Machine Learning Course - CS-433

K-Means Clustering

Nov 8, 2016

minor changes by Martin Jaggi 2016

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Clustering

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

choose K

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

means

assing n-th point to cluster k

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

K-means clustering

Assume K is known.

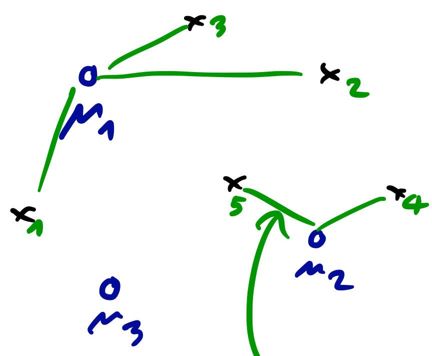
$K=3$

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

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

Z: assignment matrix

$$\text{where } \mathbf{z}_n = [z_{n1}, z_{n2}, \ldots, z_{nK}]^\top$$
$$\mathbf{z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N]^\top$$
$$\boldsymbol{\mu} = [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K]^\top$$



$z_{52} = 1$
$z_{5.} = 0$

Is this optimization problem easy?

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

update assignments

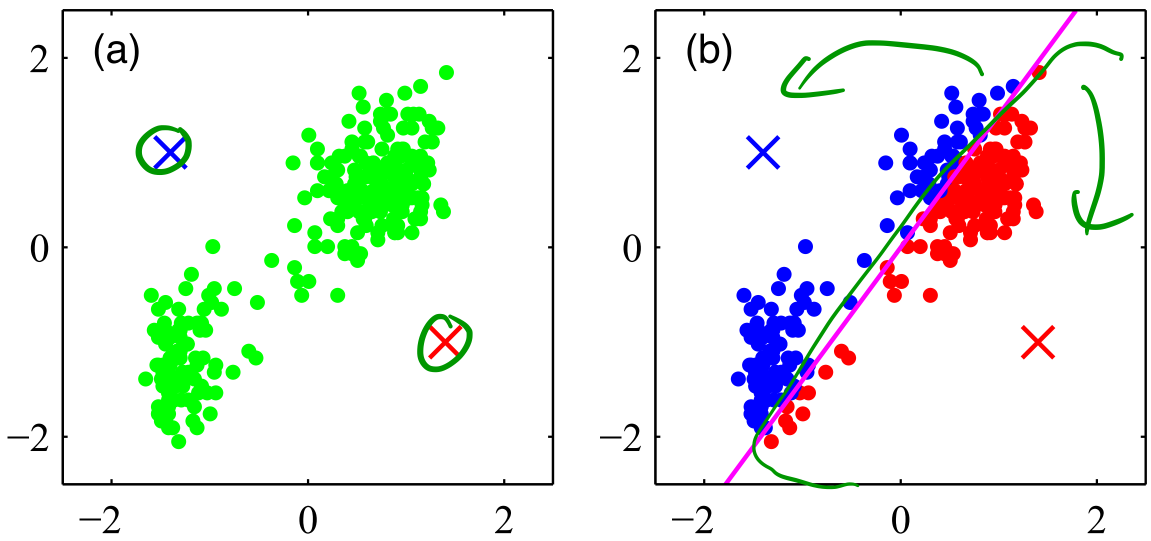
1. For all n, compute $\mathbf{z}_n$ given $\boldsymbol{\mu}$.

update means/centers

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

Example $K=2$

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



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

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

update means

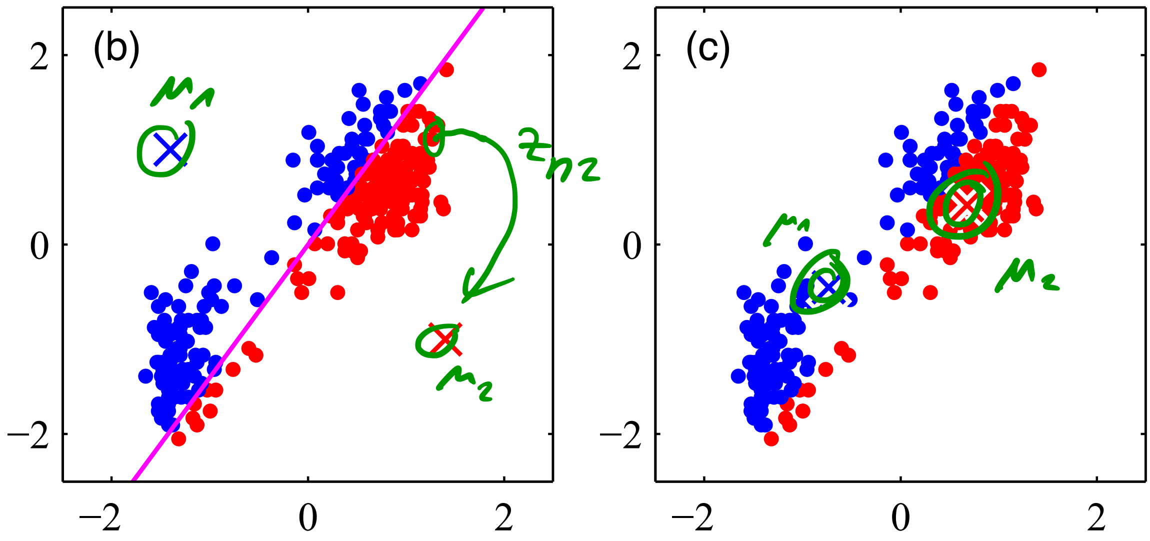
fix k

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

Hence, the name 'K-means'.

$$\mathcal{L} = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2 \qquad \mathbb{R}^D$$

$$\nabla_{\mu} \mathcal{L} = \dots \overset{!}{=} 0$$



Summary of K-means

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

update assignments z

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

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

update means μ

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

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

repeat

$\mathcal{L} \geq 0$

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

Coordinate descent

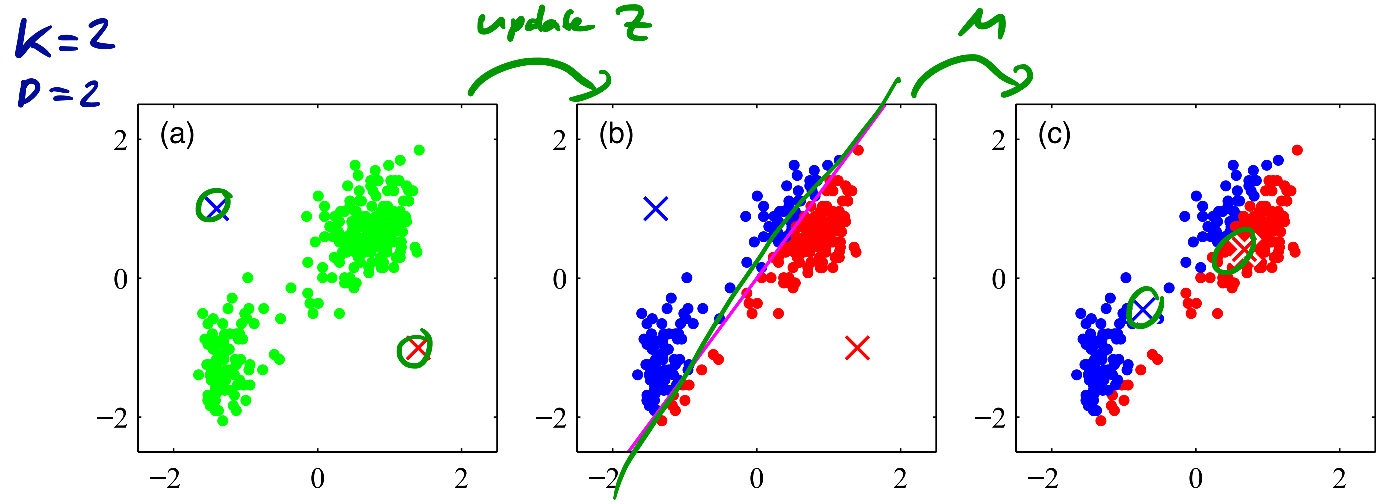
K-means is a coordinate descent algorithm, where, to find $\min_{\boldsymbol{z},\boldsymbol{\mu}} \mathcal{L}(\mathbf{z}, \boldsymbol{\mu})$, we start with some $\boldsymbol{\mu}^{(0)}$ and repeat the following:

$$\mathbf{z}^{(t+1)} := \arg\min_{\boldsymbol{z}} \mathcal{L}(\mathbf{z}, \boldsymbol{\mu}^{(t)}) \iff \text{step 1: update } z \quad z \in \{0,1\}^{\cdot\cdot}$$

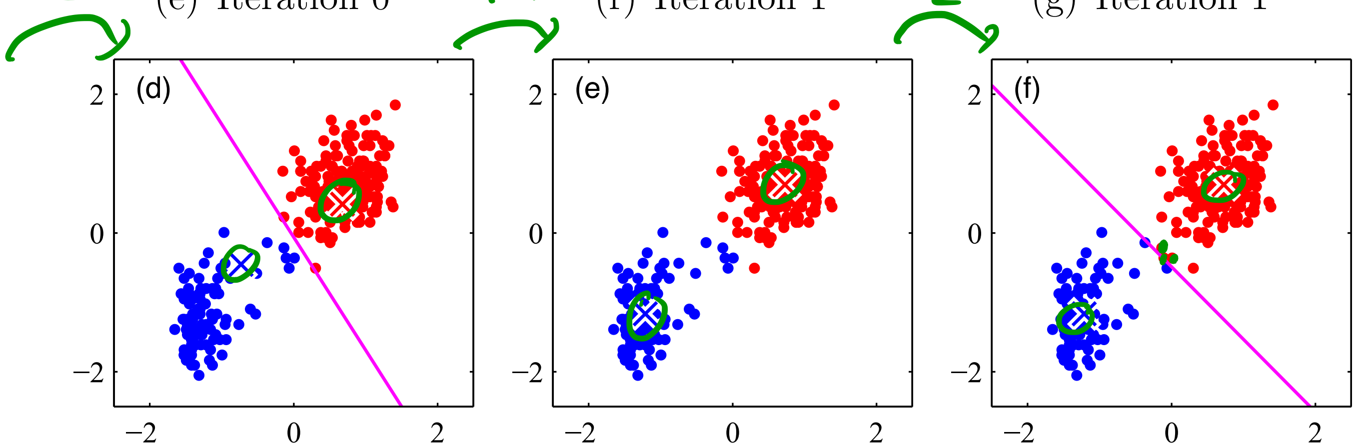
$$\boldsymbol{\mu}^{(t+1)} := \arg\min_{\boldsymbol{\mu}} \mathcal{L}(\mathbf{z}^{(t+1)}, \boldsymbol{\mu}) \iff \text{step 2: update } \mu$$

Examples

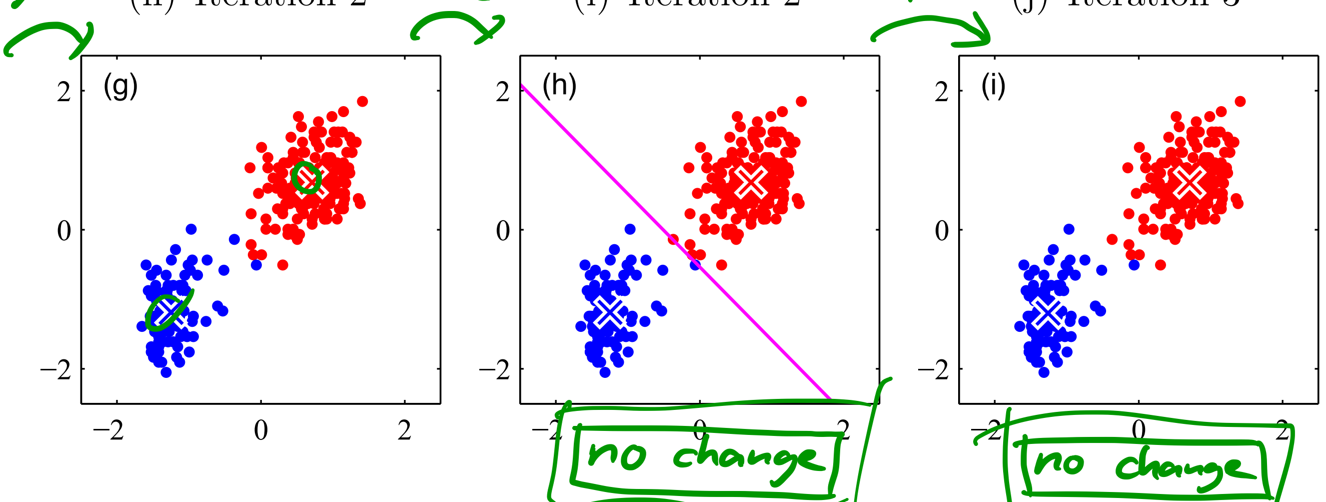
K-means for the "old-faithful" dataset (Bishop's Figure 9.1)



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

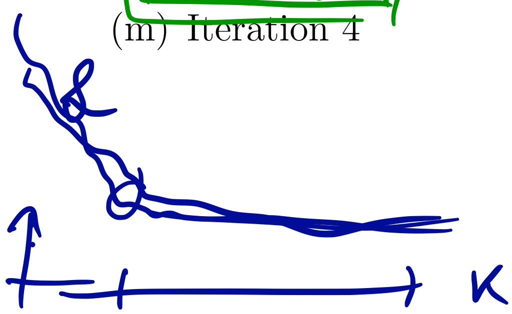


(h) Iteration 2 (i) Iteration 2 (j) Iteration 3



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

How to choose K ?



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

$N = 60'000$ $\quad x_n \in \mathbb{R}^3$: red : green : blue

Probabilistic model for K-means

likelihood of X given parameters μ, z

$$p(x_n \mid \mu, z) = \prod_n \mathcal{N}(x_n \mid \mu_k, I)$$

k s.t. $z_{nk} = 1$

$$p(\overset{\text{all}}{X} \mid \mu, z)$$

$$= \prod_n \prod_k \mathcal{N}(x_n \mid \mu_k, I)^{z_{nk}}$$

$$= \prod_n \prod_k \left(\exp - \tfrac{1}{2}\|x_n - \mu_k\|_2^2\right)^{z_{nk}}$$

$(x-\mu)^T \mathbb{1} (x-\mu)$ $\quad \Sigma$

$$-\log p(\overset{\text{all}}{X} \mid \mu, z) = \sum_n \sum_k \tfrac{1}{2}\|x_n - \mu_k\|_2^2 \, z_{nk}$$

$$= \mathcal{L}(\mu, z) + \text{const}$$

$\|A\|_{\text{Frob}}^2 = \sum_{i,j} (A_{ij})^2$

K-means as a Matrix Factorization

Recall the objective

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

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

$z_{n:} = (0, \dots, 1, \dots 0)$

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

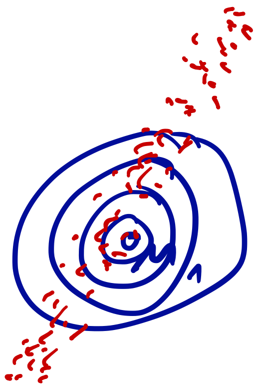
$$M = \mu = \begin{array}{c} K \\ \begin{pmatrix} | & & | \\ \mu_1 & \cdots & \mu_K \\ | & & | \end{pmatrix} \end{array} D$$

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

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

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).



ToDo

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.