

Detailed Mathematical Presentation of the Simplest Unsupervised Learning Method: K-Means Clustering

1. Introduction

K-means clustering is one of the simplest and most widely used unsupervised learning algorithms. It partitions a dataset into k distinct, non-overlapping clusters by minimizing the within-cluster variance. Unlike supervised learning, k-means requires no labeled data; it infers structure purely from the input features.

2. Problem Setup and Notation

- Let the dataset be:

$$X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}, \mathbf{x}_i \in \mathbb{R}^d$$

where n is the number of data points and d is the dimensionality of each point.

- The goal is to partition X into k clusters:

$$\mathcal{C} = \{C_1, C_2, \dots, C_k\}$$

such that each data point belongs to exactly one cluster:

$$C_i \cap C_j = \emptyset \text{ for } i \neq j, \bigcup_{j=1}^k C_j = X$$

- Each cluster C_j is represented by its centroid (mean vector):

$$\boldsymbol{\mu}_j = \frac{1}{|C_j|} \sum_{\mathbf{x} \in C_j} \mathbf{x}$$

where $|C_j|$ is the number of points in cluster C_j .

3. Objective Function: Minimizing Within-Cluster Sum of Squares (WCSS)

K-means aims to minimize the total squared Euclidean distance between points and their assigned cluster centroids:

$$J = \sum_{j=1}^k \sum_{\mathbf{x}_i \in C_j} \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2$$

where $\|\cdot\|$ denotes the Euclidean norm.

This objective function measures **inertia** or **compactness** of clusters. The smaller J , the tighter the clusters.

4. K-Means Algorithm: Lloyd's Algorithm

The algorithm iteratively performs two steps until convergence:

Step 1: Assignment Step (E-step)

Assign each data point to the cluster with the nearest centroid:

$$c_i = \arg \min_{j \in \{1, \dots, k\}} \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2$$

where c_i is the cluster index assigned to $\mathbf{x}_i$.

Step 2: Update Step (M-step)

Recalculate centroids as the mean of all points assigned to each cluster:

$$\boldsymbol{\mu}_j = \frac{1}{|C_j|} \sum_{i: c_i=j} \mathbf{x}_i$$

5. Algorithm Initialization and Convergence

- **Initialization:** Choose initial centroids $\{\boldsymbol{\mu}_1^{(0)}, \dots, \boldsymbol{\mu}_k^{(0)}\}$ randomly or by heuristic (e.g., k-means++).

- **Iteration:** Repeat assignment and update steps until cluster assignments no longer change or the decrease in J is below a threshold.
- **Convergence:** The algorithm monotonically decreases J at each iteration and converges to a local minimum.

6. Mathematical Properties

6.1 Decomposition of Objective

The objective can be rewritten using indicator variables:

Define r_{ij} as:

$$r_{ij} = \begin{cases} 1 & \text{if } \mathbf{x}_i \in C_j \\ 0 & \text{otherwise} \end{cases}$$

Then,

$$J = \sum_{i=1}^n \sum_{j=1}^k r_{ij} \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2$$

The optimization problem is:

$$\min_{\{r_{ij}\}, \{\boldsymbol{\mu}_j\}} J$$

subject to:

$$\sum_{j=1}^k r_{ij} = 1, r_{ij} \in \{0,1\}$$

6.2 Alternating Minimization

- Fix centroids $\{\boldsymbol{\mu}_j\}$, minimize J w.r.t. assignments r_{ij} by assigning each point to the closest centroid.

- Fix assignments r_{ij} , minimize J w.r.t. centroids μ_j by setting centroids to cluster means.

This alternating minimization guarantees J does not increase at each iteration.

7. Distance Metric: Euclidean Distance

The Euclidean distance between two points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ is:

$$\|\mathbf{x} - \mathbf{y}\| = \sqrt{\sum_{m=1}^d (x_m - y_m)^2}$$

In k-means, squared Euclidean distance is used to simplify optimization:

$$\|\mathbf{x} - \mathbf{y}\|^2 = \sum_{m=1}^d (x_m - y_m)^2$$

8. Computational Complexity

- Each iteration requires computing distances between every point and each centroid: $O(nkd)$.
- Typically, the number of iterations t is small, so overall complexity is:

$$O(t \cdot n \cdot k \cdot d)$$

9. Limitations and Considerations

- **Number of clusters k** must be specified a priori.
- Sensitive to **initialization**; poor initial centroids can lead to suboptimal clustering.
- Only finds **local minima** of J .
- Assumes **spherical clusters** due to Euclidean distance.
- Not suitable for clusters with complex shapes or different sizes.

10. Example: Step-by-Step

Given data points in 2D:

$$X = \{(1,2), (1,4), (1,0), (10,2), (10,4), (10,0)\}$$

and $k=2$:

- Initialize centroids randomly, e.g., $\mu_1^{(0)} = (1,2)$, $\mu_2^{(0)} = (10,2)$.
- Assign points to closest centroid.
- Update centroids as means of assigned points.
- Repeat until assignments stabilize.

11. Summary Table of Key Equations

Concept	Equation		
Objective function	$J = \sum_{j=1}^k \sum_{\mathbf{x}_i \in C_j} \ \mathbf{x}_i - \boldsymbol{\mu}_j\ ^2$		
Assignment step	$c_i = \arg\min_j \ \mathbf{x}_i - \boldsymbol{\mu}_j\ ^2$		
Update step (centroid)	$\boldsymbol{\mu}_j = \frac{1}{ C_j } \sum_{i: c_i = j} \mathbf{x}_i$	C_j	
Euclidean distance	$\ \mathbf{x} - \mathbf{y}\ = \sqrt{\sum_{m=1}^d (x_m - y_m)^2}$		
Indicator variable form	$J = \sum_{i=1}^n \sum_{j=1}^k r_{ij} \ \mathbf{x}_i - \boldsymbol{\mu}_j\ ^2$, $r_{ij} \in \{0,1\}$		

References

- Wikipedia: [K-means clustering](#)

- Fritz AI: [Mathematics behind K-means clustering](#)
- Thomas Bonald, *Clustering by k-Means*, Telecom ParisTech, 2019
- [Neptune.ai: K-Means Clustering Explained](#)

This presentation provides a rigorous mathematical understanding of k-means clustering, the simplest and most intuitive unsupervised learning method.

*
**