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.

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

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

Z_{nk}: $\begin{cases} 1 & n \text{ is assigned to cluster } k \\ 0 & \text{o.w.} \end{cases}$

→ assignment

K-means clustering

Assume K is known.

distance of x_n to μ_k with assignment $= k$

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

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

→ discrete constraint

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

Is this optimization problem easy?

NP-Hard

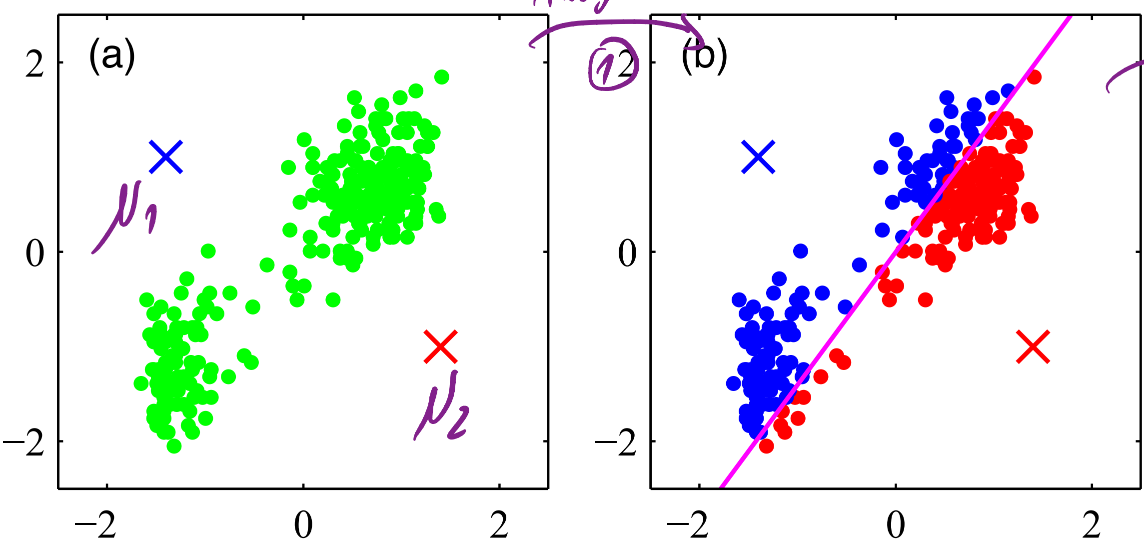
Algorithm: Initialize $\boldsymbol{\mu}_k \, \forall k$,
then iterate:

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

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

Assignment

K = 2



1 :

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

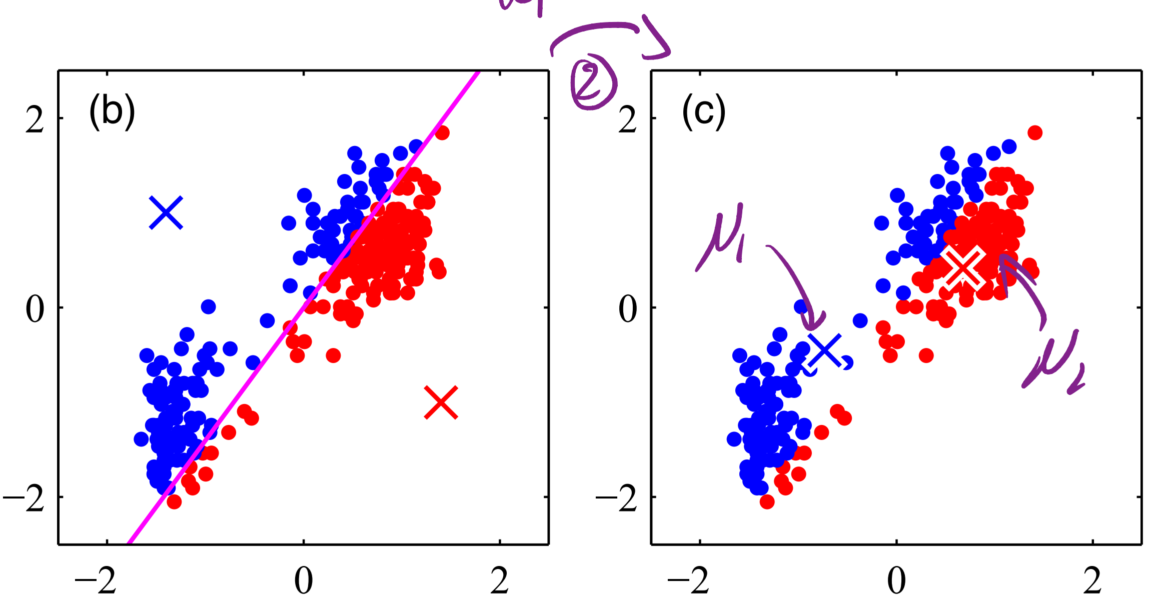
2

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

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

Hence, the name 'K-means'.

update centres



Summary of K-means

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

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

update assignment z

$$z_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}_n - \boldsymbol{\mu}_j\|_2^2 \\ 0 & \text{otherwise} \end{cases} \quad O(N.K.D)$$

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

update centres μ_k

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

$O(N.K.D)$

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

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

Coordinate descent

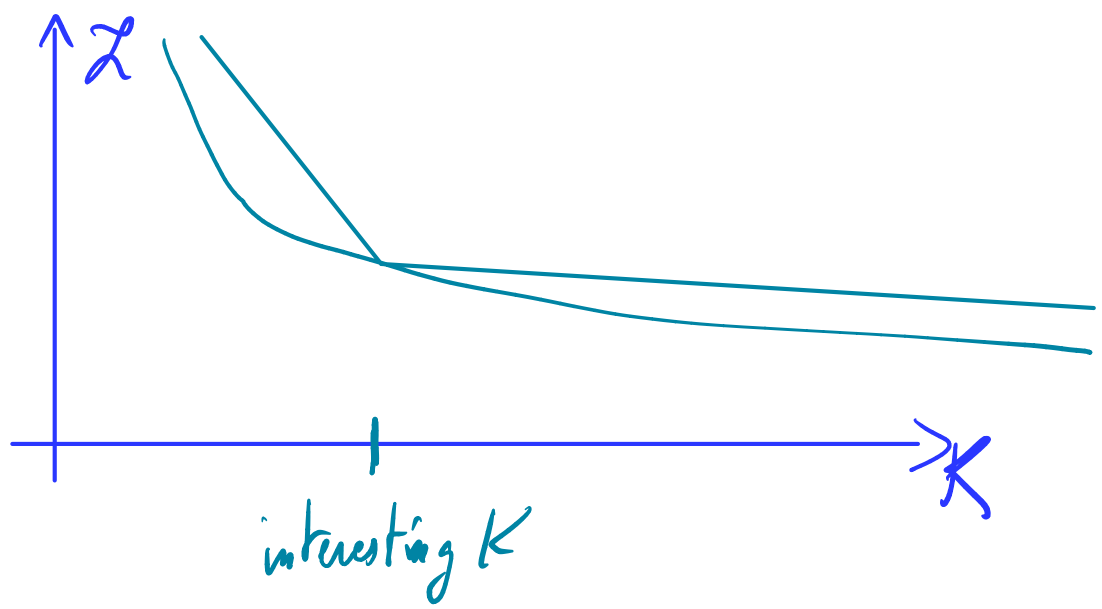
K-means is a coordinate descent algorithm, where, to find $\min_{z,\mu} \mathcal{L}(\mathbf{z}, \boldsymbol{\mu})$, we start with some $\boldsymbol{\mu}^{(0)}$ and repeat the following:

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

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

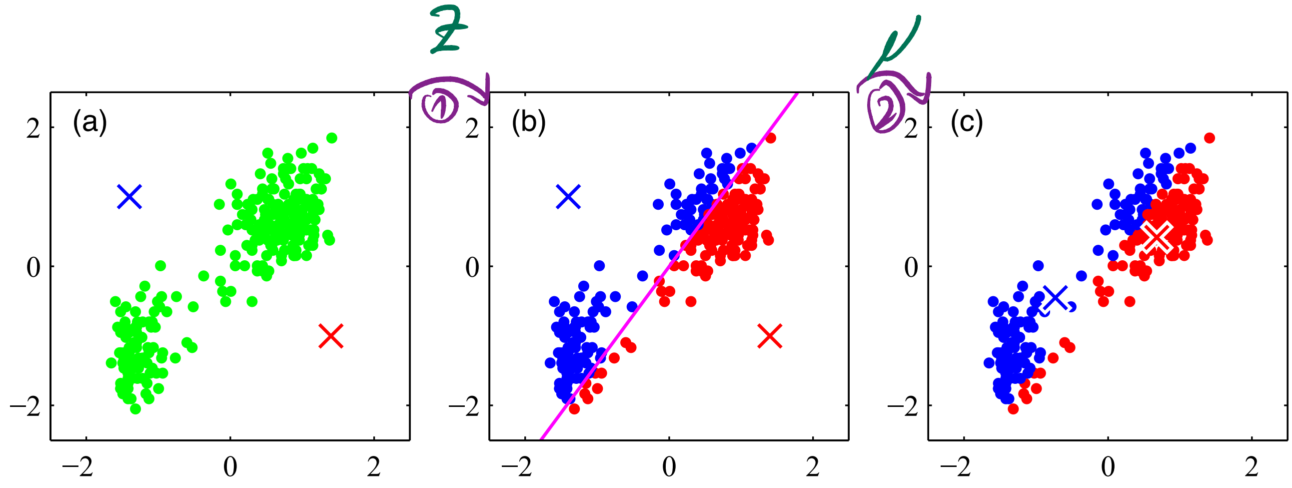
$\Leftarrow$ μ-update of K-means

How to set K?

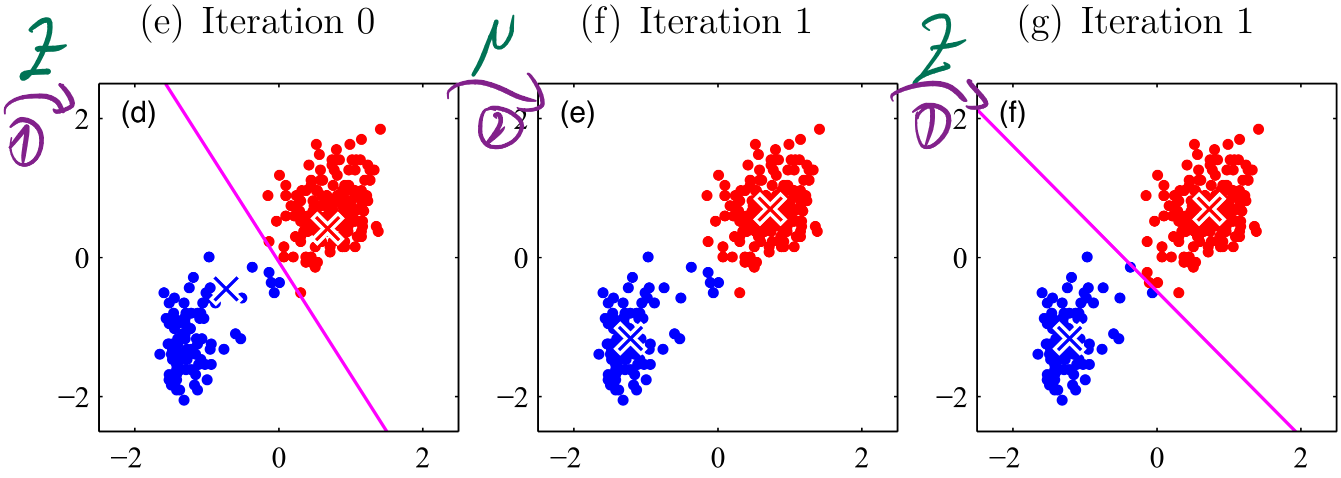


Examples

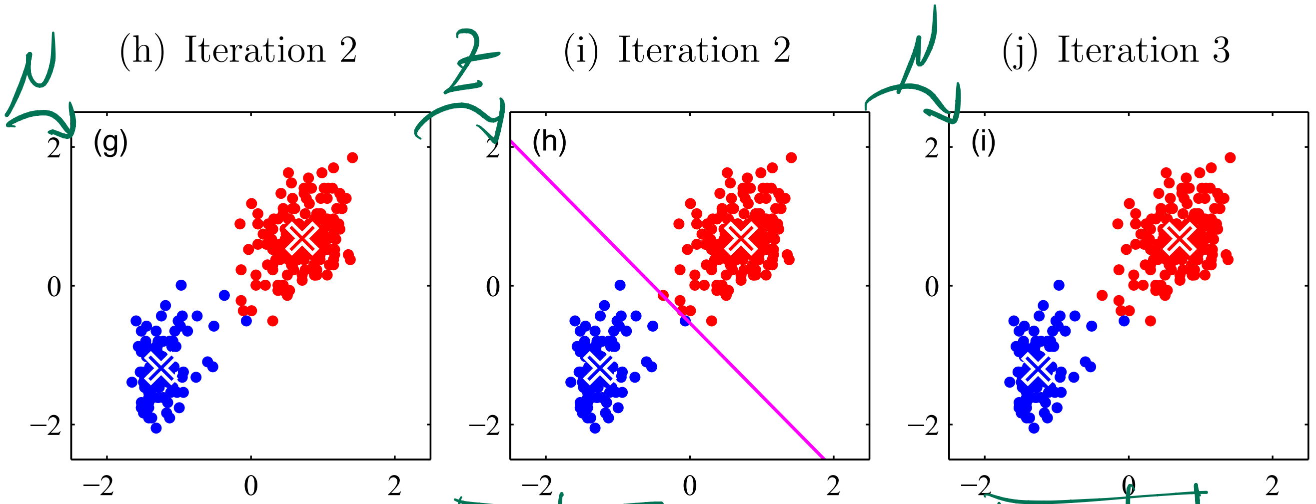
K-means for the "old-faithful" dataset (Bishop's Figure 9.1)



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



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



no change

no chge

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

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

X_n: $\begin{pmatrix}R\\G\\B\end{pmatrix} \to \mu_k \in \mathbb{R}^3$

200

500

100 k data points

Vector quantization

Probabilistic model for K-means

likelihood of X given param. μ, Z

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

spherical

assigned to k for x_n

$$P(X|\mu, Z)$$

all dataset

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

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

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

K-means as a Matrix Factorization

Recall the objective

$\min_{\mathbf{z},\boldsymbol{\mu}} f(M \cdot Z^T)$

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

$M = \begin{pmatrix} \mu_1 & \cdots & \mu_k \end{pmatrix}_{D \times K}$

$Z = \begin{pmatrix} z_1 \\ \vdots \\ z_N \end{pmatrix}_{N \times K}$

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

Issues with K-means

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

Gaussian Mixture Models solve these.