CS 470 Homework 3

Due by Thursday, March 16, 2023 at 6:00 PM

Submission instructions

Submit your assignment through the QTest system, using exam ID: **CS170.hw3**. Upload a single ZIP archive file named **hw3.zip**, containing all the files of your solution:

1. The program's source files.
2. A `README.txt` file explaining how to compile and run your program.
3. The input datasets.
4. The report in PDF format.
5. The LaTeX source files used to typeset the report.

No email submissions are accepted. No late submissions are accepted.

At the top of your solution, include a section named "Collaboration statement" in which you acknowledge any collaboration, help, or resource you used or consulted to complete this assignment.

1 K-Means Clustering (100 points)

Your task for this assignment is to implement and evaluate the k-means clustering algorithm.

1. Implement the k-means clustering algorithm either in Java or Python.
   - The program should be executable with at least 3 parameters: the name of the dataset file, k, and the name of the output file.
   - The output file should contain numerical class labels (formatted as one number per row) for all the records in the test dataset and report the sum squared error (SSE) and silhouette coefficient in the last row.
   - You only need to handle numerical attributes (categorical attributes are not required).
2. Evaluate the algorithm using SSE and silhouette coefficient with varying k on at least two datasets:

   - The Fisher Iris dataset (without class labels):
     http://archive.ics.uci.edu/ml/datasets/Iris
   - Another dataset of your choice from the UCI repository.

3. Write a report in LaTeX to present your results.
   - Describe the datasets.
   - Describe your implementation and experiment setup, e.g., any pre-processing you performed on the dataset such as normalizing the attributes, distance metrics you used, etc.
   - Present the experiment results with varying k.
   - Discuss the insights and conclusions from your experiments, and the experiences and lessons you have learned from this assignment.

Grading criteria

- 70 points for correct and complete implementation. -10 for minor mistakes; -20 if there is some mistake but the program still works in most cases; -30 for more serious mistakes but the program still works in several cases. Zero points if the program does not compile, or if the program compiles but gives mostly the wrong results or crashes.
- 30 points for a complete, clear, and well organized report. Zero points if the report is not typeset using LaTeX, or if the LaTeX source code is not provided.
- -10 points for insufficient comments in the code.
- -10 points for each deviation from the submission instructions.
- -10 points for missing collaboration statement.