annotated version

Machine Learning Course - CS-433

K-Means Clustering

Nov 19, 2020

Last updated on: November 17, 2020

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

z_n : 1-hot vector $\mathbb{R}^K$

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

assignment

K-means clustering

Assume K is known.

distance of x_n to μ_k

assigned

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

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

discrete

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

Is this optimization problem easy?

NP-hard

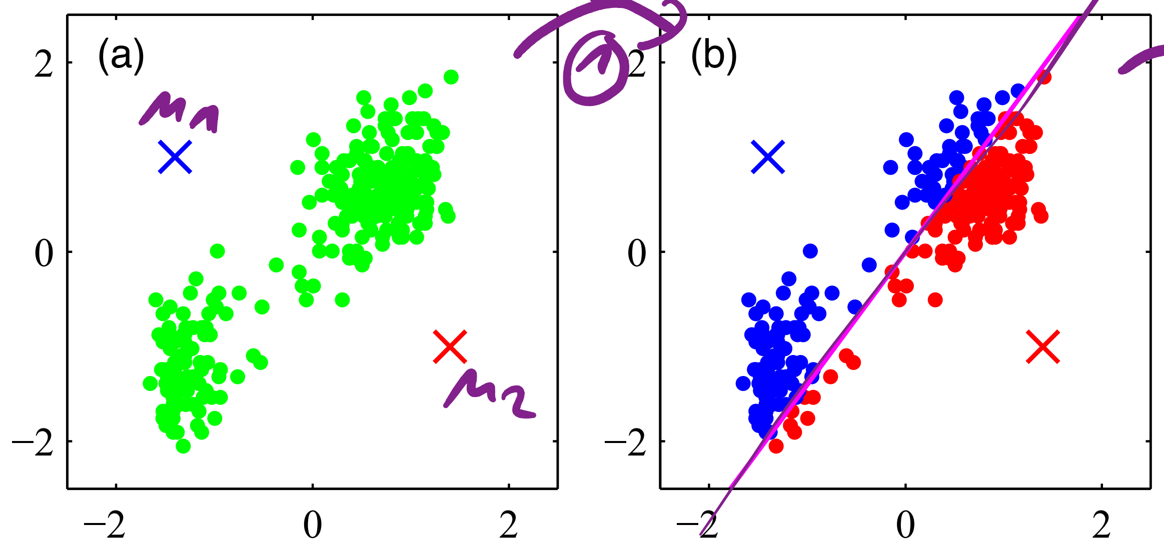
Algorithm: Initialize $\boldsymbol{\mu}_k \, \forall k$,
then iterate:

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

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

assignment z

$K=2$



1

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

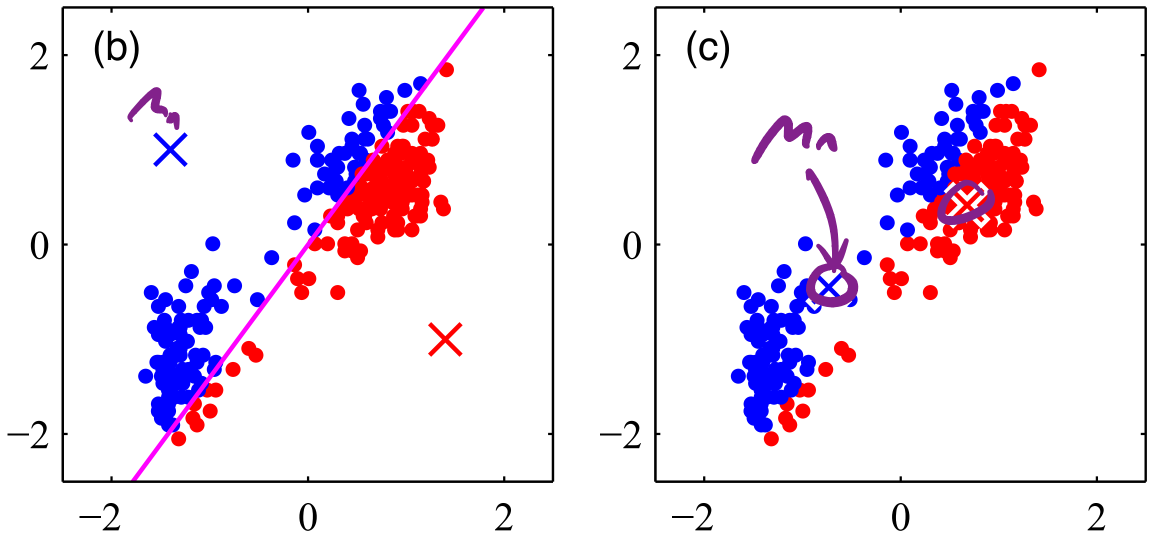
2

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

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} z_{nk}\mathbf{x}_n}{\sum_{n=1}^{N} z_{nk}} \leftarrow \text{\# points assigned to } k$$

Hence, the name 'K-means'.

update centers 2



Summary of K-means

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

update assignment z

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

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

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

update centers μ_k

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

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

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

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

$\nabla_{\mu} \mathcal{L}(z, \mu) \overset{!}{=} 0$

Coordinate descent

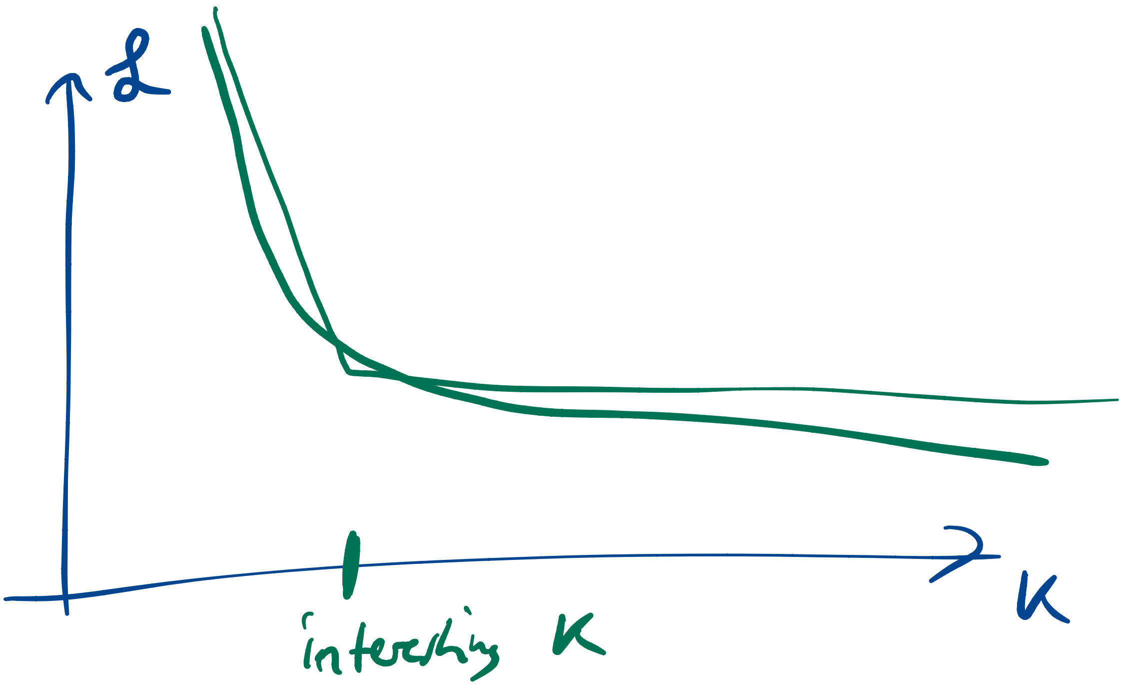
K-means is a coordinate descent algorithm, where, to find $\min_{z,\mu} \mathcal{L}(\mathbf{z}, \boldsymbol{\mu})$, we start with some $\boldsymbol{\mu}^{(0)}$ and repeat the following:

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

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

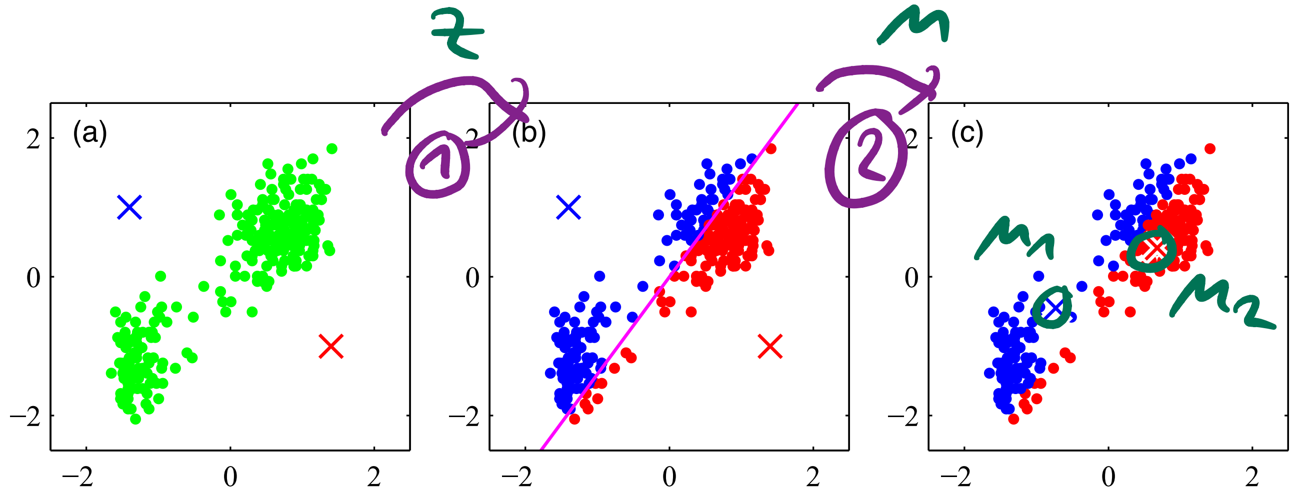
$\Leftrightarrow$ μ-update of K-means

How to set K?

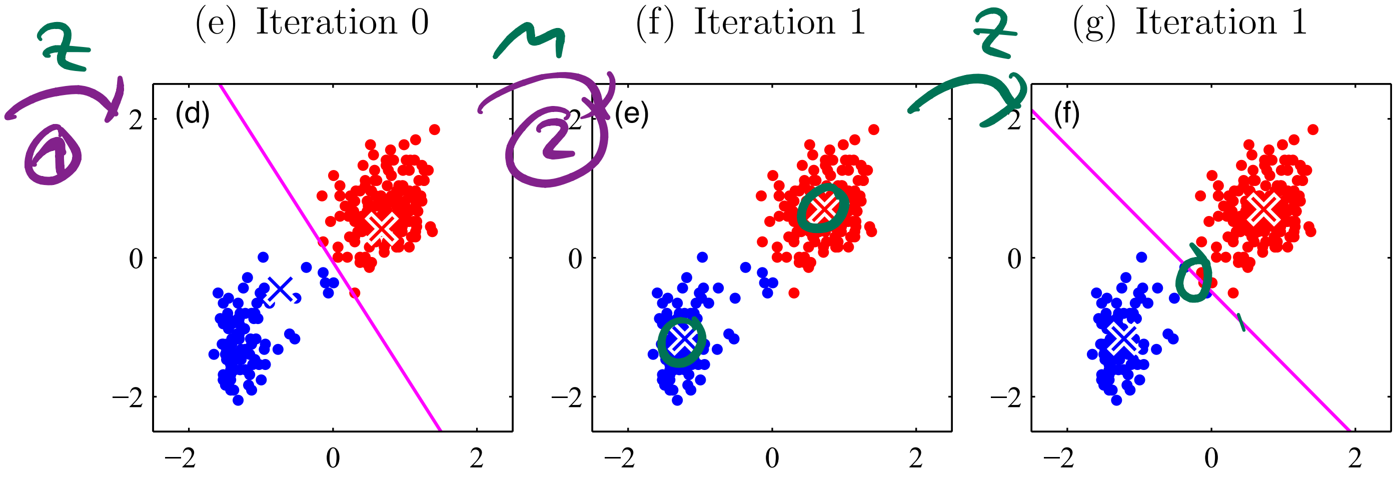


Examples

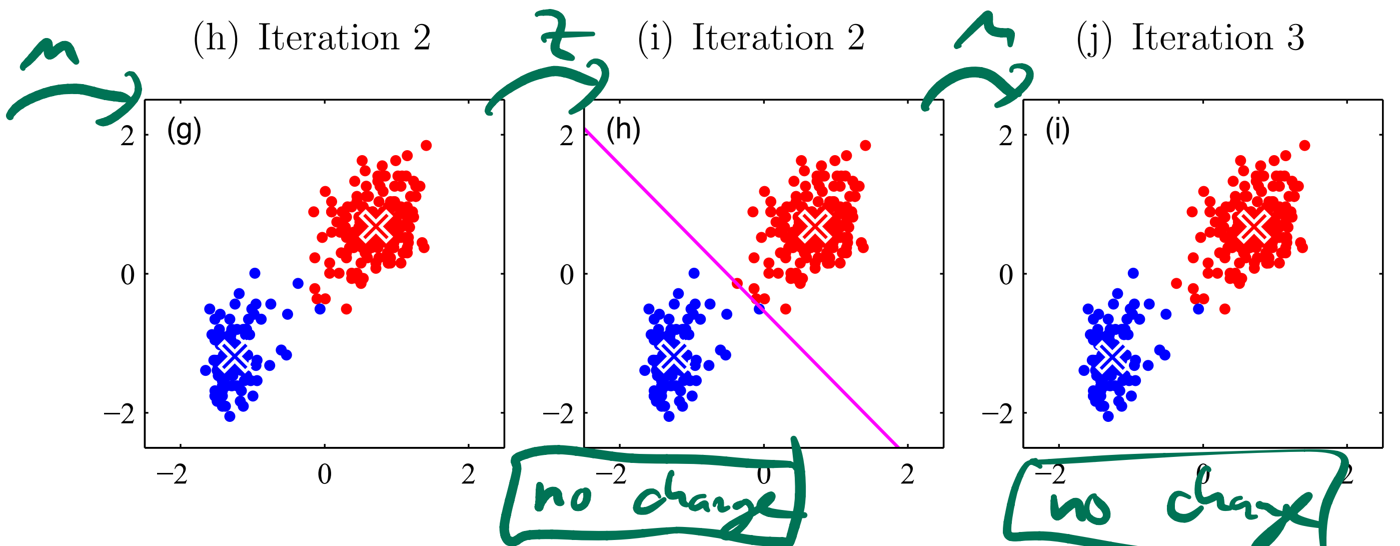
K-means for the “old-faithful” dataset (Bishop’s Figure 9.1)



(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

$\mu_k \in \mathbb{R}^3$

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

$x_n = \begin{bmatrix} R \\ G \\ B \end{bmatrix}$

200

500

100 K datapoints

vector quantization

Probabilistic model for K-means

Likelihood of x given params μ, z

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

$p(X | \mu, z)$

entire dataset

assigned for x_n

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

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

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

K-means as a Matrix Factorization

Recall the objective

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

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

$\min_{M,Z} f(M \cdot Z^\top)$

$\underset{D\times K}{M} = \begin{pmatrix} | & & | \\ m_1 & \cdots & m_K \\ | & & | \end{pmatrix}$

$\underset{N\cdot K}{Z} = \begin{pmatrix} z_{11} & - & z_1 & - & z_{1K} \\ - & & z_N & - & z_{NK} \end{pmatrix}$

Issues with K-means

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