K-means Clustering

COEN140

Santa Clara University

Example

- Consider 5 data points

Example	A	B	C	D	E
Attribute Value (X)	0.1	0.6	0.8	2.0	3.0

- Goal: group the 5 points into 2 clusters

- Each cluster has a "center"

Example

- Consider 5 data points

Example	A	B	C	D	E
Attribute Value (X)	0.1	0.6	0.8	2.0	3.0

- Initialize the cluster centers:
  - Select A as center 1: $m_1 = 0.1$
  - Select B as center 2: $m_2 = 0.6$

- Assign a point to cluster-k if it is closer to m_k, $k = 1, 2$

Example

- Consider 5 data points

Example	A	B	C	D	E
Attribute Value (X)	0.1	0.6	0.8	2.0	3.0

- Initialize the cluster centers:
  - Select A as center 1: $m_1 = 0.1$
  - Select B as center 2: $m_2 = 0.6$

- Initial clustering results:
  - Cluster 1: A
  - Cluster 2: B, C, D, E

Example

- Consider 5 data points

Example	A	B	C	D	E
Attribute Value (X)	0.1	0.6	0.8	2.0	3.0

- Initial clustering results:
  - Cluster 1: A
  - Cluster 2: B, C, D, E

- Update the cluster centers:
  - $m_1 = 0.1$
  - $m_2 = 1.6$

Example

- Consider 5 data points

Example	A	B	C	D	E
Attribute Value (X)	0.1	0.6	0.8	2.0	3.0

- Updated cluster centers:
  - $m_1 = 0.1$
  - $m_2 = 1.6$

- Update clustering results:
  - Cluster 1: A, B, C
  - Cluster 2: D, E

Example

- Consider 5 data points

Example	A	B	C	D	E
Attribute Value (X)	0.1	0.6	0.8	2.0	3.0

- Updated clustering results:
  - Cluster 1: A, B, C
  - Cluster 2: D, E

- Update cluster centers:
  - $m_1 = 0.5$
  - $m_2 = 2.5$

Example

- Consider 5 data points

Example	A	B	C	D	E
Attribute Value (X)	0.1	0.6	0.8	2.0	3.0

- Updated cluster centers:
  - $m_1 = 0.5$
  - $m_2 = 2.5$

- Update clustering results:
  - Cluster 1: A, B, C
  - Cluster 2: D, E

Example

- Consider 5 data points

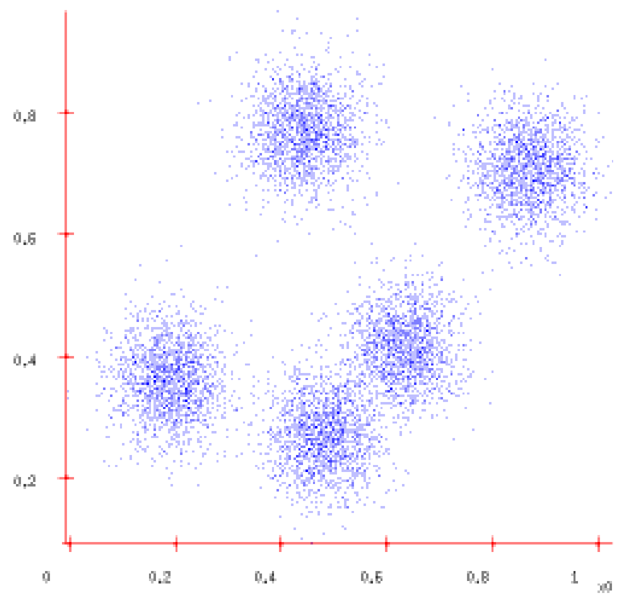
Example	A	B	C	D	E
Attribute Value (X)	0.1	0.6	0.8	2.0	3.0

- Since the clustering results are the same as in the previous iteration, we can stop the algorithm.

- What we have seen: K-means Clustering
  - The number of clusters K=2 in this example

Clustering

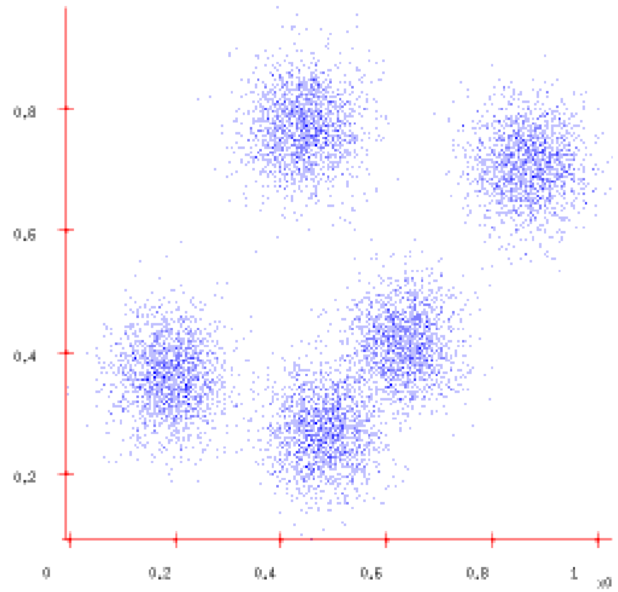
- Consider the problem of grouping N data points into K clusters



- Motivation: data compression

Clustering

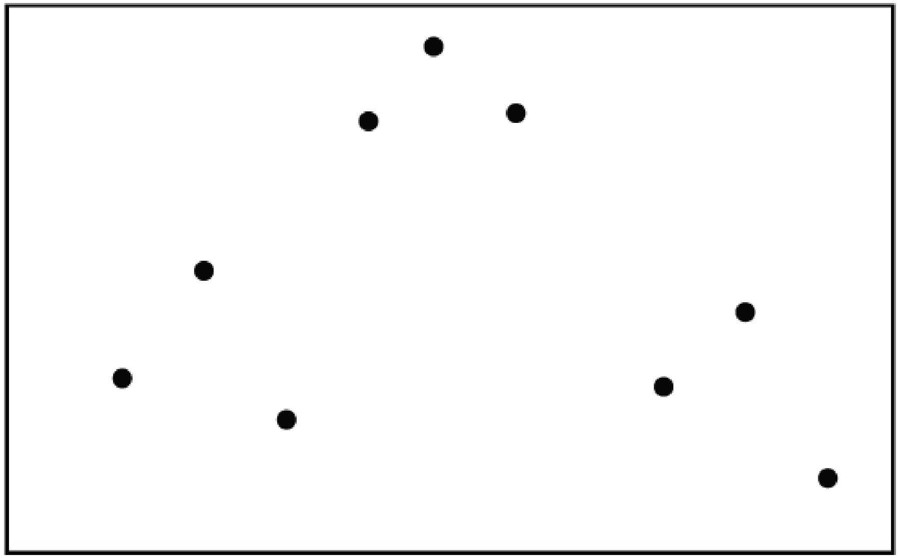
- Consider the problem of grouping N data points into K clusters



- Assume: data was generated from a number of different classes. The aim is to cluster data from the same class together.

Clustering

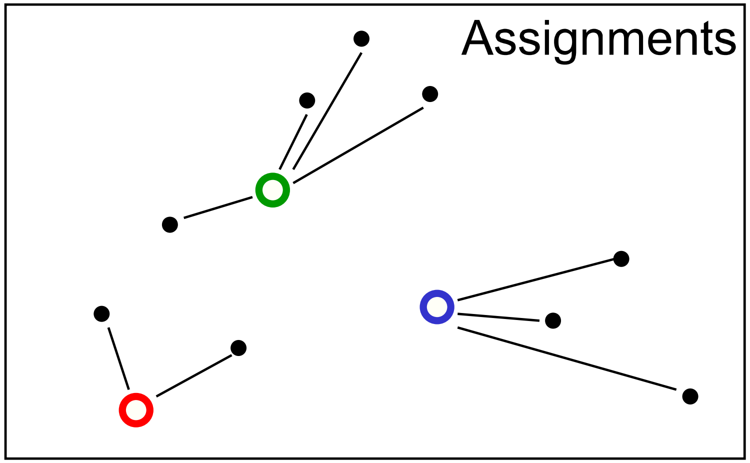
- N data points: $\mathbf{x}_n \in \mathbb{R}^d, n = 1, 2, \dots, N$
- They belong to K classes



- How to identify those classes?
- How to identify the data points that belong to each class?

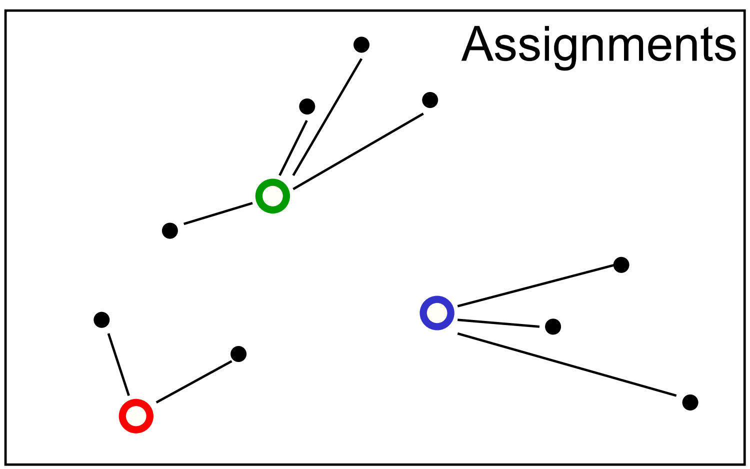
K-means Clustering

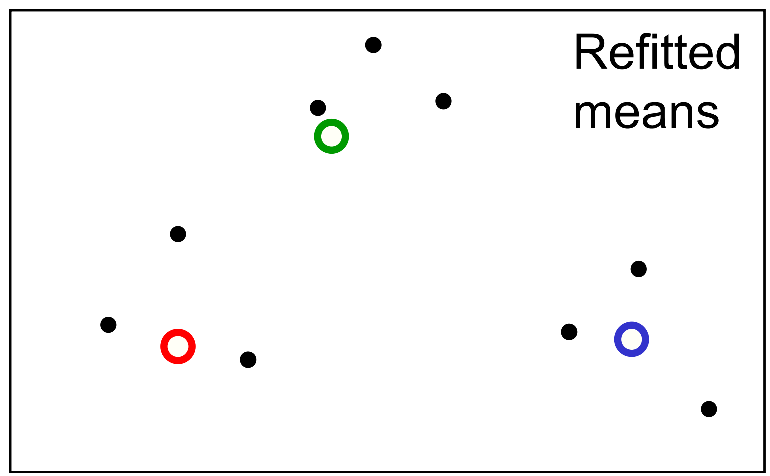
- Initialization: randomly initialize cluster centers
- Then, the algorithm iteratively alternates between two steps:
  - Step 1: Assignment - Assign each data point to the closest cluster



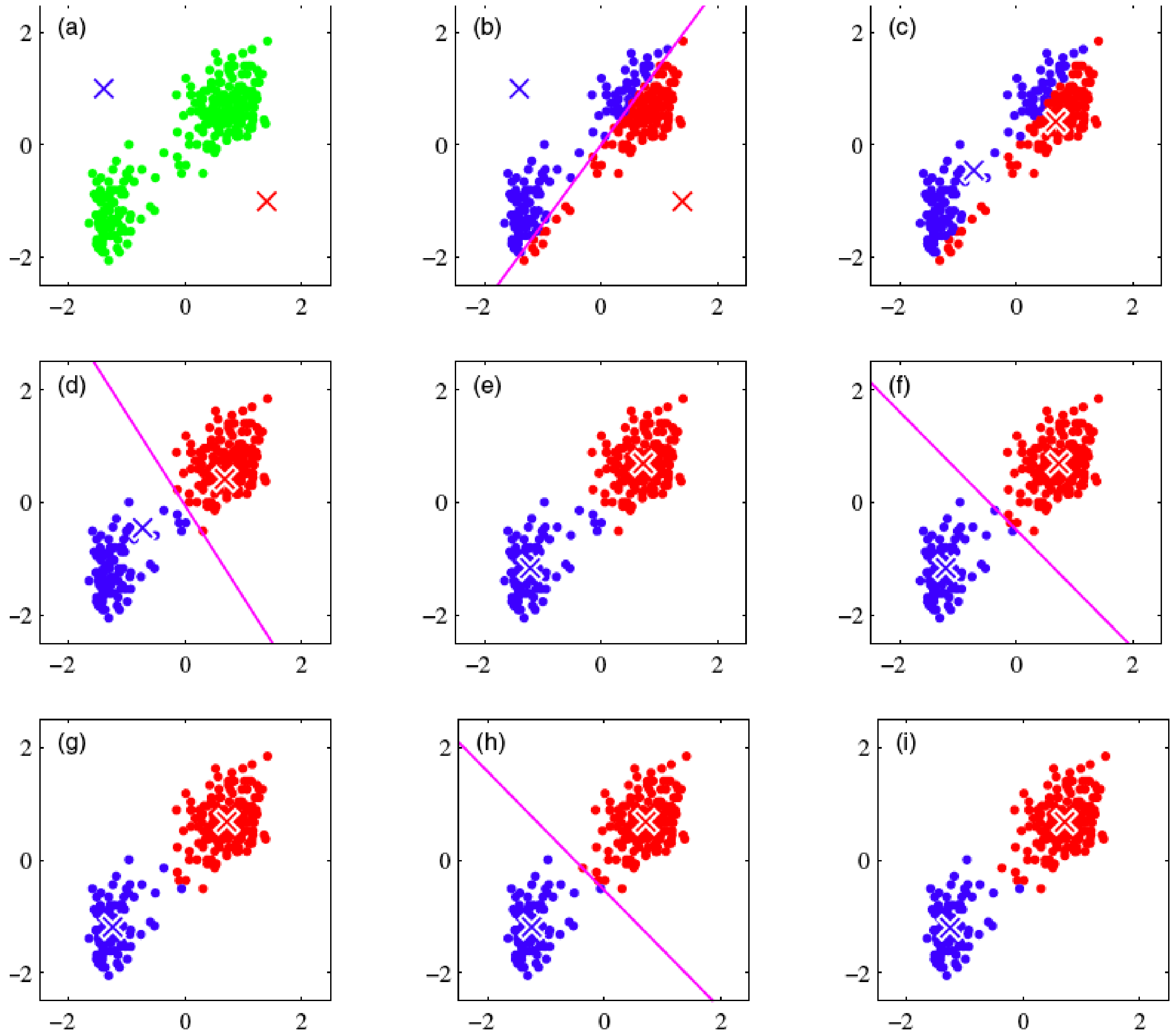
K-means Clustering

- Initialization: randomly initialize cluster centers
- Then, the algorithm iteratively alternates between two steps:
  - Step 1: Assignment - Assign each data point to the closest cluster
  - Step 2: Cluster-center Update - Move each cluster center to the center of the data points assigned to it





K-means Clustering



Initialization

- Set K cluster means $\mathbf{m}_k, k = 1, \ldots, K$ to random values

Repeat the Following Two Steps

- Step 1: Assignment
  - Each data point $\mathbf{x}_n$ assigned to the nearest mean/center

$$\hat{k}_n = \arg\min_k \|\mathbf{m}_k - \mathbf{x}_n\|_2^2$$

Responsibilities: $r_{kn} = 1 \Leftrightarrow \hat{k}_n = k$

$r_{kn} = 1$, if $\mathbf{x}_n$ is assigned to cluster k

$r_{kn} = 0$, otherwise

Repeat the Following Two Steps

- Step 2: Cluster-center Update
  - Update $\mathbf{m}_k$ by

$$\mathbf{m}_k = \frac{\sum_{n=1}^{N} r_{kn}\mathbf{x}_n}{\sum_{n=1}^{N} r_{kn}}$$

$$k = 1, 2, \ldots, K$$

Objective Function

- The sum of the squared distances of data points $\{\mathbf{x}_n\}$ to their assigned cluster centers

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{kn} \|\mathbf{m}_k - \mathbf{x}_n\|_2^2$$

where $r_{kn} = 1$ if $\mathbf{x}_n$ is assigned to cluster k, and $r_{kn} = 0$ otherwise.

- Note: $\sum_{k=1}^{K} r_{kn} = 1$

Objective Function

- The objective function value J decreases each time we execute Step 1 and Step 2.

- Also, J is non-negative.

- Therefore, the J value will converge.

Stopping Criterion

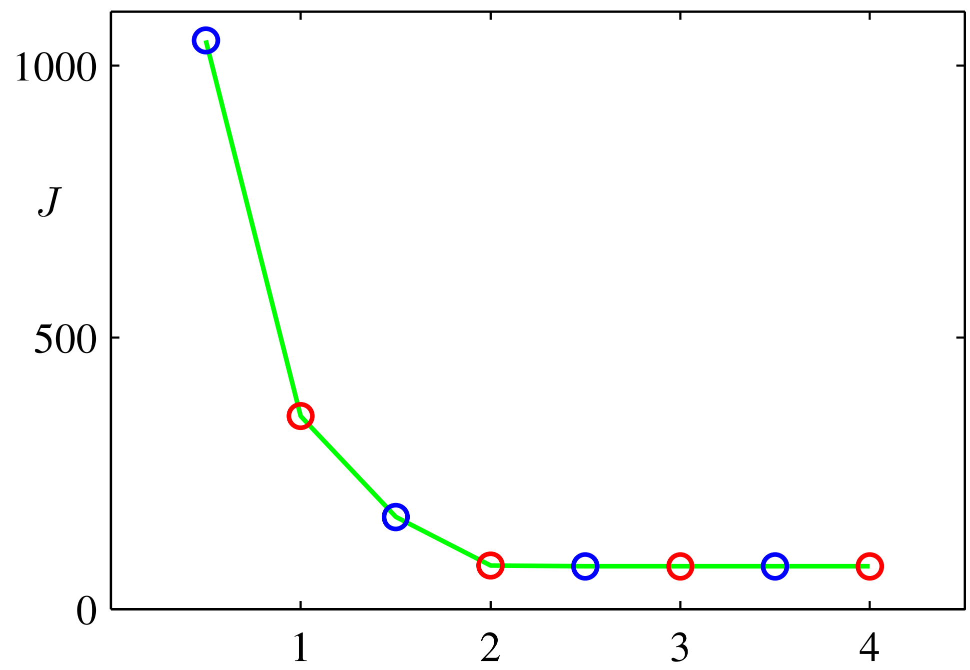
- When the decrease of the objective function J between two successive iterations is below a certain value ε, where ε is a small positive number.

$$J(Iter - 1) - J(Iter) < \varepsilon$$

- Or when the number of iterations reaches a predefined value

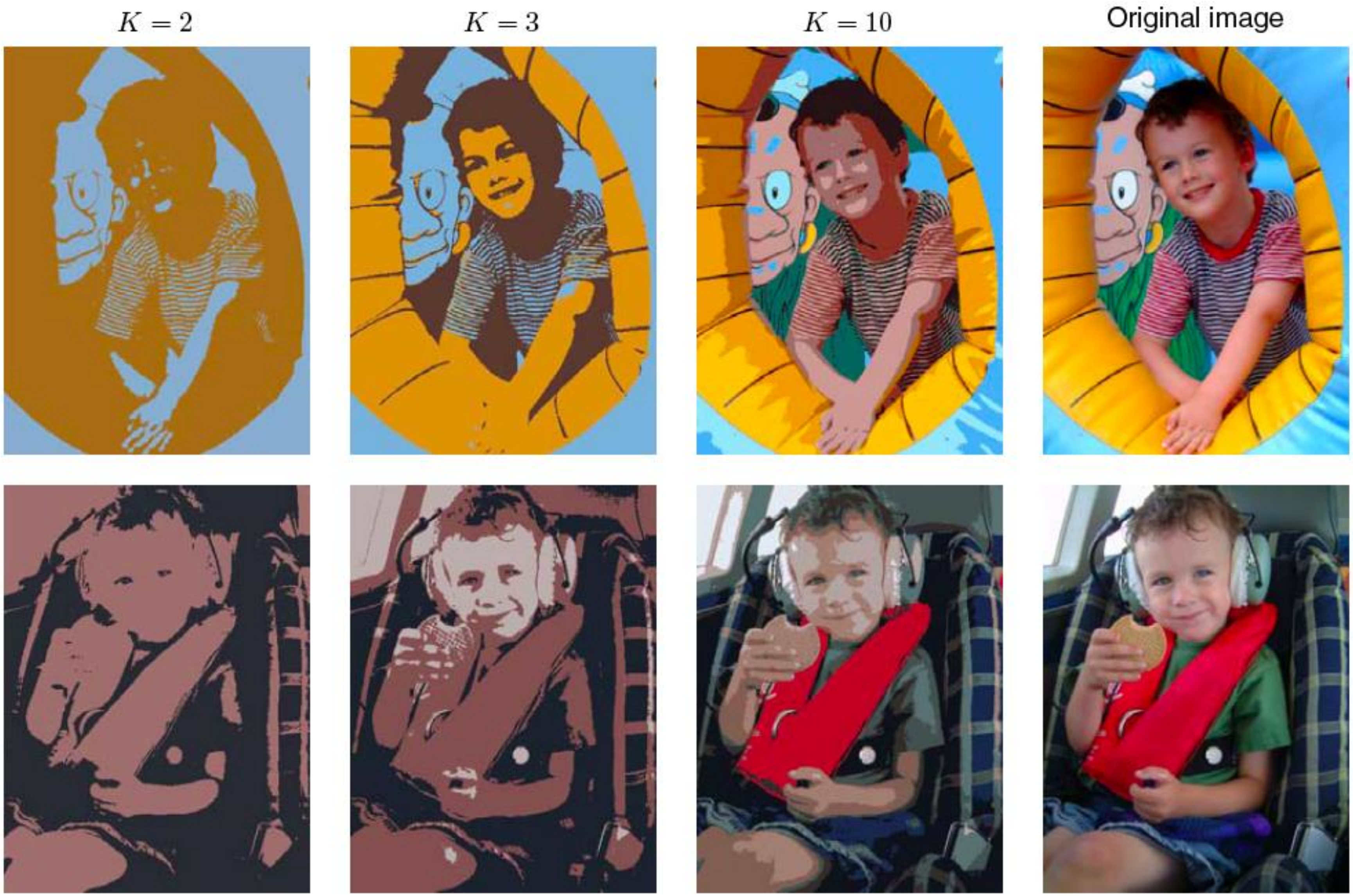
$$Iter \;==\; maxIter$$

Convergence of K-means Clustering



- Plot of the objective function value J after each Assignment Step (blue circles) and Cluster-center Update Step (red circles)

K-means Clustering for Image Compression



K-means Clustering for Image Compression

- Assume a color image is 512x512 = 2^{18} pixels
- Each pixel: 3 channels (3 Bytes)
  - Red (8 bits): value is from 0 to 255
  - Green (8 bits)
  - Blue (8 bits)
- One image: 2^{18} pixels$\times$ 3 Bytes=786432 Bytes

- K=10 clustering
  - Cluster centers: 3Bytes x (K=10) = 30 Bytes
  - Assignment: $\log_2 10 \approx (4\ bits = 0.5\ Bytes)\ per\ pixel$
  - 2^{18} pixels$\times$ 0.5 Bytes=131072 Bytes