MATH 535: Mathematical Methods in Data Science

Lecture Notes: K-Means Clustering

January 24, 27, 29, 31, 2025

K-Means Clustering

Motivation

SLIDESHOW

Problem Setup

Input: n vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ and number of clusters k
Output(?): partition of the points into k clusters

Definition 1 (Partition)**.** A partition of $[n] = \{1, \ldots, n\}$ of size k is a collection of non-empty subsets $C_1, \ldots, C_k \subseteq [n]$ that:

- are pairwise disjoint: $C_i \cap C_j = \emptyset$ for all $i \neq j$
- cover all of $[n]$: $\cup_{i=1}^k C_i = [n]$

Knowledge Check 1: Which of these is a valid partition of $\{1, 2, 3, 4\}$ into $k = 2$ clusters?

A) $\{1, 2\}, \{3, 4\}$

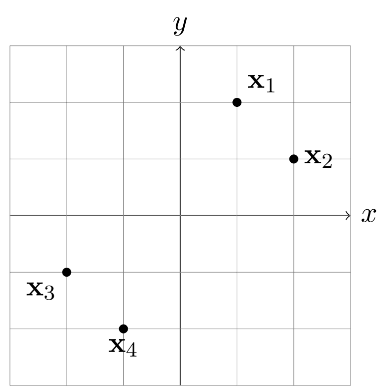
B) $\{1, 2\}, \{2, 3, 4\}$

C) $\{1, 2\}, \{1, 3, 4\}$

D) $\{1, 2, 3\}, \{4, 5\}$

Example: Consider points in $\mathbb{R}^2$:

$$\mathbf{x}_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \mathbf{x}_2 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \mathbf{x}_3 = \begin{pmatrix} -2 \\ -1 \end{pmatrix}, \mathbf{x}_4 = \begin{pmatrix} -1 \\ -2 \end{pmatrix}$$



For $k = 2$, a natural partition would be:

$$C_1 = \{1, 2\}, \quad C_2 = \{3, 4\}$$

The K-Means Objective

Goal(?): partition "with small within-cluster distances"

For a partition $C_1, \ldots, C_k$, define:

$$\mathcal{G}(C_1, \ldots, C_k) = \min_{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k \in \mathbb{R}^d} \sum_{i=1}^{k} \sum_{j \in C_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

where $\boldsymbol{\mu}_i$ is the representative (center) of cluster i. Our goal is to find a partition $C_1, \ldots, C_k$ that minimizes $\mathcal{G}(C_1, \ldots, C_k)$, i.e., solves the problem

$$\min_{C_1, \ldots, C_k} \mathcal{G}(C_1, \ldots, C_k) = \min_{C_1, \ldots, C_k} \min_{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k \in \mathbb{R}^d} \sum_{i=1}^{k} \sum_{j \in C_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

over all partitions of $[n]$ of size k.

Remark 1 (Why squared distances?)**.** While using distances $\|\mathbf{x}_j - \boldsymbol{\mu}_i\|$ might seem more natural, squared distances offer a key advantage: it decomposes into a sum over coordinates (as we will see in proving Theorem 1). Each term in the sum depends on a single component of a single representative. This property make the optimization problem more tractable.

Knowledge Check 2: What happens to the K-means objective value if we increase k from 2 to 3?

A) Always increases

B) Always decreases

C) Could increase or decrease

D) Stays the same

Theorem 1 (Optimal Representatives)**.** Fix a partition $C_1, \ldots, C_k$. The optimal representatives are the centroids:

$$\boldsymbol{\mu}_i^* = \frac{1}{|C_i|} \sum_{j \in C_i} \mathbf{x}_j$$

Proof. Let's prove this in general while working through our example where $k = 2$, $d = 2$, $C_1 = \{1, 2\}$, $C_2 = \{3, 4\}$.

General case:

1. Write objective as sum over coordinates m and clusters i:

$$\sum_{i=1}^{k}\sum_{j\in C_i}\|\mathbf{x}_j-\boldsymbol{\mu}_i\|^2=\sum_{i=1}^{k}\sum_{m=1}^{d}\left[\sum_{j\in C_i}(x_{jm}-\mu_{im})^2\right]$$

2. For fixed i and m, get quadratic function:

$$q_{im}(\mu_{im})=\sum_{j\in C_i}x_{jm}^2-2\mu_{im}\sum_{j\in C_i}x_{jm}+|C_i|\mu_{im}^2$$

3. Take derivative, set to zero:

$$\frac{d}{d\mu_{im}}q_{im}=-2\sum_{j\in C_i}x_{jm}+2|C_i|\mu_{im}=0$$

4. Solve:

$$\mu_{im}^*=\frac{1}{|C_i|}\sum_{j\in C_i}x_{jm}$$

Example: 1. Objective:

$$\|\mathbf{x}_1-\boldsymbol{\mu}_1\|^2+\|\mathbf{x}_2-\boldsymbol{\mu}_1\|^2$$
$$+\|\mathbf{x}_3-\boldsymbol{\mu}_2\|^2+\|\mathbf{x}_4-\boldsymbol{\mu}_2\|^2$$

2. For first cluster and first coordinate (i.e., $i=1$, $m=1$):

$$(1-\mu_{11})^2+(2-\mu_{11})^2$$

Expand:

$$(1^2+2^2)-2\mu_{11}(1+2)+2\mu_{11}^2$$
$$=5-6\mu_{11}+2\mu_{11}^2$$

3. Take derivative:

$$-6+4\mu_{11}=0$$

4. Solve:

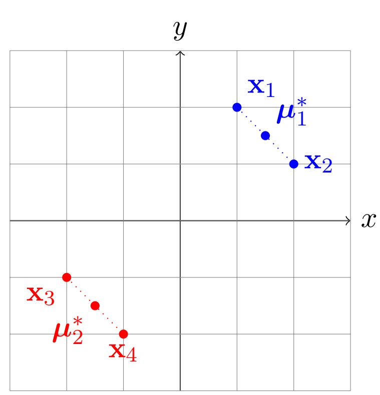
$$\mu_{11}^*=\frac{6}{4}=1.5$$

This proves the centroids minimize the objective. □

Example (continued): For partition $C_1=\{1,2\},C_2=\{3,4\}$:

Optimal representatives:

$$\boldsymbol{\mu}_1^*=\frac{1}{2}\begin{pmatrix}1\\2\end{pmatrix}+\frac{1}{2}\begin{pmatrix}2\\1\end{pmatrix}=\begin{pmatrix}1.5\\1.5\end{pmatrix}$$
$$\boldsymbol{\mu}_2^*=\frac{1}{2}\begin{pmatrix}-2\\-1\end{pmatrix}+\frac{1}{2}\begin{pmatrix}-1\\-2\end{pmatrix}=\begin{pmatrix}-1.5\\-1.5\end{pmatrix}$$



Knowledge Check 3: If a cluster contains points $(0,0)$, $(2,0)$, and $(4,0)$, what is its centroid?

A) $(0,0)$

B) $(2,0)$

C) $(4,0)$

D) $(3,0)$

Theorem 2 (Optimal Clustering)**.** Fix representatives $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$. The optimal partition assigns each point to its closest representative:

$$j \in C_i \iff \|\mathbf{x}_j - \boldsymbol{\mu}_i\| = \min_{\ell} \|\mathbf{x}_j - \boldsymbol{\mu}_\ell\|$$

Proof. Let's prove this using our example where $k = 2$, $d = 2$, and we have representatives $\boldsymbol{\mu}_1 = (1.5, 1.5)$, $\boldsymbol{\mu}_2 = (-1.5, -1.5)$.

General case:
1. Write objective as sum over points:

$$\sum_{i=1}^{k} \sum_{j \in C_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2 = \sum_{j=1}^{n} \left\|\mathbf{x}_j - \boldsymbol{\mu}_{c(j)}\right\|^2$$

where $c(j)$ is cluster assignment of point j (i.e., $c(j) = i$ is same as $j \in C_i$).
2. Since sum is independent across points, minimize each term separately:

$$\min_{c(j)} \left\|\mathbf{x}_j - \boldsymbol{\mu}_{c(j)}\right\|^2$$

3. Since square root is monotone:

$$\arg\min_i \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2 = \arg\min_i \|\mathbf{x}_j - \boldsymbol{\mu}_i\|$$

Therefore assign point j to closest representative.

Example for point $\mathbf{x}_1 = (1, 2)$:
1. Need to minimize one of:

$$\|\mathbf{x}_1 - \boldsymbol{\mu}_1\|^2 \text{ or } \|\mathbf{x}_1 - \boldsymbol{\mu}_2\|^2$$

2. Compute both distances:

$$\|\mathbf{x}_1 - \boldsymbol{\mu}_1\| = \sqrt{(1-1.5)^2 + (2-1.5)^2}$$
$$= \sqrt{0.5} \approx 0.71$$
$$\|\mathbf{x}_1 - \boldsymbol{\mu}_2\| = \sqrt{(1-(-1.5))^2 + (2-(-1.5))^2}$$
$$= \sqrt{20.5} \approx 4.53$$

3. Since $0.71 < 4.53$, assign to cluster 1

This proves point $\mathbf{x}_1$ should be assigned to cluster 1. Similar calculations for other points confirm the clustering is optimal. □

Example (continued): For point $\mathbf{x}_2$:

$$\|\mathbf{x}_2 - \boldsymbol{\mu}_1^*\| = \sqrt{(2-1.5)^2 + (1-1.5)^2} = \sqrt{0.5}$$
$$< \|\mathbf{x}_2 - \boldsymbol{\mu}_2^*\| = \sqrt{(2-(-1.5))^2 + (1-(-1.5))^2} = \sqrt{20.5}$$

For point $\mathbf{x}_3$:

$$\|\mathbf{x}_3 - \boldsymbol{\mu}_1^*\| = \sqrt{(-2-1.5)^2 + (-1-1.5)^2} = \sqrt{20.5}$$
$$> \|\mathbf{x}_3 - \boldsymbol{\mu}_2^*\| = \sqrt{(-2-(-1.5))^2 + (-1-(-1.5))^2} = \sqrt{0.5}$$

Similar calculations confirm optimality for the other point.

Implementation

SLIDESHOW

Convergence Result

Theorem 3 (Convergence of k-means)**.** The sequence of objective function values produced by the k-means algorithm is non-increasing. That is, if we denote by $\mathcal{G}^{(t)}$ the objective value at iteration t, then

$$\mathcal{G}^{(t+1)} \leq \mathcal{G}^{(t)}$$

Proof. Let's see why each iteration cannot increase the objective value.

General case:
Let $C'_1, \ldots, C'_k$ be current clusters with representatives $\boldsymbol{\mu}'_1, \ldots, \boldsymbol{\mu}'_k$.
After Step 1, new representatives $\boldsymbol{\mu}''_1, \ldots, \boldsymbol{\mu}''_k$ satisfy:

$$\sum_{i=1}^{k} \sum_{j \in C'_i} \left\|\mathbf{x}_j - \boldsymbol{\mu}''_i\right\|^2 \leq \sum_{i=1}^{k} \sum_{j \in C'_i} \left\|\mathbf{x}_j - \boldsymbol{\mu}'_i\right\|^2$$

by Theorem 1 (optimal representatives).
After Step 2, new clusters $C''_1, \ldots, C''_k$ satisfy:

$$\sum_{i=1}^{k} \sum_{j \in C''_i} \left\|\mathbf{x}_j - \boldsymbol{\mu}''_i\right\|^2 \leq \sum_{i=1}^{k} \sum_{j \in C'_i} \left\|\mathbf{x}_j - \boldsymbol{\mu}''_i\right\|^2$$

by Theorem 2 (optimal clustering).

Example:
Start with $C_1 = \{1, 2\}$, $C_2 = \{3, 4\}$ and representatives:

$$\boldsymbol{\mu}'_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \boldsymbol{\mu}'_2 = \begin{pmatrix} -2 \\ -1 \end{pmatrix}$$

Step 1: New optimal representatives:

$$\boldsymbol{\mu}''_1 = \begin{pmatrix} 1.5 \\ 1.5 \end{pmatrix}, \boldsymbol{\mu}''_2 = \begin{pmatrix} -1.5 \\ -1.5 \end{pmatrix}$$

reduce objective value from

$$0^2 + 2^2 + 2^2 + 0^2 = 8$$

to

$$0.5 + 0.5 + 0.5 + 0.5 = 2$$

Step 2: Check distances to $\boldsymbol{\mu}''_1$, $\boldsymbol{\mu}''_2$ for each point. Points stay in same clusters, no further improvement.

Combining the inequalities shows objective cannot increase. Since it's bounded below by 0, it converges. □

Matrix Representation

Stack data vectors into matrix:

$$X = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{bmatrix}$$

Stack representatives similarly:

$$U = \begin{bmatrix} \boldsymbol{\mu}_1^T \\ \boldsymbol{\mu}_2^T \\ \vdots \\ \boldsymbol{\mu}_k^T \end{bmatrix} = \begin{bmatrix} \mu_{11} & \mu_{12} & \cdots & \mu_{1d} \\ \mu_{21} & \mu_{22} & \cdots & \mu_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{k1} & \mu_{k2} & \cdots & \mu_{kd} \end{bmatrix}$$

Encode cluster assignments in matrix $Z = [Z_{j\ell}]_{j,\ell}$ where:

$$Z_{j\ell} = \begin{cases} 1 & \text{if point } j \text{ assigned to cluster } \ell \\ 0 & \text{otherwise} \end{cases}$$

Representative of cluster assigned to point j:

$$\boldsymbol{\mu}_{c(j)}^T = \sum_{\ell=1}^{k} Z_{j\ell} \boldsymbol{\mu}_\ell^T = (ZU)_{j,\cdot}$$

K-means objective in matrix form:

$$G(C_1, \ldots, C_k; \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k) = \|X - ZU\|_F^2$$

where $\|\cdot\|_F$ is the Frobenius norm:

$$\|A\|_F = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{m} A_{ij}^2}$$

Key Insight: K-means finds a low-rank matrix factorization ZU approximating data matrix X.

Example (continued): For our simple example with partition $C_1 = \{1, 2\}, C_2 = \{3, 4\}$:
Assignment matrix:

$$Z = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

With optimal representatives:

$$U = \begin{bmatrix} 1.5 & 1.5 \\ -1.5 & -1.5 \end{bmatrix}$$

Product ZU gives representative for each point:

$$ZU = \begin{bmatrix} 1.5 & 1.5 \\ 1.5 & 1.5 \\ -1.5 & -1.5 \\ -1.5 & -1.5 \end{bmatrix}$$

Back to the dataset

SLIDESHOW