annotated version

Machine Learning Course - CS-433

K-Means Clustering

Nov 23, 2022

Martin Jaggi
Last updated on: November 20, 2022

credits to Mohammad Emtiyaz Khan & Rüdiger Urbanke

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 $\mathbb{R}^K$

The goal is to find "prototype" points $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2,, \dots, \boldsymbol{\mu}_K$ and cluster assignments $z_n \in \{1, 2, \dots, K\}$ for all $n = 1, 2, \dots, N$ data vectors $\mathbf{x}_n \in \mathbb{R}^D$.

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

K-means clustering

Assume K is known.

distance of x_n to μ_k

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

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

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

Is this optimization problem easy?

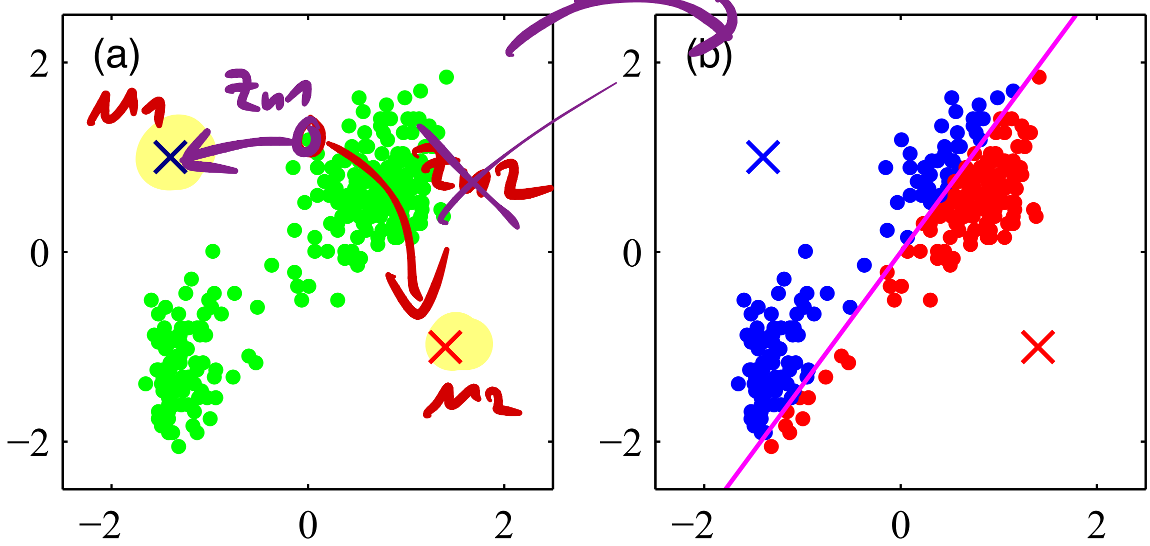
NP-hard

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 representative μ

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

assignment step 1



assignment

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

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

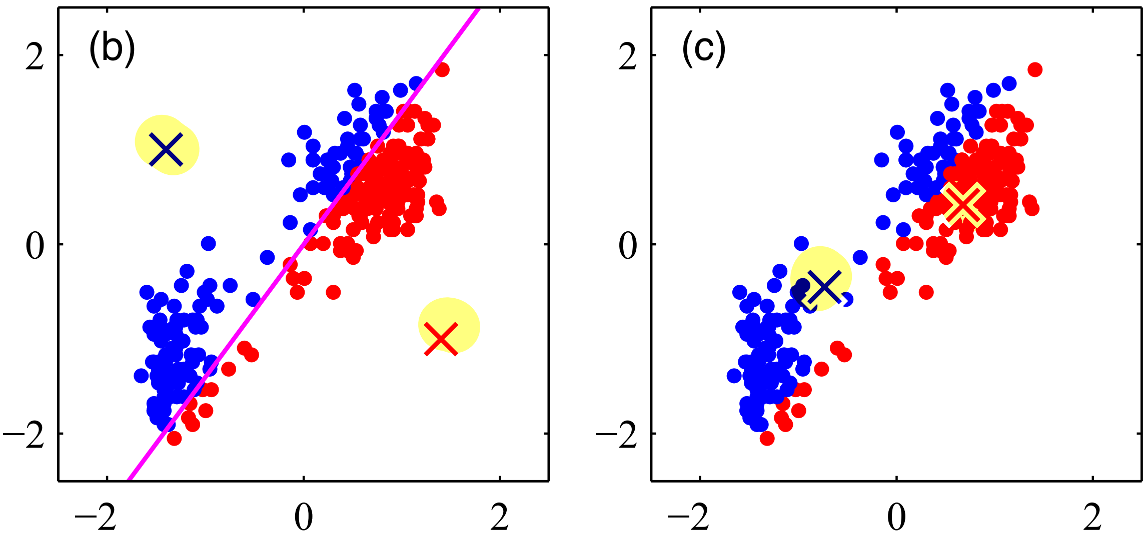
$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} z_{nk}\mathbf{x}_n}{\sum_{n=1}^{N} z_{nk}}$$

points assigned to group k

\# points . . .

Hence, the name 'K-means'.

② cluster center update



Summary of K-means

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

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

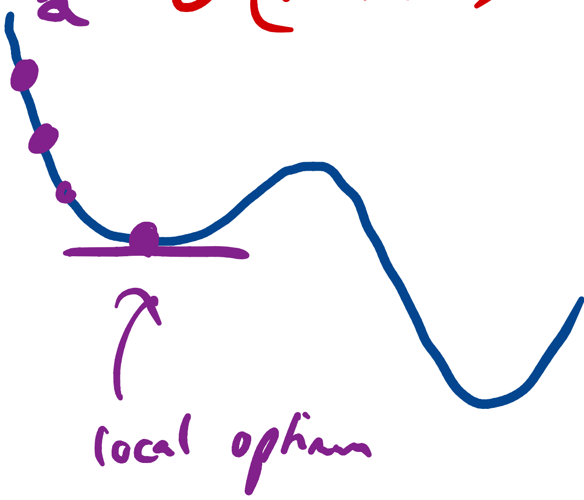
$O(N \cdot K \cdot D)$

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

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

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

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



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

$\mathcal{L}(\mu, z)$

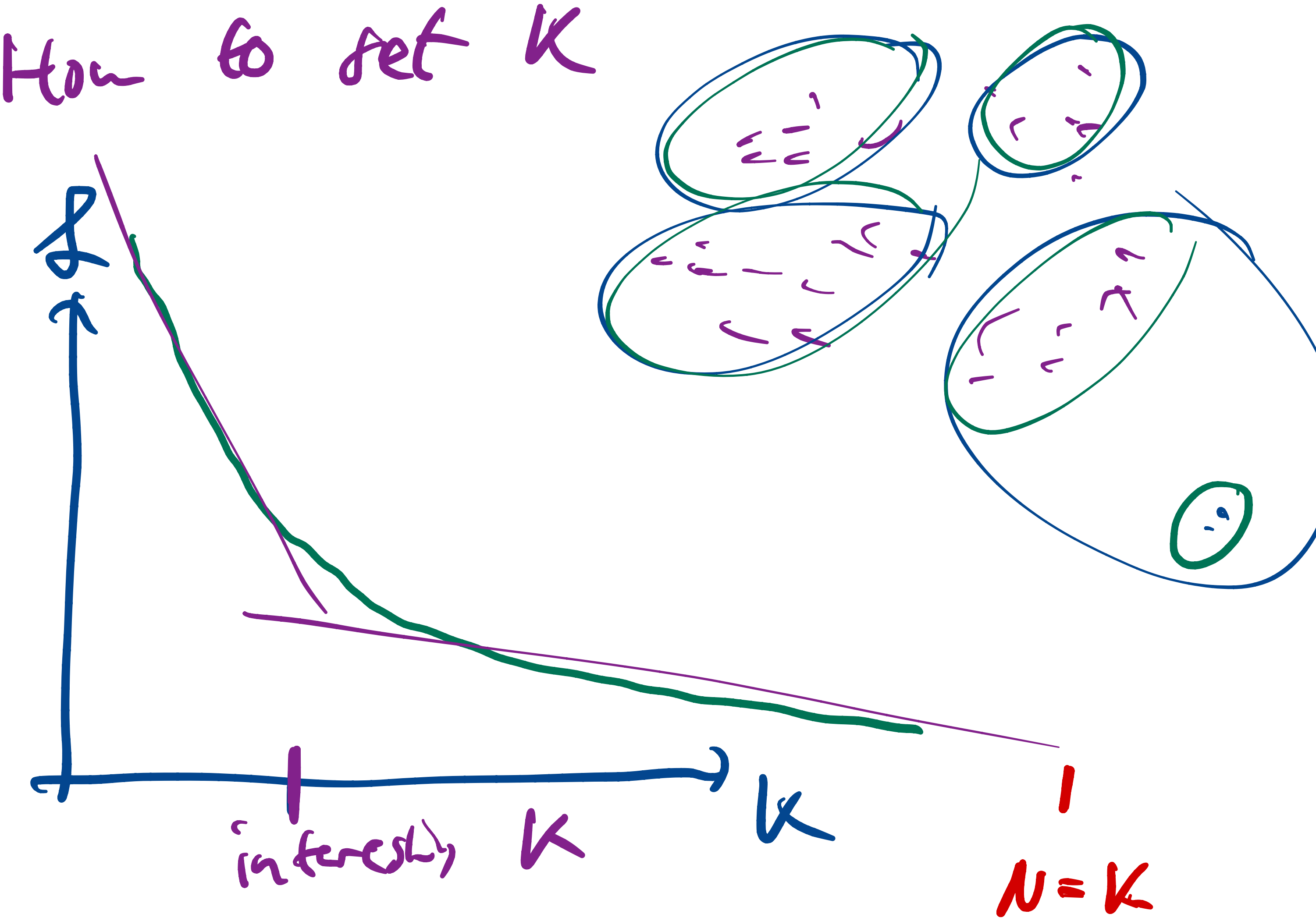
Coordinate descent

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:

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

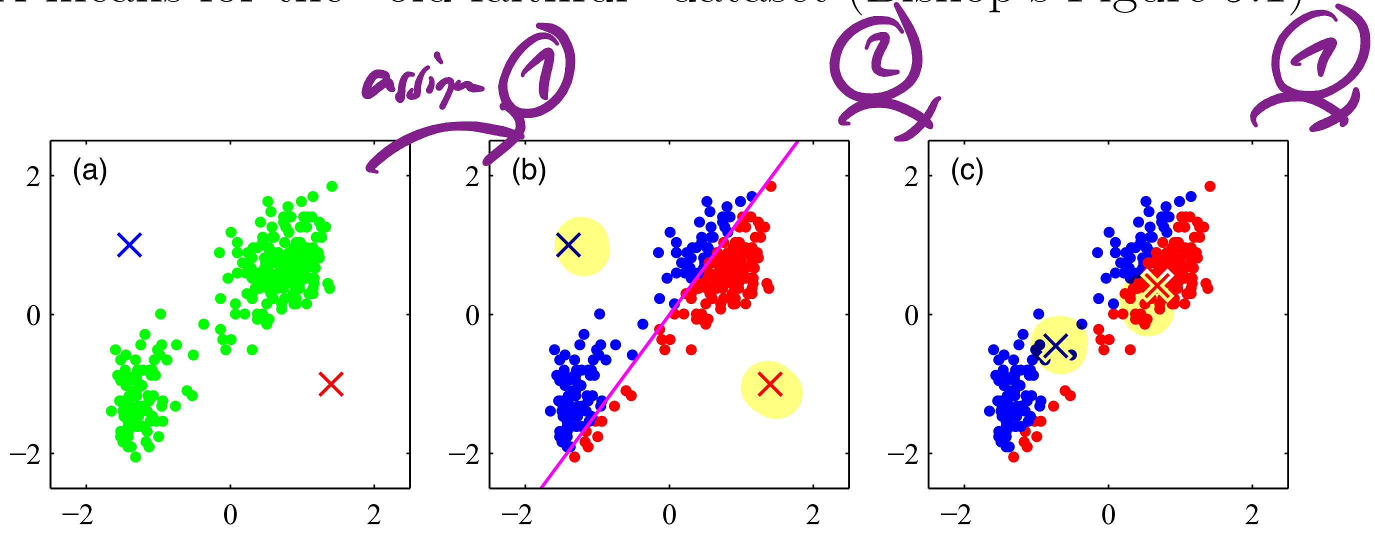
$$\boldsymbol{\mu}^{(t+1)} := \arg\min_{\mu} \mathcal{L}(\mathbf{z}^{(t+1)}, \boldsymbol{\mu})$$ equivalent to ②

How to set K

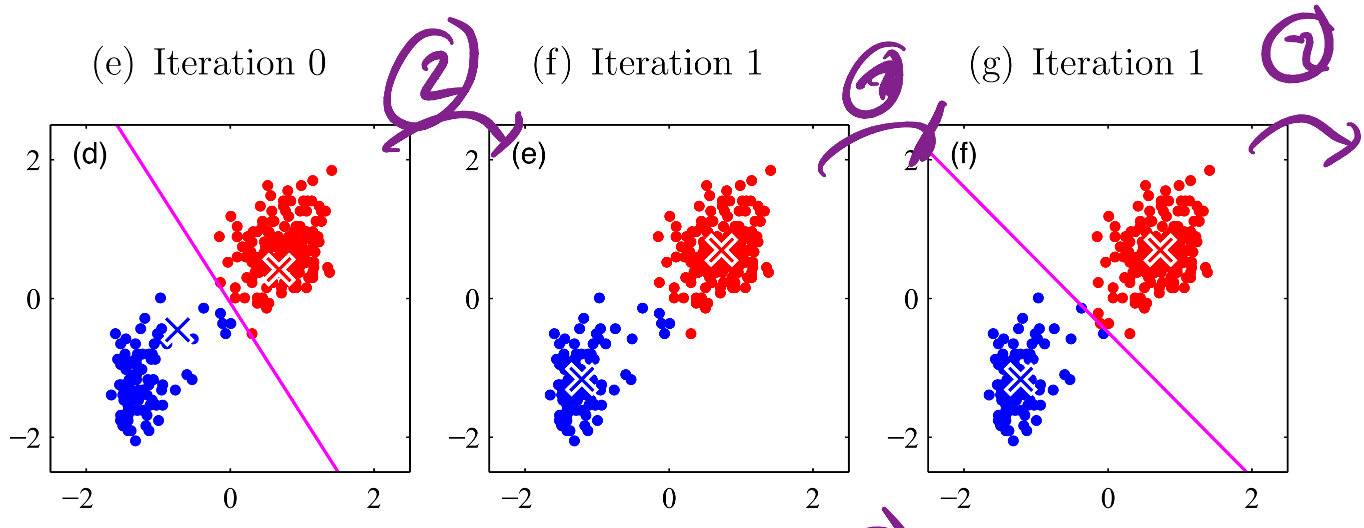


Examples

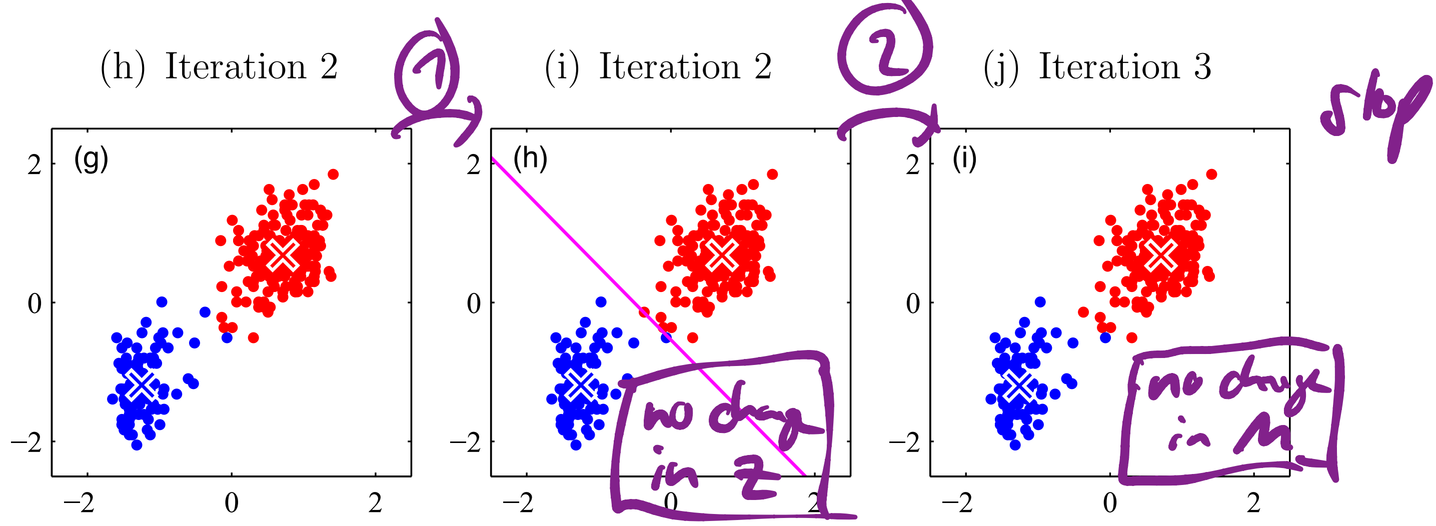
K-means for the “old-faithful” dataset (Bishop’s Figure 9.1)



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



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



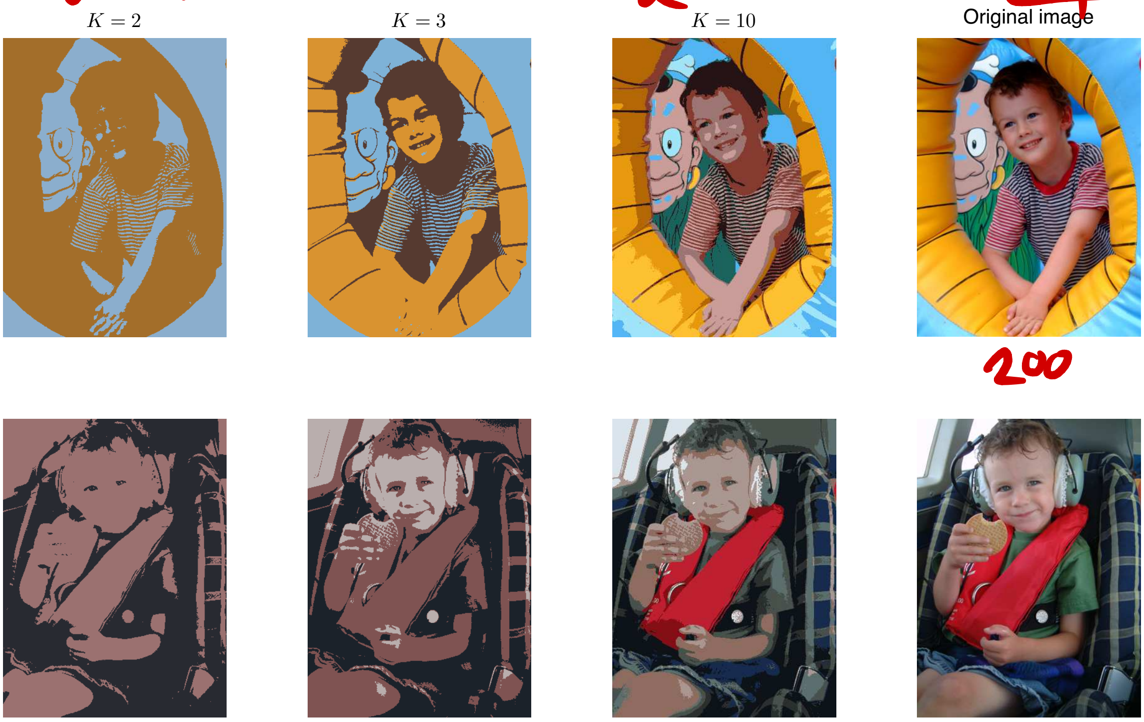
(k) Iteration 3 (l) Iteration 4 (m) Iteration 4

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

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

$x_n = \begin{bmatrix} r \\ g \\ b \end{bmatrix} \in \mathbb{R}^3$

$K=2$ $K=10$



500

200

100k datapoints

„vector quantization"

Probabilistic model for K-means

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

$\uparrow$ group of x_n

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

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

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

$$= -\sum_{n} \sum_{k} \frac{1}{2} \|x_n - \mu_k\|^2 \cdot z_{nk} + c'$$

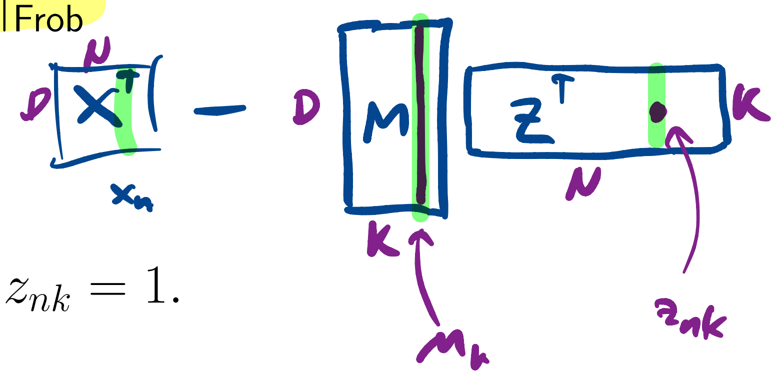
$$= \mathcal{L}(\mu, z)$$

K-means as a Matrix Factorization

Recall the objective

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

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



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).