Second Semester Examination
2022/2023 Academic Session

Julai/Ogos 2023

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 **KESEMUA EMPAT (4)** soalan.]*

…2/-

1. [a] Explain the similarities between biological neural networks and artificial neural networks in terms of its function, structure and components.

(30 marks)

[b] A study uses a feedforward multilayer perceptron neural network to predict whether a patient is likely to have a cancer based on two inputs, I_1 and I_2. The neural network contains TWO (2) nodes, N_a and N_b in the hidden layer and ONE (1) node (N_c) in the output layer. Two samples, listed in Table 1[b] are used to provide the training to the neural network. Suppose that "Yes" is represented by "1" and "No" is represented by "0". Using sigmoid function as the activation function, calculate TWO (2) iterations of training for both samples. The learning rate is 0.2. The threshold is 0.5 for all nodes. The initial weights for the hidden layer and the output layer are as follows: w_{1a} = 0.3; w_{1b} = 0.1, w_{2a}=0.1, w_{2b} = 0.2, w_{ac} = 0.5, and w_{bc} = 0.6.

Table 1[b]

Patient ID	Input 1, I_1: Weigh loss?	Input 2, I_2: Sudden bleeding?	Cancer?
S1	Yes	Yes	Yes
S2	No	Yes	No

(70 marks)

2. In Langkawi, a Genetic Algorithm is used to schedule the ferrymen in a ferry operating enterprise. The enterprise has sixteen ferrymen to operate all five ferries every day (seven days per week). While two ferrymen are required to work on any ferry on any given day, company regulations state that each ferryman cannot work more than two days in a row.

[a] Suggest a suitable binary string solution in the Genetic Algorithm to this problem.

(40 marks)

[b] Propose a fitness function with a static penalty for this problem, in the form of either description, pseudo-code or flow diagram.

(60 marks)

…3/-

3. [a] Explain the “Curse of Dimensionality” with ONE(1) example.

(40 marks)

[b] State the approach to overcome the “Curse of Dimensionality” phenomenon.

(20 marks)

[c] Based on the descriptions as well as the attribute values given in the Table 3[c]:

- Genre: Contains various names of movie styles.
- Rating: Movies are rated on a 5-point scale from very bad to very good.
- Gross: Money that the movie made.
- Cinema: If the movie was shown in cinemas.

Table 3[c]

ID	Genre	Rating	Gross	Cinema
1	Horror	Very bad	5000	1
2	Drama	Good	8000	1
3	Comedy	Very good	9000	1

(i) Determine the data types that match each of the attributes.
(ii) Discuss whether data preprocessing approach is required for this dataset.

(40 marks)

4. [a] Table 4[a] consists of training data from an employee database. The data have been generalized. For example, “31 . . . 35” for age represents the age range of 31 to 35. For a given row entry, count represents the number of data tuples having the values for department, status, age, and salary given in that row.

Table 4[a]

department	*status*	*age*	*salary*	*count*
sales	senior	31...35	46K...50K	30
sales	junior	26...30	26K...30K	40
sales	junior	31...35	31K...35K	40
systems	junior	21...25	46K...50K	20
systems	senior	31...35	66K...70K	5
systems	junior	26...30	46K...50K	3
systems	senior	41...45	66K...70K	3
marketing	senior	36...40	46K...50K	10
marketing	junior	31...35	41K...45K	4
secretary	senior	46...50	36K...40K	4
secretary	junior	26...30	26K...30K	6

Let “status” be the class attribute.

(i) Explain how would you modify the table to load into the WEKA software.

(ii) Construct a decision tree from the given data.

(60 marks)

[b] Figure 4[b] shows the confusion matrix output on J-48 algorithm using WEKA.

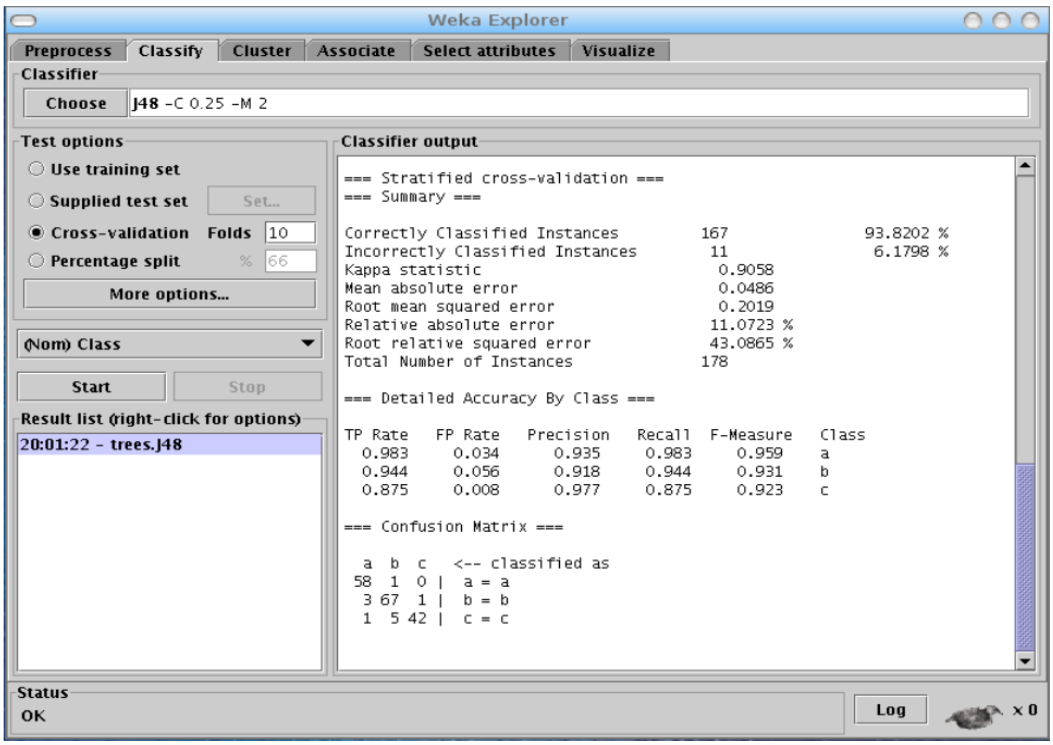

Figure 4[b]

…5/-

(i) Calculate the number of incorrectly classified instances.

(ii) Explain whether the results reflect a supervised or unsupervised learning problem.

(40 marks)

-oooOOooo-

Formulas given for feedforward multilayer perceptron neural networks

Activation functions

$$X = \sum_{i=1}^{n} x_i w_i - \emptyset$$

$$Y^{step} = \begin{cases} 1, if\ X \geq 0 \\ 0, ifX < 0 \end{cases}$$

$$Y^{sign} = \begin{cases} +1, if\ X \geq 0 \\ -1, ifX < 0 \end{cases}$$

$$Y^{sigmoid} = \frac{1}{1 + e^{-X}}$$

$$Y^{linear} = X$$

Error back propagation functions

$$e(p) = Y_d(p) - Y(p)$$
$$w_i(p+1) = w_i(p) + \alpha \times x_i(p) \times e(p)$$
$$\Delta w_{jk}(p) = \alpha \times \gamma_j(p) \times \delta_k(p)$$

Error gradient for neurons in the output layer

$$\delta_k(p) = y_{k(p)} \times [1 - y_k(p)] \times e_k(p)$$
$$e_k(p) = y_{d,k}(p) - y_k(p)$$
$$\Delta w_{jk}(p) = \alpha \times \gamma_j(p) \times \delta_k(p)$$
$$w_{jk}(p+1) = w_{jk}(p) + \Delta w_{jk}(p)$$

Error gradient for neurons in the hidden layer

$$\delta_j(p) = y_{j(p)} \times [1 - y_j(p)] \times \sum_{k=1}^{l} \delta_k(p) \times w_{jk}(p)$$

$$\Delta w_{ij}(p) = \alpha \times x_i(p) \times \delta_j(p)$$

$$w_{ij}(p+1) = w_{ij}(p) + \Delta w_{ij}(p)$$