THE UNIVERSITY OF HONG KONG

SCHOOL OF COMPUTING AND DATA SCIENCE
DEPARTMENT OF COMPUTER SCIENCE

COMP7409A Machine Learning in Trading and Finance
Final Examination

Date: December 17, 2024 Time: 6:30pm-8:30pm

INSTRUCTIONS:

a. Answer ALL questions. They are all COMPULSORY.
b. Total mark is 100. The mark value of each question (or part of a question) is indicated before the question (or part of the question)
c. Write your university number clearly at the beginning of your answer script. DO NOT write down your name.
d. Only approved calculators as announced by the Examinations Secretary can be used in the examination. It is the candidates' responsibility to ensure that their calculator operates satisfactorily, and candidates must record the name and type of the calculator used on the front page of your answer script.

1. (24%) This question may require you to either: (i) determine the output of a given Python code snippet, or (ii) write simple Python programs to perform designated tasks. Assume that the libraries 'pandas' as 'pd' and 'numpy' as 'np' have already been imported. If additional libraries are needed for your code, ensure they are imported at the start. Each sub-question carries equal marks.

   (a) What is the output of the following snippet:

```
a = np.array([[1, 2, 3], [2.5, 3.6, 7.3]])
print(type(a), a.ndim, a.size, a.shape)
```

   (b) Insert the missing statements in the following snippet

```
n = int(input())
y = np.arange(1, n)
#missing statement for transforming y
```

   so that when a user inputs 25, y will be transformed to the following numpy array

```
array([[ 0.5,  1. ,  1.5,  2. ,  2.5,  3. ],
       [ 3.5,  4. ,  4.5,  5. ,  5.5,  6. ],
       [ 6.5,  7. ,  7.5,  8. ,  8.5,  9. ],
       [ 9.5, 10. , 10.5, 11. , 11.5, 12. ]])
```

   (c) What is the output of the following snippet:

```
a = np.linspace(1, 10, 10)[8::-2]
print(a)
```

   (d) Write a Python code snippet that generates a graph for the function f(x) = x * log (x), where x ranges from 0 to 10 (inclusive). You should include the necessary library.

   (e) What is the output of the following snippet?

```
a = np.arange(0,24).reshape(2,-1,2)
print(a[np.newaxis,...])
```

   (f) Give the Python statement that products following Pandas Series:

	0
A1	100
A2	25
A3	65
A4	45
Midterm	0

(g) Given the following data frame df:

	Univ	Win	Loss
0	HKU	10	2
1	CUHK	15	7
2	HKUST	4	3

Write a single Python statement that rearranges the rows of df in ascending order of Win, i.e., modify df as follows:

	Univ	Win	Loss
2	HKUST	4	3
0	HKU	10	2
1	CUHK	15	7

(h) Consider the Excel file named "educ.uoe.fine06.xls". This file contains some missing values, represented by ":". Provide the Python statement that imports this file into a Pandas dataframe, ensuring that the missing values are replaced with the NaN value.

2. (26%) The following comma-separated-value file "boston_house_prices.csv" contains information about houses in the suburbs of Boston collected in 1978:

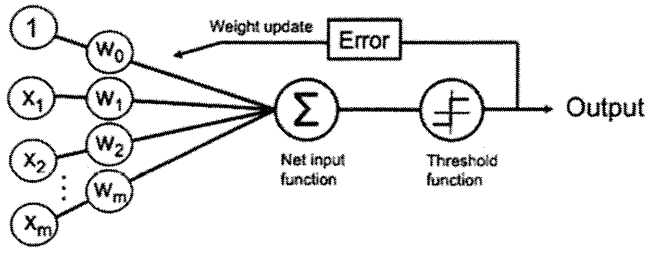
```
CRIM,ZN,INDUS,CHAS,NOX,RM,AGE,DIS,RAD,TAX,PTRATIO,B,LSTAT,MEDV
0.00632,18,2.31,0,0.538,6.575,65.2,4.09,1,296,15.3,396.9,4.98,24
0.02731,0,7.07,0,0.469,6.421,78.9,4.9671,2,242,17.8,396.9,9.14,21.6
0.02729,0,7.07,0,0.469,7.185,:,4.9671,2,242,17.8,392.83,4.03,34.7
0.03237,0,2.18,0,0.458,6.998,45.8,6.0622,3,222,18.7,394.63,2.94,33.4
0.06905,0,2.18,0,0.458,7.147,54.2,6.0622,3,222,18.7,396.9,5.33,36.2
0.02985,0,2.18,0,0.458,6.43,58.7,6.0622,3,222,18.7,394.12,5.21,28.7
```

Note that there are some missing values in the file which are represented by the character ":". The meaning of each column is given as follows:

- CRIM : Per capita crime rate by town
- ZN : Proportion of residential land zoned for lots over 25,000 sq. ft.
- INDUS : Proportion of non-retail business acres per town
- CHAS : Charles River dummy variable (= 1 if tract bounds river and 0 otherwise)
- NOX : Nitric oxide concentration (parts per 10 million)
- RM : Average number of rooms per dwelling
- AGE : Proportion of owner-occupied units built prior to 1940
- DIS : Weighted distances to five Boston employment centers
- RAD : Index of accessibility to radial highways
- TAX : Full-value property tax rate per $10,000
- PTRATIO : Pupil-teacher ratio by town
- B : $1000(Bk - 0.63)^2$, where Bk is the proportion of [people of African American descent] by town
- LSTAT : Percentage of lower status of the population
- MEDV : Median value of owner-occupied homes in $1000s

This question asks you to design a system that makes use of "boston_house_prices.csv" to develop a machine learning system to predict asset prices (or more precisely, the column MEDV) in suburban Boston. You should give details of the necessary steps for the implementation of this system, including data cleaning, exploratory data analysis, feature selection, model training, and testing. This question is open-ended, allowing for the inclusion of any supplementary steps you consider appropriate.

3. (25%) Consider the following Perception neural network



(a) (13%) Develop a Python module that defines the class Perceptron, containing the methods __init__(), fit() and predict(). The Perceptron class should utilize the Perceptron learning rules for data training and Perceptron prediction rules for making predictions. You are encouraged to incorporate additional auxiliary functions to support your implementation.

(b) (12%) You are asked to utilize your Perceptron class to train a model for predicting the gender of a high school student based on their height. You can access the dataset "heights.csv," which contains the heights (in inches) of 1050 students along with their gender.

```
› heights
       sex   height
1     Male 75.00000
2     Male 70.00000
3     Male 68.00000
4     Male 74.00000
5     Male 61.00000
6   Female 65.00000
7   Female 66.00000
8   Female 62.00000
9   Female 66.00000
10    Male 67.00000
```

You should use part of the data set to train a model and the remaining data to test the performance of your model.

4. (a) Write a Python program that develops and assesses a Linear Regression Machine Learning model to predict Bitcoin prices for the following day based on the current day's prices. Your program should import the required Bitcoin price data from Yahoo Finance (using the ticker "BTC-USD"), specifically from the "Close" column of the dataframe provided by Yahoo Finance. Your program should correctly partition the data into a training set and a testing set. Use the training set to train your Linear Regression model, and then use the testing set to evaluate the performance of your model.

(b) Despite your efforts to improve your model using more complex and advanced ML models, you find that the performance remains unsatisfactory. This could be attributed to the fundamental nature of Bitcoin. Please provide an explanation as to why ML models, based solely on Bitcoin prices, might not yield accurate predictions

END OF PAPER