THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

MSCBDT 5002:

Fall 2018 Final Examination

Exam Type: open book

Exam Duration: 00:00 am Dec 8, 2018 to 12:00 pm Dec 10, 2018

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

Turn In: mscbdt5002fall18@gmail.com

This exam contains 8 questions in 9 pages (not including the cover page).

You have 60 hours to complete this exam.

Question	Max Points
1	15
2	8
3	5
4	8
5	18
6	18
7	10
8	18
Total	100

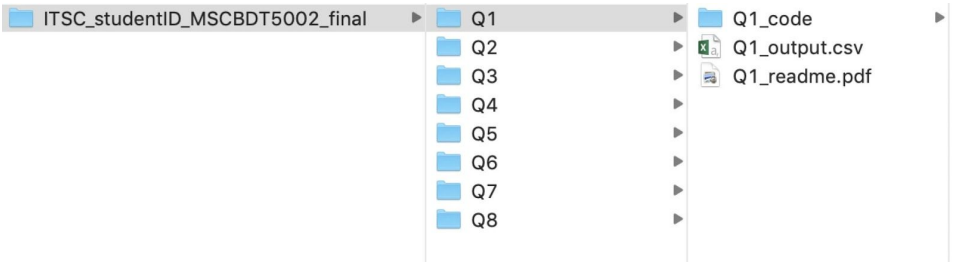
Notes!!!

Exam Data

1. Data link: https://drive.google.com/drive/folders/195txyfFHJ3oqVpA0zWccggisJK1ETbaW?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 each question, for example, the data of Question 1 is in folder Data_Q1.

Submission

1. Turn in: mscbdt5002fall18@gmail.com.
2. Students should submit files/folders for each question in corresponding folder named as QX (X represent question number). For example, students must put files for Question 1 in Q1 folder.
3. Students should pack all the folders together in one folder name as ITSC_studentID_MSCBDT5002_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 shown as below (Note that the contents in QX changes according to the requirements of each question):



4. Compress ITSC_studentID_MSCBDT5002_final folder and sent to specified email address. Don't use any web link in email, you must submit it **as an attachment.**
5. If you store training models or use some data resources, please put the them into the network disk, and then write the links in the readme.pdf. Please do not add them directly into your attachment.
6. You **MUST** submit all your files before 12:00 pm Dec 10, 2018. In order to avoid network congestion and submission failure, please submit your attachment in advance. **Any late submissions will lead to zero points.**

Other

1. There is noise in real data, please remove noise first.
2. For programming language, in principle, python is preferred. Codes 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.
3. If your codes or answer refer to any blog, github, paper and so on, please report their links in corresponding readme.pdf.
4. Computation of some questions is very large, students might use cloud computing platform, such as azure, AWS, aliyun.
5. This exam must be completed by individual students. Students cannot collaborate with anyone.
6. Please arrange your time reasonably and try to answer every question, since report also takes part of the score.
7. If students have any question about this exam, please sent email to mscbdt5002fall18@gmail.com during the examination.
8. **Plagiarism will lead to zero points.**

Q1. Supervised Outlier Detection (15 points)

In this question, you need to use a supervised classification model to find outliers from our given image data set. The data set will contain two types of tags: outliers and inliers. And the main content of the data set is some random scenes with text as the main body.

Data Descriptions:

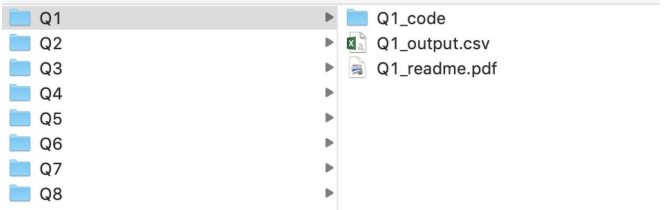
1. All the data is in Data_Q1.
2. Folder Outlier_train contains all training data labeled as outlier.
3. Folder Inlier_train contains all training data labeled as inlier.
4. Folder test contains all the testing data.

Submissions:

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 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	Result
0	0
1	1
…	…
n	1

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.*
2. *You are allowed to use any of the methods we mentioned in class or methods and libraries you searched from the Internet.*
3. *We will grade according to the code, the experiment steps and methods you mentioned in the report and the recall and precision of the your model's prediction.*

Q2. Grid-Based Outlier Discovery Approach (8 points)

In this question, you should implement a grid-based outlier detection method to find outliers in a large data set.

Data Descriptions:

1. Relevant data is in folder Data_Q2.
2. X.csv: Testing data, as input.

submissionSample.csv: sample of submission, 0 indicate inlier, 1 indicate outlier.

Requirements:

1. No relevant third-party packages, you must implement the algorithm by yourself.

Submissions:

1. Please report your main experimental steps in **Q2_readme.pdf**. If your codes refer to any blog, github, paper and so on, please report their links in it.
2. Output your results in **Q2_output.csv**. The format refer to submissionSample.csv or below. Note that the .csv file should contain one column.

result
0
1
…
1

3. Pack all code files in folder **Q2_code**.
4. Pack all files/folders above in folder **Q2**.

Notes:

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

Q3. Data Augmentation (5 points)

We all know that adequate training data is a precondition for training machine learning models. But in real-world problems, the data that can be used to train the model is often not enough. Suppose you are doing a classification task and your training dataset is extremely insufficient. Please explain how you will expand the amount of data.

Notes:

You do **NOT** need to code in this question, but you need to answer in detail. Please give at least **two** specific examples to illustrate, such as image classification, text classification and so on. You can also refer to other materials to answer this question, if you do so, please also list your references.

Submissions:

1. Put your answer and references in **Q3_readme.pdf**, and put it in folder **Q3**.
2. No page limit for the answer.

Q4. Expectation-Maximization Algorithm (8 points)

In this question, you are required to code by yourself to complete the EM algorithm.

Data Descriptions:

1. The data is in Data_Q4 folder.
2. The test data is shown in **Q4_Data.csv**. There are 6 attributes, which are 'A','B'…'F', and totally 626 instances in the dataset. You need to cluster all the instances into two classes. Assume the initial centers are c1=(0,0,0,0,0,0) and c2=(1,1,1,1,1,1).

Requirements:

1. Report the updated centers and SSE for the first two iterations.
2. Report the overall iteration step when your algorithm terminates.
3. Report the final converged centers for each cluster.

Submissions:

1. Put all reports in requirements in **Q4_readme.pdf**.
2. Submit your source code in folder **Q4_code**.
3. Put files/folder above in folder **Q4**.

Notes:

Please use the terminate condition below:

Terminate condition: the EM algorithm will terminate when:

1). The sum of L1-distance for each pair of old-new center

$$\sum_{each\ center} \|C_{old} - C_{new}\|_1$$

is smaller than 0.0001, or

2). The iteration step is greater than the maximum iteration step 100.

Q5. Sentiment Analysis and Opinion Mining (18 points)

Generally speaking, sentiment analysis aims to determine the attitude of a speaker, writer, or other subject with respect to some topic or the overall contextual polarity or emotional reaction to a document, interaction, or event. The attitude may be a judgment or evaluation (see appraisal theory), affective state (that is to say, the emotional state of the author or speaker), or the intended emotional communication (that is to say, the emotional effect intended by the author or interlocutor).

Recently, the birth of genetically edited babies has created a huge controversy. People have different opinions on the development of genetic technology. Now you are asked to do a Sentiment Analysis Task based on topics such as "gene editing", "genetic engineering", and "transgene".

In this task, you need to implement a series of processes from background investigation to collecting data to determining the solution to implementing the

algorithm to get the results.

***Requirements*:**

➢ About training:

1. You can use any algorithm that you know, supervised learning and unsupervised learning are both ok.
2. You can use any data resource. You need to find your own data resources such as some corpus or lexical resource.
3. You **can not** directly use complete models that others have already trained to do classification without any detailed process.
4. You can use some basic word vector models to build your algorithm, such as word2vec.

➢ About testing:

1. You need to collect **100 pieces** of news/comments/articles related to the above topic, then use your algorithm or model to divide them into two categories——positive or negative. (You may need some knowledge of Crawler, in Python, BeautifulSoup is a very useful crawler tool.)
2. You can get the test text from any website or social media.
3. The text you collect must be **in English**.

Submissions:

1. Please write down your algorithm details and all links of the model/data resources you used in the **Q5_readme.pdf**. If your code refer to any blog, github, paper and so on, please write the their links in it.
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 3 columns as shown below. In "Result", 0 represents negative and 1 represents positive.

ID	**Contents**	**Result**
0	text0	0
1	text1	1
…	…	…
99	text99	1

4. Put all files/folders above in folder **Q5**.

Notes:

1. Crawler is not required and will be not included in the scoring criteria. You can also get the text manually or by other tools.
2. Your grade will be based on your report, code and accuracy of the results.

Q6. Short Video Classification (18 points)

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 Descriptions:

1. Data is in Data_Q6 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.

Requirements:

➢ 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.

➢ About grading rule

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

Submissions:

1. Please write down your algorithm details in the **Q6_readme.pdf**. If your code refer to any blog, github, paper and so on, please write the their links in it.
2. Please put all the code of this question in the **Q6_code** folder.
3. You need submit **Q6_output.csv.** Your .csv file should contain 2 columns as shown below.

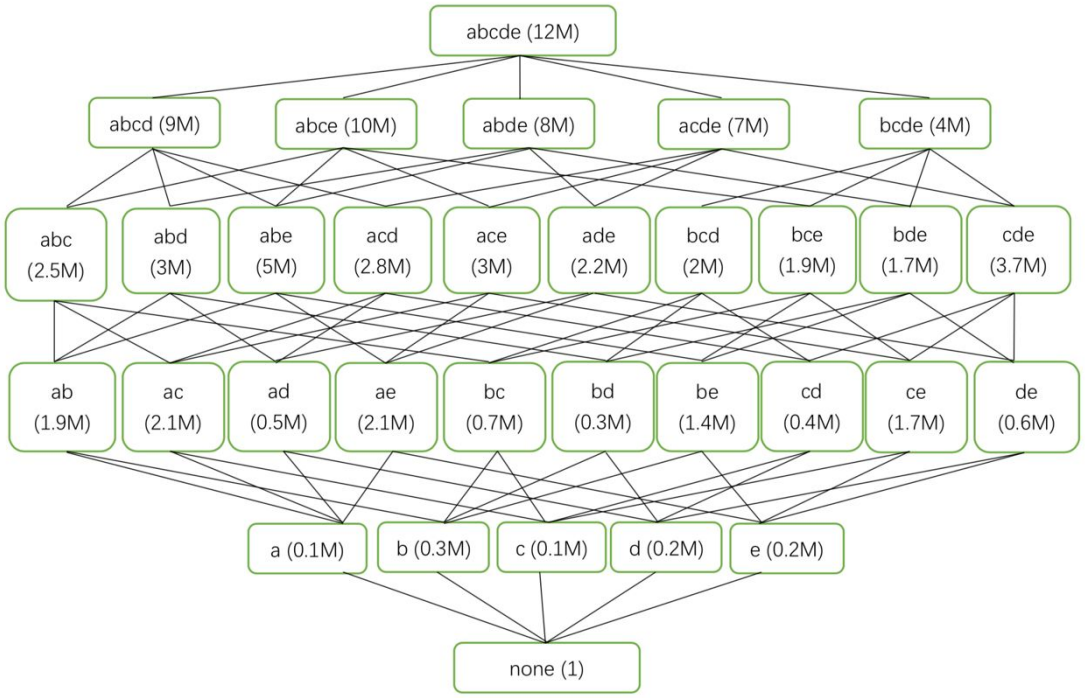
file_name	label

861108106.mp4	0
…	…
801454381_11_21.mp4	13

4. Put all files/folders in **Q6** folder.

Q7. Selective Materialization Problem (10 points)

(1) Can you select a set V of k views such that Gain (V U {top view}, {top view}) is maximized? Set k=3. Please give your answer. **(7 points)**
(2) The lecture note shows how greedy algorithm perform badly. Please give a complete proof of the lower bound of this greedy algorithm. (Maybe you need some references.) **(3 points)**



Requirements:

1. For (1), you must code by yourself rather than calculate by hand.

Submissions:

1. Put your codes in **Q7_code** folder.
2. For (1), you should give the answer in **Q7_readme.pdf**.
3. For (2), you should give the proof in **Q7_readme.pdf**.
4. Put all files/folders in **Q7** folder.

Q8. Recommendation System (18 points)

You have learned some basic models including user-based and item-based collaborative filtering methods in class. However, some features of items or users can also help to improve the performance of recommendation system.

In this question, you are given a movie rating dataset which contains basic rating information, movie titles, movie genres and user information. You should try to figure out how to utilize these features to construct a recommendation system.

You need to:

Based on rating_train.csv and other relevant data in this question, build a recommendation system to predict user ratings for movies in rating_test.csv.

Data Descriptions:

1. Data is in Data_Q8 folder.
2. Data descriptions are shown in Data_Q8.

Submissions:

1. Put all you codes in **Q8_code** folder.
2. Your prediction result named as ***Q8_output.csv***. (**Notes:** Each line represents the user's rating of the movie, which means your final output should contain 3 columns: 'UserID', 'MovieID' and 'Rating')

Bonus:

There will be some bonus score if you use some creative or the state-of-arts models . Please report the advantages of your methods and list all your references in ***Q8_readme.pdf.***