Phase Two Report

Seth Garner
Dr. Emre Celebi
CSCI 4372 - Data Clustering
September 27, 2025

Introduction

This report summarizes my implementation of the **standard k-means** for Phase 2. The goal is to read a dataset, initialize the `K` centers uniformly at random from the data points, then iterate assignment and update steps until convergence, printing the SSE for each iteration. The program uses 64-bit floating points for all attributes and avoids using the `sqrt()` and `pow()` as required from the requirements of the assignment.

Data Structures

I store the dataset as a `List<double[]>`, where each array is a dimensional point. The first line of the file gives the `N` (points) and `D` (dimensions). Using a list of primitive arrays keeps the memory compact and iterations fast.

Algorithm

Initialization: I select distinct indices uniformly at random (via shuffled index list) and copy those rows as the initial centers (no removal from the dataset).

Assignment: For each point `x`, compute squared Euclidean distance to every center `c_k`:

$$d(x, c_k) = \sum_{j=1}^{D} (x_j - c_{k,j})^2$$

Assign to the nearest center. Ties break to the smallest center index.

Update: For each cluster, set the center to the mean of it's assigned points. If a cluster is empty, I keep the previous center.

Convergence: After each iteration, compute SSE:

$$\text{SSE}(t) = \sum_{i=1}^{N} \|x_i - c_{a(i)}\|^2$$

Stop when `(SSE(t1)SSE(t))/SSE(t1) < T` or when the iteration count reaches limit. It runs the algorithm the amount of runs with different randon initialization and report the per-iteration SSE and the best run's final SSE.

Complexity

Each iteration costs $O(NKD)$. Across the number of runs and iterations: $O(R \cdot I \cdot N \cdot K \cdot D)$. Memory is $O(ND + KD)$.

Results

- Printed SSE decreased per iteration iteration (never increased).
- **Iris Bezdek (K=3):** over `R=100` runs, I observed no run with SSE < 78.8514, consistent with the checks in the assignment.
- Timing was done using the `time` function in Linux command line, typical runs on medium datasets completed on average at 14.4 seconds. Random seeds were set as `baseSeed + runIndex` for reproducibility.

Notes/Limitations

This submission uses uniform random initialization only. Empty clusters retain their previous centers.