

Wright State University
Computer Science and Engineering

CS4710-6710— Assignment 2
Introduction to Data Mining

Instructor: Dr. Tomojit Ghosh

Due by 2025/11/14

This assignment is mandatory for both sections: CS4710 and 6710. Good luck and happy coding work!

Distribution of Marks

Question	Points	Score
1	100	
Total:	100	

1. (100 points) **Objective:** In this question you will implement k-Mean algorithm from the scratch. **You are not allowed to use scikit-learn package or any online code, or other online package.** The k -Mean steps are already provided in lecture material. You will use the following two strategies to initialize the initial centers.

Strategy 1: For each feature find out the minimum and maximum values. Then randomly select k values (using uniform distribution) from the min-max range to initialize the centers.

Strategy 2: For each feature arrange the values in ascending order. Partition the values in five quartiles. Discard the first and last quartile and find out the minimum and maximum values from rest of the quartiles. Now randomly select k values (using uniform distribution) from the min-max range to initialize the centers.

To pick the value of k , you can use the elbow method which was discussed in the class. Use the Wine Quality data set (red and white wine) for your experiment. After clustering, you can project the data in 2/3 dimensional space using PCA and use appropriate legends for each class. Apart from visualization, you must use purity measure to quantify the clustering performance.

Submission Instructions:

- Your submission must contain a pdf file, python scripts. The report (pdf file) must contain a discussion on the results.
- Submit all files via the course portal by **November 14, 2025**.
- Late submissions will incur penalties as per course policy.

Note:

- You are allowed to use Pandas to load data set.
- Do not use any machine learning libraries such as **scikit-learn**.
- You may use **numpy** and **matplotlib** for numerical computation and visualization.
- You are welcome to discuss ideas with your peers, but your submitted code and analysis must be entirely your own work.