THE HONG KONG UNIVERSITY OF SCIENCE & TECHNOLOGY
Department of Computer Science and Engineering
COMP4331:Introduction to Data Mining
Fall 2019 Assignment 2
Due time and date: 1:20 pm, Nov 11 (Mon), 2019. (End of the lecture)

Name: Email: Student ID:

IMPORTANT NOTES

- **Your grade will be based on the correctness, efficiency and clarity.**
- **Both the report and the code will be graded.**
- **Late submission: 25 points will be deducted for every 24 hours after the deadline.**

Q1. (40 points) You are given the following dataset.

	I_1	I_2	T
e_1	2	1	0
e_2	-1	3	1
e_3	5	1	0
e_4	1	3	1

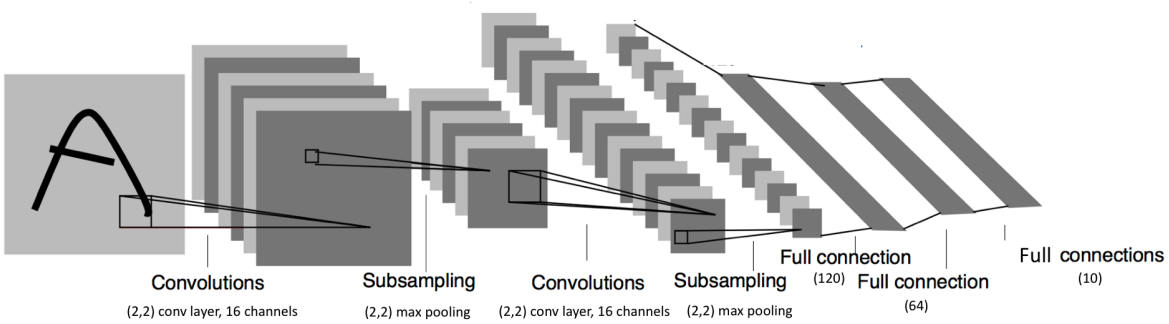
Run the perceptron algorithm (p19 of chap3e_ann1x.pdf) by hand, and write the results in the following format.

iteration	W^{old}	I	T	O	T=O?	W^{new}
1						
2						
3						
4						
⋮						
⋮						
⋮						

Q2a. (40 points) In this question, you are required to implement (using Keras) a 10-output CNN with the following layers:

1. 16 channels of 2×2 convolution, with ReLU activation;
2. max-pooling layer with stride 2;
3. 16 channels of 2×2 convolution, with ReLU activation;

4. max-pooling layer with stride 2;
5. fully connected layer with 120 ReLU hidden units;
6. another fully connected layer with 64 ReLU hidden units.



In this experiment, we use the MNIST (`https://hkustconnect-my.sharepoint.com/:u:/g/personal/hzhangal_connect_ust_hk/ERzaZJwExepPlf92u_1cCPABLyuC21lBggcZ9GHx0mpyPQ?e=YklceJ`) dataset. The first column is the class label. The other columns are the intensity values for each individual pixel in each MNIST image. Each student will have his/her own test set (which is based on your student id). Run your code using adam as the optimizer, train your model for 10 epochs on the training set, and report the accuracy on your test set.

b. Now we vary the architecture. For each of the variations below, train your model for 10 epochs on the training set, and report the accuracy on your test set.

(a) Change all the convolution filters to 4×4;

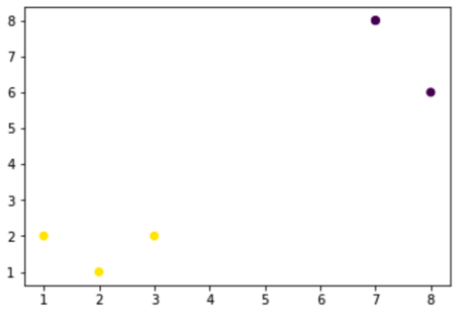
(b) Change the number of channels in all convolution layers to 32;

(c) After the last max-pooling layer, add one more layer of 16 channels of 2×2 convolution with ReLU activation, and a max-pooling layer of stride 2.

Q3. (20 points) In this question, you use the code segments provided in the tutorial to implement the k-means and k-medoid algorithms as follows:

```
kmeans k input-file output-file
kmedoids k input-file output-file
```

k-means and k-medoid algorithms as follows: The input-file is the dataset (with one sample per line), and the output-output is a png file, which use different colors to show the different clusters. An example is shown below.



Run the algorithms (with $k = 2$) on the following data sets, and output the png files (use colors yellow and purple for the two clusters).

(a) Dataset:

```
1,2
3,2
5,1
3,5
8,7
12,8
10,9
0,0
```

(b) Dataset:

```
1,2
3,2
5,1
3,5
8,7
12,8
10,9
0,0
40,40
```

(c) Based on the results from parts (a) and (b), what do you observe?

Submission Guidelines Print your codes for Q2 and Q3 (please use the jupyter notebook and print the ipynb file with all the **results**. Please remember to write your name, email address, and student id on your codes). For Q2b, please print the code and output for different hyper-parameter settings using jupyter notebook such that the TA can verify your result. Write all the required answers in this report and print the report. You should put the code and report together and submit it to TA after the lecture on Nov 11. The TA(s) will be waiting there.

Note that the assignment should be clearly legible, otherwise you may lose some points if the assignment is difficult to read. Plagiarism will lead to zero point on this assignment.