K-means Clustering

Mohammad Emtiyaz Khan
EPFL

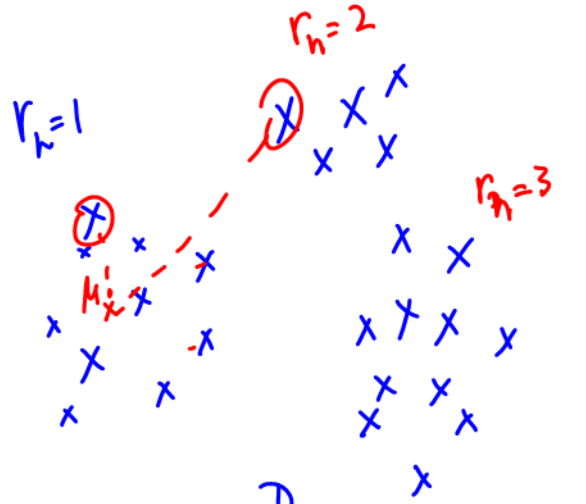
Nov 3, 2015

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

$\mu_1, \mu_2, \mu_3 \in \mathbb{R}^D$, $r_n \in \{1,2,3\}$

$r_n=1$ $\quad$ $r_n=2$ $\quad$ $r_n=3$

$\underline{r}_n = \begin{bmatrix}1\\0\\0\end{bmatrix} \quad \underline{r}_n = \begin{bmatrix}0\\1\\0\end{bmatrix} \quad \underline{r}_n = \begin{bmatrix}0\\0\\1\end{bmatrix}$

$$\underbrace{\sum_{n: r_n=1} \|\underline{x}_n - \underline{\mu}_1\|^2}_{} + \underbrace{\sum_{n: r_n=2} \|\underline{x}_n - \underline{\mu}_2\|^2}_{} + \sum_{n: r_n=3} \|\underline{x}_n - \underline{\mu}_3\|^2$$

K-means clustering

Assume K is known.

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

$\|\underline{x}_n - \underline{\mu}_k\|_2^2$ is L_2 norm

$= (\underline{x}_n - \underline{\mu}_k)' (\underline{x}_n - \underline{\mu}_k)$

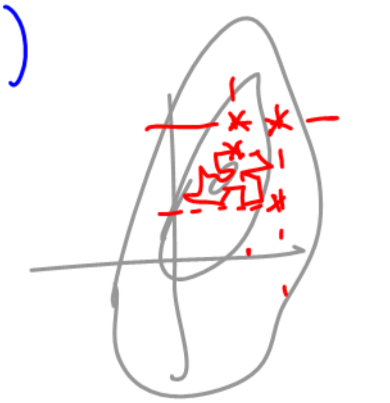
$$\text{s.t.} \boldsymbol{\mu}_k \in \mathbb{R}^{N}, r_{nk} \in \{0, 1\}, \sum_{k=1}^{K} r_{nk} = 1,$$

$$\text{where } \mathbf{r}_n = [r_{n1}, r_{n2}, \ldots, r_{nK}]^T$$
$$\mathbf{r} = [\mathbf{r}_1^T, \mathbf{r}_2^T, \ldots, \mathbf{r}_N^T]^T$$
$$\boldsymbol{\mu} = [\boldsymbol{\mu}_1^T, \boldsymbol{\mu}_2^T, \ldots, \boldsymbol{\mu}_K^T]^T$$

Is this optimization problem easy?

$\mathcal{L}(\underline{r}, \underline{\mu})$

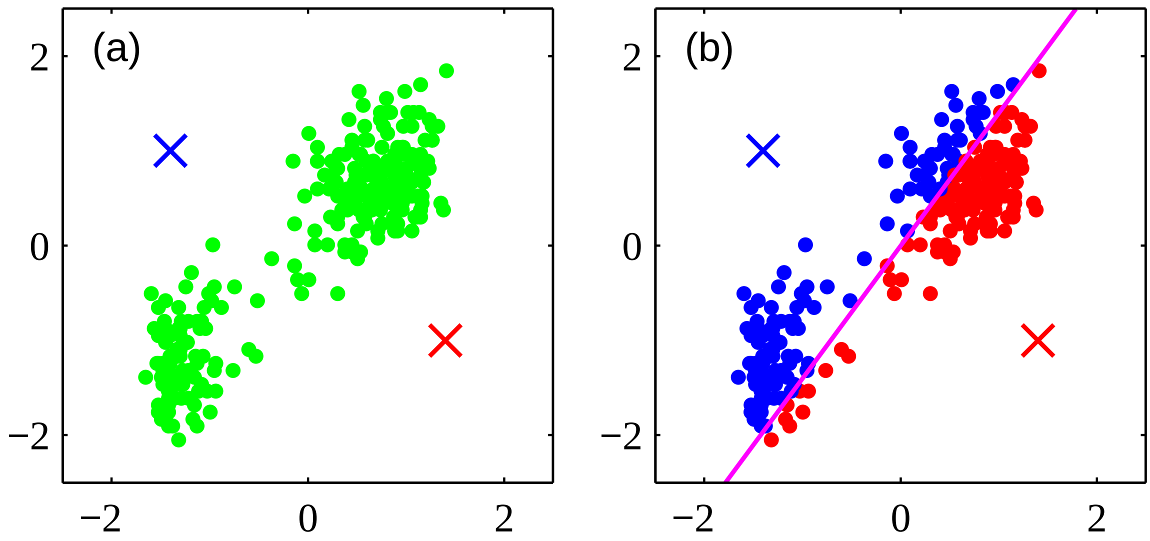
$\uparrow$ fix $\underline{r}$



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

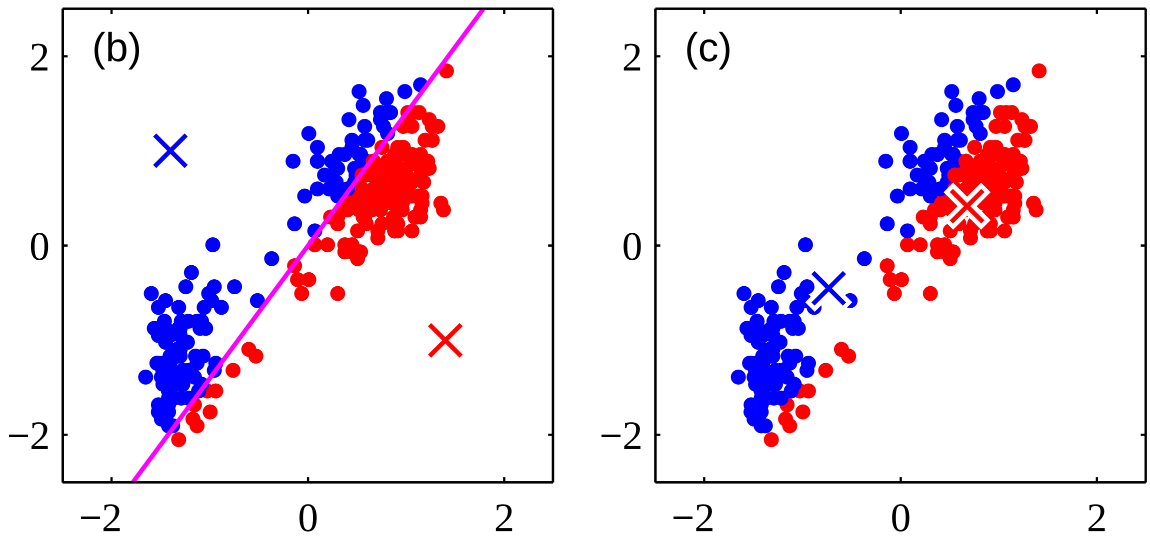
1. For all n, compute $\mathbf{r}_n$ given $\boldsymbol{\mu}$.
2. For all k, compute $\boldsymbol{\mu}_k$ given $\mathbf{r}$.

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: $\mathcal{L}(\mu_1) = \sum_{n=1}^{N} r_{n1}(\underline{x}_n - \underline{\mu}_1)^T(\underline{x}_n - \underline{\mu}_1)$

$+ \text{cnst}$

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

$$\frac{\partial \mathcal{L}}{\partial \underline{\mu}_1} = -\sum_{n=1}^{N} r_{n1}(\underline{x}_n - \underline{\mu}_1) = 0$$

$$\Rightarrow -\sum_n r_{n1}\underline{x}_n + \sum_n r_{n1}\underline{\mu}_1 = 0$$

Hence, the name 'K-means'.

What is K-median?

Summary of K-means

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

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

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

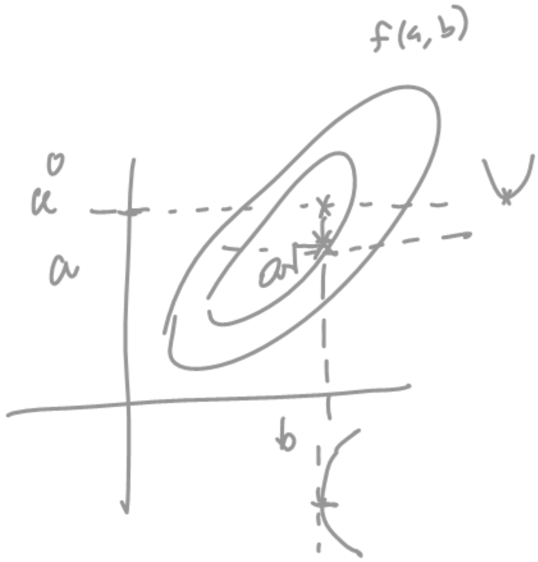
2. For all k, compute $\boldsymbol{\mu}_k$ given $\mathbf{r}$.

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

Example

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

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

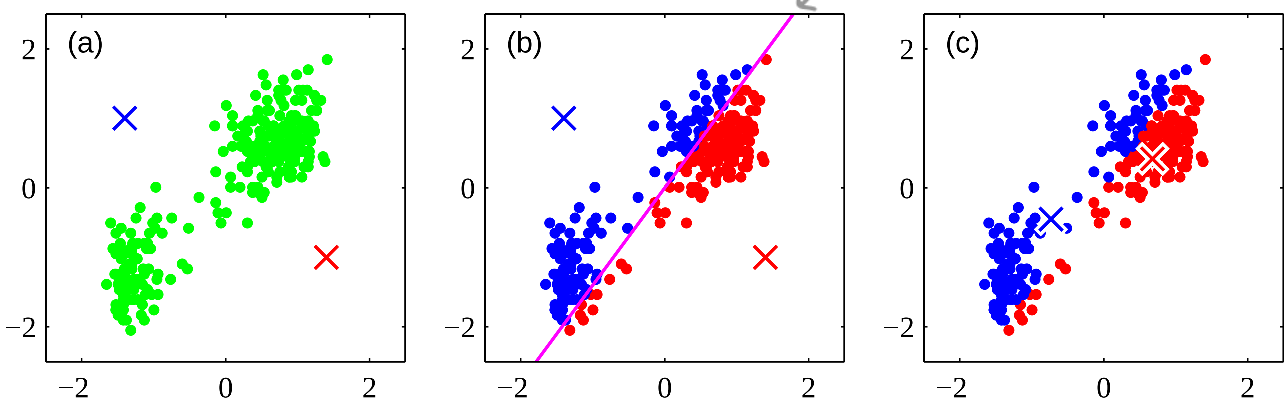
$$\mathbf{b}_{k+1} = \arg\min_{a} f(\mathbf{a}_{k+1}, \mathbf{b})$$

Convergence is assured when both subproblems have a unique minimum.

Examples

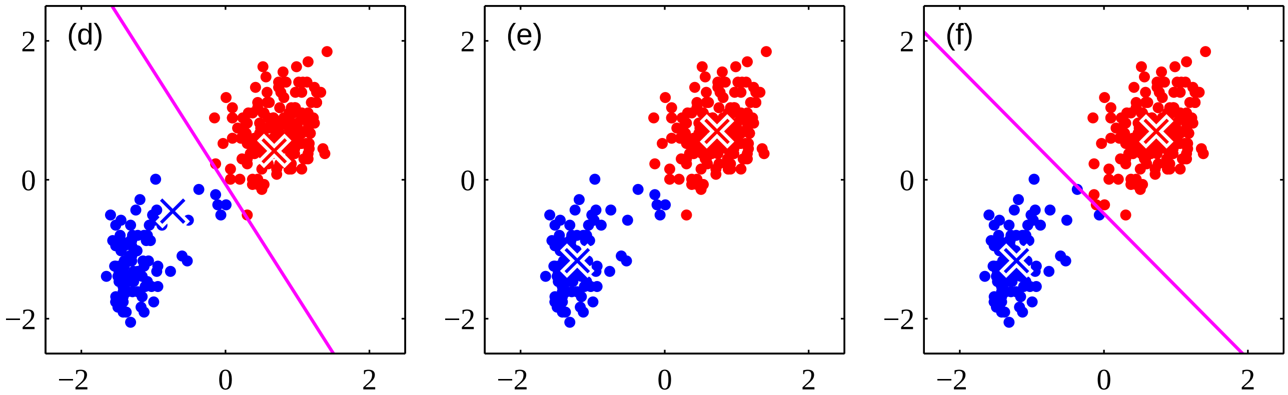
K-means for the "old-faithful" dataset (Bishop's Figure 9.1)

Initialize μ_1, μ_2 | r_n given μ | μ given r_n

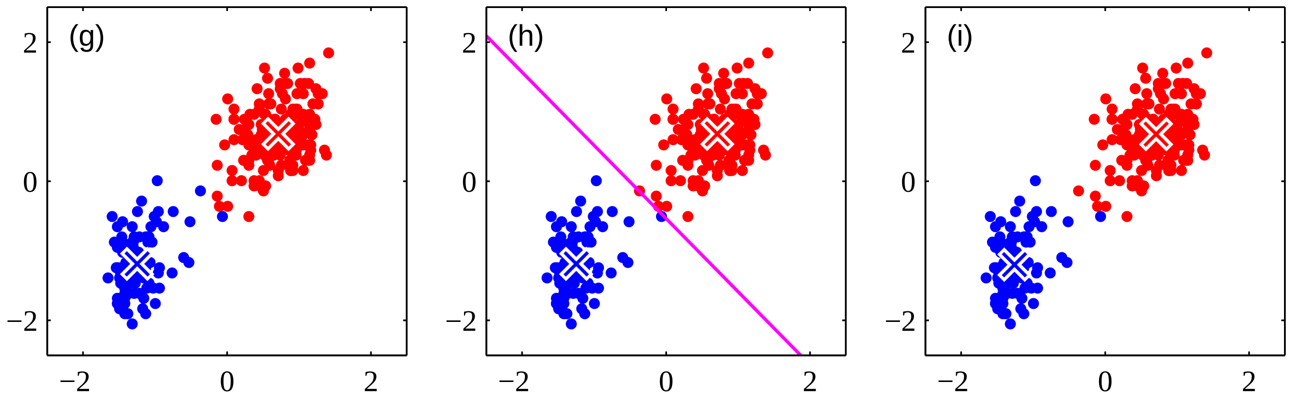


(e) Iteration 0 (f) Iteration 1 (g) Iteration 1

r_n given μ

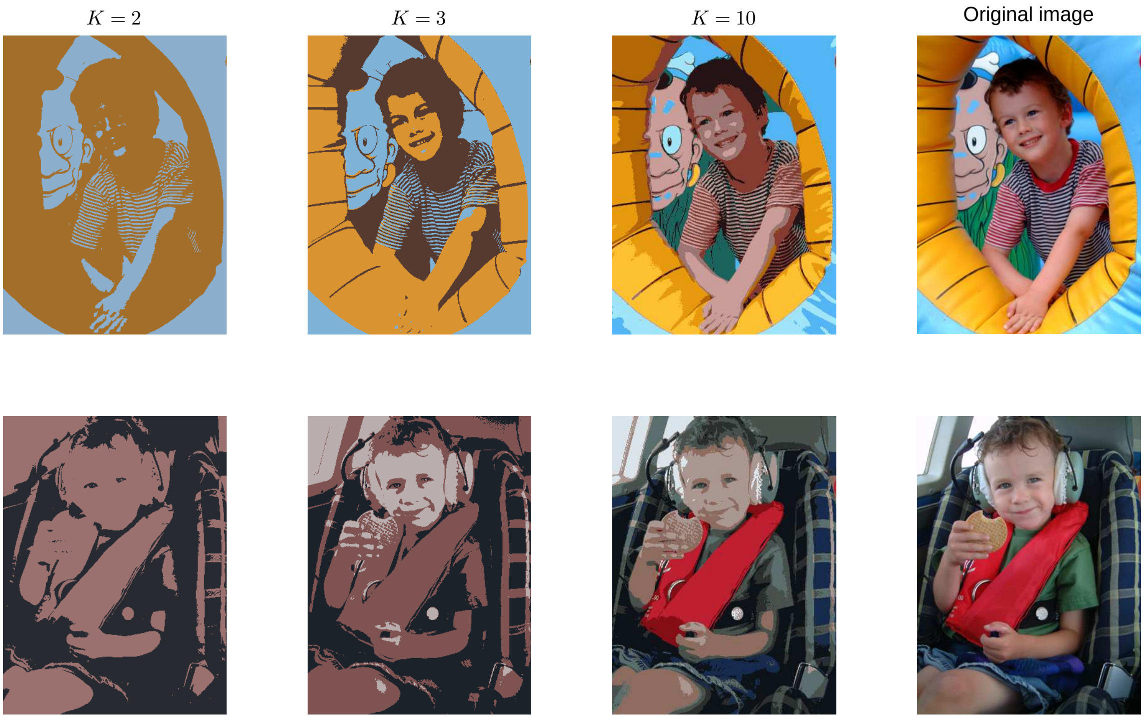


(h) Iteration 2 (i) Iteration 2 (j) Iteration 3



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

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



$$-\frac{1}{2}\log p(\underline{r}, \underline{\mu}) = \mathcal{L}(\underline{r}, \underline{\mu}) + \text{cnst}$$

Probabilistic model for K-means

$$\mathcal{L}(\underline{r}, \underline{\mu}) = \sum_{k=1}^{K} \sum_{n=1}^{N} r_{nk} \underbrace{(\underline{x}_n - \underline{\mu}_k)^2}$$

$$p(\underline{r}, \underline{\mu}) \propto \exp -\sum_k \sum_n r_{nk}(\underline{x}_n - \underline{\mu}_k)^2$$

$$p(\underline{r}, \underline{\mu}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \left[\mathcal{N}(\underline{x}_n / \underline{\mu}_k, I) \right]^{r_{nk}}$$

$$= \prod_n \prod_k \exp -\frac{1}{2} r_{nk} (\underline{x}_n - \underline{\mu}_k)^2$$

$$= \prod_n \prod_k \left[\exp -\frac{1}{2} (\underline{x}_n - \underline{\mu}_k)^2 \right]^{r_{nk}}$$

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).

Q: What is the computational complexity?
(Do it as homework)

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?