Annotated version

Machine Learning Course - CS-433

K-Means Clustering

Nov 22, 2023

Martin Jaggi
Last updated on: November 20, 2023

credits to Mohammad Emtiyaz Khan & Rüdiger Urbanke

EPFL

Clustering

Clusters are groups of points whose inter-point distances are small compared to the distances outside the cluster.

The goal is to find "prototype" points $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K$ and cluster assignments $z_n \in \{1, 2, \ldots, K\}$ for all $n = 1, 2, \ldots, N$ data vectors $\mathbf{x}_n \in \mathbb{R}^D$.

K-means clustering

Assume K is known.

$$\min_{\mathbf{z}, \boldsymbol{\mu}} \mathcal{L}(\mathbf{z}, \boldsymbol{\mu}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|_2^2$$

$$\text{s.t. } \boldsymbol{\mu}_k \in \mathbb{R}^D, z_{nk} \in \{0, 1\}, \sum_{k=1}^{K} z_{nk} = 1,$$

$$\text{where } \mathbf{z}_n = [z_{n1}, z_{n2}, \ldots, z_{nK}]^\top$$

$$\mathbf{z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N]^\top$$

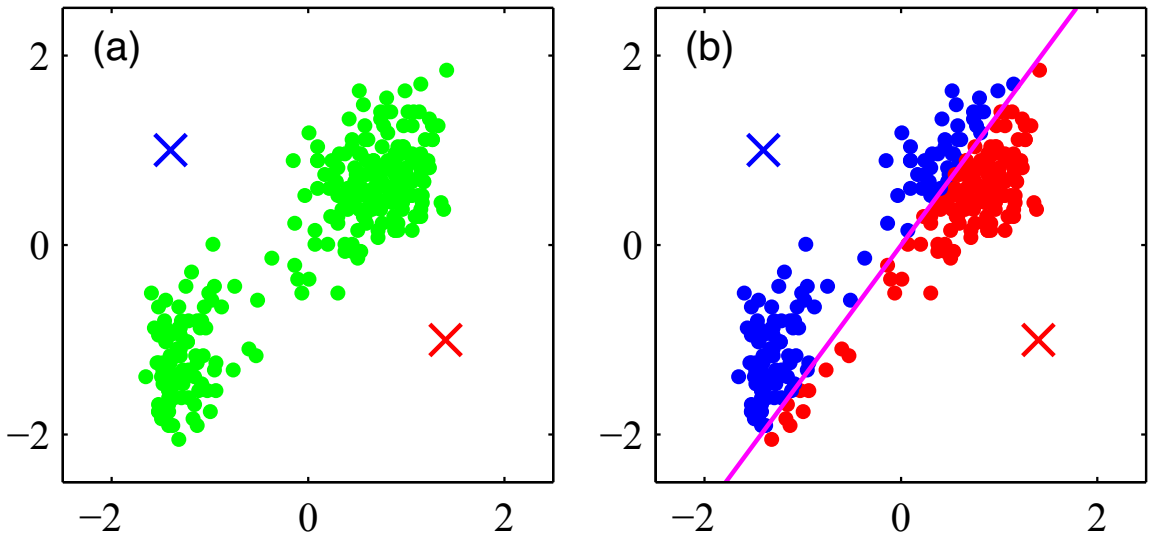
$$\boldsymbol{\mu} = [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K]^\top$$

Is this optimization problem easy?

Algorithm: Initialize $\boldsymbol{\mu}_k \, \forall k$,
then iterate:

1. For all n, compute $\mathbf{z}_n$ given $\boldsymbol{\mu}$.
2. For all k, compute $\boldsymbol{\mu}_k$ given $\mathbf{z}$.

Step 1: For all n, compute $\mathbf{z}_n$ given $\boldsymbol{\mu}$.

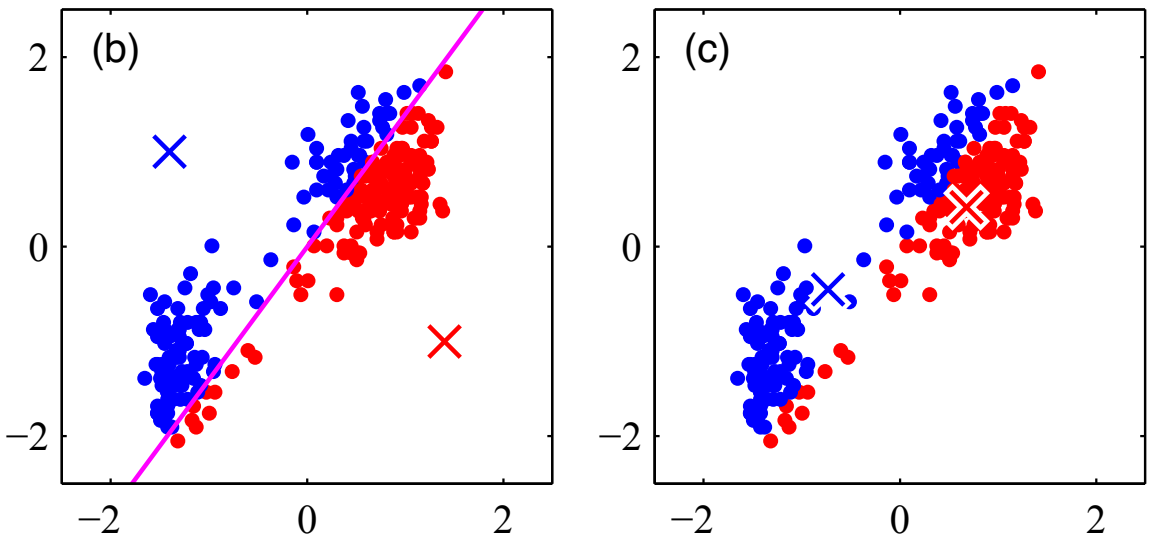


$$z_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_{j=1,2,\ldots K} \|\mathbf{x}_n - \boldsymbol{\mu}_j\|_2^2 \\ 0 & \text{otherwise} \end{cases}$$

Step 2: For all k, compute $\boldsymbol{\mu}_k$ given $\mathbf{z}$.
Take derivative w.r.t. $\boldsymbol{\mu}_k$ to get:

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} z_{nk} \mathbf{x}_n}{\sum_{n=1}^{N} z_{nk}}$$

Hence, the name 'K-means'.



Summary of K-means

Initialize $\boldsymbol{\mu}_k \, \forall k$, then iterate:

1. For all n, compute $\mathbf{z}_n$ given $\boldsymbol{\mu}$.

$$z_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}_n - \boldsymbol{\mu}_j\|_2^2 \\ 0 & \text{otherwise} \end{cases}$$

2. For all k, compute $\boldsymbol{\mu}_k$ given $\mathbf{z}$.

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^N z_{nk} \mathbf{x}_n}{\sum_{n=1}^N z_{nk}}$$

Convergence to a local optimum is assured since each step decreases the cost (see Bishop, Exercise 9.1).

Coordinate descent

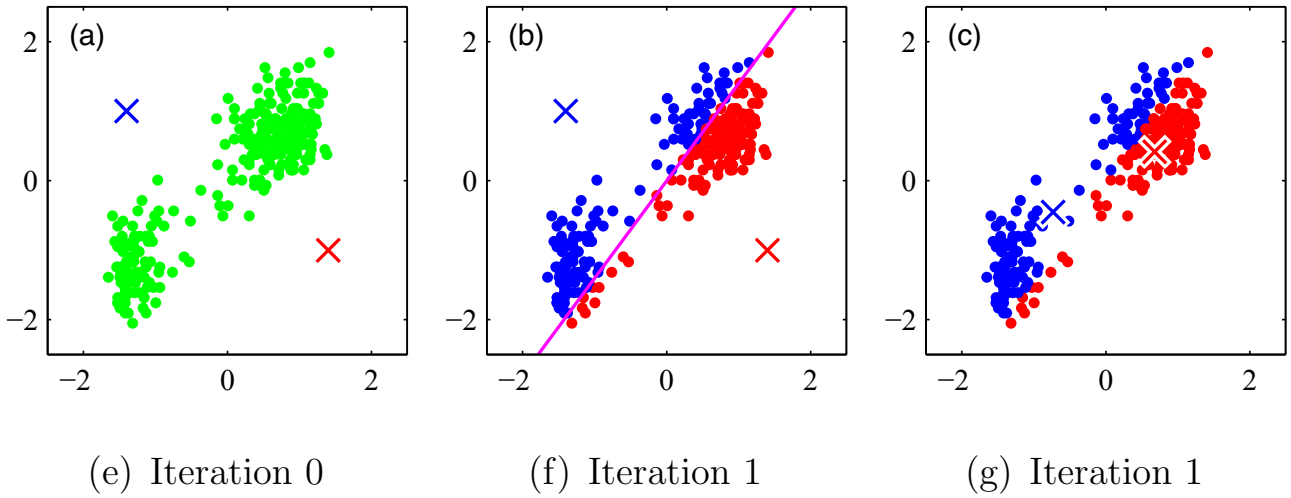
K-means is a coordinate descent algorithm, where, to find $\min_{\boldsymbol{z},\boldsymbol{\mu}} \mathcal{L}(\mathbf{z}, \boldsymbol{\mu})$, we start with some $\boldsymbol{\mu}^{(0)}$ and repeat the following:

$$\mathbf{z}^{(t+1)} := \arg\min_{\boldsymbol{z}} \; \mathcal{L}(\mathbf{z}, \boldsymbol{\mu}^{(t)})$$

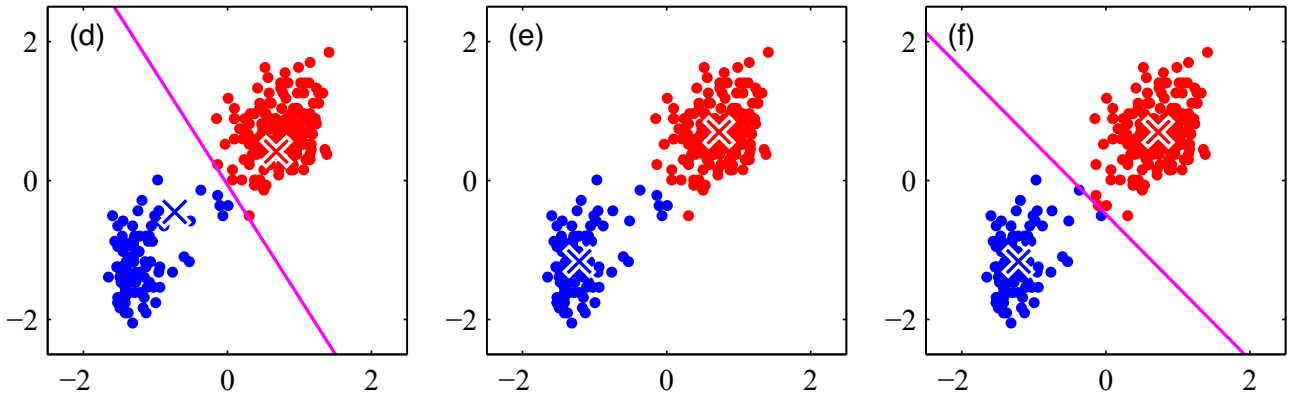
$$\boldsymbol{\mu}^{(t+1)} := \arg\min_{\boldsymbol{\mu}} \; \mathcal{L}(\mathbf{z}^{(t+1)}, \boldsymbol{\mu})$$

Examples

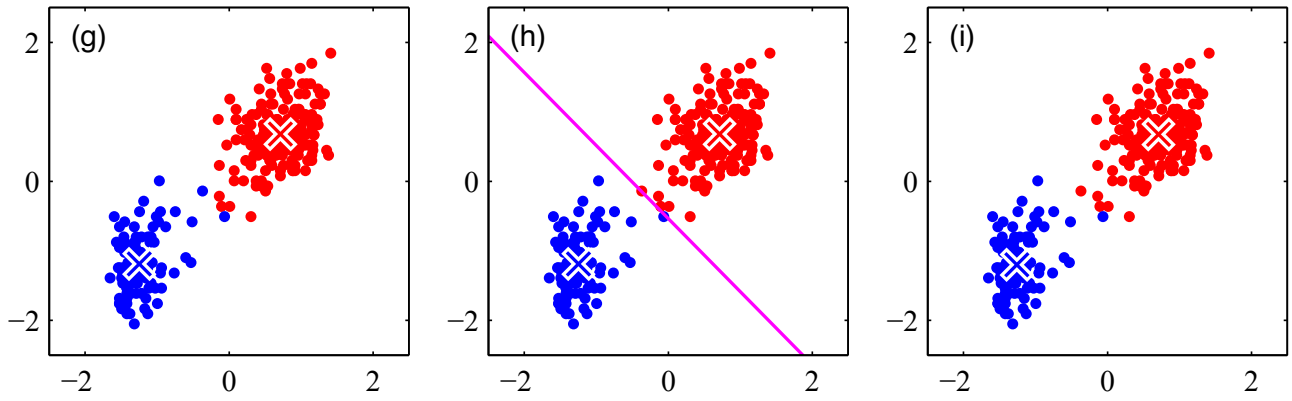
K-means for the "old-faithful" dataset (Bishop's Figure 9.1)



(e) Iteration 0 (f) Iteration 1 (g) Iteration 1



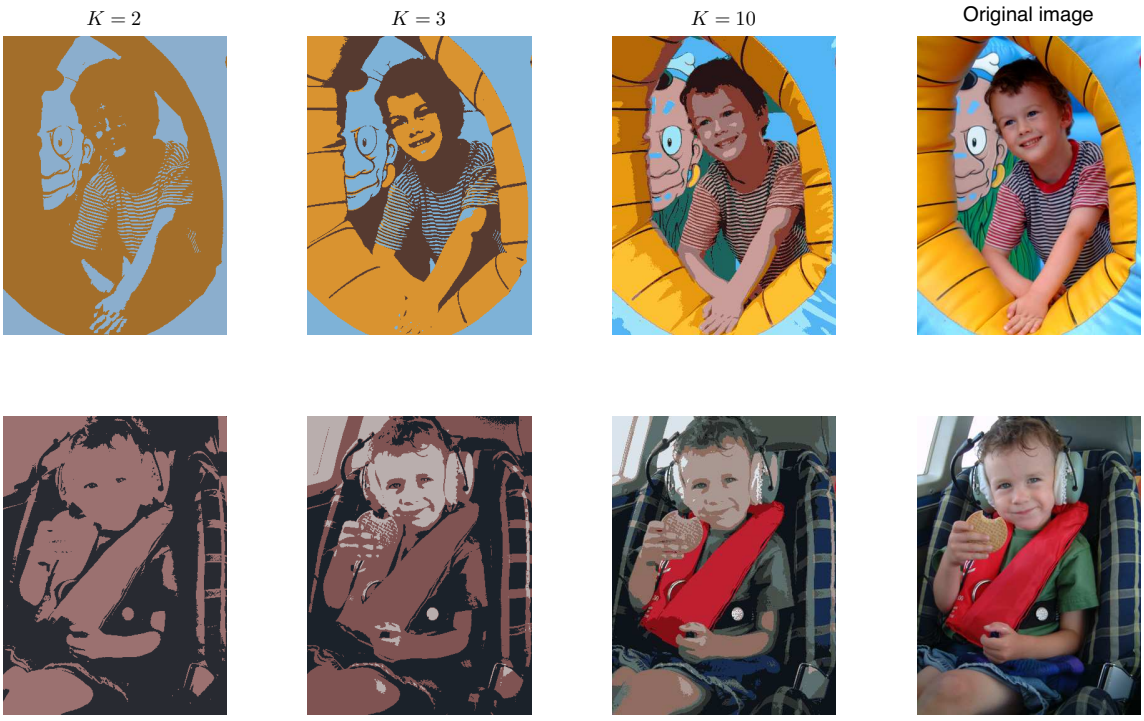
(h) Iteration 2 (i) Iteration 2 (j) Iteration 3



(k) Iteration 3 (l) Iteration 4 (m) Iteration 4

Data compression for images (this is also known as vector quantization).

$x_n = \begin{pmatrix} r \\ g \\ b \end{pmatrix} \in \mathbb{R}^3$



Probabilistic model for K-means

$$\log \prod_{n=1}^{N} p(x_n \mid \mu, z) = \log \prod_{n} \mathcal{N}(x_n \mid \mu_{k'}, I_0)$$

k': group of x_n

$$\log p(X \mid \mu, z) \quad (X = (x_1, \dots, x_N))$$

$$= \log \prod_{n} \prod_{k=1}^{K} \mathcal{N}(x_n \mid \mu_k, I_0)^{z_{nk}}$$

$$= \log \prod_{n} \prod_{k=1}^{K} c \cdot e^{-\frac{1}{2}\|x_n - \mu_k\|^2 \cdot z_{nk}}$$

$$= -\sum_{n=1}^{N} \sum_{k=1}^{K} \frac{1}{2}\|x_n - \mu_k\|^2 z_{nk} + c'$$

$$= -L(\mu, z)$$

K-means as a Matrix Factorization

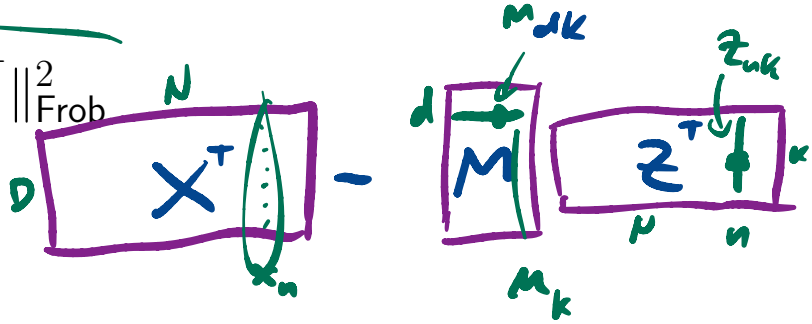
Recall the objective

$$\min_{\mathbf{z},\boldsymbol{\mu}} \mathcal{L}(\mathbf{z},\boldsymbol{\mu}) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\|\mathbf{x}_n - \boldsymbol{\mu}_k\|_2^2$$

$$= \|\mathbf{X}^\top - \mathbf{M}\mathbf{Z}^\top\|_{\text{Frob}}^2$$

$$\text{s.t. } \boldsymbol{\mu}_k \in \mathbb{R}^D,$$

$$z_{nk} \in \{0,1\}, \sum_{k=1}^{K} z_{nk} = 1.$$



Issues with K-means

1. Computation can be heavy for large N, D and K.

$O(N \cdot D \cdot K)$

2. Clusters are forced to be spherical (e.g. cannot be elliptical).

3. Each example can belong to only one cluster ("hard" cluster assignments).