**NANYANG TECHNOLOGICAL UNIVERSITY**

SEMESTER I EXAMINATION 2019-2020

**CM4044: ARTIFICIAL INTELLIGENCE IN CHEMISTRY**

November 2019 TIME ALLOWED: 2 HOURS

INSTRUCTIONS TO CANDIDATES

1. This examination paper contains **FOUR (4)** questions and comprises **SIX (6)** pages.
2. Answer **All FOUR (4)** questions.
3. Answer each question beginning on a **FRESH** page of the answer book.
4. This **IS NOT** an **OPEN BOOK** exam.

5. The examination paper **CANNOT** be removed from the exam hall.

**Question 1 (30 marks)**

Explain, by using diagrams and/or equations if necessary, the following concepts.

a) Machine Learning

(5 marks)

b) Training Data Set

(5 marks)

c) Classification

(5 marks)

d) Clustering

(5 marks)

e) Neural Network

(5 marks)

f) Bias-Variance Tradeoff

(5 marks)

**Question 2 (30 marks)**

a) Write a Python 3.0 program to use numpy array for these purposes:

i). Create a 2D 5 by 5 array with random numbers in the range of 0.0 and 1.0;

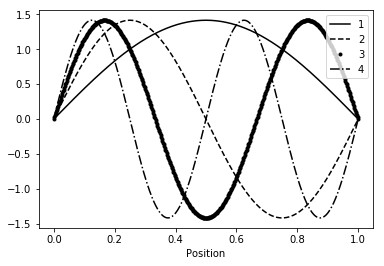
ii). Find the position indexes of maximum and minimum;

iii). Sort the elements along their column axis.

(15 marks)

b) Write a Python 3.0 program to display the one-dimensional particle-in-box (PIB) wavefunctions for the first four states. The one-dimensional PIB wavefunction is:

In your program, the box length L can be set to be 1 (atomic unit). The display of the wavefunctions should be the same as below:

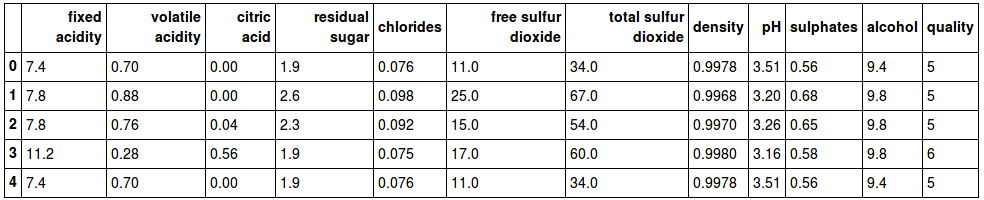


(Hint: use the matplotlib module to generate the display.)

(15 marks)

**Question 3 (25 marks)**

Wine quality analysis by machine learning is a text book example to demonstrate the value of computational modelling in data science. When data sets become very large, machine learning helps to make faster and unbiased decisions as human beings are not equipped with the mental faculty to pour over large data-sets filled with copious information. A data file “winequality-red.csv” has 1600 data instances with 12 columns: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, quality (score between 0 and 10). The first five rows of data are displayed in the table below:



a) Describe the general work flow to build a supervise learning model to analyse the wine quality. If necessary, use a diagram to show the concept.

(5 marks)

b) What is the learning target in this supervise learning model?

(3 marks)

Question 3 continues on page 5

c) Write the python code to load the data and randomly display ten data instances.

(5 marks)

d) Write the python code to draw a scatter plot of fixed acidity vs. volatile acidity.

(5 marks)

e) Write the python code to train a random forest machine learning model and validate its prediction accuracy.

(7 marks)

**Question 4 (15 marks)**

a) Draw the structures represented by the following SMILES strings:

i) c1ccccc1

ii) Cl/C=C/Br

iii) CCOc1ccccc1

(6 marks)

b) What is QSAR? Discuss the advantages and disadvantages of applying the QSAR technique to study chemicals.

(6 marks)

c) What is a chemical descriptor?

(3 marks)

**End of Paper**