annotated version

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 7, 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 $\mathbf{x}_n \in \mathbb{R}^D$.

$\mathbb{R}^D$

means

Specify # groups K

K-means clustering

Assume K is known.

within-cluster distances

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

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

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

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

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

Is this optimization problem easy?

non-convex

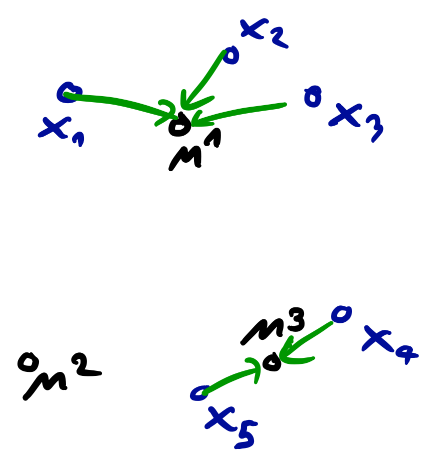
Binary Assignment

$$z_{nk} = \begin{cases} 1 & \text{if } n \text{ assigned to } k \\ 0 & \text{otherwise} \end{cases}$$

$z_n = (0, \ldots, 1, \ldots 0)$

actual k

K=3



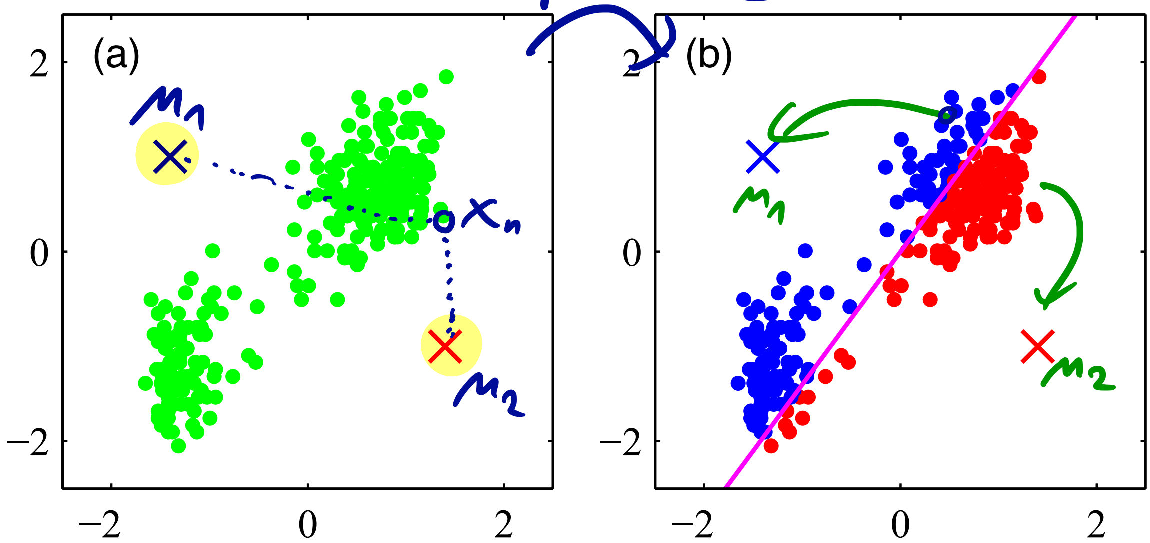
Algorithm: Initialize $\boldsymbol{\mu}_k \, \forall k$,
then iterate:

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

$K=2$

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

update z



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

for every n

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

Fix k

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

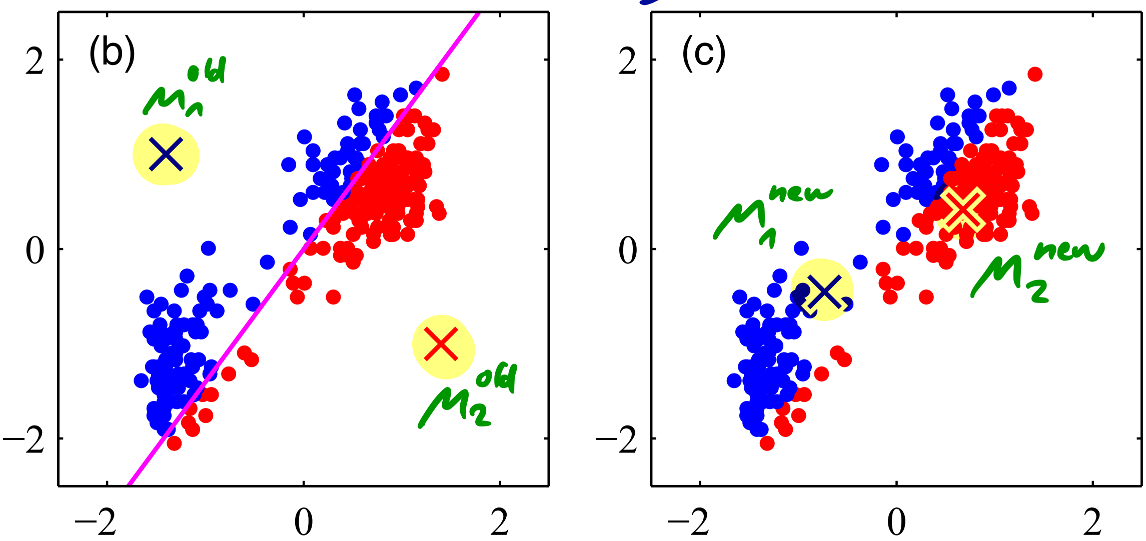
= Average Mean of all points of group k

Hence, the name 'K-means'.

$\min_{\mu} \mathcal{L}(\mu, z)$: $\quad \nabla_\mu \mathcal{L}(\mu, z) \stackrel{!}{=} 0$

step 2

update μ



Summary of K-means

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

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

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

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

repeat

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

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

$\forall k \quad O(N \cdot \cancel{K} \cdot D)$

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

$\mathcal{L} \geq 0$

Coordinate descent

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

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

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

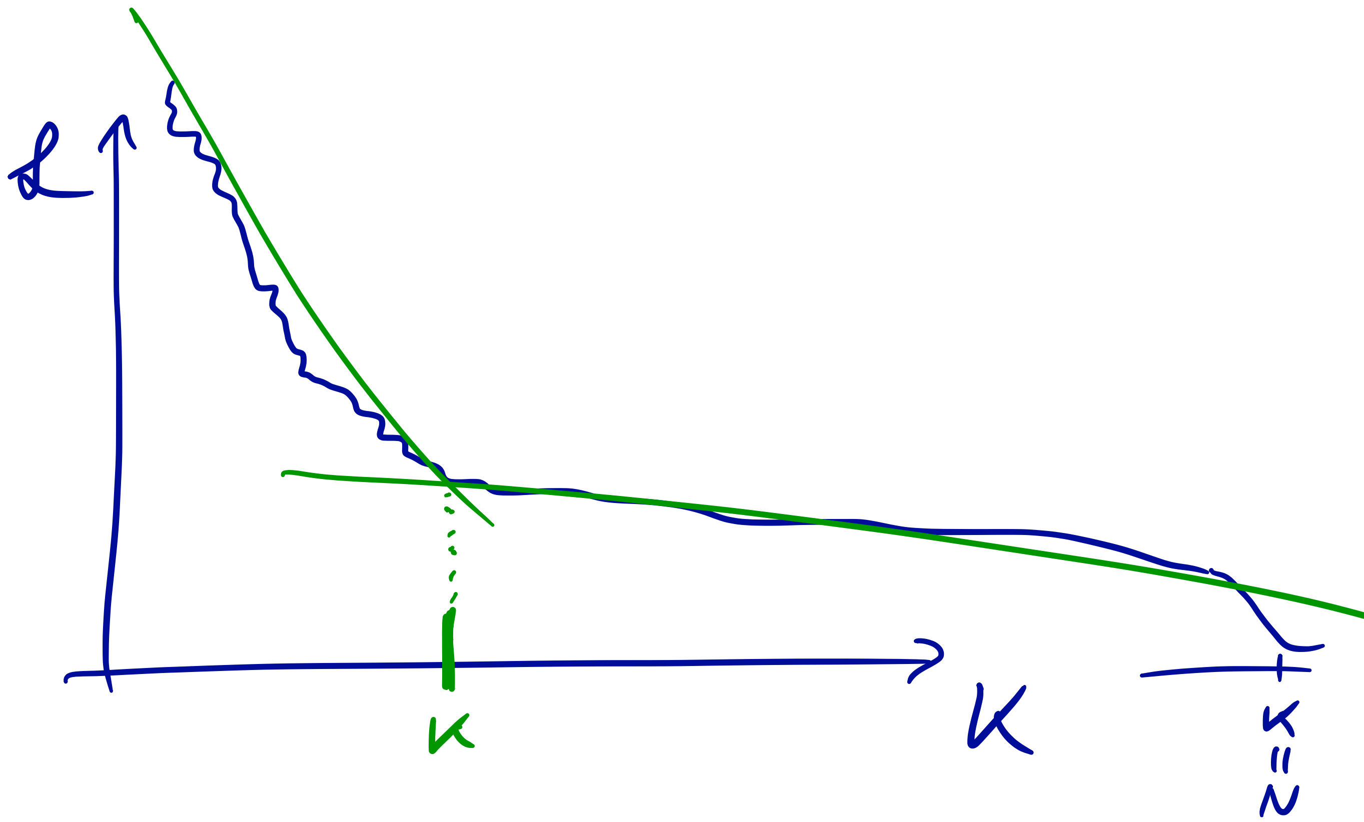
$\cancel{\nabla_z \mathcal{L}}$

update assignments

update means

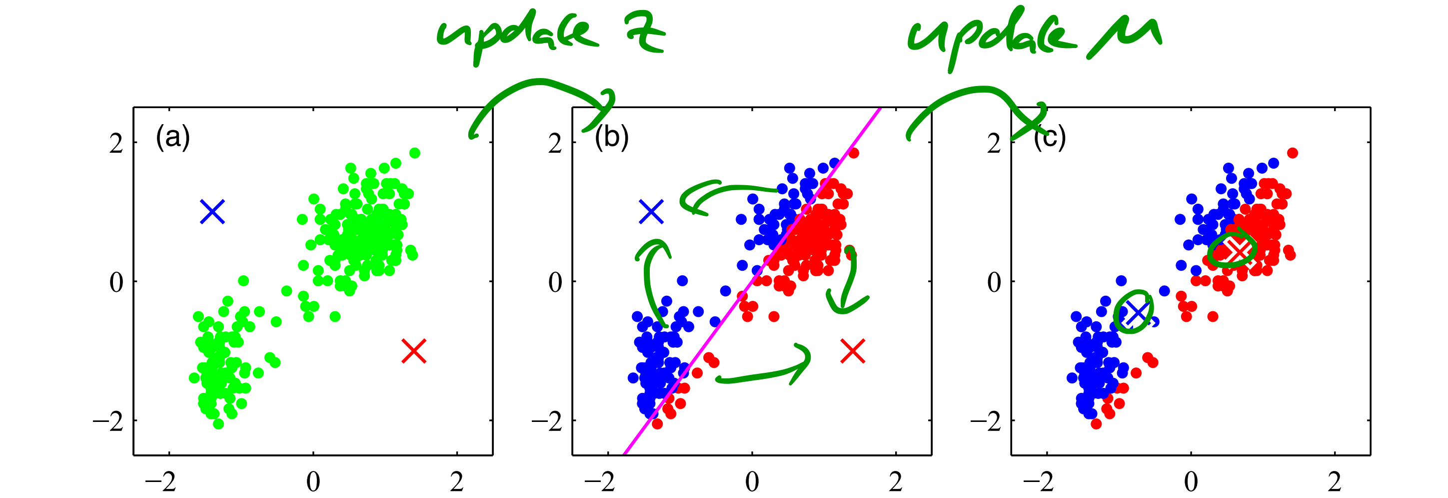
$\nabla_\mu \mathcal{L} \stackrel{!}{=} 0$

How to set K

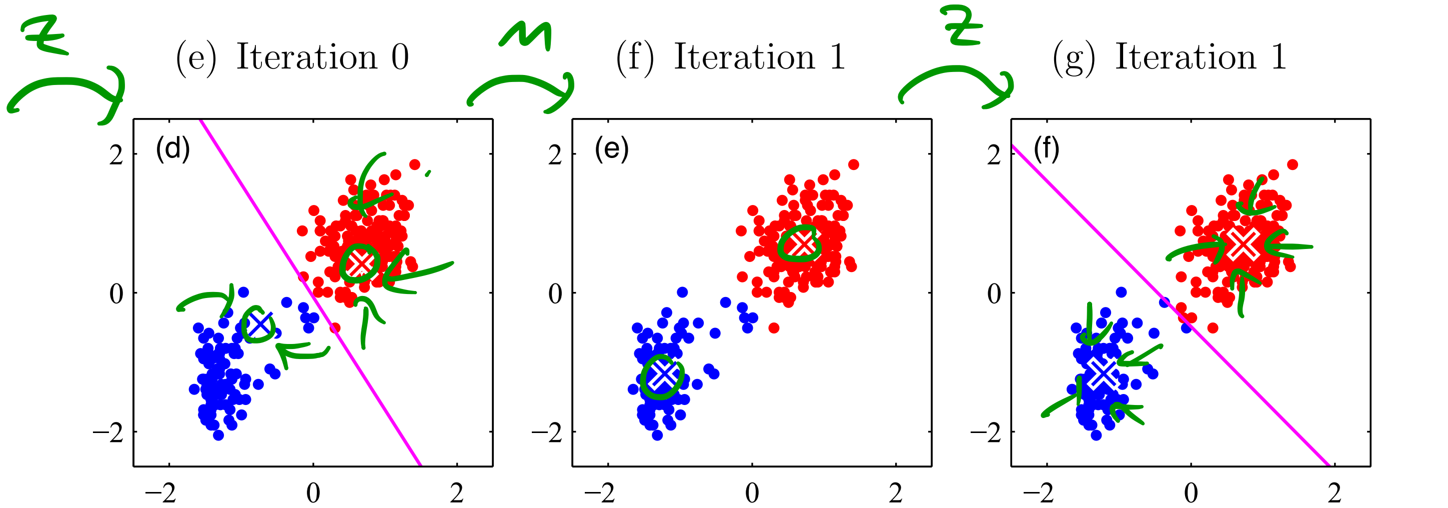


Examples

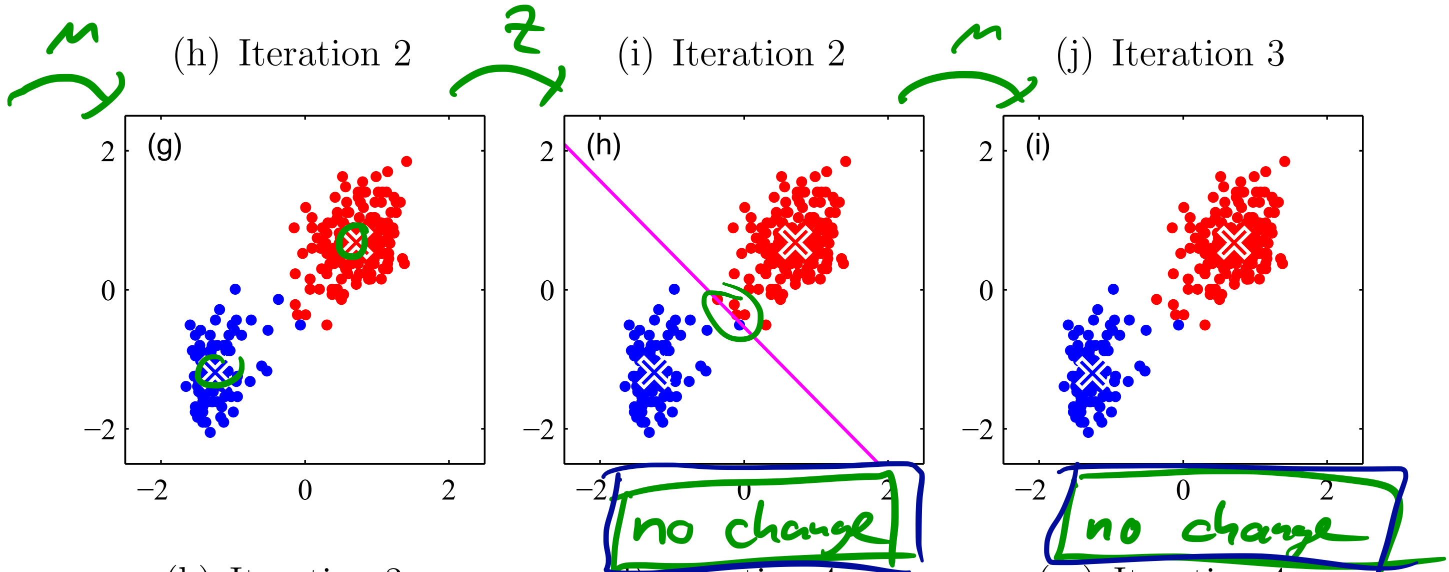
K-means for the “old-faithful” dataset (Bishop’s Figure 9.1)



(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

stop

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

200

300

60k

Probabilistic model for K-means

Likelihood of X given parameters μ, z

$$p(x_n \mid \mu, z) = \prod_{n=1}^{N} \mathcal{N}(x_n \mid \mu_k, I)$$

$$p(X \mid \mu, z)$$

k for which $z_{nk} = 1$

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

$$= \prod_{n=1}^{N} \prod_{k=1}^{K} c \cdot \left(e^{-\frac{1}{2} \underbrace{\|x_n - \mu_k\|^2}_{= (x_n - \mu)^T (x_n - \mu)}} \right)^{z_{nk}}$$

$$-\log p(x_n \mid \mu, z) = \sum_{n=1}^{N} \sum_{k=1}^{K} \frac{1}{2} \|x_n - \mu_k\|^2 z_{nk} + c'$$

$$-\log p(X \mid \mu, z) = \mathcal{L}(\mu, z) + \text{const}$$

K-means as a Matrix Factorization

Recall the objective

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

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

$$\begin{pmatrix} | & & | \\ x_1 & \cdots & x_N \\ | & & | \end{pmatrix} - MZ^T$$

Data:

$$X = \begin{pmatrix} - & x_1 & - \\ & \vdots & \\ - & x_N & - \end{pmatrix} \quad N \times D$$

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

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

$$M = \begin{pmatrix} | & & | \\ \mu_1, & \ldots, & \mu_K \\ | & & | \end{pmatrix} \quad D \times K$$

$z_{nk} = 1$

$$Z = \begin{pmatrix} - & z_{1:} & - \\ & & \\ - & z_{N:} & - \end{pmatrix} \quad N \times K$$

$$z_{n:} = (0, \ldots, 1, \ldots, 0) \in \mathbb{R}^K$$

$$\|A\|_{\text{Frob}}^2 = \sum_{i,j} A_{ij}^2 = \sum_j \|A_{:j}\|^2 = \sum_i \|A_{i:}\|^2$$

Issues with K-means

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

problematic for very large 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.