NATIONAL UNIVERSITY OF SINGAPORE

IT5005 – ARTIFICIAL INTELLIGENCE

FINAL ASSESSMENT

(Semester 2: AY2024/25)

Time Allowed: 2 Hours

INSTRUCTIONS

1. This assessment paper contains **SIX (6)** questions and comprises **SIX (6)** printed pages.
2. This is a **CLOSED BOOK** assessment.
3. Only an A4 cheat sheet is allowed.
4. Approved calculators are allowed.
5. Answer **ALL** questions and write your answers only on the **ANSWER SHEET** provided.
6. Do **not** write your name on the ANSWER SHEET.
7. The maximum mark of this assessment is 100.

Question	**Max. mark**
Q1	15
Q2	15
Q3	20
Q4	20
Q5	20
Q6	10
Total	**100**

——— END OF INSTRUCTIONS ———

There are 6 questions [Total: 100 marks]

1. [Total: 15 marks] Markov Chains

Consider a Markov chain model that represents the status of a customer service representative in a bank. The possible states for the representative at time t (i.e., R_t) are:

- Available (A): Ready to serve the customer.
- On Call (C): Currently speaking with a customer.
- Break (B): Taking a scheduled break.
- Offline (O): Temporarily unavailable due to other reasons.

The status of the representative is monitored every hour to determine their availability and workload. The transition matrix (T) is as follows:

$$T = \begin{array}{c} \\ A \\ C \\ B \\ O \end{array} \begin{array}{c} \begin{array}{cccc} A & C & B & O \end{array} \\ \begin{bmatrix} 0.5 & 0.4 & 0.1 & x_1 \\ 0.3 & 0.6 & 0.1 & x_2 \\ 0 & 0.6 & 0.2 & x_3 \\ 0.4 & 0.3 & 0.2 & x_4 \end{bmatrix} \end{array}$$

a. Identify the missing values in the transition probabilities. [4 marks]

b. Assuming that the representative is currently available, i.e., $R_t = A$, what is the probability distribution of the representative's state one hour from now, i.e., $\boldsymbol{P}(R_{t+1})$? You need to simplify the relevant expressions for this problem. [6 marks]

c. The bank allows the representatives to take a lunch break after a 4 hour shift. Aiken, the manager, realized that the representative is offline at the start of the shift, i.e., $R_1 = O$. Aiken wants to know the probability distribution of the representative's state for the last hour of the shift, i.e., $\boldsymbol{P}(R_4)$? Closed-form expression with values assigned to each variable is sufficient. [5 marks]

2. [Total: 15 marks] Hidden Markov Models

Satellite communications are not completely deterministic - the quality of transmission depends on many factors that are uncontrollable. In a simplified experiment to study the effects of interference, our satellite transmits a sequence of 2 binary bits – that is, the sequence has 2 bits, where each bit is either -1 or 1. E.g., the sequence could be [1, 1] or [1, -1], etc. Let X_t be the signal transmitted by the satellite. The satellite chooses the first bit to transmit randomly with equal probability, i.e., $P(X_0 = -1) = P(X_0 = 1) = 0.5$. Furthermore, the probability that the next transmission X_{t+1} has been pre-programmed, as shown in the table. Ideally, we should receive the same bits as transmitted. However, due to interference, each bit may be flipped during transmission – e.g., the satellite transmitted a "1", but we received "-1" instead. Let E_t be the observed bits. Through experiments, we have estimated the probability of bits flipping to be 0.2:

$\boldsymbol{P(X_{t+1} \vert X_t)}$	$X_{t+1} = -1$	$X_{t+1} = 1$
$X_t = -1$	0.4	0.6
$X_t = 1$	0.7	0.3

$P(E_t \neq x_t \mid X_t = x_t) = 0.2$, and $P(E_t = x_t \mid X_t = x_t) = 0.8$

In an experiment, a sequence two bits are received. Answer the following.

a. Identify the prior distribution $\boldsymbol{P}(X_0)$ and transition matrix T. [2 marks]

b. Let the first received bit be -1. What is the probability that the first transmitted bit was indeed -1. [5 marks]

c. Let the second received bit be +1. What is the probability that the second transmitted bit was also +1? (**Note**: Model it as a filtering problem.) [8 marks]

3. [Total: 20 marks] Linear and Logistic Regression

You are developing advanced machine learning models to analyze and predict wildlife behavior using environmental sensor data. The dataset includes the following features:

- x_1: time of day (in hours)
- x_2: ambient temperature (°C)
- x_3: motion count (number of sensor detections per 5-minute window)

The goal is to design models that can predict both *activity levels* (continuous output) and *alert status* (binary output). Each sub-question explores a specific algorithmic perspective.

a. **Linear Regression for Activity Prediction** [12 marks]

You use ridge regression to predict an animal's activity score $\hat{y} \in [0, 50]$. The predictive model for the ridge regression is similar to linear regression. The only difference lies in the loss function. The loss function for the ridge regression model is defined as:

$$L = \frac{1}{2}(\hat{y} - y)^2 + \frac{\lambda}{2}\sum_{i=1}^{3} w_i^2,$$

where $\lambda = 0.5$ is the regularization coefficient. The above loss function is also known as regularized loss. Given:

- Weight vector: $\boldsymbol{w} = \begin{bmatrix} 2.0 \\ 0.3 \\ -0.1 \\ 0.05 \end{bmatrix}$
- True label: $y = 15$
- Input vector: $\boldsymbol{x} = \begin{bmatrix} 8 \\ 22 \\ 120 \end{bmatrix}$

(i) Compute the predicted value $\hat{y}$ for the above model parameters ($\boldsymbol{w}$ and λ) and input $\boldsymbol{x}$. [1 mark]

(ii) Compute the regularized loss L. [3 marks]

(iii) Derive and compute the gradient of the regularized loss *w.r.t.* each weight $\frac{\partial L}{\partial w_i}$. [4 marks]

(iv) Perform one iteration of gradient descent. Assume a learning rate of $\alpha = 0.01$. Provide the updated weights. [4 marks]

b. **Logistic Regression for Alert Classification** [8 marks]

You are building a logistic regression classifier to detect whether an animal is in an alert state (1) or calm state (0), based on the same input features. You consider the following activation functions:

- Sigmoid: $\sigma(z) = \frac{1}{1+e^{-z}}$
- ReLU: $\mathrm{ReLU}(z) = \max(0, z)$

(i) Which activation function is most appropriate for binary classification? Justify your choice clearly. [2 marks]

(ii) A logistic regression model predicts probabilities for five animals. Given a classification threshold of 0.5, identify TP, FP, FN, TN, and calculate precision and recall. [4 marks]

(iii) If scientists want to ensure all potentially alert animals are identified (even if this increases false alarms), should the model prioritize precision or recall? How should the classification threshold be adjusted? [2 marks]

Animal	$\hat{y}$	True label
A	0.91	1
B	0.76	1
C	0.62	0
D	0.43	1
E	0.22	0

4. [Total: 20 marks] Deep Neural Network

a. **Issues in neural network**

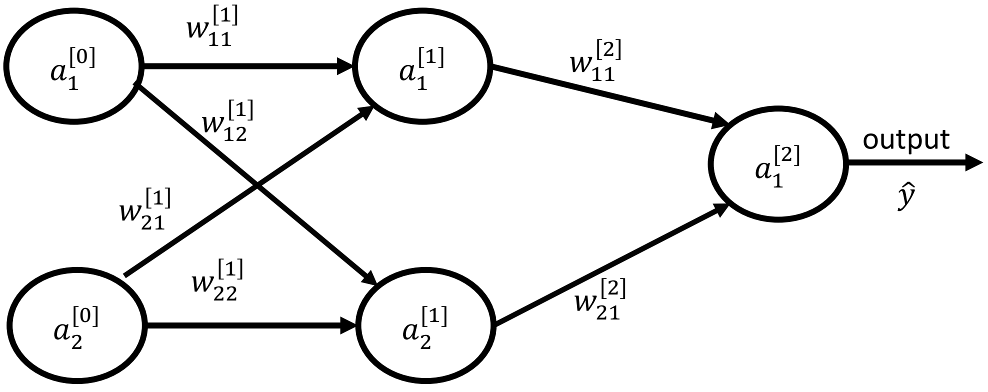


Figure 1: Two-Layer Neural Network

Consider the neural network shown in Figure 1. The weights of this network are initialized as

$$w_{11}^{[1]} = 0.1, \quad w_{21}^{[1]} = -0.2, \quad w_{12}^{[1]} = -0.2, \quad w_{22}^{[1]} = 0.1, \quad w_{11}^{[2]} = 0.1, \; w_{21}^{[2]} = -0.1$$

Assuming ReLU activation calculate the following for the input vector $\begin{bmatrix} a_1^{[0]} \\ a_2^{[0]} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$.

(i) $a_1^{[1]}$, $a_2^{[1]}$ and $a_1^{[2]}$ [6 marks]

(ii) Comment on the values calculated in Q4a(i). Have you noticed potential issue with the activation function in the above network? If so, recommend an alternative activation function to overcome the issue. You may assume that the objective is to predict a continuous target variable. [3 marks]

(iii) Mr. Aiken commented that the above network can be trained to do classification with three classes. Could Aiken use it for classification of data with three classes? If yes, provide the rationale. Otherwise, what changes would you recommend to Aiken? [3 marks]

b. **Simple Two-Layer Neural Network**

Aiken believes in simplicity and consequently proposes the following two-layer network with single neuron in each layer as shown in Figure 2. The objective is to learn the mapping between the input (x) and output (y). Answer the following queries.

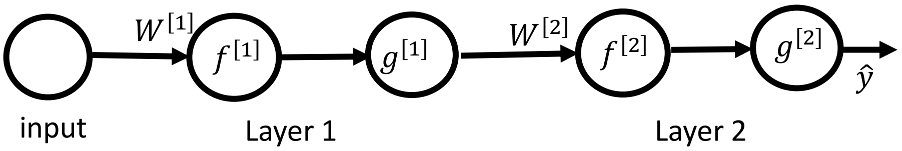


Figure 2: Simple Two-layer Neural Network

(i) Aiken proposed to use Sigmoid function as activation function. Highlight the limitation of Sigmoid function from the perspective of backpropagation. [2 marks]

(ii) Aiken wants to train this network through gradient descent algorithm. To this end, the initial weights are selected as $W^{[1]} = 0.001$ and $W^{[2]} = 0.005$. Identify the potential issue with this weight initialization. You have to provide the rationale. [3 marks]

(iii) Dueet argued that network proposed in Figure 3 can be trained better than the network in Figure 2 if the weights are initialized as $W^{[1]} = 0.001$ and $W^{[2]} = 0.005$. Do you agree with this argument? If you agree with the argument, provide the rationale. [3 marks]

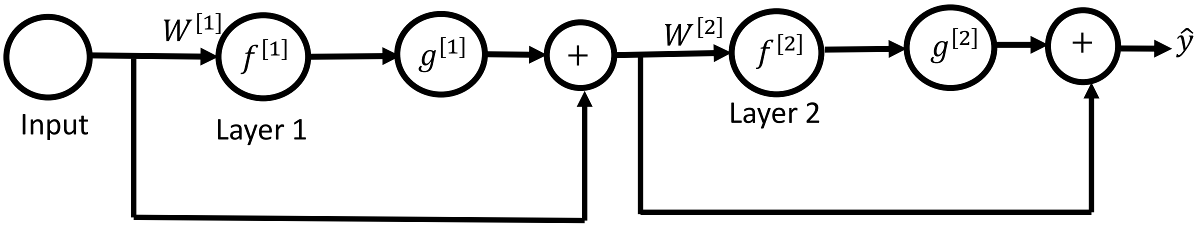


Figure 3: Simple two-layer Neural Network with Skip Connection

5. [Total: 20 marks] Convolutional Neural Networks

a. Consider a CNN architecture with the following sequential layers and skip connection. The input to this block is a 32x32x64 feature map. There are two paths: direct path and skip connection. The direct path has two convolution layers:

Convolution Layer A: 128 filters of size 3x3, stride 1, and padding 1
Convolution Layer B: 128 filters of size 3x3, stride 1, and padding 1

The input of **Convolution Layer A** is also directly added to the output of **Convolution Layer B**.

(i) What are the dimensions of the outputs of the **Convolution Layer A** and **Convolution Layer B**? [4 marks]

(ii) Comment on the projection along skip layer: is it identity mapping or projection mapping? [1 mark]

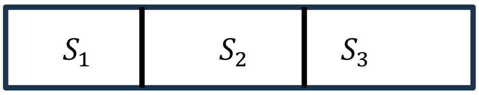
(iii) Design the skip connection. No need to draw the figure. State the parameters of basic building blocks of skip connection along with the intuition. [5 marks]

b. Aiken wants to perform convolution on an input with 256 channels. The objective is to get an output with 256 channels with same height and width as input. To this end, he proposed a convolution layer with 3x3 kernel with 256 channels. The stride and padding are selected to maintain the same width and height as input. However, Dueet argued that this architecture is parameter-inefficient and instead he proposed an alternative architecture named bottleneck:

1. A 1x1 convolution that reduces the channels to 64
2. A 3x3 convolution that maintains 64 channels
3. A 1x1 convolution that expands the channels to 256.

A network is called parameter-efficient if it has a smaller number of parameters. Prove/disprove that Dueet's bottleneck architecture is efficient. [10 marks]

6. [Total: 10 marks] Markov Decision Process



State	Actions	Reward
S_1	$\{stay, right\}$	-1
S_2	$\{left, right\}$	-1

Figure 4: 1-D Maze and State-Action-Reward Details

Consider a robot navigating a 1D environment with 3 states $S_1, S_2, and\ S_3$ as shown in Figure 4. The state S_1 is the start state and state S_3 is the terminal state. The table on the right side of the figure highlights the actions and rewards for different states and actions. In state S_1, the action $right$ takes the agent to state S_2 and action $stay$ results in agent staying in the same state. For both actions, the agent collects a reward of -1. In state S_2, the action $right$ takes the agent to state S_3 and action $left$ takes the agent to state S_1. Both actions in state S_2 would yield a reward of -1. State S_3 is the terminal state and the utility of the terminal state is 10. Answer the following questions for this MDP. There is no uncertainty in the model, i.e. the given problem is a deterministic MDP. Assume a discount factor of $\gamma = 0.9$.

a. Assuming that $\pi(S_1) = \pi(S_2) = right$, find the utilities of the states. [4 marks]

b. Assuming initial utilities $U_0(S_1) = U_0(S_2) = 0$, perform one iteration of the Value Iteration algorithm to find the updated utilities for states S_1 and S_2. [6 marks]

=== END OF PAPER ===