.6, 0.4, 0], and the bias is initialized to 1. Assume you are given a training data sample having the feature vector [1, 5, 8, 4].

i) Show how the output value is calculated for this feature vector through forward propagation.

ii) Explain how to convert this simple perceptron into a sigmoid perceptron.

[3.0 Marks]

c) A commonly used error function for binary classification with a NN is,

$$-[y^{(i)} \log(y'^{(i)}) + (1-y^{(i)}) \log(1-y'^{(i)})]$$

where, $y^{(i)}$ is the output for the i^{th} data sample, and $y'^{(i)}$ is the predicted output. Show that the given error function is optimal.

[2.0 marks]

d) i) "tanh function is better than the sigmoid function to be used as the activation function of the output layer of a NN used for binary classification". Do you agree with this statement? Justify your answer.

ii) "During the training process of a NN, the input layer should never be trained". Do you agree with this statement? Justify your answer.

iii) Discuss the uses of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs).

[3.0 marks]

Q5 a) i) Briefly explain two main uses of Association Rule Mining (ARM).

ii) Explain the Apriori algorithm used in ARM using the example given below. State any assumptions you make.

Example

order 1: apple, egg, milk, carrot
order 2: milk
order 3: apple, egg, carrot
order 4: apple, egg
order 5: apple

iii) Define the following terms in the context of frequent item set mining.

1. Support
2. Confidence
3. Lift

iv) Using the example given in part a) ii), compute the following support, confidence and lift percentages.

1. support {apple,egg}
2. confidence{apple->egg}
3. confidence{egg->apple}
4. lift {apple,egg}
5. lift{egg,apple}

[6.0 marks]

b) Briefly explain the concept of Support Vector Machines (SVM).

[1.0 mark]

c) i) What does it mean by Genetic Algorithms (GA)?

ii) Explain the GA process using a pseudo code.

[2.0 marks]

d) Explain reproduction process (i.e. selection, crossover, and mutation) using an example.

[1.0 mark]

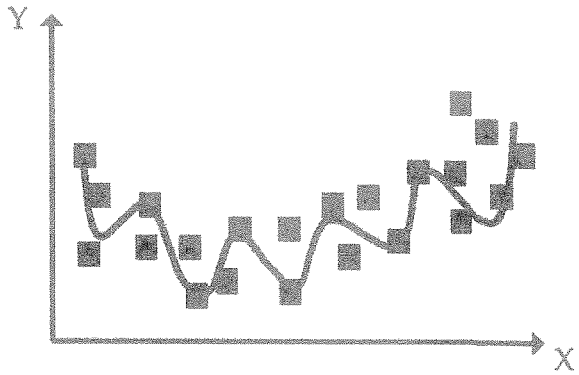

Figure Q2 c)

Table Q3 a)

Insect ID	Abdomen Length (cm)	Antenna Length (cm)	Insect Class
1	2.7	5.5	Grasshopper
2	8.0	9.1	Katydid
3	0.9	4.7	Grasshopper
4	1.1	3.1	Grasshopper
5	5.4	8.5	Katydid
6	2.9	1.9	Grasshopper
7	6.1	6.6	Katydid
8	0.5	1.0	Grasshopper
9	8.3	6.6	Katydid
10	8.1	4.7	Katydids

Table Q3 b)

Name	Sex
Drew	Male
Claudia	Female
Drew	Female
Drew	Female
Alberto	Male
Karin	Female
Nina	Female
Sergio	Male