NANYANG TECHNOLOGICAL UNIVERSITY

SEMESTER 1 EXAMINATION 2025-2026

MA6514 – MACHINE LEARNING AND DATA SCIENCE

November/December 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. Your department in smart manufacturing of a local medium-size enterprise has just hired an intern who has just completed Year 2 Manufacturing Engineering in a local university, and you are assigned by your department manager to guide the intern for six months.

   (a) As the intern is still learning and has many doubts about your department operations and responsibilities, the intern continually seeks clarifications from you.

   (i) Briefly explain the similarities and differences between data science and big data to the intern.

   (7 marks)

   (ii) Briefly explain to the intern what are the four (4) most difficult challenges in data science. Clearly state the justifications for your answer.

   (8 marks)

   (b) Explain to the intern what are the five (5) key trends in data analytics and rank them according to importance to your department. Clearly state any assumptions you have made.

   (10 marks)

2. You are a reliability engineer at a manufacturing plant that produces automotive components. The plant has multiple CNC machines that are critical to production. Unexpected machine failures cause significant downtime and production losses. Your team has installed vibration sensors on 12 machines, collecting vibration data across multiple axes (X, Y, Z) along with temperature, RPM, and power consumption.

Management wants to implement a predictive maintenance system, but you need to determine the sensor variables that are most important for detecting potential failures. With the limited computational resources in the edge devices for monitoring these machines, you need to reduce the dimensionality of the data while preserving the most important information.

(a) There are missing data due to various downtimes.

(i) Explain how, by deploying Pandas and Principal Component Analysis (PCA) may be used to impute the missing values.

(4 marks)

(ii) Discuss the situation in which Pandas or PCA should be used.

(4 marks)

(b) You are provided with a dataset inside a folder with files containing 6 months of sensor readings from one of the CNC machines, one file for each time period. The data in each file includes a timestamp (ts), 3-axis vibration measurements (in mm/s^2) (x, y, z), bearing temperature (°C) (br), spindle RPM (rot), power consumption (kW) (pow), and a "machine_state" column (state) indicating whether the machine was operating normally [NORM], showing early warning signs [WARN], or experiencing failure [FAIL]. At times, some of the data from the sensors may not have been captured.

By following the code structure (you may write pseudo-code)

(i) Load and preprocess the sensor data.

(4 marks)

(ii) Apply PCA to reduce dimensionality.

(5 marks)

(iii) Determine how many principal components are needed to capture at least 95% of the variance.

(4 marks)

(iv) Referring to the code blocks below, and making any reasonable assumptions, interpret and provide recommendations on which sensor variables are most important to monitor for predictive maintenance.

(4 marks)

Note: Question No. 2 continues on Page 3.

```
# Import necessary libraries
import pandas as pd
import numpy as np
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
import matplotlib.pyplot as plt
import seaborn as sns
# Load and preprocess the data
# [Your code here]
# Apply PCA
# [Your code here]
# Determine optimal number of components
# [Your code here]
# Visualize results
# [Your code here]
# Interpret principal components
# [Your code here]
# Recommendations based on PCA results
# [Your recommendations here]
```

3. (a) Using a machine learning method or algorithm as an example of supervised learning and unsupervised learning, discuss overfitting and underfitting based on the following:

(i) The difference in concepts in detecting them for supervised and unsupervised learning.

(4 marks)

(ii) Explain the techniques for detecting or overcoming them.
Illustrate your answer where appropriate with diagrams or formulas.

(6 marks)

(b) Explain Mercer's Theorem in relation to the following:

(i) Kernel functions and their validity
(ii) Kernel Trick by illustrating with a kernel function and its impact on the input space-feature space mapping.

(7 marks)

(c) With diagram(s), explain K-fold Cross Validation in terms of the following:

(i) The key steps in selecting the optimal tuning or regularization parameter of a machine learning model.
(ii) For a regression model, select a performance metric of your choice and give a reason for your choice. Explain how the cross-validation score is calculated. Discuss the generalization ability of a regression model in relation to the outcome of the performance metric.

(8 marks)

4. (a) Explain categorical attributes and what is the purpose in data pre-processing for these attributes. Name one approach and explain its concept in implementing categorical attributes in Machine Learning.

(5 marks)

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

```
# Table1, Table2 and Table3, each records the data for the different ship types namely
# Container, Multi-purpose Container and Bulk carrier respectively.

1  import os
2  import pandas as pd
3  import numpy as np
4  import pymc as pm

5  sheet_names = ['Table1', 'Table2', 'Table3']

6  all_ships_data = pd.DataFrame()
7  usecols = [ 'Dwt', 'L', 'B', 'T', 'Speed', 'Capacity', 'Type' ]
8  cwd =  os.path.abspath('/content/gdrive/MyDrive/All_Ship_Database_rev(2).xlsx')
9  for sheetname in sheet_names:
10    ships_data = pd.read_excel(cwd, sheetname, usecols = usecols)

11    all_ships_data = pd.concat([all_ships_data, ships_data])
12    print(sheetname)
13    print(ships_data.head())

14 K = 3 # Number of ship types

15 df_All_ships_data = all_ships_data.dropna(subset=['Speed'])

16 X = df_All_ships_data['Speed'].values
17 Y = df_All_ships_data['Type']

18 with pm.Model() as model_ship_types:
19    p = pm.Dirichlet('p', a=np.ones(K))
20    means = pm.Normal('means', mu=X.mean(), sigma=10.9, shape=K)
21    sd = pm.HalfNormal('sd', sigma=10.9)

22    y = pm.NormalMixture('y', w=p, mu=means,sigma=sd, observed=X)
23  idata_mg=pm.sample(draws=2000, tune=500,target_accept = 0.99)
```

(i) Explain the purpose of Dirichlet and the related arguments in line 19.
(ii) State the Bayesian formula in terms of the probability terms involved.
(iii) Explain how line 22 implements the Bayesian formula and state the purpose of the use of Normal Mixture in the code.

(7 marks)

Note: Question No. 4 continues on Page 5.

(c) (i) Explain how an agent in Reinforcement Learning (RL) makes decisions.
(3 marks)

(ii) You are designing an RL system to teach a robot to navigate through a maze. Identify the agent, environment, states, actions, and rewards in this scenario. Discuss why RL should be used for this task.
(5 marks)

(iii) Recommend the algorithm to use assuming that the grid is a 10 x 10 matrix with fixed walls and obstacles. Explain the reasons for the recommendation.
(5 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.