

# K-means Clustering

Mohammad Emtiyaz Khan  
EPFL

Nov 3, 2015



ÉCOLE POLYTECHNIQUE  
FÉDÉRALE DE LAUSANNE

©Mohammad Emtiyaz Khan 2015

# 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  $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_K$  and cluster assignments  $r_n \in \{1, 2, \dots, K\}$  for all  $n = 1, 2, \dots, N$  data vectors.

## K-means clustering

Assume  $K$  is known.

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

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

$$\text{where } \mathbf{r}_n = [r_{n1}, r_{n2}, \dots, r_{nK}]^T$$

$$\mathbf{r} = [\mathbf{r}_1^T, \mathbf{r}_2^T, \dots, \mathbf{r}_N^T]^T$$

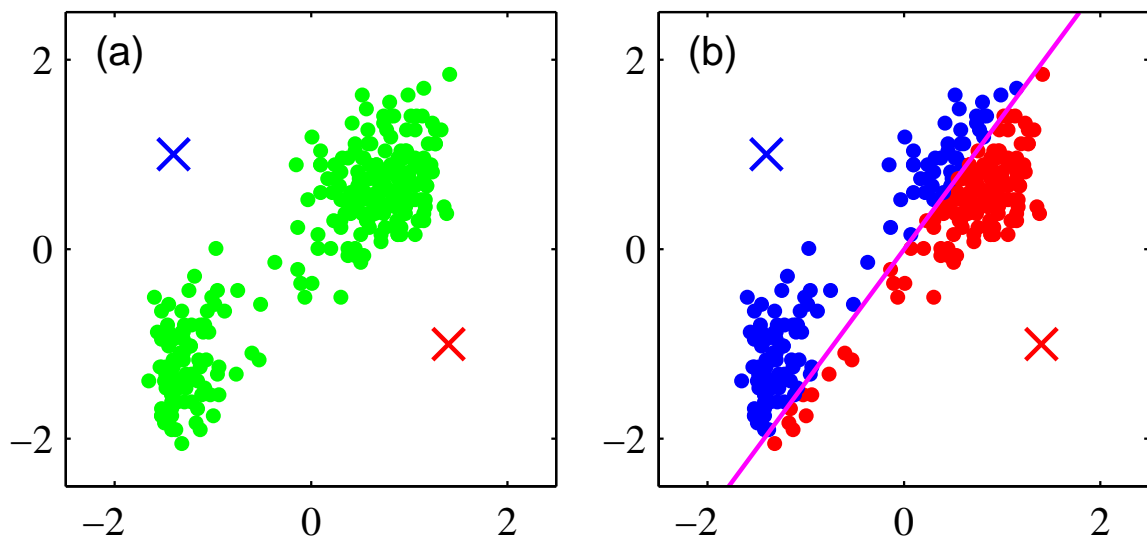
$$\boldsymbol{\mu} = [\boldsymbol{\mu}_1^T, \boldsymbol{\mu}_2^T, \dots, \boldsymbol{\mu}_K^T]^T$$

Is this optimization problem easy?

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

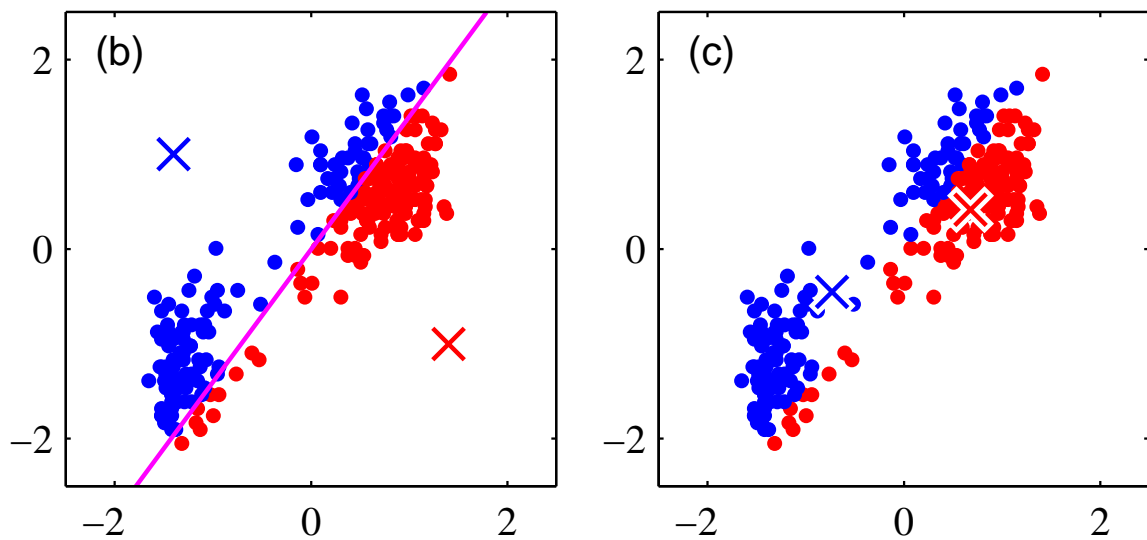
1. For all  $n$ , compute  $\mathbf{r}_n$  given  $\boldsymbol{\mu}$ .
2. For all  $k$ , compute  $\boldsymbol{\mu}_k$  given  $\mathbf{r}$ .

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



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

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



Take derivative w.r.t.  $\boldsymbol{\mu}_k$  to get:

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

Hence, the name 'K-means'.

# Summary of K-means

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

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

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

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

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

Convergence 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  $\min_{\mathbf{a}, \mathbf{b}} f(\mathbf{a}, \mathbf{b})$ , we start with some  $\mathbf{b}_0$  and repeat the following:

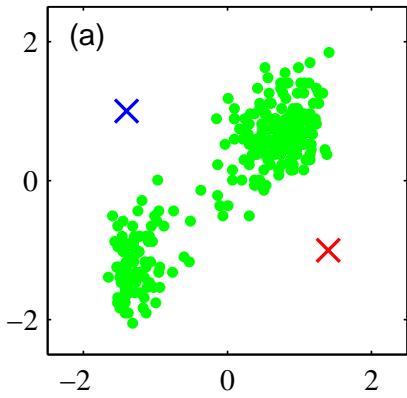
$$\mathbf{a}_{k+1} = \arg \min_b f(\mathbf{a}, \mathbf{b}_k)$$

$$\mathbf{b}_{k+1} = \arg \min_a f(\mathbf{a}_{k+1}, \mathbf{b})$$

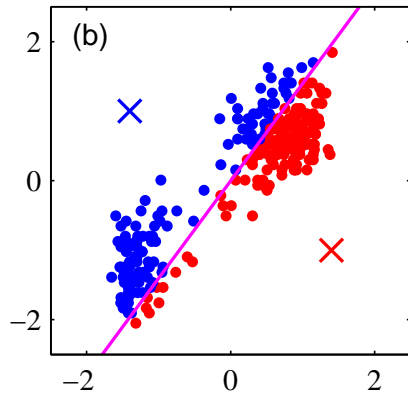
Convergence is assured when both subproblems have a unique minimum.

# Examples

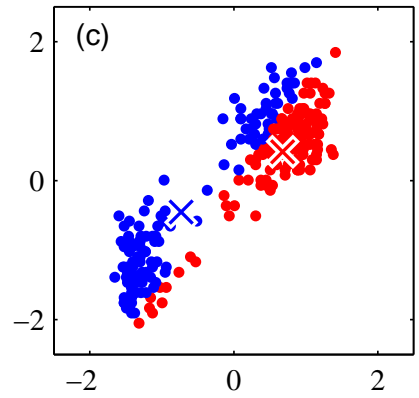
K-means for the “old-faithful” dataset (Bishop’s Figure 9.1)



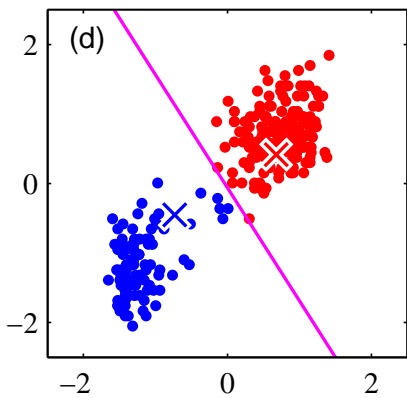
(a) Iteration 0



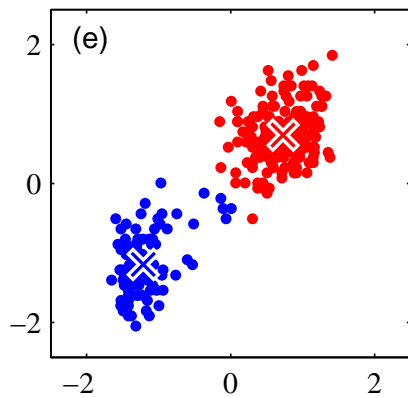
(b) Iteration 1



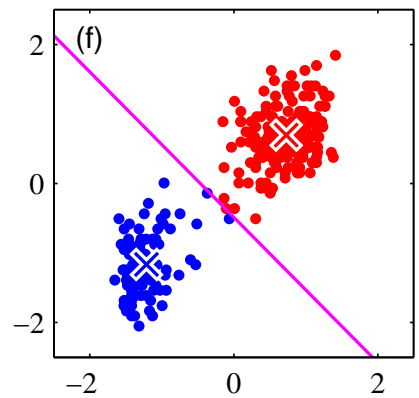
(c) Iteration 1



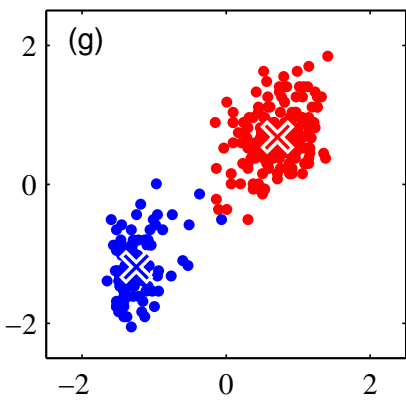
(d) Iteration 2



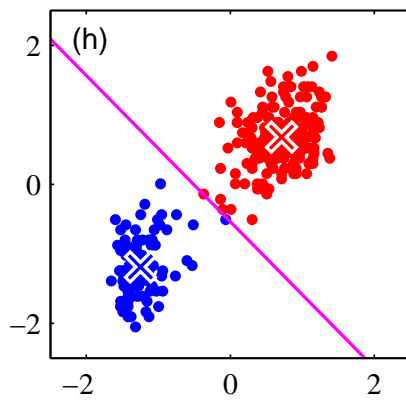
(e) Iteration 2



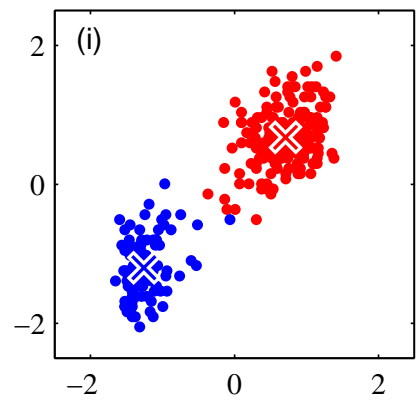
(f) Iteration 3



(g) Iteration 3

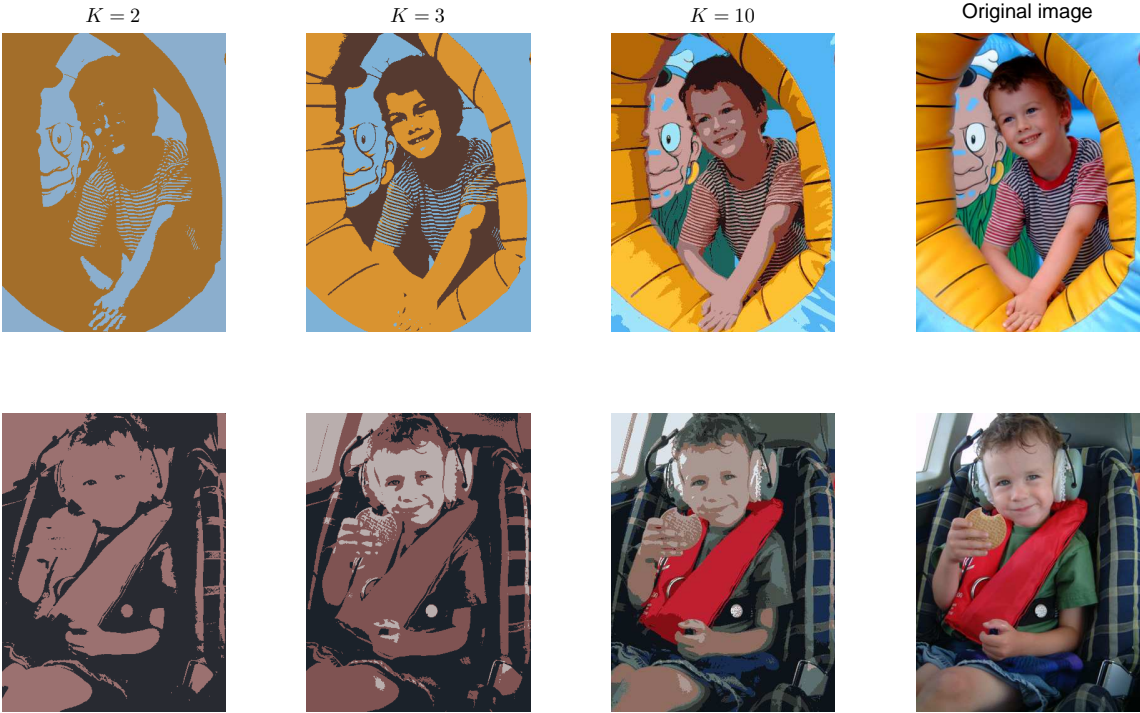


(h) Iteration 4



(i) Iteration 4

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



# Probabilistic model for K-means

## Issues with K-means

1. Computation can be heavy for large  $N$  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).

## To do

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. Derive the probabilistic model associated with the cost function.
3. What is the computational complexity of K-means?