Lecture 5: Introduction to Clustering & k-Means

ECE 2410 – Introduction to Machine Learning

Farzad Farnoud

University of Virginia

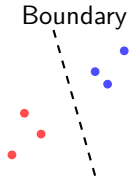
Spring 2026

Outline

Supervised vs. Unsupervised Learning

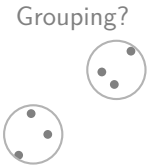
Supervised Learning (L1-L4)

- **Data:** Pairs $(\boldsymbol{x}, y)$
- **Goal:** Predict y given $\boldsymbol{x}$
- **Feedback:** Correct labels provided



Unsupervised Learning (L5-L7)

- **Data:** Just $\boldsymbol{x}$! No labels.
- **Goal:** Find *structure* or patterns
- **Feedback:** None (exploratory)



Why Clustering?

Definition

Clustering: Grouping data points so that items in the same group are more similar to each other than to those in other groups.

Real-World Applications:

- 🛒 **Customer Segmentation:** Group customers by purchasing behavior (e.g., "Big Spenders", "Bargain Hunters")
- 🎨 **Image Compression:** Reduce millions of colors to a small palette
- 🧬 **Genetics:** Cluster genes with similar expression patterns
- 📰 **Document Analysis:** Organize articles by topic without reading them

Today's Algorithm: k-Means

We'll focus on one of the simplest and most widely used clustering algorithms: **k-Means**.

Why k-Means?

- Simple to understand and implement
- Fast (scales to large datasets)
- Often works surprisingly well!

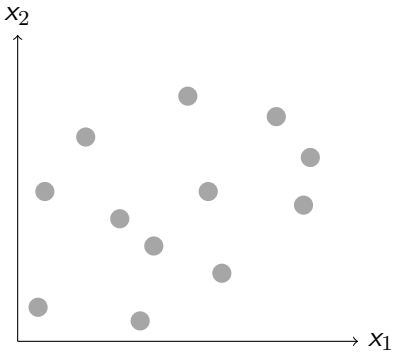
The Clustering Problem

Input:

- A dataset $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ where each $\boldsymbol{x}_i \in \mathbb{R}^D$
- A number of clusters K (chosen by us)

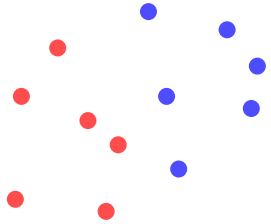
Output:

- K cluster **centers** (centroids): $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$
- Cluster **assignments** for each point: $c_1, \ldots, c_N$ where $c_i \in \{1, \ldots, K\}$

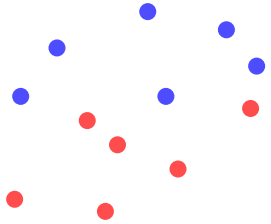


$N = 10$ data points

How Do We Measure a "Good" Clustering?



Clustering A



Clustering B

Key Question

We need a **number** to quantify how good a clustering is.
But how?

Think of it Like This...

The Post Office Problem:

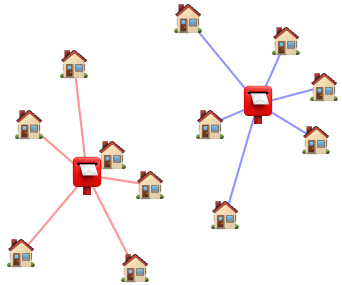
- You have N houses (data points).
- You want to build K post offices (centers).
- Each house is served by its nearest post office.
- The cost of traveling distance d is d^2.
- **Goal:** Minimize total cost!

The Objective Function

We want to minimize the **total squared distance** from each point to its assigned center:

$$J = \sum_{i=1}^{N} ||\boldsymbol{x}_i - \boldsymbol{\mu}_{c_i}||^2$$

- $\boldsymbol{x}_i$: Data point i
- c_i: Cluster assignment of point i
- $\boldsymbol{\mu}_{c_i}$: Center of the cluster assigned to point i



Key Insight

This is an **optimization problem**: find assignments $\{c_i\}$ and centers $\{\boldsymbol{\mu}_k\}$ that minimize J.

The k-Means Algorithm

An iterative algorithm that alternates between two steps:

1. **Initialization**: Pick K random points as initial centers $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$.
2. **Repeat until convergence**:
   - **Step A (Assignment):** Assign each point $\boldsymbol{x}_i$ to the *nearest* center. **(If equal, keep last assignment)**
   $$c_i = \arg\min_k ||\boldsymbol{x}_i - \boldsymbol{\mu}_k||^2$$
   - **Step B (Update):** Move each center $\boldsymbol{\mu}_k$ to the *mean* of its assigned points.
   $$\boldsymbol{\mu}_k = \frac{1}{|C_k|} \sum_{i \in C_k} \boldsymbol{x}_i$$

Convergence: Stop when assignments don't change.

Why Does This Work?

Step A (Assignment):

- Given fixed centers $\boldsymbol{\mu}$, the best assignment for each point is obviously its *nearest* center.
- If any assignment changes, J decreases.

Step B (Update):

- Given fixed assignments, the best center for a cluster is its *mean* (minimizes squared error).
- If any center changes, J decreases.

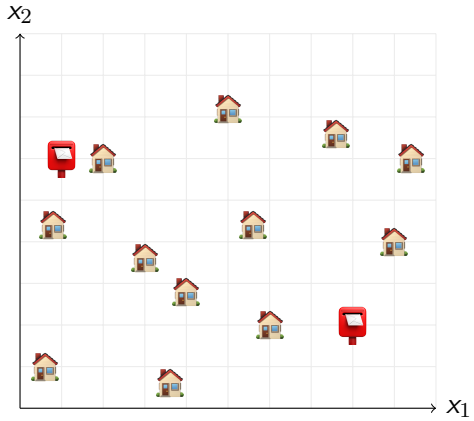
Stopping Criterion

If a step doesn't change assignments or centers, we stop.

Guarantee

J decreases at every step $\Rightarrow$ **Algorithm always converges!**

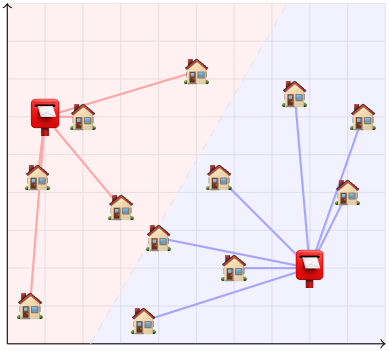
Step 0: The Data + Random Initialization



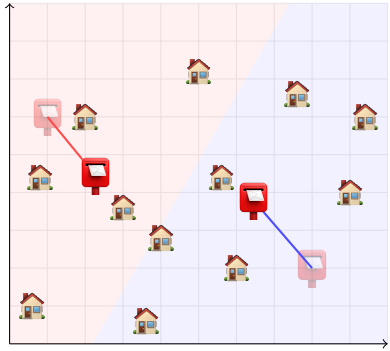
$N = 12$ houses, $K = 2$ post offices (random placement)

Iteration 1: Assign & Update

Assignment

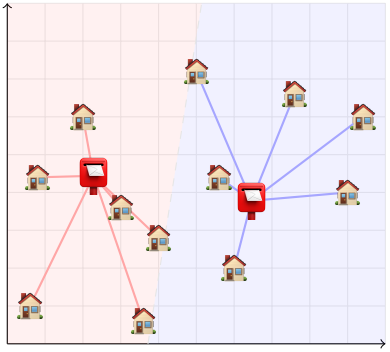


Update Centers

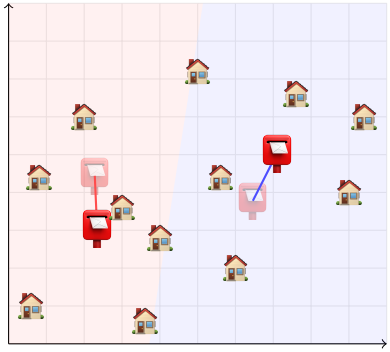


Iteration 2: Assign & Update

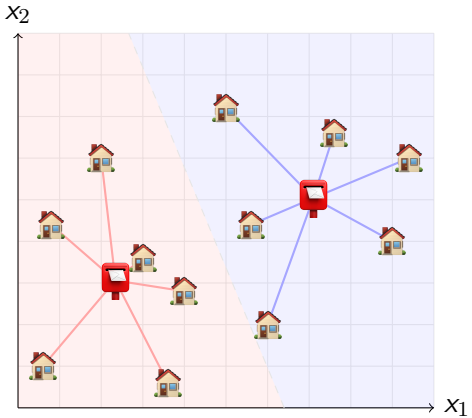
Re-Assignment



Update Centers



Iteration 3: Converged!



Converged! Assignments unchanged from Iteration 2.

Summary

Key Concepts

1. **Unsupervised Learning:** Finding structure without labels.
2. **k-Means Algorithm:**
   - **Initialize** K centers randomly.
   - **Assign** each point to its nearest center.
   - **Update** centers to the mean of their points.
   - **Repeat** until convergence.
3. **Simple but Powerful:** Used everywhere from image compression to genetics.

Questions for Next Time

- Does it always find the *best* clustering? 🤔
- How do we choose K?
- What are its limitations?