THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY(GUANGZHOU)

MSC DSAA 5002

Fall 2023 Final Examination

Exam Type: open book

Exam Duration: 7:00pm Dec 14, 2023 to 7:00pm Dec 17, 2023

Exam Rule: Must be completed by individual students. Students cannot collaborate with anyone.

This exam contains 7 questions.

Problem	Task	Max Points
1	Supervised Outlier Detection	15
2	Weather Recognition	15
3	Short Video Classification	15
4	Recommendation and Business Analysis	15
5	Smoke Status Recognition	15
6	Bank Customer Clustering	10
7	Social Media Network Analysis	15
Total	/100	

Notes!!!

Exam Data

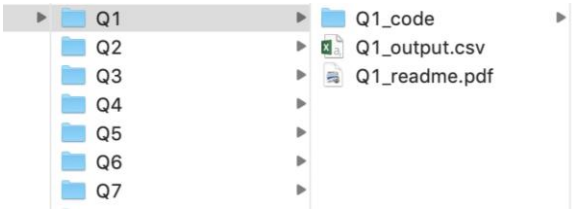
1. Data link:

   https://drive.google.com/drive/folders/1up0Gz-zIzpt73iaXHE_wmJY66GkzMRve?usp=sharing

2. Since the amount of data on some questions is relatively large, in order not to affect the progress of the exam, students can try to figure out the questions without using data first.
3. The data is in the corresponding folder, for example, the data of question1 is in folder Data_Q1.

Submission

1. Submit via Canvas.
2. Students should submit files for each question in corresponding folder named as QX (X represent question number). For example, students must put files for Question1 in Q1 folder.
3. Students should pack all the folders together in one folder name as DSAA5002_studentID_name_final. If there are codes, student should pack all your code files in a folder named as QX_code(X represent question number). The example of directory structure is as below(Note that the files in QX changes according to the requirements for each question):



4. Compress DSAA5002_studentID_name_final folder and submit.
5. Students MUST submit all your files before 7:00pm(19:00) Dec 17, 2023. In order to avoid network congestion and submission failure, please submit your attachment in advance. Any late submissions will not be accepted.

Others

1. For programming language, in principle, python is preferred. Code MUST be runnable, and code comments are necessary. Missing the necessary comments will be deducted a certain score. For programming question, your grade will be based on the corresponding metrics, efficiency and clarity.
2. If your code or answer refer to any blog, github, paper and so on, please write the their url in corresponding readme.pdf.
3. Computation of some questions is very large, students might use cloud computing platform, such as azure, AWS, aliyun.
4. This exam must be completed by individual students. Students cannot collaborate with anyone.
5. Please arrange your time reasonably and try to answer every question, since report also takes part of the score.

6. If students have any question about this exam, please sent email to jli226@connect.hkust-gz.edu.cn during examination.
7. Plagiarism will lead to zero points.

Q1. Supervised Outlier Detection (15 pts)

In our community, many households have cats as pets, and we have equipped our feline friends with a special device designed to detect falls. It's important to note that we are not only concerned about falls from great heights, but also instances where the cat may fall from tables or similar surfaces. This task can be treated as an outlier detection problem. The cat's wearable device collects multiple signals every millisecond, corresponding to the variables x, y, z, a, b, c, and d in the dataset. The occurrence of a cat falling is labeled as 'Is_falling,' where 0 indicates normal behavior and 1 indicates a fall.

Data Description:

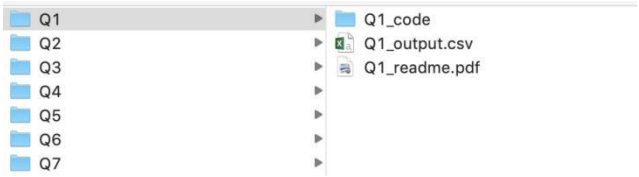
1. All the data is in Data_Q1.
2. Folder train contains all training data
3. Folder test contains all the testing data
4. Column 'Is_Falling' is the label

Submission:

1. Please write your main experimental steps and the methods to a report in **Q1_readme.pdf**. If your code refer to any blog, github, paper and so on, please write the their links in it.
2. Output your testing results in **Q1_output.csv.** Your .csv file should contain 2 columns as shown below. In "Result", 0 represents negative and 1 represents positive.

ID	Is_Falling
1	0
…	…
n	0

3. Pack all code files in folder **Q1_code**.
4. Pack all files/folders above in folder **Q1** like below:



Notes:

1. *Because the number of outliers and inlier is extremely uneven, you need to deal with the problem of data imbalance in the given dataset. And the recall rate of the class 1 is much important.*
2. *You are allowed to use any of the methods we mentioned in class or methods and libraries you searched from the Internet.*

Grading:

We will grade according to the code, the experiment steps and methods you mentioned in the report, and the recall and precision of your model's prediction.

Q2. Weather Recognition (15 pts)

Suppose you are preparing to train an image classifier for weather recognition, with the goal of predicting five categories: Sunny, Snowy, Cloudy, Rainy, and Foggy. We provide only the raw image data, and you are required to handle the data processing and train a deep learning model on your own. In this task, you do not need a test set, but you must 1) design a dataloader and successfully construct a well-format training set using ALL the 250 images. 2) You are also required to construct a deep learning model, and successfully train it for at least 50 epochs. 3) After training the model, you need to test the model on ALL the training data and report the accuracy.

Data Description:

1. All the data is in Data_Q2.
2. The category labels are implied in the image naming.

Submission:

1. Please write a report 'Q2_readme.pdf' of up to 3 pages, including the following:
    a. The design of your dataloader(e.g., how you standardized input data, defined labels, whether shuffling was applied, and the chosen batch size).
    b. Code screenshots related to (a)
    c. A brief introduction to your model along with relevant code screenshots.
    d. Screenshots depicting the training process.
    e. **Accuracy** on the training set (**with screenshots**).
2. Pack all code files in folder **Q2_code**.
3. Pack all files/folders above in folder **Q2**:

Notes:

1. *If you are unfamiliar with what a dataloader is in deep learning, please take the initiative to study it on your own.*
2. *Your labels should be in the form of one-hot encoding, such as [0, 0, 1, 0, 0].*
3. *When the sizes of input images are inconsistent, you can use resizing.*

Grading:

The scoring will be based on the following criteria: TA can successful running your code(MUST). the quality of your report (10 pts), and whether the accuracy of your model on these 250-image training set can exceed 89.5% (5 pts).

Q3. Short Video Classification (15 pts)

Short video applications are becoming more and more popular among the young. In reality, internet companies generally use automatic classification algorithms to process large amounts of short video uploaded by users. Now you are asked to implement a short video classification algorithm.

Data Description:

1. Data is in Data_Q3 folder:
2. In our data set, there are a total of 2063 training videos (in the "train_video" folder) and 896 test videos (in the "test_video" folder). They belong to the following 15 categories:

Label ID	Video Content
0	dog
1	boy selfie
2	seafood
3	snack
4	doll catching
5	Ballroom dance
6	origami
7	weave
8	ceramic art
9	Zheng playing
10	fitness
11	parkour
12	diving
13	billiards
14	eye makeup

"train_tag.txt" stores the label information. For example, in the line "873879927.mp4,3", "873879927.mp4" represents the file name of the video, "3" is the label of the video.

Requirments:

➢ About training:

1. You can use any algorithm that you know.
2. You **can not** directly use complete models that others have already trained to do classification without any detailed process.

Submissions:

1. Please write down your algorithm details in the **Q3_readme.pdf**.
2. Please put all the code of this question in the **Q3_code** folder.
3. You need submit **Q3_output.csv.** Your .csv file should contain 2 columns as shown below.

file_name	label
861108106.mp4	0

…	…
801454381_11_21.mp4	13

4. Put all files/folders in **Q3** folder.

Grading:
Your grade will be based on your report, code and accuracy of the results.

Q4. Recommendation and Business Analysis(15 pts)

We offer you 1M retail transaction records from a specific city covering the period from 2009 to 2011. Assuming the role of a Business Analyst(BA), your objective is to compose a report based on this dataset to provide business insights for a retail supermarket. You should write a report encompassing the following two components: 1) Utilize visual methods to create a minimum of 10 pictures, delivering no fewer than 10 business insights to the supermarket owner. 2) Utilize **association rule analysis** to offer the supermarket owner no fewer than 10 sales (recommendation) suggestions.

Data Description:

1. Data is in Data_Q4 folder.
2. In our data set, there are a total of 1M transaction records.

Submission:

1. Please write a report 'Q4_readme.pdf' of up to 8 pages, including the following:
    a. At least 8 pictures and at least 8 business insights
    b. The algorithm details, process, and results of association rule analysis, along with providing no fewer than 5 sales/recommendation suggestions.
2. Pack all code files in folder **Q4_code**.
3. Pack all files/folders above in folder **Q4**:

Grading:

Your grade will be based on your report(12 pts) and code(3 pts). It's worth noting that you must use Python to plot your pictures.

Hints:

The business insights could include but is not limited to:

1. Total sales volume per product per year …
2. The average spending per customer and the overall distribution …
3. …

Q5. Smoke Status Recognition (15 pts)

Smoking is one of the major health problems. From a biomedical point of view, we can determine whether a patient smokes from certain biometric information. Now you are required to implement a binary algorithm to predict a patient's smoking status given information about various other health indicators.

Data Description:

1. Data is in Data_Q5 folder.
2. In the dataset, there are a total of 159,256 training samples (in the "Q5_train.csv" file) and 106,171 test samples (in the "Q5_text.csv" file).

Submissions:

1. Please write down your algorithm details in the **Q5_readme.pdf**.
2. Please put all the code of this question in the **Q5_code** folder.
3. You need submit **Q5_output.csv.** Your *.csv file should **contain 2 columns** (as shown in the "Q5_sample_submission.csv" file).
4. Put all files/folders in **Q5** folder.

Grading:

Test set score (50%) + Your report (50%). Your test set score will be evaluated based on the ROC metric on the test set. If your ROC on the test set is greater than 0.87, you will receive full marks; otherwise, you will only receive half of the marks. TA will test your Q5_sample_submission.csv file.

Q6. Bank Customer Clustering (10 pts)

Banks classify or cluster customers for several reasons, aiming to enhance their operational efficiency, tailor services, and manage risks effectively. You are required to offer insights as a data analyst. Furthermore, you are expected to employ at least three distinct clustering algorithms to categorize customers.

Data Description:

1. Data is in Data_Q6 folder.

Submission:

1. Please write a report 'Q6_readme.pdf' of up to 6 pages, including the following:
    a. Please explore the data table using visualization techniques. Please provide at least 10 pictures and at least 10 business insights
    b. Please use at least three different clustering algorithms to cluster customers. You can use any variables you consider valuable.
    c. Please explain the common characteristics shared by customers within the same cluster after your clustering, as well as the differences among customers in different clusters. Please provide evidence.
2. Pack all code files in folder **Q6_code**.
3. Pack all files/folders above in folder **Q6**:

Grading:

Your grade will be based on your report(8 pts) and code(2 pts). It's worth noting that you must use Python to plot your pictures.

Q7. Social Media Network Analysis (15 points)

Different users can become friends on social networks, forming a vast social network. In this question, you will receive user social network data from a certain platform. You are required to conduct various graph analyses and provide insights into social relationships.

Data Description:

1. Data is in Data_Q7 folder.
2. In the dataset, there are 1,134,890 users and 2,987,624 edges.

Submissions:

1. Complete the following tasks using code and present the process, results, and **analysis** in the report. **Q7_readme.pdf:**

   a. Please calculate and plot the *clustering coefficient and degree distribution* of the network.

   b. Identify the most influential nodes in a network and analyze them. Please use *centrality* metrics such as degree centrality and then visualize.

   c. Identify *Isolated Nodes* in the Network

   d. Recognize *Connected Components* in the Network

   e. Compute *Average Shortest Path Length* of the Network.

   f. Calculate the *Diameter* of the Network

   g. Detect Community Structures in the Network. Please employ community detection algorithms (e.g., Louvain algorithm) to find community structures within the network. Please analyze the detection results, including factors such as the number of communities and statistics within the community. Assess the impact of relevant parameters(if any). Try your best to visualize the detection results.

2. Pack all code files in folder **Q7_code**.
3. Pack all files/folders above in folder **Q7**

Grading:

Your grade will be based on your report(12 pts) and code(3 pts).

Notes:

1. The edges are undirected.
2. Python package *networkx* may be useful. No restrictions on the use of packages.