k-Means - Derivations & Proofs

Paul F. Roysdon, Ph.D.

Contents

1 Mathematical Derivations & Proofs

1.1 Introduction

k-Means (a.k.a. Lloyd–Max quantization) is a prototype-based clustering method that partitions n points in $\mathbb{R}^d$ into K clusters by minimizing the total within-cluster sum of squared Euclidean distances. The model is defined by *assignments* of points to clusters and *centroids* (cluster representatives). Alternating minimization over assignments and centroids yields a monotone descent algorithm that converges in finitely many steps to a local optimum.

1.2 Data and Notation

Let the training data be

$$\mathcal{D} = \{(\mathbf{x}_i)\}_{i=1}^{n}, \qquad \mathbf{x}_i \in \mathbb{R}^d \text{ (column vectors)},$$

and let the data matrix be $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$. Fix the desired number of clusters $K \in \{1, \ldots, n\}$.

Assignments. For each point i, introduce a one-hot assignment vector

$$\mathbf{z}_i \in \{0,1\}^K, \qquad \mathbf{1}^\top \mathbf{z}_i = 1,$$

and stack them as $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n] \in \{0,1\}^{K\times n}$. Write z_{ki} for the kth entry of $\mathbf{z}_i$.

Centroids. Let the centroid (prototype) of cluster k be $\boldsymbol{\mu}_k \in \mathbb{R}^d$, and stack them as

$$\boldsymbol{\mu} = [\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K] \in \mathbb{R}^{d\times K}.$$

The cluster size is $n_k = \sum_{i=1}^{n} z_{ki}$; in matrix form, $\mathbf{N} = \mathbf{Z}\mathbf{1} \in \mathbb{N}^K$ and $\mathbf{D} = \operatorname{diag}(n_k) = \mathbf{Z}\mathbf{Z}^\top$.

1.3 Model Formulation (Objective)

The k-Means *distortion* (within-cluster sum of squares) is

$$\mathcal{J}(\boldsymbol{\mu}, \mathbf{Z}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ki} \left\| \mathbf{x}_i - \boldsymbol{\mu}_k \right\|_2^2 \quad \text{s.t.} \quad \mathbf{Z} \in \{0,1\}^{K \times n},\ \mathbf{1}^\top \mathbf{z}_i = 1 \ \forall i. \tag{1}$$

Equivalently, in matrix form,

$$\mathcal{J}(\boldsymbol{\mu}, \mathbf{Z}) = \left\| \mathbf{X} - \boldsymbol{\mu} \mathbf{Z} \right\|_F^2. \tag{2}$$

Dimensions. $\mathbf{X} \in \mathbb{R}^{d \times n}$, $\boldsymbol{\mu} \in \mathbb{R}^{d \times K}$, $\mathbf{Z} \in \{0,1\}^{K \times n}$, and $\mathcal{J} \in \mathbb{R}_{\geq 0}$.

1.4 Blockwise Minimizers (Centroid and Assignment Updates)

We derive the exact minimizers of Eqn. (1) when one block ($\boldsymbol{\mu}$ or $\mathbf{Z}$) is held fixed.

Centroid update (given assignments)

For fixed $\mathbf{Z}$, the objective separates over k:

$$\mathcal{J}(\boldsymbol{\mu}, \mathbf{Z}) = \sum_{k=1}^{K} \sum_{i=1}^{n} z_{ki} \left\| \mathbf{x}_i - \boldsymbol{\mu}_k \right\|_2^2.$$

For a particular k, expand and differentiate w.r.t. $\boldsymbol{\mu}_k$:

$$\sum_{i=1}^{n} z_{ki} \left\| \mathbf{x}_i - \boldsymbol{\mu}_k \right\|_2^2 = \sum_{i=1}^{n} z_{ki} \big(\|\mathbf{x}_i\|_2^2 - 2\boldsymbol{\mu}_k^\top \mathbf{x}_i + \|\boldsymbol{\mu}_k\|_2^2 \big).$$

Ignoring the constant $\sum_i z_{ki} \|\mathbf{x}_i\|^2$, the minimization problem is quadratic with gradient

$$\nabla_{\boldsymbol{\mu}_k} = -2 \sum_{i=1}^{n} z_{ki} \mathbf{x}_i + 2 n_k \, \boldsymbol{\mu}_k.$$

Setting the gradient to zero gives the unique minimizer (if $n_k > 0$):

$$\boldsymbol{\mu}_k^\star = \frac{1}{n_k} \sum_{i=1}^{n} z_{ki} \, \mathbf{x}_i \quad \Longleftrightarrow \quad \boldsymbol{\mu}^\star = \mathbf{X} \, \mathbf{Z}^\top \, (\mathbf{Z}\mathbf{Z}^\top)^{-1}. \tag{3}$$

Thus, *each centroid is the mean of the points assigned to its cluster.* (If $n_k = 0$, the objective does not depend on $\boldsymbol{\mu}_k$; see the "Empty clusters" note below.)

Assignment update (given centroids)

For fixed $\boldsymbol{\mu}$, the problem decouples over i:

$$\min_{\mathbf{z}_i \in \{0,1\}^K,\ \mathbf{1}^\top \mathbf{z}_i = 1} \sum_{k=1}^{K} z_{ki} \left\| \mathbf{x}_i - \boldsymbol{\mu}_k \right\|_2^2.$$

Since $\mathbf{z}_i$ is one-hot, the minimizer sets $z_{k^\star i} = 1$ where

$$k^\star \in \arg \min_{k \in \{1, \ldots, K\}} \ \|\mathbf{x}_i - \boldsymbol{\mu}_k\|_2^2, \tag{4}$$

and zero elsewhere. Hence, *each point is assigned to its nearest centroid* (Voronoi rule).

1.5 Lloyd–Max Algorithm and Monotone Descent

Alternating Eqn. (3) and Eqn. (4) yields the classical Lloyd–Max algorithm:

1. **Initialize** centroids $\boldsymbol{\mu}^{(0)}$ (e.g., random points or k-means++ seeding).
2. For $t = 0, 1, 2, \ldots$
   (a) **Assignment (E-step):** For each i, set $\mathbf{z}_i^{(t+1)}$ by Eqn. (4) using $\boldsymbol{\mu}^{(t)}$.
   (b) **Centroid (M-step):** For each k with $n_k^{(t+1)} > 0$, set $\boldsymbol{\mu}_k^{(t+1)}$ by Eqn. (3) using $\mathbf{Z}^{(t+1)}$.
3. **Stop** when assignments no longer change or when the decrease in $\mathcal{J}$ is below tolerance.

Monotonicity and finite convergence proof At iteration t:

- With $\boldsymbol{\mu}^{(t)}$ fixed, the assignment step chooses, for each i, the minimizer of a finite set of values $\{\|\mathbf{x}_i - \boldsymbol{\mu}_k^{(t)}\|^2\}_k$, hence
$$\mathcal{J}(\boldsymbol{\mu}^{(t)}, \mathbf{Z}^{(t+1)}) \leq \mathcal{J}(\boldsymbol{\mu}^{(t)}, \mathbf{Z}^{(t)}).$$
- With $\mathbf{Z}^{(t+1)}$ fixed, the centroid step computes the unique minimizer Eqn. (3), hence
$$\mathcal{J}(\boldsymbol{\mu}^{(t+1)}, \mathbf{Z}^{(t+1)}) \leq \mathcal{J}(\boldsymbol{\mu}^{(t)}, \mathbf{Z}^{(t+1)}).$$

Therefore, $\mathcal{J}$ is nonincreasing across iterations. Since there are only finitely many distinct assignment matrices (K^n), the algorithm reaches a fixed point in finitely many steps (ties aside), which is a local minimum of Eqn. (1). ■

1.6 Alternative Forms and Identities

Matrix factorization view. Using Eqn. (2),

$$\|\mathbf{X} - \boldsymbol{\mu}\mathbf{Z}\|_F^2 = \operatorname{tr}(\mathbf{X}^\top \mathbf{X}) - 2\operatorname{tr}(\mathbf{Z}\,\mathbf{X}^\top \boldsymbol{\mu}) + \operatorname{tr}(\mathbf{Z}^\top \boldsymbol{\mu}^\top \boldsymbol{\mu}\mathbf{Z}).$$

At the minimizer for fixed $\mathbf{Z}$, substituting $\boldsymbol{\mu}^\star$ from Eqn. (3) yields

$$\min_{\boldsymbol{\mu}} \|\mathbf{X} - \boldsymbol{\mu}\mathbf{Z}\|_F^2 = \|\mathbf{X}\|_F^2 - \operatorname{tr}\Big(\mathbf{X}\,\mathbf{Z}^\top (\mathbf{Z}\mathbf{Z}^\top)^{-1} \mathbf{Z}\,\mathbf{X}^\top\Big) = \sum_{i=1}^{n} \|\mathbf{x}_i\|_2^2 - \sum_{k=1}^{K} n_k \|\bar{\mathbf{x}}_k\|_2^2, \tag{5}$$

where $\bar{\mathbf{x}}_k = \frac{1}{n_k} \sum_{i:z_{ki}=1} \mathbf{x}_i$ is the cluster mean. Hence minimizing $\mathcal{J}$ is equivalent to maximizing the sum of squared centroid norms weighted by cluster sizes.

Within/between scatter decomposition. Let the global mean be $\bar{\mathbf{x}} = \frac{1}{n} \sum_i \mathbf{x}_i$. One has the ANOVA-like identity

$$\sum_{i=1}^{n} \|\mathbf{x}_i - \bar{\mathbf{x}}\|_2^2 = \underbrace{\sum_{k=1}^{K} \sum_{i:z_{ki}=1} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|_2^2}_{\text{within-cluster (minimized)}} + \underbrace{\sum_{k=1}^{K} n_k \, \|\boldsymbol{\mu}_k - \bar{\mathbf{x}}\|_2^2}_{\text{between-cluster (maximized)}} .$$

Thus k-Means simultaneously minimizes within-cluster scatter and maximizes between-cluster scatter.

1.7 Connections to Probabilistic Models

Consider an isotropic Gaussian mixture with equal priors $\pi_k = \frac{1}{K}$ and common covariance $\sigma^2\mathbf{I}$:

$$p(\mathbf{x} \mid k) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \sigma^2\mathbf{I}).$$

The complete-data negative log-likelihood (with hard assignments) is, up to constants,

$$-\sum_{i,k} z_{ki} \log p(\mathbf{x}_i \mid k) = \frac{1}{2\sigma^2} \sum_{i,k} z_{ki} \, \|\mathbf{x}_i - \boldsymbol{\mu}_k\|_2^2 + \text{const.}$$

Maximizing likelihood over $(\boldsymbol{\mu}, \mathbf{Z})$ is equivalent to minimizing $\mathcal{J}$ in Eqn. (1). Lloyd's algorithm coincides with hard-EM for this model; taking $\sigma^2 \downarrow 0$ recovers the nearest-centroid E-step.

1.8 Variants and Extensions

Weighted k-Means. With nonnegative sample weights w_i, minimize $\sum_{i,k} w_i z_{ki} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2$ subject to one-hot assignments. The centroid update becomes

$$\boldsymbol{\mu}_k^\star = \frac{\sum_i w_i z_{ki} \mathbf{x}_i}{\sum_i w_i z_{ki}},$$

and the assignment step uses weighted nearest-centroid only through the decision rule if class-dependent costs are added.

Mahalanobis/whitened k-Means. Replacing $\|\cdot\|_2^2$ by $\|\mathbf{x}-\boldsymbol{\mu}\|_{\boldsymbol{\Sigma}^{-1}}^2 = (\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$ yields ellipsoidal Voronoi cells. If $\boldsymbol{\Sigma}$ is common and positive definite, a linear whitening transform reduces to standard k-Means.

Initialization and empty clusters. k-Means is sensitive to initialization. Popular seeding such as *k-Means++* selects the first center randomly, then picks subsequent centers with probability proportional to squared distance to the current set. If a cluster becomes empty ($n_k = 0$), a common fix is to re-seed $\boldsymbol{\mu}_k$ to the farthest point from its current nearest centroid.

1.9 Algorithm (Lloyd–Max)

1. **Input:** data $\{\mathbf{x}_i\}_{i=1}^n \subset \mathbb{R}^d$, clusters K, tolerance ε, max iters $T_{\max}$.
2. **Initialize** $\boldsymbol{\mu}^{(0)}$ (random data points or k-Means++); set $t \leftarrow 0$.
3. **Repeat** until convergence or $t = T_{\max}$:
   (a) **Assign:** $z_{ki}^{(t+1)} \leftarrow \mathbf{1}\{k \in \arg\min_j \|\mathbf{x}_i - \boldsymbol{\mu}_j^{(t)}\|^2\}$.
   (b) **Update:** $\boldsymbol{\mu}_k^{(t+1)} \leftarrow \frac{\sum_i z_{ki}^{(t+1)} \mathbf{x}_i}{\sum_i z_{ki}^{(t+1)}}$ (if denominator > 0; otherwise re-seed).
   (c) **Check:** stop if $\mathcal{J}(\boldsymbol{\mu}^{(t+1)}, \mathbf{Z}^{(t+1)})$ decreases by $< \varepsilon$ or assignments unchanged.
4. **Output:** $\boldsymbol{\mu}^{(*)}, \mathbf{Z}^{(*)}$, and $\mathcal{J}^{(*)}$.

1.10 Complexity and Geometry

Per iteration, computing all distances costs $O(nKd)$ and centroid updates cost $O(nd)$; with T iterations the total is $O(nKdT)$. The assignment rule induces a Voronoi tessellation of $\mathbb{R}^d$ with sites at $\{\boldsymbol{\mu}_k\}$; at stationarity, each centroid equals the mean of its Voronoi cell intersected with the data.

1.11 Summary of Variables and Their Dimensions

- $\mathbf{x}_i \in \mathbb{R}^d$: the ith data vector; $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$.
- $K \in \mathbb{N}$: number of clusters; n: number of samples; d: number of features.
- $\boldsymbol{\mu}_k \in \mathbb{R}^d$: centroid of cluster k; $\boldsymbol{\mu} \in \mathbb{R}^{d \times K}$ stacks centroids.
- $\mathbf{z}_i \in \{0,1\}^K$: one-hot assignment of point i; $\mathbf{Z} \in \{0,1\}^{K \times n}$ stacks assignments.
- $n_k = \sum_i z_{ki}$: size of cluster k; $\mathbf{D} = \mathbf{Z}\mathbf{Z}^\top = \mathrm{diag}(n_k)$.
- $\mathcal{J}(\boldsymbol{\mu}, \mathbf{Z}) = \sum_{i,k} z_{ki} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2 = \|\mathbf{X} - \boldsymbol{\mu}\mathbf{Z}\|_F^2 \in \mathbb{R}_{\geq 0}$: objective value.

1.12 Summary

From first principles, k-Means minimizes the within-cluster sum of squared Euclidean distances over one-hot assignments and centroids. The centroid step yields arithmetic means (closed form), and the assignment step is nearest-centroid (Voronoi). Alternating these exact blockwise minimizers monotonically decreases the objective and converges in finitely many steps to a local optimum. The objective also arises as the hard-EM limit of spherical Gaussian mixtures with equal priors, linking k-Means to probabilistic clustering and matrix factorization viewpoints.