Second Semester Examination
2023/2024 Academic Session

July/August 2024

EPE 472 – Artificial Intelligence and Data Mining
(Kecerdasan Buatan dan Pelombongan Data)

Duration: 2 hours
(Masa: 2 Jam)

Please check that this examination paper consists of SIX (6) pages of printed material before you begin the examination.

[Sila pastikan bahawa kertas peperiksaan ini mengandungi ENAM (6) muka surat yang bercetak sebelum anda memulakan peperiksaan ini.]

Instructions: Answer ALL **FOUR (4)** questions.

*[**Arahan**: Jawab **EMPAT (4)** soalan]*

...2/-

1. The fixed-demand economy order quantity (EOQ) is applied to determine the optimal order quantity in production, with formula given as below.

$$Q = \sqrt{\frac{2000 \times S}{H}}$$

Q = Optimum number of units per order

S = Setup cost for each order

H = Holding cost per unit per year

[a] Sketch a feedforward artificial neural network system consists of three layers of neurons to generate results approximate to EOQ.

(20 marks)

[b] Create TWO (2) training data sets, with the first and second sets of S and H values being [3,5] and [7,1], respectively. Based on the training data sets, compute one cycle of training iteration using the design described in 1[a], with all neuron and threshold weights set to 0.5, thresholds set to -1, and learning rate set to 0.2.

(50 marks)

[c] Plot the ReLu (rectified linear unit) and Sigmoid activation functions roughly between 1 and 0, with 0 as the center point. Provide one (1) advantage of ReLu over the Sigmoid activation function.

(30 marks)

2. A genetic algorithm program is used to solve an optimization problem. There are four jobs (A, B, C, and D) that an operator must do. The operator can only process one job at a time and within the specific processing time. Upon completing the job, the operator can immediately move to the next job without any delay.

The processing time (*PT*) and due date (*DD*) of each job, *j* are presented in Table 2[a]. The operator needs to maximize the overall number of early completion days compared to the due date, as defined by $f(x) = \sum_{j=1}^{4}(DD_j - CT_j)$, for all jobs where *CT* represents the completion time in days. To illustrate *DD*-*CT*, let assume Job A with *DD* = 70 begins on Day 5 and ends on Day 20 (*CT*=20), therefore it is completed 50 days sooner than its due date. Due to the nature of the job, Job B and D cannot be processed one after another, e.g. B→D or D→B. Upon such occurrence, a static penalty of 20 days is deducted to the original fitness.

...3/-

Table 2[a]

Job	Processing time, *PT* (Days)	Due date, *DD* (Days)
A	15	70
B	35	98
C	5	50
D	20	80

In a solution, the chromosomes represent the order of jobs, and each job is represented by a real number ranging from 0 to 1. The job with the lowest value is scheduled first, followed by the job with the next lowest value. If two jobs have the same value, the one with the earliest due date will be scheduled first. Table 2[b] shows that after an iteration, the population arrived at three solutions, S1, S2 and S3.

Table 2[b]

	Job			
Solution	A	B	C	D
S1	0.15	0.56	0.67	0.89
S2	0.75	0.62	0.33	0.01
S3	0.23	0.55	0.44	0.87

[a] Evaluate the fitness of each solution and arrange them in descending order, with the fittest first and the least fit last.

(35 marks)

[b] Given two Rolette wheel selections and the order of solutions set in 2[a], P(0.2) and P(0.8) were obtained to select two parents from the population. Perform one-point crossover at the middle of the chromosomes of these parents to create TWO(2) offsprings. Provide the order of jobs in these offsprings (no need to calculate the offspring fitness).

(35 marks)

[c] If dynamic penalty is implemented to S3 , build ONE(1) case to explain the effect to the solution fitness over two different iterations.

(30 marks)

3. [a] Discuss the concept of decision trees in classification and provide an overview of how they work.

(30 marks)

[b] Discuss ONE (1) advantage and ONE (1) disadvantage of decision trees as classification models.

(20 marks)

...4/-

[c] The following Table 3[c] contains information about customers who visited a retail website, including features such as age, gender, browsing duration, number of pages they visited, device type, history of previous purchase, and whether or not they made a purchase during the visit.

Table 3[c]. Customer purchase behaviour dataset

Age	Gender	Browsing Duration (minutes)	Pages Visited	Device Type	Previous Purchase	Purchased
20	Female	15	5	Mobile	No	0
25	Male	10	3	Desktop	No	0
30	Female	20	7	Tablet	Yes	1
35	Male	25	4	Mobile	Yes	1
40	Female	30	6	Desktop	Yes	1
45	Male	5	2	Tablet	No	0
50	Female	35	8	Mobile	Yes	1
55	Male	40	9	Desktop	Yes	1
60	Female	45	10	Tablet	No	0
65	Male	50	1	Mobile	No	0

(i) Identify and explain the features that are obviously redundant to the prediction task.

(10 marks)

(ii) Construct a simple decision tree using "Browsing Duration $\leq$ 30 min" as the root node to predict whether a customer will make a purchase based on their demographic and browsing behaviour.

(40 marks)

4. [a] In Figure 4[a], you are given the Iris dataset in the ARFF (Attribute-Relation File Format) format containing measurements of sepals and petals of three different species of Iris flowers: Iris-setosa, Iris-versicolor, and Iris-virginica. Your task is to build a classification model to predict the species of Iris flowers based on their sepal and petal measurements.

...5/-

```
@relation Iris

@attribute sepallength numeric
@attribute sepalwidth numeric
@attribute petallength numeric
@attribute petalwidth numeric
@attribute class {Iris-setosa, Iris-versicolor, Iris-virginica}

@data
5.1, 3.5, 1.4, 0.2, Iris-setosa
4.9, 3.0, 1.4, 0.2, Iris-setosa
4.7, 3.2, 1.3, 0.2, Iris-virginica
4.6, 3.1, 1.5, 0.2, Iris-setosa
5.0, 3.6, 1.4, 0.2, Iris-versicolor
5.4, 3.9, 1.7, 0.4, Iris-setosa
4.6, 3.4, 1.4, 0.3, Iris-versicolor
5.0, 3.4, 1.5, 0.2, Iris-setosa
4.4, 2.9, 1.4, 0.2, Iris-setosa
4.9, 3.1, 1.5, 0.1, Iris-virginica
```

Figure 4[a]

(i) Based on the input data shown, identify the attributes included in the Iris dataset and the class attribute.

(25 marks)

(ii) Determine the number of instances in the dataset.

(5 marks)

[b] You are tasked with evaluating the performance of a multi-class classification model trained to classify instances into three different categories: Alpha, Beta, and Gamma. The model has been assessed using a confusion matrix generated from WEKA, as shown in Figure 4[b].

```
=== Confusion Matrix ===

   Alpha   Beta   Gamma   <-- classified as
     100      5       5   |  X = Class Alpha (True Positives)
      10     80      10   |  Y = Class Beta (True Positives)
      15     10      75   |  Z = Class Gamma (True Positives)
```

Figure 4[b]

(i) Calculate the overall accuracy of the model.

(ii) Compute how many instances were incorrectly classified by the model.

(iii) Identify which class was mostly misclassified by the model.

...6/-

(iv) If the baseline ZeroR algorithm accuracy is known to be 0.6 (60%), interpret the performance of the model in terms of its ability to correctly classify instances belonging to each class, considering the provided confusion matrix.

(70 marks)

- oooOOOooo -