NANYANG TECHNOLOGICAL UNIVERSITY

SEMESTER 2 EXAMINATION 2024-2025

MA6514 – MACHINE LEARNING AND DATA SCIENCE

April/May 2025 Time Allowed: 3 hours

INSTRUCTIONS

1. This paper contains **FOUR QUESTIONS (4)** and comprises **FIVE (5)** pages.
2. **COMPULSORY** to answer **ALL** questions.
3. Marks for each question are as indicated.
4. This is a **CLOSED-BOOK** examination.

---

1. You are a consultant engaged by a local enterprise involved in manufacturing smart tele-communication equipment for a major Multi-National Company (MNC) headquartered in Singapore. The MNC has just informed the local enterprise that to continue the business with the MNC, the local enterprise needs to implement and upgrade its Industry 4.0 capabilities.

   (a) One of the senior management members of the local enterprise mentioned that she knew data science is one of the key technologies of Industry 4.0 but she wanted to know more about it.

   (i) Briefly explain what is data science and what broadly are its three (3) components. (5 marks)

   (ii) Briefly explain what you consider are the five (5) most critical challenges in data science. Clearly state the justifications for your choices. (10 marks)

   (b) Write an executive summary to the senior management of this local enterprise on what are the other five (5) important key technologies that must be included to meet the expectations of the MNC. Justify why you consider these 5 technologies as most important in implementing Industry 4.0. Clearly state your assumptions made. (10 marks)

2. (a) Three historical essays have been discovered discussing logistic management policy. You are provided with the following term (word)-frequency matrix, where rows represent essays and columns represent the frequency of common words after preprocessing:

Table 1

Essay	market	trade	fiscal	policy	reform
Essay 1	15	12	8	10	5
Essay 2	14	11	7	9	6
Essay 3	8	6	12	14	11

(i) Explain why Principal Components Analysis (PCA) would be valuable for analysing potential common authorship between these essays.

(3 marks)

(ii) Describe the specific steps you would take to prepare this data for PCA analysis, including any necessary normalization.

(5 marks)

(iii) What is the shape of the covariance matrix and the diagonal matrix obtained from doing the PCA on this data using eigen-decomposition? Explain the differences between eigen-decomposition and singular vector decomposition.

(5 marks)

(b) Answer the following:

(i) If the first principal component accounts for 83% of the variance and the second accounts for 16%, what insights might this provide about potential common authorship?

(ii) Draw the scree plot.

(6 marks)

(c) The term-frequency matrix was extracted from the original essays using the following script

```
essays = [
  'file1.txt',
  'file2.txt',
  'file3.txt',
]

def clean_words(words):
  ...

essay_index = []
for i, es in enumerate(essays):
  with open(es, 'r') as f:
    data = f.read()
  words = data.split(' ')
  cleaned = clean_words(words)
  essay_index.append(Counter({x: count for x, count in Counter(cleaned).items() }))
```

Note: Question No. 2 continues on Page 3

Sample text from file1.txt

The nature of market economies relies fundamentally upon principles of free trade and measured policy reform. Our examination of commercial prosperity must acknowledge how capital deployment shapes the development of industry. The banking system, through its extension of credit and regulation of monetary exchange, serves as the principal mechanism for economic growth. International trade, when properly balanced with domestic market considerations, promotes the sustained development of commercial enterprise.

(i) Explain how this data can be loaded as a DataFrame into Pandas.
(ii) Write the body of the clean words function to prepare the raw data for analysis.

(6 marks)

3. (a) Discuss the difference between Bagging and Pasting used in Ensemble Learning in terms of the following:

(i) The training dataset.
(ii) Variance Reduction and overfitting.
(iii) Bagging and Voting.

(6 marks)

(b) The Python program listed below implements the main body of a Bayesian model:

```
# L = Length of Ship
# B = Breadth of Ship
# Draft = Measured the depth of the hull in the water
# Speed = Ship's Speed
# TEU = Number of containers carried by the ship

1  import pymc as pm
2  path = "/content/gdrive/MyDrive/Ship_data/ Container_Ships_Data.xlsx"
3  Container_Ships_Data=pd.read_excel(path,usecols=['L', 'B', 'Draft', 'Speed', 'TEU'])

4  list1 = Container_Ships_Data['L']
5  list2 = Container_Ship_Data['TEU']
6  x_3 = list1.values
7  y_3 = list2.values

8  with pm.Model() as all_ship_model_neg:
9          α = pm.Normal("α", mu=0, sigma=1)
10         β = pm.Normal("β", mu=0, sigma=10)
11         σ = pm.HalfNormal("σ", 10)
12         μ = pm.Deterministic("μ", pm.math.exp(α + β * x_3))
13         y_pred=pm.NegativeBinomial('y_pred', mu=μ, alpha=σ, observed=y_3)
14         idata_neg=pm.sample(tune=2000,target_accept=0.95, return_inferencedata=True)
15         idata_neg.extend(pm.sample_posterior_predictive(idata_neg))
```

Note: Question No. 3 continues on Page 4

(i) In line 12, explain the Deterministic variable μ.
(ii) In the Bayesian model described above, explain line 13 in relation to Bayesian Inference.
(iii) Explain the purpose of the Negative Binomial in line 13.

(2 marks each for each item,Total: 6 marks)

(iv) In line 14, explain its purpose of the idata_neg and that of its arguments (i.e., tune and target_accept) in preventing sampling divergence. If sampling divergence occurs, what three steps would you take to prevent its occurrence?

(8 marks)

(c) Name the TWO shrinkage models in Regression Analysis. Compare their strengths and weaknesses.

(5 marks)

4. (a) A certain company called 'Self-Driving-Car Innovation' (SDC-Innovation) involved in the development of self-driving cars considers Deep Reinforcement Learning as the promising solution to achieve learning of the environment as a self-driving car (SDC) navigates the environment. Knowing that you have taken a course in Machine Learning, you are tasked to design and explain the implementation of Deep Q Network (DQN) to the management. The self-driving car is required to execute the following operations:

1. Lane following.
2. Lane changing.
3. Obstacle avoidance.
4. Start, Stop and Reverse.

The SDC is fully equipped with sensors for the above operations. Assume that there is no issue with the sensor inputs of the varied data from the suite of sensors.

The steps in the DQN Algorithm are given as follows:

1: Initialize learning rate α
2: Initialize temperature parameter τ
3: Initialize number of batches per training step, B
4: Initialize number of updates per batch, U
5: Initialize batch size N
6: Initialize experience replay memory with max size K
7: Randomly initialize the network parameters θ
8: **for** $m = 1 \ldots MAX\_STEPS$ **do**
9: Gather and store h experiences (s_i, a_i, r_i, s_i') using the current policy
10: **for** $b = 1 \ldots B$ **do**
11: Sample a batch b of experiences from the experience replay memory
12: **for** $u = 1 \ldots U$ **do**
13: **for** $i = 1 \ldots N$ **do**
14: # Calculate target Q-values for each example
15: $y_i = r_i + \delta_{s_i'} \gamma \max_{a_i'} Q^{\pi_\theta}(s_i', a_i')$ where $\delta_{s_i'} = 0$ if s_i' is terminal,
↪ 1 otherwise

Note: Question No. 4 continues on Page 5

16: **end for**
17: # Calculate the loss, for example using MSE
18: $L(\theta) = \frac{1}{N}\sum_i \left(y_i - Q^{\pi_\theta}(s_i, a_i)\right)^2$
19: # Update the network's parameters
20: $\theta = \theta - \alpha \nabla_\theta L(\theta)$
21: **end for**
22: **end for**
23: Decay τ
24: **end for**

(No definition of the other terms or symbols used in the above algorithm are given as these are already defined in the Lecture Notes on Reinforcement Learning).

(i) Explain the purpose of the experience replay memory and the target Q network in reinforcement learning and its input and output relationships in relation to the DQN architecture.

(4 marks)

(ii) Explain experience replay memory in terms of state, action and reward and how these are related to improve the learning efficiency of DQN algorithm.

(4 marks)

(iii) Explain how the loss function is related to the experience replay memory, the learning rate and the discount factor.

(4 marks)

Credit will be given to students who use the appropriate equations and diagrams to illustrate their answers

(b) While analysing some datasets, a Data Scientist noticed that the data can be classified as missing at random (MAR). He decides to use Expectation Maximization (EM) algorithm to deal with the missing data. Explain the following in the EM algorithm:

(i) The EM algorithm in terms of the two key steps.
(ii) The importance of the log-likelihood in the E-Step.

(7 marks)

(c) Detecting and identifying outliers are important steps in the training and performance of Machine Learning algorithms.

Answer the following questions with flow charts or illustrations where appropriate:

(i) Name TWO Machine Learning (ML) algorithms that either detect or resolve the issue of outliers.
(ii) Discuss TWO criteria on how outliers can be determined.
(iii) For one Machine Learning algorithm, discuss in point form the key steps of the algorithm to detect these outliers.

State the assumptions used in your answer.

(6 marks)

END OF PAPER

CONFIDENTIAL

MA6514 MACHINE LEARNING & DATA SCIENCE

Please read the following instructions carefully:

1. **Please do not turn over the question paper until you are told to do so. Disciplinary action may be taken against you if you do so.**

2. You are not allowed to leave the examination hall unless accompanied by an invigilator. You may raise your hand if you need to communicate with the invigilator.

3. Please write your Matriculation Number on the front of the answer book.

4. Please indicate clearly in the answer book (at the appropriate place) if you are continuing the answer to a question elsewhere in the book.