Machine Learning Course - CS-433

K-Means Clustering

Nov 9, 2017

minor changes by Martin Jaggi 2016

minor changes by Martin Jaggi 2017

Last updated: November 9, 2017

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

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.

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

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

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

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

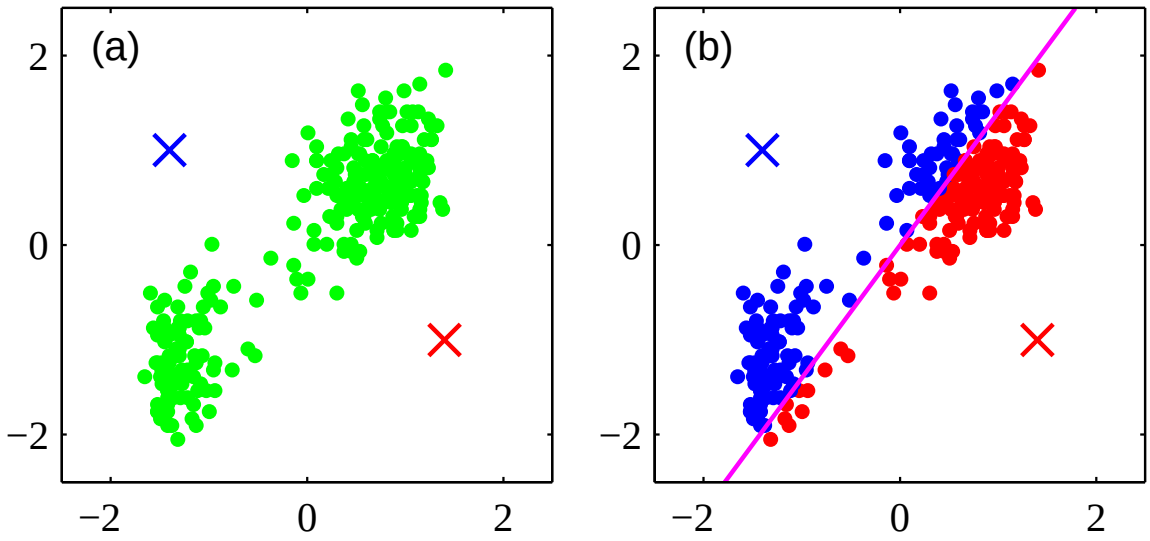
$$\boldsymbol{\mu} = [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K]^\top$$

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 $\boldsymbol{\mu}$.

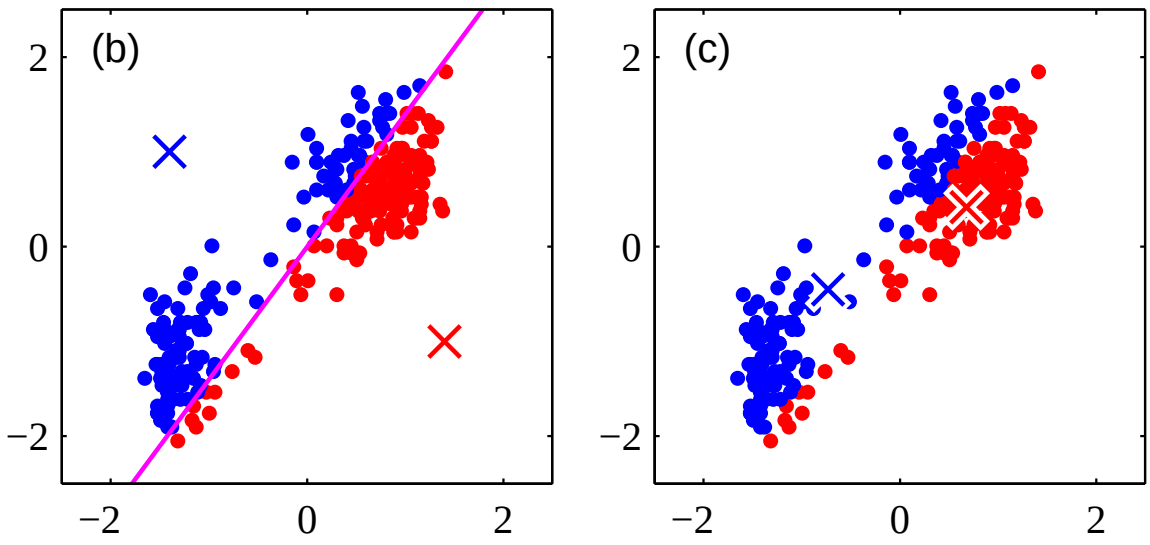


$$z_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_{j=1,2,\ldots K} \|\mathbf{x}_n - \boldsymbol{\mu}_j\|_2^2 \\ 0 & \text{otherwise} \end{cases}$$

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

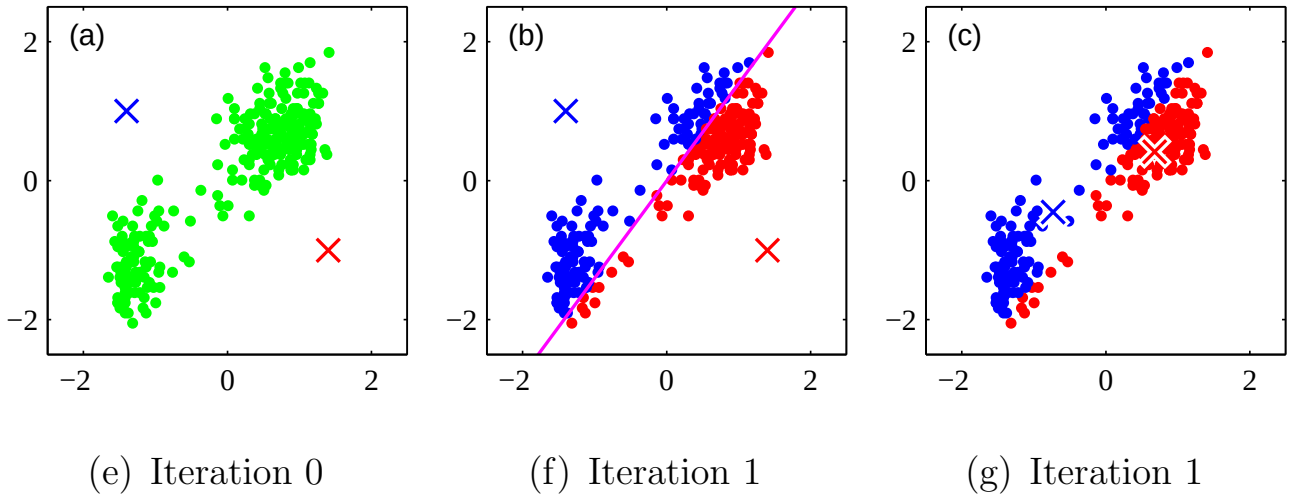
K-means is a coordinate descent algorithm, where, to find $\min_{\boldsymbol{z},\boldsymbol{\mu}} \mathcal{L}(\mathbf{z}, \boldsymbol{\mu})$, we start with some $\boldsymbol{\mu}^{(0)}$ and repeat the following:

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

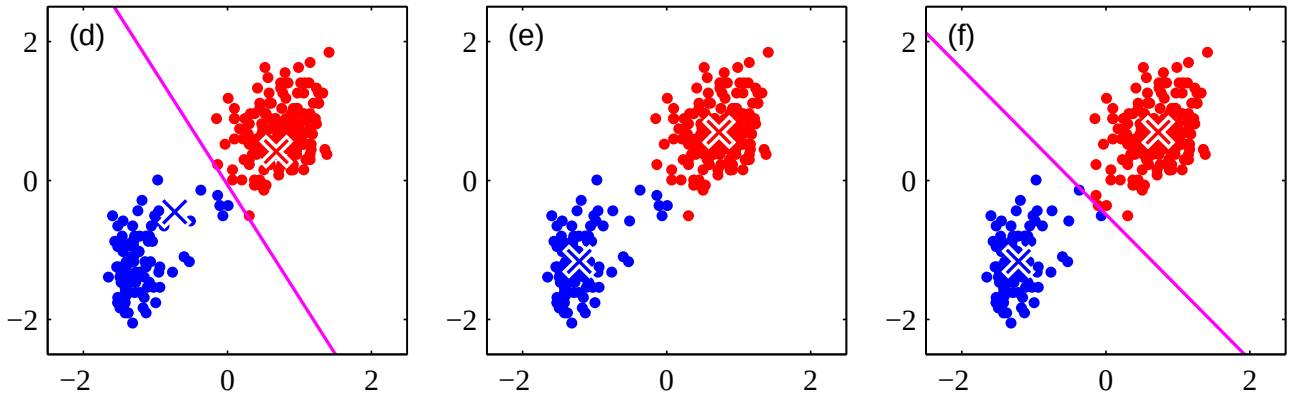
$$\boldsymbol{\mu}^{(t+1)} := \arg\min_{\boldsymbol{\mu}} \; \mathcal{L}(\mathbf{z}^{(t+1)}, \boldsymbol{\mu})$$

Examples

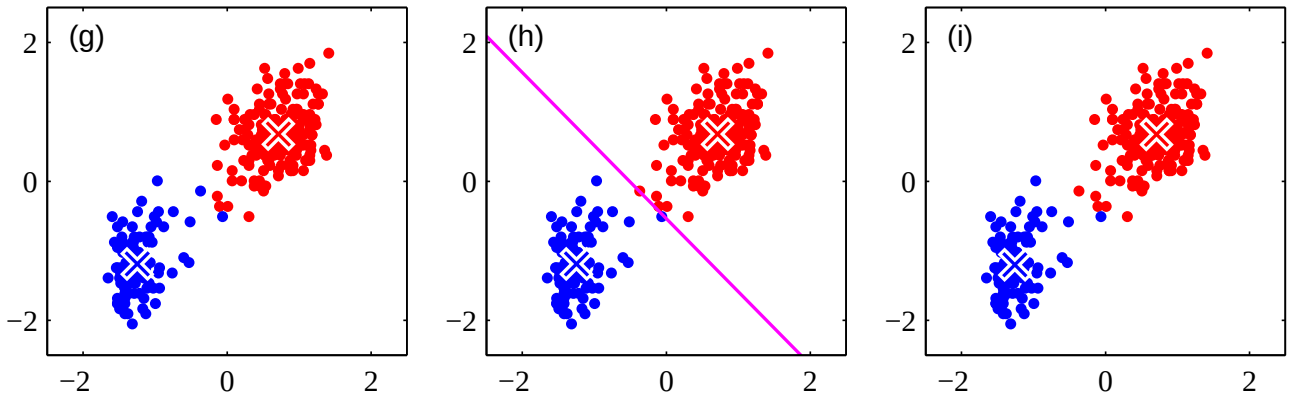
K-means for the "old-faithful" dataset (Bishop's Figure 9.1)



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

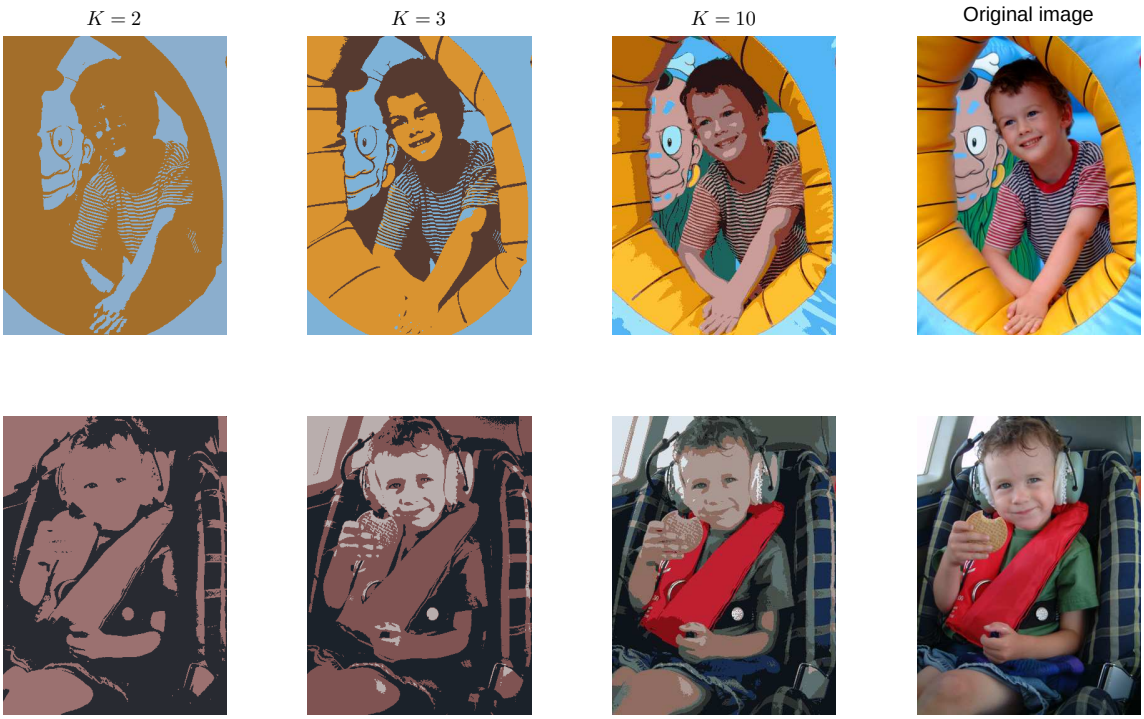


(h) Iteration 2 (i) Iteration 2 (j) Iteration 3



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

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



Probabilistic model for K-means

K-means as a Matrix Factorization

Recall the objective

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

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

Issues with K-means

1. Computation can be heavy for large N, D and K.
2. Clusters are forced to be spherical (e.g. cannot be elliptical).
3. Each example can belong to only one cluster ("hard" cluster assignments).

Exercises

1. Understand the iterative algorithm for K-means. Why is the problem difficult to optimize and how does the iterative algorithm make it simpler?
2. What is the computational complexity of K-means?
3. Derive the probabilistic model associated with the cost function.