2025 CSE304 Assignment Manual

March, 2025

1 Introduction

- This document provides guidelines for the assignment, which constitutes 50% of the CSE304 course. The goal of this assignment is to work on a self-selected topic related to **clustering**. Students are expected to develop innovative approaches to problem-solving and summarise their findings in the form of a research paper.
- It is recommended to utilize Large Language Models (LLMs) throughout various steps of the assignment, including topic selection, implementation, algorithm design, and experimentation. However, rather than relying on LLMs to draft the entire content, students should take the initiative and use them as a supplementary tool to aid their work.
- This assignment is conducted individually. However, students will be grouped to facilitate discussions. There will be no group evaluations, but group discussions are encouraged, and scheduled meetings with the instructor and group members will be arranged.
- Attending some classes are mandatory for align sub-topics. Please check the blackboard regularly.

2 Schedule

The key deadlines and events for this assignment are as follows:

- **Research Paper Submission Deadline**: May 25, 2025 (by 23:59)
- **Poster File Submission Deadline**: May 28, 2025 (by 23:59)
- **Peer Review Evaluation & Presentation**: May 30, 2025 (13:00 - 15:00, Room I112-110)
  - **Session 1**: 13:00 - 14:00 (21 students, expected)
  - **Session 2**: 14:00 - 15:00 (21 students, expected)
- The session assignments for students will be announced later on **Blackboard**.

3 Report Writing (40%)

3.1 Topic

The core focus of this assignment is to explore the **clustering problem**. Clustering is a fundamental technique in **data mining and machine learning**, widely applied across various domains. In this assignment, students are required to study existing clustering techniques and propose novel problems, methodologies, or applications. By referencing the latest research, students should suggest **original solutions** to clustering challenges and validate their approaches through experiments.

The report should clearly describe the **research background, methodology, experimental design, result analysis, and conclusions**. Additionally, the **poster presentation** will serve as an opportunity to share research findings and receive feedback from peers.

3.2 Topic Selection

Students should not feel overwhelmed when selecting a topic. Various clustering-related topics can be explored by reviewing existing research, brainstorming ideas, and discussing potential directions with the **instructor and peers**. A few sample topics have been added to the **Appendix** for reference, so please review them.

3.3 Report Writing Guidelines

Please refer to the following guidelines when preparing your report:

- **Format**: Use the provided template. The report should include the following six sections: `Introduction, Related Work, Problem Statement, Algorithm, Experiments, Conclusion`. (The font size must not be modified.)
- **Language**: English
- **Author Information**: Include your student ID, name, and affiliation at the top of the first page.
- **Experimental Procedures and Results**: Clearly describe the experiments conducted and the obtained results.
- **Page Limit**: The report must not exceed **4 pages** (excluding references).
- **Formatting Restrictions**: Do not alter the font size or page margins.
- **Originality**: Clearly articulate how your work differs from existing research.
- **Code and Data Sharing**: Upload the code and data used in your experiments to GitHub or another publicly accessible repository.

3.4 Evaluation Criteria

The assignment will be evaluated based on the following criteria:

- **Novelty**: How original is the contribution compared to existing work? Does it go beyond simple applications and propose new ideas or techniques?
- **Technical Soundness**: Is the proposed method technically valid and well-grounded?
- **Experimental Validation**: Are the claims supported by appropriate experiments?
- **Impact**: Does the research address meaningful problems in the field of clustering?
- **Presentation & Writing**: Is the report clearly and logically written?
- **Related Work**: Are the contributions well-positioned relative to prior work?

4 Presentation Evaluation (10%)

The presentation is an important opportunity to effectively communicate your research and receive feedback from peers and the instructor. All students must adhere to the presentation format and evaluation procedures below.

4.1 Presentation Format

- Each student must prepare a presentation consisting of **two A3-sized slides**, both in **portrait orientation (vertical layout)**.
- These two slides will be printed separately on **two A3 sheets**, and presented side-by-side during the poster session. Students should design their slides accordingly to ensure visual coherence across both sheets.
- A sample layout is shown in Figure 1. Each slide should stand alone but together convey a cohesive research story.
- The instructor will print all posters in **grey-scale**. Make sure all visuals (e.g., charts, diagrams) are clearly distinguishable in black and white.
- The presentation must be conducted **in English**.
- Students who are unable to attend on the designated day must arrange to give their presentation during a regular class session.
- Refer to Figure 1.

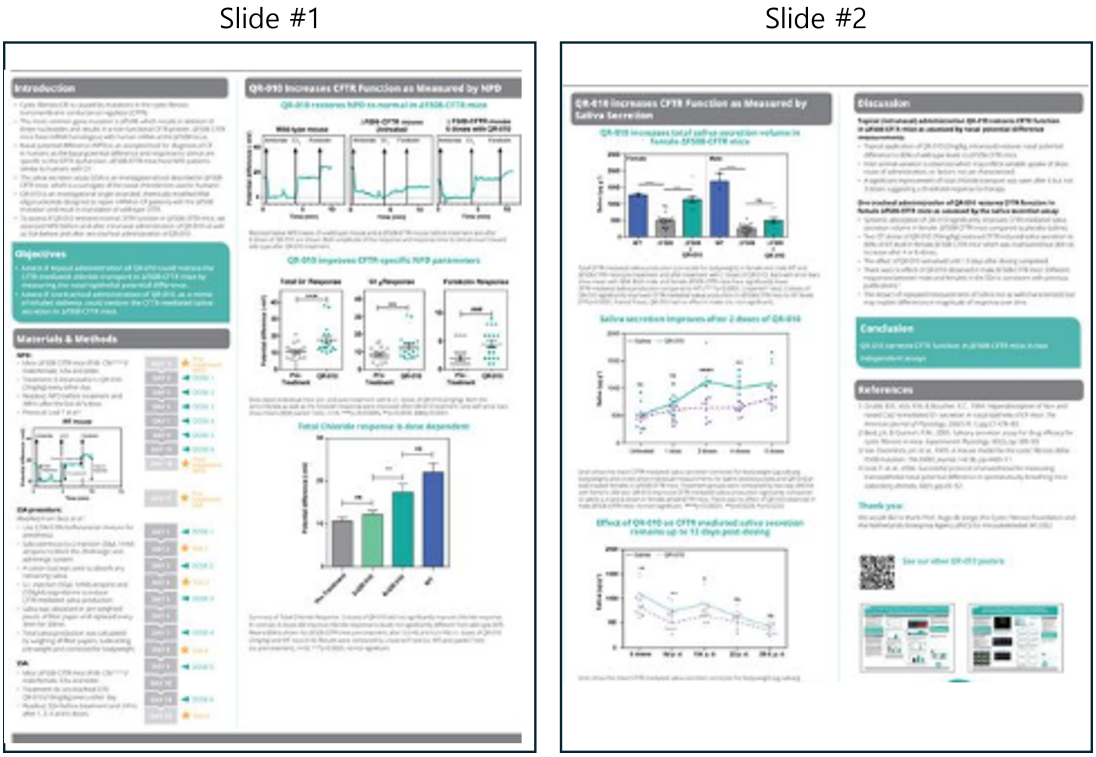

Figure 1: Example Poster Layout

4.2 Evaluation Procedure

Each student will receive a presentation score based on **peer evaluation and TA evaluation**. The procedure is as follows:

- **Peer Evaluation**: Each presenter will be evaluated by **5 students** (subject to change). Each student evaluates 5 peers within the session.
- **Scoring Method**:
  - The highest and lowest peer evaluation scores will be discarded. The **average of the remaining 3 scores** will be used as the final peer evaluation score.
  - The TA evaluation score will account for **50% of the total score**.
  - **Final Score** = (Average of 3 peer evaluations: 50%) + (TA evaluation score: 50%)

4.3 Evaluation Criteria

Presentations will be evaluated based on the following criteria:

- **Clarity**: Is the research clearly and logically communicated?
- **Contribution**: Does the research make a meaningful contribution to the field of clustering?
- **Poster Quality**: Is the poster well-structured and effective in conveying the research?

- **Q&A**: Are the responses to questions logical and accurate?
- **Presentation Skills**: Was the presentation delivered smoothly and professionally?

A Sample Topics

Below are several sub-topics and specific examples that can inspire your project. These are for reference only and are not intended to be restrictive.

A.1 Improvements to Classical Algorithms

- **Enhancing DBSCAN Efficiency**: Can we optimize DBSCAN to avoid re-scanning the entire dataset when parameters (μ, ε) are modified?
- **Automatic Determination of Cluster Count**: Is it possible to develop a method that estimates the optimal number of clusters without manual tuning?
- **Improved Initialization for K-means**: Can we use nearest-neighbor-based methods to improve the initialization of centroids in K-means or K-means++?
- **Clustering with Prior Knowledge**: How can prior knowledge—such as the number or approximate size of clusters—be incorporated to enhance clustering performance?
- **Constraint-based Clustering**: Can we design algorithms to identify clusters that contain a specific subset of data points $X' \subseteq X$?
- **Extending K-means Beyond Spherical Clusters**: Can we adapt K-means to effectively handle non-spherical data distributions?

A.2 Graph Clustering

- **Scalable Graph Clustering in Large Networks**: Investigate scalable methods for detecting overlapping, non-overlapping, or hierarchical communities.
- **GNN-based Clustering**: Explore the potential of graph neural networks (GNNs) and graph embeddings for identifying meaningful clusters.
- **Streaming Graph Clustering**: Design efficient clustering algorithms for dynamic graph environments that evolve over time.
- **Generalizable Clustering Across Graph Types**: Can we develop clustering approaches that apply to bipartite, dynamic, and attributed graphs?
- **Structure-Constrained Graph Clustering**: How can we detect clusters with specific structural constraints?

A.3 AI-Assisted and Generative Approaches

- **LLM-Assisted Clustering**: Use large language models (LLMs) to support clustering tasks such as feature construction or representation learning for specific types of data.
- **Generative Models for Embedding and Clustering**: Leverage models like VAEs or diffusion models to generate embeddings that improve clustering performance or interpretability.
- **Reinforcement Learning for Clustering Optimization**: Employ reinforcement learning to dynamically tune clustering strategies or parameters.
- **Meta-Learning for Algorithm Selection**: Design systems that learn to select the most suitable clustering algorithm based on the characteristics of the dataset.

A.4 Clustering on High-Dimensional or Complex Data

- **Dimensionality Reduction + Clustering**: Evaluate how PCA, t-SNE, or UMAP affect clustering performance in different domains.
- **Text Embedding + Clustering**: Perform clustering over dense representations of text, such as BERT embeddings.
- **Multi-modal Data Clustering**: Develop clustering methods that operate on data composed of multiple modalities such as text, image, and graph.
- **Compression-Aware Clustering**: Investigate whether data compression techniques can improve computational efficiency without sacrificing clustering quality.

A.5 Real-Time and Scalable Processing

- **Clustering for Streaming Data**: Design online algorithms that continuously update clusters as new data arrives.
- **Distributed Clustering for Large-Scale Data**: Use distributed frameworks such as Spark or MapReduce to scale clustering to massive datasets.
- **Approximate Clustering with Sampling**: Explore sampling-based techniques that strike a balance between speed and accuracy.

A.6 Visualization, Compression, and Preprocessing

- **Visual Clustering**: Leverage visual analytics techniques to enhance understanding and interpretation of clustering results.
- **Compression-Based Clustering**: Explore clustering methods that exploit data compression for performance gains.

- **Robustness to Preprocessing**: Analyze the impact of normalization, outlier removal, and missing data on clustering performance.

A.7 Application-Oriented Clustering

- **Customer Segmentation for Marketing**: Cluster users based on behavioral and demographic features to support targeted marketing.
- **Clustering of Academic Papers or Patents**: Use citation networks, co-authorship graphs, or textual embeddings to group similar documents.
- **IoT and Sensor Data Clustering**: Detect patterns and anomalies in environmental or industrial sensor data from IoT systems.
- **Code Clustering**: Construct code graphs and apply clustering to uncover meaningful development or usage patterns.
- **Task-Oriented Clustering**: Use clustering results to enhance performance in downstream tasks such as classification, recommendation, or anomaly detection.