Nilai UNIVERSITY

BACHELOR IN SOFTWARE ENGINEERING (HONOURS)

FINAL EXAMINATION
MAY 2024

Course : EC3357 (Machine Learning)   Time: 9.00 am – 12.00 noon (3 hours)

Lecturer : Ms. Rajesvary Rajoo   Date : 16 August 2024

Instructions:

Answer **ALL** questions in the Answer Booklet provided.

The maximum number of marks is 100.

This question paper consists of **4** printed pages.
(excluding front cover)

DO NOT OPEN THIS BOOKLET UNTIL YOU ARE TOLD TO DO SO

1. a. Discuss the relationship between Artificial Intelligence, Machine Learning and Deep Learning. **(6 MARKS)**

    b. Every Machine Learning algorithm has three components which are Representation, Optimization and Evaluation. Select an application of your interest and identify a method for each of this component. **(4 MARKS)**

    c. Discuss any **TWO (2)** limitations of machine learning algorithms. For each limitation, suggest an approach to mitigate it. **(6 MARKS)**

    d. Explain the significance of the use of matrices in machine learning. Illustrate with an example. **(9 MARKS)**

2. a. i.   Sketch the graph of single linear regression. Label the axes and draw the regression line. Show its mathematical representation.   **(5 MARKS)**

   ii   Interpret the slope and intercept in a single linear regression model. Explain how the slope and intercept both define the simple linear regression line.   **(6 MARKS)**

   b.   Illustrate how does multiple linear regression extend single linear regression. Show its mathematical representation.   **(5 MARKS)**

   c.   Removing outliers in independent variables can be a sensitive process, as outliers can contain valuable information. However, there are circumstances where it is acceptable, and sometimes necessary, to remove outliers to improve the quality and interpretability of the analysis. Explain **TWO (2)** circumstances in which it is acceptable to remove outliers in independent variables.   **(4 MARKS)**

   d.   Discuss any two deciding factors to consider when choosing between logistic regression and linear regression in an application   **(5 MARKS)**

3. a. Assuming that you are working for an email service provider, and your task is to develop a system that can automatically classify incoming emails as either "spam" or "not spam" (also known as "ham"). The dataset you have includes a large number of emails, each labeled as spam or not spam, with features like the presence of certain keywords, the frequency of specific words, the length of the email and the presence of suspicious links.

   Argue the suitability of the Support Vector Machine (SVM) for the given scenario by highlighting any **TWO (2)** key aspects of SVM.

   (6 MARKS)

   b. Determine the use of kernel trick in SVM.

   (4 MARKS)

   c. Describe briefly the key metrics used to evaluate the performance of a classification model

   (4 MARKS)

   d. Imagine you have a model designed to detect whether a patient has a certain disease (positive class) or not (negative class). After running your model on a test dataset of 100 patients, you get the following results:

   - 40 patients actually have the disease (positive class).
   - 60 patients do not have the disease (negative class).

   The model made the following predictions:
   - Correctly identified 35 patients with the disease (True Positives).
   - Correctly identified 50 patients without the disease (True Negatives).
   - Incorrectly identified 10 patients without the disease as having the disease (False Positives).
   - Incorrectly identified 5 patients with the disease as not having the disease (False Negatives).

   i. Draw the confusion matrix for this scenario.

   (3 MARKS)

   ii. Calculate Accuracy, Precision, Recall and F1-Score

   (8 MARKS)

4. a. Explain the concept of clustering in unsupervised learning and relate it to a real-world application where clustering is useful. **(9 MARKS)**

   b. Determine the importance of proper initialization of centroids in k-means algorithm. **(4 MARKS)**

   c. Compare the use of dimensionality reduction in supervised learning and unsupervised learning. **(6 MARKS)**

   d. Evaluate **TWO (2)** key factors to consider when selecting a machine learning algorithm for handling large datasets. **(6 MARKS)**

-END OF QUESTION PAPER-