**NANYANG TECHNOLOGICAL UNIVERSITY**

SEMESTER I EXAMINATION 2020-2021

**CM4044: ARTIFICIAL INTELLIGENCE IN CHEMISTRY**

November 2020 TIME ALLOWED: 2 HOURS

INSTRUCTIONS TO CANDIDATES

1. This examination paper contains **FOUR (4)** questions and comprises **TEN (10)** 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 **CAN** be removed from the exam hall.

**Question 1 (30 marks)**

Answer the following questions in your answer book in a concise way.

a) What is the difference between supervised learning and unsupervised learning? Name one machine learning algorithm for supervised learning and one for unsupervised learning.

(6 marks)

b) What is the difference between classification and regression? Describe one machine learning task for classification and one for regression.

(6 marks)

c) Describe bias and variance in machine learning model. Comment on bias-variance trade off in building machine learning model.

(6 marks)

d) How is KNN different from KMeans clustering?

(6 marks)

e) What is Z-scoring data normalization? Explain why it is important to normalize data sets in the process of building a machine learning model.

(6 marks)

**Question 2 (35 marks)**

Use Python 3 programming language to write your codes for the questions below.

a) Develop your python codes for the purpose below:

i) Initialize a 4 by 5 two-dimensional numpy array of values **randomly**

in the range of [0.0, 20.0);

ii) Find the position index of the maximum and minimum elements in this array;

iii) Calculate the mean, average and standard deviation of the data in this array;

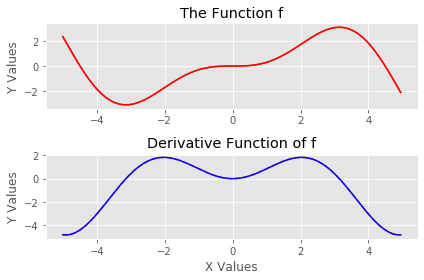
vi) Normalize the array so that the values of the array’s elements are

in the range of [0, 1).

(Hint: Choose the appropriate methods from the Table 1 in Appendix in page 9 for your codes. )

(20 marks)

b) Use matplotlib package to plot the math function and its derivative to generate the figures below:



Question 2 continues on page 4

The starting of the program is given below:

import numpy as np

import matplotlib.pyplot as plt

# define the figure size

plt.figure(figsize=(6, 4))

# use the ggplot display style

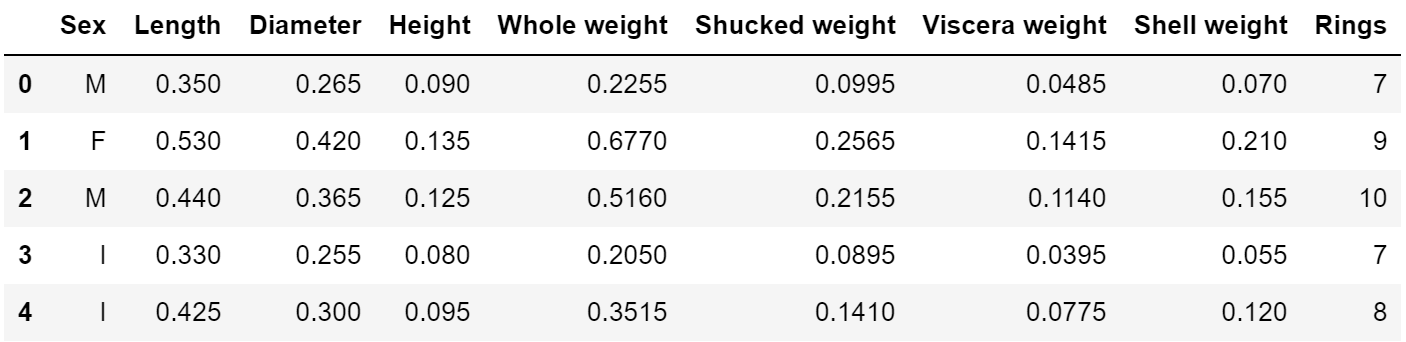
plt.style.use('ggplot')

(Hint: Choose the appropriate methods from the Table 2 in Appendix in page 10 for your codes.)

(15 marks)

**Question 3 (15 marks)**

Abalone is a common name of a very large group of sea snail. The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope -- a time-consuming task. Once the number of rings is known, the age equals to the number of rings plus 1.5. Machine learning can be an alternative way to predict the age of abalone if there are sufficient data collected from physical measurements on the samples of abalones of different ages. The abalone data set from UCI data repository (http://archive.ics.uci.edu/ml/datasets/Abalone) is one of the most popular data set used to develop machine learning models to predict the age of abalone. This data set has 4117 data samples with nine attributes. The first five rows of data are displayed in the table below:



a) What is the label of this data set?

(2 marks)

b) What is the general workflow to build a machine learning model?

(3 marks)

c) Propose two methods to check whether the data set has missing data.

(3 marks)

Question 3 continues on page 6

d) A correlation analysis on the data set is collected in the table below:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Length | Diameter | Height | Whole Weight | Shucked Weight | Viscera Weight | Shell Weight | Rings |
| Length | 1.00 | 0.99 | 0.90 | 0.93 | 0.90 | 0.90 | 0.90 | 0.56 |
| Diameter | 0.99 | 1.00 | 0.91 | 0.93 | 0.89 | 0.90 | 0.91 | 0.57 |
| Height | 0.90 | 0.91 | 1.00 | 0.89 | 0.84 | 0.87 | 0.89 | 0.61 |
| Whole Weight | 0.93 | 0.93 | 0.89 | 1.00 | 0.97 | 0.97 | 0.96 | 0.54 |
| Shucked Weight | 0.90 | 0.89 | 0.84 | 0.97 | 1.00 | 0.93 | 0.88 | 0.42 |
| Viscera Weight | 0.90 | 0.90 | 0.87 | 0.97 | 0.93 | 1.00 | 0.91 | 0.50 |
| Shell Weight | 0.90 | 0.91 | 0.89 | 0.96 | 0.88 | 0.91 | 1.00 | 0.63 |
| Rings | 0.56 | 0.57 | 0.61 | 0.54 | 0.42 | 0.50 | 0.63 | 1.00 |

Comment on the outcomes of the correlation analysis. How do the correlations among these attributes affect the feature selections in building machine learning model?

(7 marks)

**Question 4 (20 marks)**

a) Name three common ways to describe molecules in a computer information system.

(3 marks)

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

i) CC(C(=O)O)N

ii) C1=CC=C(C=C1)O

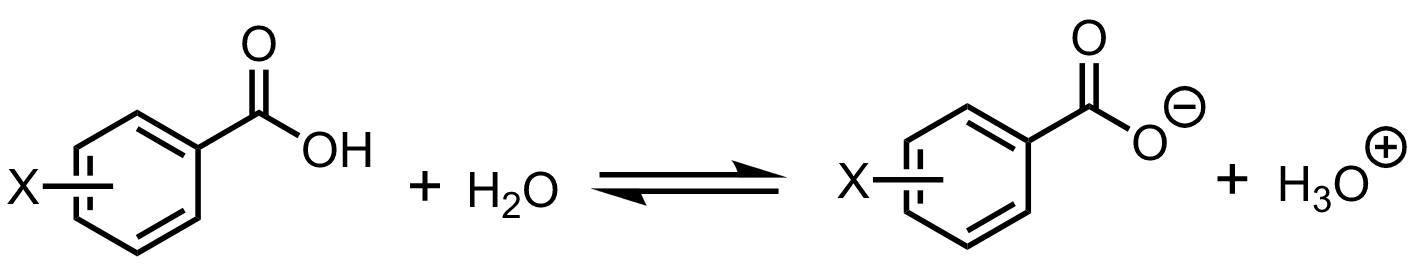
iii) Br[C@@](Cl)(I)F

(6 marks)

c) Many cheminformaticians agree that the establishment of the periodic table of elements marks the start of cheminformatics science. Give the reason.

(3 marks)

d) For the reaction below:



The acid dissociation constants are collected for X = (H, CH3, OCH3, F, Cl and NO2) in Table 1.

Table 1. Dissociation Constants (105×Ka) of Some Substituted Benzoic Acids in Water at 25oC

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| X | H | CH3 | OCH3 | F | Cl | NO2 |
| *ortho* | 6.27 | 12.3 | 8.06 | 54.1 | 11.4 | 671 |
| *meta* | 6.27 | 5.35 | 8.17 | 13.6 | 14.8 | 32.1 |
| *para* | 6.27 | 4.24 | 3.38 | 7.22 | 10.5 | 37.0 |

i) What are the major electronic effects that affect the acid dissociation constant?

Question 4 continues on page 8

(4 marks)

ii) The Hammett equation correlates substitute effect with rate constant and equilibrium constant of benzoic acid and its derivatives in their acid dissociation reactions.

What are and in the Hammett equation?

(4 marks)

Appendix

Table 1. The methods in numpy package in CM4044

|  |  |  |
| --- | --- | --- |
| numpy.arange | numpy.linspace | numpy.logspace |
| numpy.empty | numpy.zeros | numpy.ones |
| numpy.eye | numpy.empty\_like | numpy.zeros\_like |
| numpy.ones\_like | numpy.sort | numpy.argsort |
| numpy.searchsorted | numpy.nonzero | numpy.count\_nonzero |
| numpy.reshape | numpy.transpose | numpy.flatten |
| numpy.ravel | numpy.concatenate | numpy.newaxis |
| numpy.loadtxt | numpy.genfromtxt | numpy.savetxt |
| numpy.save | numpy.savez | numpy.savez\_compressed |
| numpy.sum | numpy.prod | numpy.max |
| numpy.min | numpy.argmax | numpy.argmin |
| numpy.mean | numpy.average | numpy.std |
| numpy.var | numpy.clip | numpy.dot |
| numpy.inner | numpy.linalg.det | numpy.linalg.eig |
| numpy.random.rand | numpy.random.randint |  |

Table 2. Some selected matplotlib methods in 2D plotting

|  |  |  |
| --- | --- | --- |
| plt.figure | plt.subplots |  |
| figure.add\_axes | figure.tight\_layout |  |
| axes.plot | axes.scatter | axes.bar |
| axes.step | axes.fill\_between | axes.pie |
| axes.hist |  |  |
| axes.set\_xlabel | axes.set\_ylabel | axes.set\_xlim |
| axes.set\_ylim | axes.set\_xticks | axes.set\_xticklabels |
| axes.set\_yticks | axes.set\_yticklabels | axes.set\_title |
| axes.legend | axes.annotate | axes.spines |
| axes.xaxis.set\_ticks\_position | axes.yxais.set\_ticks\_position |  |
| axes.xaxis.set\_label\_coords | axes.yaxis.set\_label\_coords |  |
| \*axes.spines[‘Top’].set\_color | \*axes.spines[‘Top’].set\_position |  |
| axes.get\_xticklabels | axes.get\_yticklabels |  |
| plt.label.set\_fontsize | plt.label.set\_bbox |  |
|  |  |  |

\* The value ‘Top’ can be replaced by ‘Left’, ‘Right’, or ‘Bottom’

**End of Paper**