Course Project Specification of CSC6004 (Data Mining)

This document presents the detailed specification of the course projects of CSC8004 (Data Mining).

The project assessment of the course contains two components, **a small project** and **a major project**. Students are required to complete **both of them**.

Part I Small Course Project (20 Marks)

Given the following dataset which contains 14 two-dimensional data objects (each data object has its own *x* and *y* coordinates) and we want to carry out a mini data mining project to perform *k*-Means clustering on this dataset to generate clusters. We use the **Euclidean distance** to calculate the distance between a pair of data objects in the clustering.

Index	**x**	**y**
1	1.5	1
2	1	1.5
3	1.7	1.2
4	0.7	1.1
5	4	9
6	3.9	7.7
7	4.1	8
8	4.3	7.9
9	3.7	7
10	3.7	1
11	4.2	0.4
12	4	1.5
13	4.3	1.8
14	0.8	7.2

Students are encouraged to **manually** produce the clustering results in this project to fully understand how *k*-Means clustering works.

Answer the following questions:

Q1. How many iterations are needed for *k*-Means clustering (In other words, how many times you have to run *k*-Means before it can be terminated) on the dataset when *k*=2 with the initial cluster centres being (0.8, 7.2) and (0.7, 1.1). Please present the result of each iteration of the clustering process as well as visualize the final clustering result using a scatterplot that highlights clusters using different colors or shapes; (5 Marks)

Q2. How many iterations are needed for *k*-Means clustering on the dataset when *k*=3 with the initial cluster centres being (0.8, 7.2), (0.7, 1.1) and (4.3, 1.8). Please present the result of each iteration of the clustering process as well as visualize the final clustering result using a scatterplot that highlights clusters using different colors or shapes; (5 Marks)

Q3. How many iterations are needed for *k*-Means clustering on the dataset when *k*=3 with the initial cluster centres being (4, 9), (0.7, 1.1) and (4.3, 1.8). Please present the result of each iteration of the clustering process as well as visualize the final clustering result using a scatterplot that highlights clusters using different colors or shapes; (5 Marks)

Q4. Based on your observations on the clustering results of Questions 1-3, please comment on the impact of the values of *k* and the initial selection of the cluster centres on the efficiency (i.e., the number of iterations taken) and accuracy of clustering of the *k*-Means clustering on the dataset; (4 Marks)

Q5. Which data in the dataset are more likely considered as outlier(s) and why? (1 Marks)

Note:

- For Q1-3, the result of each iteration of clustering is presented in the following format which contains the information of the centroid as well as the data in each cluster:

    - ▪ Iteration *i*: Centroid of Cluster *k (x, y)*
      Data in Cluster *k* (data1, data 2, …data n)

  - For example:
    - ▪ Iteration 1: Centroid of Cluster 3 *(0.5, 0.7)*
      Data in Cluster 3 (#1, #4, #7)

    means in the first iteration, the centroid of the 3rd cluster is (0.5, 0.7) and the cluster contains three data objects which are the 1st, 4th and 7th data in the dataset, respectively.

- Please kindly note that **you do not need to do any programming or implementation to complete the small project**. As you can see, the small project is actually a few questions around *k*-Means on a toy dataset. Given the extremely small number of data in the dataset, it is expected that you go through each iteration of the clustering process manually so that you can gain a better hands-on understanding on how *k*-Means works. Of course, there is no way for me to stop you from using any existing tools or systems. However, please be mindful that many of the existing tools/systems only provide the last result of the clustering without giving you the results of the intermediate iterations.

- To present clusters using different colors in Excel, you need to create different data series corresponding to different clusters. Right-click your initial scatterplot and choose to select data from the spreadsheet for creating different data series. Then, Excel will use different colors automatically to show different data series in the scatterplot. You can also use any other visualisation tools or packages to present the clusters.

Part II Major Course Project (80 Marks)

2.1 Project options

There are two possible types of major projects that you can conduct in this course based on your personal preference.

Option 1: Implementation and application of the existing data mining algorithms

You can choose to implement **at least ONE** of the following mainstream data mining algorithms:

- **k-Means algorithm for clustering**
- **Apriori algorithm for associate rule mining**
- **LOF algorithm for outlier detection**

Besides implementing one of the above algorithms, you need to apply the algorithm to at least one real-life dataset to perform the corresponding data mining function, either clustering, association rule mining or outlier detection. You need to present the results obtained by applying the algorithm together with any useful, interesting findings, such as the patterns or knowledge, discovered from the dataset.

Option 1 is appropriate for the students with strong programming background who are interested in implementing mainstream data mining algorithms in order to understand how those techniques work under the hood. Necessary graphical user interface should be developed to allow friendly human-computer interaction and result visualisation. It's preferred that the algorithms be implemented using one of the popular programming languages such as C/C++, Java or Python, but we accept other programming languages for implementation as well.

It is important to note that in implementing the algorithm(s), **you are not allowed to directly call the algorithm if it has already been well encapsulated in the programming language**. In other words, you are not allowed to call *k*-Means algorithm (Apriori or LOF algorithm) directly as a function from the library if it has already been implemented by the programming language. Source codes will be checked by the marker to ensure this requirement is reinforced.

Option 2: Conduct your own project

You can also opt in doing a research project in data mining if you already have some problems which can be solved using data mining methods based on your working or studying experience.

It is expected that the project is consistent with the first option in terms of the difficulty level and workload. Overly simple projects may be subject to a great loss of marks. Students choosing this option are strongly encouraged to approach the examiner for consultation to clear any doubts about their project topics.

The projects in this category should encompass the following ingredients, though your report may not be organised in the exact same way:

- **Motivations**: the students should establish the motivations for pursuing this project. What are the problems you want to solve using data mining methods?
- **Methodology**: what are the data mining techniques and/or systems you want to use to solve the problem and how to do this?
- **Datasets**: what are the dataset(s) to be used for this project and how do you acquire them?
- **Experimental evaluation**: how well the applied data mining methods/systems solve your problem in terms of efficiency (speed) and effectiveness or against other performance metrics if appropriate?

You can carry out programming and/or use the existing data mining software or tools such as *WEKA* or *Rapidminer* to complete the project.

Option 2 is appropriate to the students who already have some practical problems in mind to solve using data mining techniques and may have some previous experience in data analytics.

2.2 Deliverables of the major project

You need to deliver the following several items in the major project, regardless of the option you choose.

a) Proposal of the major project

A proposal (1–2 pages) for the final major project needs to be submitted which outlines the background, motivations or aims, problem formulation, possible data mining solution, system architecture design, datasets and a timeline for completing the project. What you need to submit is a standalone PDF file.

It is expected that your proposal is consistent with the final project you will conduct, even though opportunities are provided for updating your proposal in the final submission.

b) Report of the major project

Typically, the report should be over 20-page long (single spacing with reasonable margins) and must at least contain the following sections:

- **A 1-page cover page** at the beginning of the report which contains the following basic information：
    - Your name;
    - Your USQ ID；
    - Your project option (1 or 2);
    - Abstract of your project (Briefly present the problem you want to solve, the data mining method(s) you have used and the final mining results)
- **Background and motivation**. Discuss the background of the project and motivation behind your decision for carrying out the project;
- **Problem formulation.** Present the formulation of the problem you are going to resolve using data mining techniques and point out the possible challenges in solving the problem;
- **Literature review.** A short literature review that you have conducted to survey the existing work related to the data mining problem that you are going to tackle and the existing methods in literature. This review is supposed to be brief and the coverage of

over 10 related and recent (e.g., published in the last five years) papers is deemed sufficient. Complete and correct reference information need to be provided at the end of the report for the papers cited;

- **Design and architecture**. Present your design as to how the problem can be solved. For example, you can present the workflow to show the steps or procedures to solve the problem. A diagram showing the different functional modules involved in the system and their interaction and relationships can also be presented;
- **Dataset(s).** A description of the dataset(s) that you have used in your project for validating the performance of your data mining method;
- **Snapshots**. The complete set of snapshots of the system interfaces. You can capture the screen using the screen capture software;
- **Use instructions**. A document contains instructions as to how to compile and execute your program. If you are conducting your own project in Option 2, you should provide instruction on how to use your data mining methods/systems to solve your problem. Please be as specific as possible in the instructions;
- **Findings, lessons and experiences**. Please discuss any useful, interesting knowledge and patterns discovered from your project as well as the lessons you learn from this project and any experiences you would like to share from this work;
- **Conclusion**. Conclude the whole project and possibly identify the limitation of the current system and the possible future work.

c) Video demonstration

One of the most important assessment items for the major project that you need to submit is a **short (3-5 min.)** video clip that shows how your system is executed and working. This video is important as it provides a good opportunity for training your skills in presenting your good work to others and significantly facilitate the evaluation of your project by the marker. Your video should cover the whole operation of your system, from the start of the system until finishing going through all the major functions/features that are developed**.** Please refer to Section 3 for the details regarding the video recording.

Please kindly note that you will lose **a significant portion** of your mark (up to 20 marks) if you fail to provide this video demonstration for assessment.

d) Source program files

You need to provide the source program files for your implementation in your major project. You do not need to do so if you choose Option 2 and use some existing data mining systems to complete the project. **All the source files should be zipped to a compressed file for submission.**

e) Datasets used

Regardless of your project choice, you need to also submit the datasets used in your project as a zipped file called "Dataset.zip". If the dataset file is too big, you can alternatively provide a URL address in a text file called "Dataset link.txt" for us to download the datasets.

f) Further remarks about the deliverables for Option 1 projects

Some further remarks are provided here for option 1 projects to further clarify the general expectation for the background, problem formulation, system architecture and literature review.

- **Background** - you need to discuss the problem that the data mining technique you choose solved. For example, if you choose to implement k-Means, then you need to give some background about the clustering problem as well as k-means itself. You can also briefly discuss the background of the dataset that you're going to apply your implemented data mining technique to.
- **Problem formulation** - you need to define more formally the data mining problem that you're going to deal with when implementing the technique. For example, if you want to implement K-means, then give a formulation of the clustering problem. Again, you can go one step further to define the real problem that can be solved by the implemented technique based on the dataset.
- **System architecture** - the final deliverable is a data mining system which may contain different functional modules such as input, data preprocessing, data mining, user interface, output, etc. So the system architecture basically is a diagram to show what are the functional modules within your data mining system as well as how they are working with each other together, as a whole system, to produce the data mining results.
- **Literature review** - again if you choose k-means, then you may survey papers that proposed clustering methods which are based on k-means, i.e., its variants, as well as related application of k-means.

2.3 Instruction for the video recording

A video recording is required for **all students** to present the demonstration of the major project.

You can use **ANY** existing on-screen activity capturing software to produce this video. One of the good candidates is Camtasia Studio or Camtasia Mac (*www.techsmith.com/Camtasia*). This software is not free though, but you can download a trial version for producing your video. You can also find and use some free screen capture software such as Free Online Screen Recorder *(https://www.apowersoft.com/free-online-screen-recorder*).

Your video doesn't need to be fancy and have any (sophisticated) editing. A simple recording of running your system is sufficient. It is required to **record voice** in the video at the same time for any explanations you may want to provide. Please export the video using the **low resolution** in order to keep its size manageable.

Please submit the video in some commonly used formats such as **MP4 or AVI**. You can submit the video together with other documentations and source code of your project to Studydesk or you can upload your video to some other websites such as YouTube and provide a link in your submission.

Part III Project Submission Instruction

The deliverable for both the small and major projects needs to be made electronically through the submission link created on the Studydesk course page. The due dates of the assessment items are given on the course page.

4.1 Major Project proposal

Project proposal is submitted as a single PDF format in **Assignment 1** (together with your answers to other questions in Assignment 1).

4.2 Final deliverables of the Small and Major Projects

The final project deliverables for both the small and major projects are submitted as the following two files:

(1) **A single PDF file** containing the answers to the questions of the small project; and
(2) **A single zipped file** containing the following mandatory items for the major project:

- The report of the major project (in a single PDF file);
- A video demonstration for the major project (in a mainstream video format);
- A folder containing all the source files for the major project;
- A compressed file or a link for the datasets used for the major project.

Please strictly stick to the following naming convention for your small and major project submissions: **"Small_project + Name + student ID"** and **"Major_project + Name + student ID"**, respectively.

Part IV Important Reminders

- It is important that you attach the approval of extension from the examiner in your submission, if you have requested, to avoid the late submission penalty (**5% per day**) from the marker.

- To prevent plagiarism effectively in the project work, you must show **your full name and student ID** on the screen interface of your system developed in the major project, regardless of which project options you would choose. This can prove that the system that you demonstrate is indeed your own work. The only exception to this requirement is for the students who undertake a project in Option 2 which uses some existing data mining systems/software to solve their own problems.

- Please note that, for the submission which may involve large files, the attachment size of the assignment is set to be **100M** (which is the maximum size that can be possibly set for assignments in the system), so please compress the file in order to meet the size limit. You can also consider submitting multiple smaller files rather than a large one or uploading the large files to your Dropbox or other cloud storage space and submit the link through the assignment submission page.

- If you choose to use **Google Drive** to store your project files for your final submission, please make sure that you provide the necessary access to my Gmail account, i.e., zhangji77@gmail.com, rather than my USQ email account.

- You will carry the liability for using your own real datasets for the course project. It is your responsibility to seek necessary approval from relevant authority for using the datasets if they are classified or confidential and perform necessary anonymization operations on the datasets to remove the sensitive/confidential information, if any, from the datasets before submission.