# **Homework #1** - Machine Learning

Srikanth Pai, Asst. Professor, MSE

January 12, 2026

## Instructions

In this assignment you will analyze, implement, and apply the K-means algorithm. You may use standard numerical, data-handling, and visualization libraries (e.g. NumPy, Pandas, OpenCV, Matplotlib). You must not use any built-in implementation of K-means or related clustering routines. All clustering must be performed using our own implementation of Lloyd's algorithm.

Unless stated otherwise, use squared Euclidean distance. Answers should be concise and mathematically precise.

## 1 Analysis

1. Show that the K-means objective

$$J(r, \mu) = \sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik} \|x_i - \mu_k\|^2$$

is not jointly convex in $(r, \mu)$. Your argument should exhibit two feasible points and show failure of convexity directly. [Hint: Construct the counterexample on a line with very few points.]

2. Let $(r^*, \mu^*)$ be a fixed point of Lloyd's algorithm. Show that $(r^*, \mu^*)$ is a coordinate-wise minimizer of $J$, but need not be a global minimizer. Give a concrete example illustrating this distinction.

3. Let $(r^*, \mu^*)$ be a fixed point of Lloyd's algorithm. Assume no distance ties.

   Show that there exists $\varepsilon > 0$ such that if the data points $\{x_i\}$ are perturbed by at most $\varepsilon$ in norm, then the assignment $r^*$ remains unchanged. Explain why this implies local stability but not global optimality.

4. In our implementation of Lloyd's algorithm, empty clusters are handled by reseeding the corresponding centroid to a randomly chosen data point. Our proofs assumed there are no empty clusters. So there is a gap and a natural question arises. Is the algorithm guaranteed to converge when we reseed? Either exhibit a concrete dataset and initialization for which the algorithm does not terminate, or rigorously prove why termination still occurs with reseeding.

5. Consider the dataset in $\mathbb{R}^2$

$$\mathcal{D} = \{(\pm 1, 0), (0, \pm 1), (\pm 3, 0), (0, \pm 3)\}.$$

   Implement Lloyd's algorithm for K-means with $K = 2$.

(a) Run the algorithm using the initialization

$$\mu_1^{(0)} = (1, 0), \qquad \mu_2^{(0)} = (-1, 0),$$

and report the final centroids and the final value of the K-means objective.

(b) Repeat the experiment with the initialization

$$\mu_1^{(0)} = (0, 1), \qquad \mu_2^{(0)} = (0, -1).$$

(c) Compare the two outcomes and explain why both results are consistent with the convergence guarantees of Lloyd's algorithm.

# 2 Project

In this project you will apply your own implementation of Lloyd's algorithm to a real dataset. Your work should clearly describe the feature representation, the choice of $K$, the effect of initialization, and at least one limitation of K-means revealed by the application. The emphasis is on understanding the behavior of the algorithm rather than achieving optimal performance.

### Project A: Image Segmentation and Color Quantization

Treat each pixel of an image as a data point. Typical feature representations include $(R, G, B)$ or $(R, G, B, x, y)$. After clustering, reconstruct the image by replacing each pixel with the centroid of its assigned cluster.

You should experiment with different values of $K$ and different feature choices, and analyze how these choices affect the resulting segmentation.

Datasets:
`https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/`
`https://www.kaggle.com/datasets`

### Project B: Customer Segmentation

Cluster customers using demographic or behavioral data. Each customer is represented as a vector in $\mathbb{R}^d$. Feature scaling should be justified.

Suggested datasets:
`https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial`
`https://archive.ics.uci.edu/ml/datasets/online+retail`

# Deliverables

You must submit your source code together with figures or tables summarizing your experimental results. In addition, submit a short report (1–2 pages) describing your setup, observations, and conclusions.