K-means Clustering

J.-S. Roger Jang (張智星)
CSIE Dept., National Taiwan Univ., Taiwan
http://mirlab.org/jang
jang@mirlab.org

Problem Definition

Input:

- $X = \{x_1, x_2, \ldots, x_n\}$: A data set in d-dim. space
- m: Number of clusters (we avoid using k here to avoid confusion with other summation indices…)

Output:

- m cluster centers: $c_j, 1 \le j \le m$
- Assignment of each x_i to one of the m clusters:

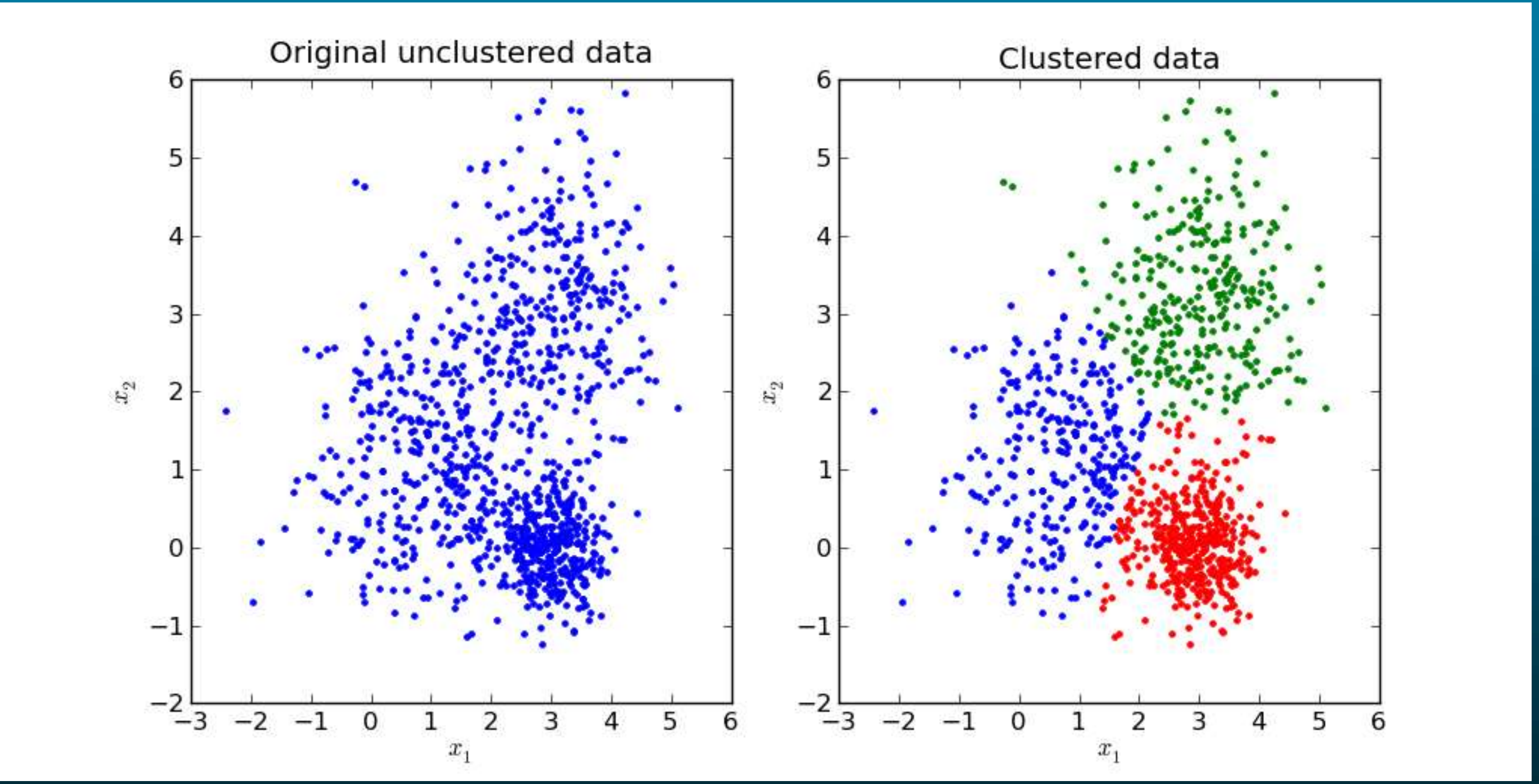
$$a_{ij} \in \{0,1\}, 1 \le i \le n, 1 \le j \le m$$

$$\sum_{j=1}^{m} a_{ij} = 1, \forall i$$

Requirement:

- The output should minimize the objective function…

Goal of K-means Clustering



$X = [x_1 \cdots x_{100}]$ 2, $C = [c_1\ c_2\ c_3]$ 2×3

Objective Function

又叫 C.X有误差

Objective function (aka. distortion)

A 100×3 共306参数

Quiz!

$$e_j = \sum_{x_i \in G_j} \left\| x_i - c_j \right\|^2$$

距

每个group和中心距离平方和 好微分

目標函數

$$J(X;C,A) = \sum_{j=1}^{m} e_j = \sum_{j=1}^{m} \sum_{x_i \in G_j} \left\| x_i - c_j \right\|^2 = \sum_{j=1}^{m} \sum_{i=1}^{n} a_{ij} \left\| x_i - c_j \right\|^2, where$$

只有 $x_i \in G_j$ $a_{ij}=1$ ∴还是剩在 G_j 内的 X 其它是 0

$$X = \{x_1, x_2, \ldots, x_n\}$$

$$C = \{c_1, c_2, \ldots, c_m\}$$

$$a_{ij} = 1\ iff\ x_i \in G_j,\ with \sum_{j=1}^{m} a_{ij} = 1, \forall i$$

- **d*m (for matrix C) plus n*m (for matrix A) tunable parameters with certain constraints on matrix A**
- **Np-hard problem if exact solution is required**

Strategy for Minimization

Observation

- J(X; C, A) is parameterized by C and A
- Joint optimization is hard, but separate optimization with respect to C and A is easy

Strategy

先定C優化A. 再定A優C → iteratly

- Fix C and find the best A to minimize J(X; C, A)
- Fix A and find the best C to minimize J(X; C, A)
- Iterate the above two steps until convergence

Properties

- The approach is also known as "coordinate optimization"

Task 1: How to Find Assignment A?

Goal

- **Find A to minimize J(X; C, A) with fixed C**

Facts

argument→取q.不是取min

- **Analytic (close-form) solution exists:**

Quiz!

$$\hat{a}_{ij} = \begin{cases} 1 \ \mathit{if} \ \ j = \arg\min_{q} \left\| x_i - c_q \right\|^2 \\ 0, \ \mathit{otherwise} \end{cases}$$

- $$\hat{A} = \arg\min_{A} J(X;C,A) \Leftrightarrow J(X;C,A) \geq J(X;C,\hat{A}), \forall C$$

若是 $\sum |y-x_i|$，没平方

$x_1 \quad x_2 \quad x_3 \quad x_4 \quad x_5 \quad x_6$

$x_1\ x_6$ 的 min 都在①，$x_2\ x_5$ 在②

$x_3\ x_4$ 在②，所以在②

Task 2: How to Find Centers in C?

$x=[x_1 \cdots x_n] \Rightarrow \arg\min_y \sum_{i=1}^{n} |y-x_i|^2 = \text{mean}(x) = \frac{\sum x_i}{n}$

$\hookrightarrow J(y)=\sum (y-x_i)^2 \quad J'=\sum 2(y-x_i)=0 \quad \sum y=\sum x_i \quad ny=\sum x_i$

Goal

- **Find C to minimize J(X; C, A) with fixed A**

Facts

- **Analytic (close-form) solution exists:**

$$\hat{c}_j = \frac{\sum_{i=1}^{n} a_{ij} x_i}{\sum_{i=1}^{n} a_{ij}}$$

Quiz!

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ & \vdots & \end{bmatrix}$$

直行 $\sum$

= group 个数

- $$\hat{C} = \arg\min_C J(X;C,A) \Leftrightarrow J(X;C,A) \ge J(X;\hat{C},A), \forall A$$

Algorithm

2.3 对調

1 要先切区塊

1. **Initialize**
   - Select initial m cluster centers
2. **Find clusters**
   - For each x_i, assign the cluster with nearest center
   - ➔ Find A to minimize J(X; C, A) with ~~fixed~~ given C
3. **Find centers**
   - Recompute each cluster center as the mean of data in the cluster
   - ➔ Find C to minimize J(X; C, A) with ~~fixed~~ given A
4. **Stopping criterion**
   - Stop if clusters stay the same. Otherwise go to step 2.

Stopping Criteria

Two stopping criteria

- **Repeating until no more change in cluster assignment**
- **Repeat until distortion improvement is less than a threshold**

objective function

Facts

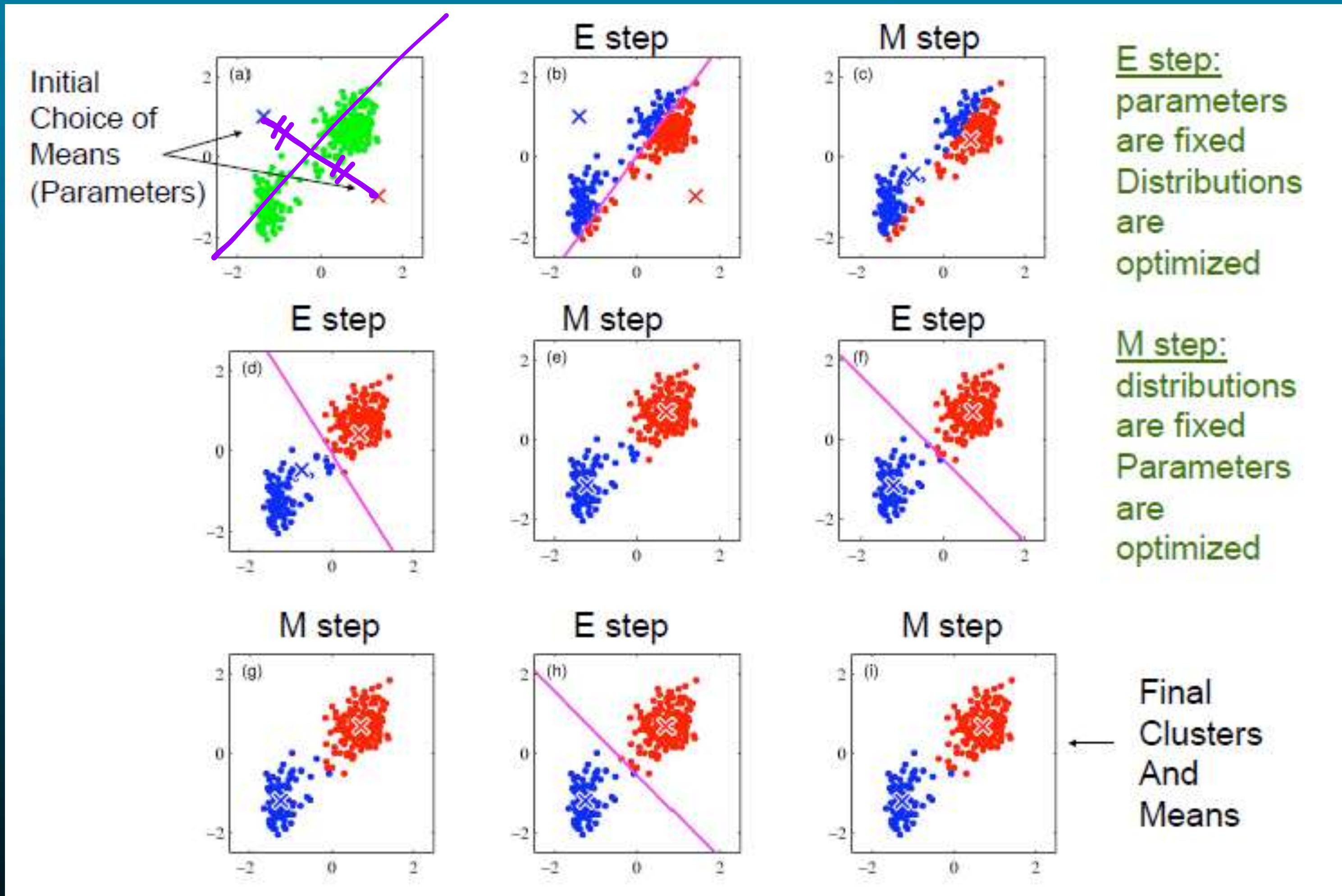
- **Convergence is assured since J is reduced repeatedly.**

$$J(X;C_1,\_) \geq J(X;C_1,A_1) \geq J(X;C_2,A_1) \geq J(X;C_2,A_2) \geq J(X;C_3,A_2) \geq J(X;C_3,A_3) \geq \cdots$$

Properties of K-means Clustering

- K-means can find the approximate solution efficiently.
- The distortion (squared error) is a monotonically non-increasing function of iterations.
- The goal is to minimize the square error, but it could end up in a local minimum.
- To increase the probability of finding the global minimum, try to start k-means with different initial conditions.
- "Cluster validation" refers to a set of methods which try to determine the best value of k.
- Other distance measures can be used in place of the Euclidean distance, with corresponding change in center identification.

K-means Snapshots

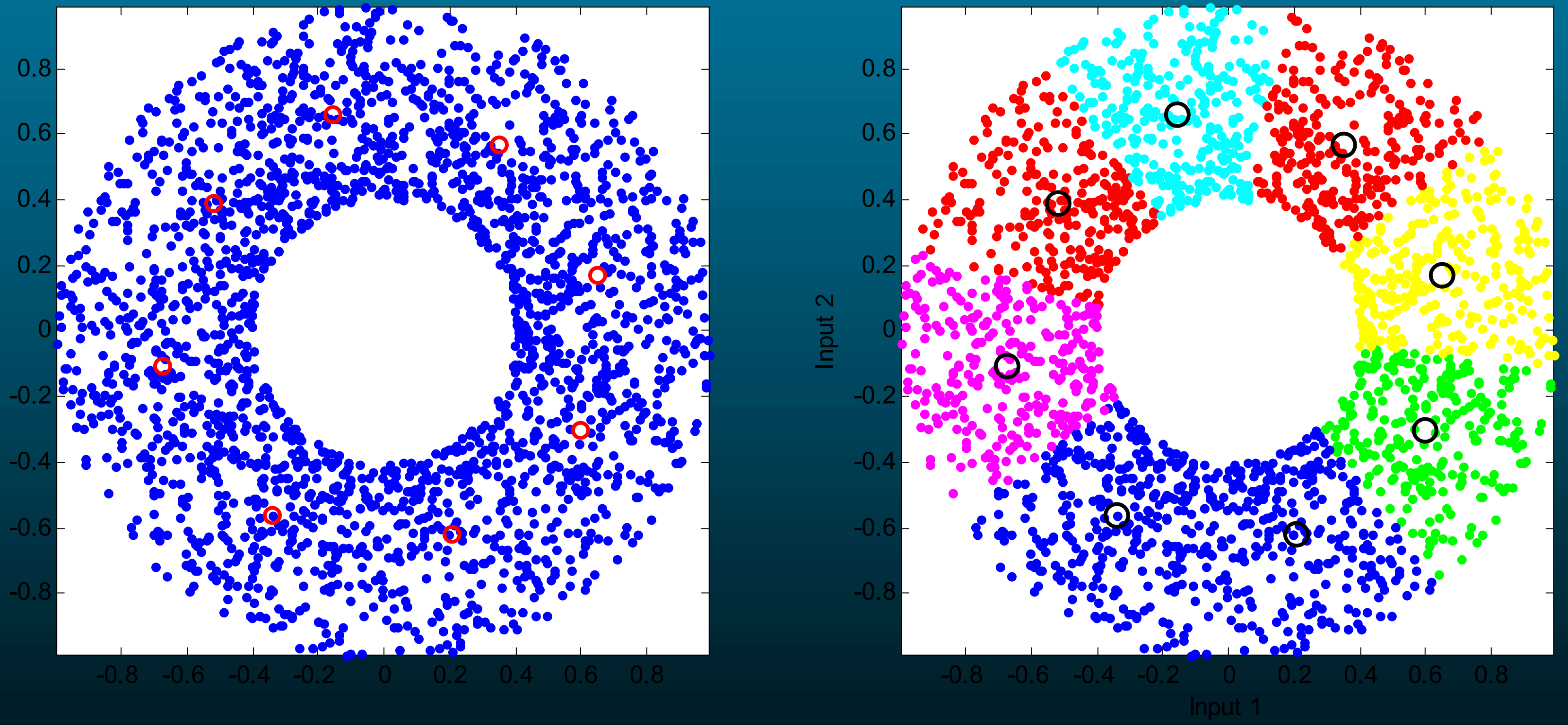


Demo of K-means Clustering

- Toolbox download
  - Utility Toolbox
  - Machine Learning Toolbox
- Demos
  - kMeansClustering.m
  - vecQuantize.m

Demos of K-means Clustering

kMeansClustering.m



Application: Image Compression

Goal

- Convert a image from true colors to indexed colors with minimum distortion.

Steps

- Collect data from a true-color image
- Perform k-means clustering to obtain cluster centers as the indexed colors
- Compute the compression rate

$$before = m * n * 3 * 8\ bits$$

c=0~255 → lgc 个 bit 存

$$after = m * n * \underline{\log_2(c)} + \underbrace{c * 3 * 8\ bits}_{\text{map size}}$$

$$\rho = \frac{before}{after} = \frac{m * n * 3 * 8}{m * n * \log_2(c) + c * 3 * 8} = \frac{24}{\log_2(c) + \frac{24c}{m * n}} \approx \frac{24}{\log_2(c)}$$

Quiz!

True-color vs. Index-color Images

True-color image

- Each pixel is represented by a vector of 3 components [R, G, B]

Index-color image

- Each pixel is represented by an index into a color map

Example: Image Compression

Date: 1998/04/05

Dimension: 480x640

Raw data size: 480*640*3 bytes = 900KB

File size: 49.1KB

Compression ratio = 900/49.1 = 18.33

Example: Image Compression

Date: 2015/11/01

Dimension: 3648x5472

Raw data size: 3648*5472*3 bytes = 57.1MB

File size: 3.1MB

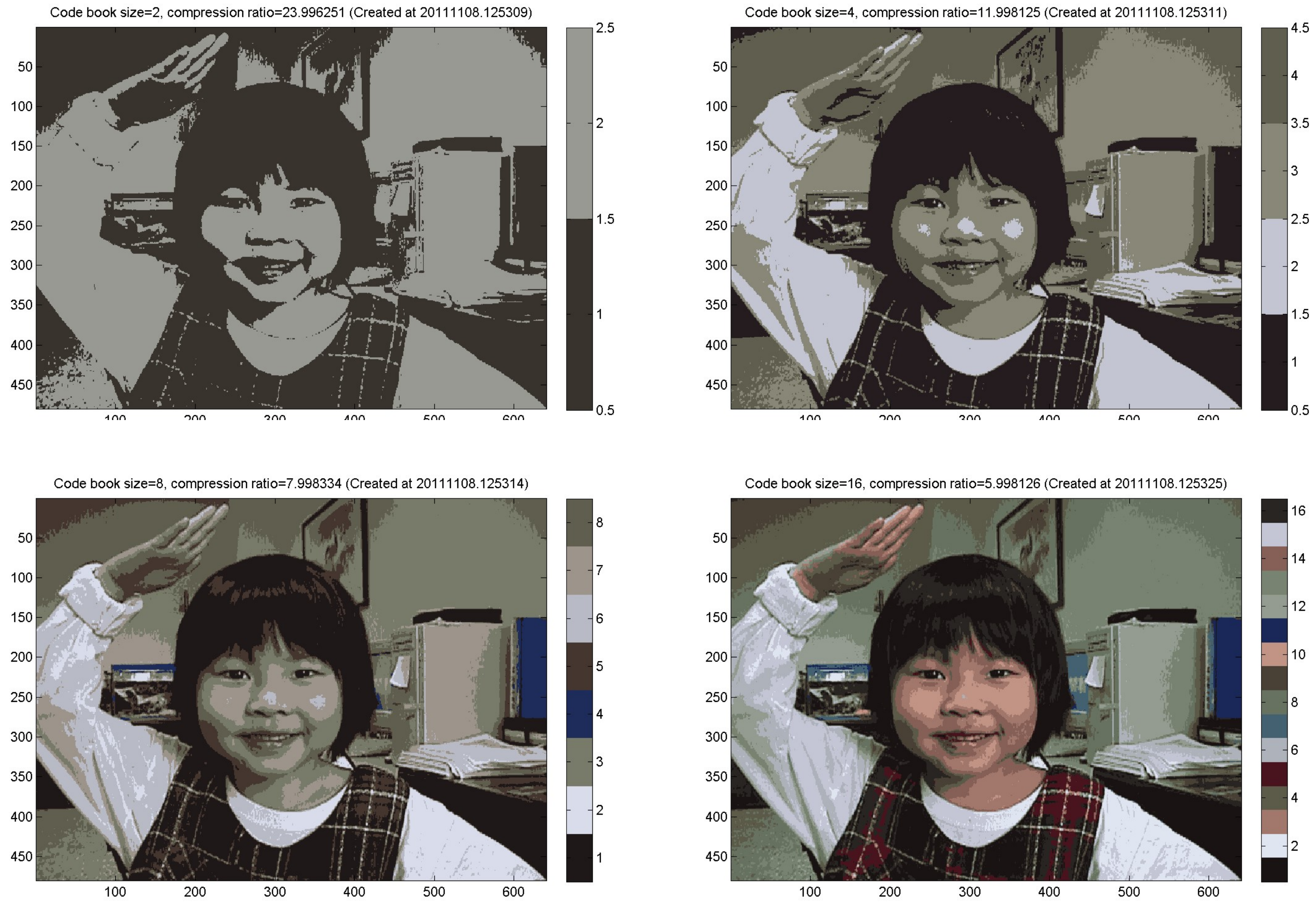
Compression ratio = 57.1/3.1 = 18.42

Example: Image Compression

Some quantities of the k-means clustering

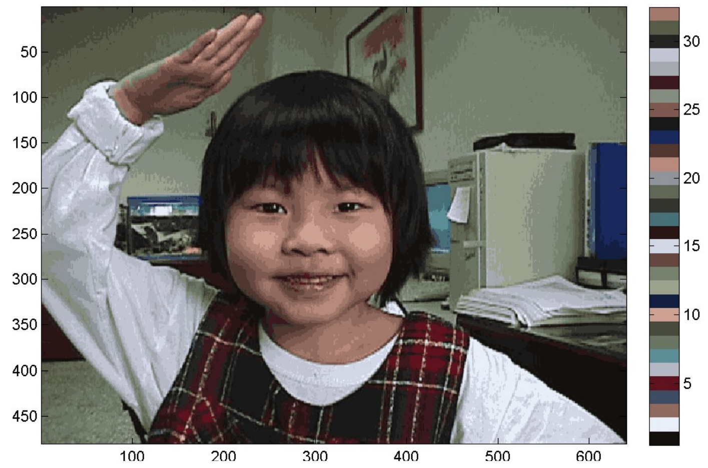
- n = 480x640 = 307200 (no of vectors to be clustered)
- d = 3 (R, G, B)
- m = 256 (no. of clusters)

Example: Image Compression

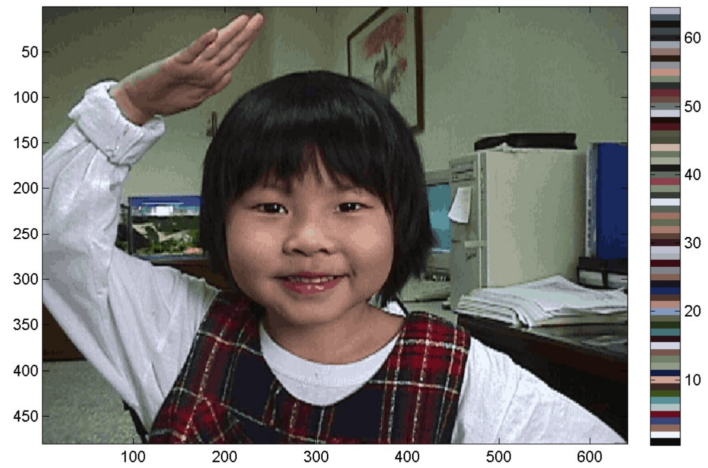


Example: Image Compression

Code book size=32, compression ratio=4.797601 (Created at 20111108.125401)



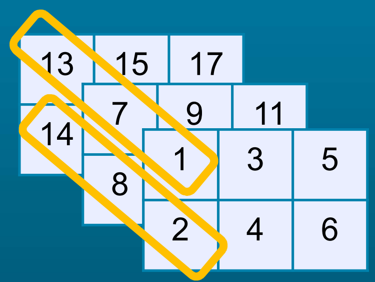
Code book size=64, compression ratio=3.996669 (Created at 20111108.125537)



Code book size=128, compression ratio=3.423680 (Created at 20111108.141727)

Indexing Techniques

Indexing of pixels for an 2*3*3 image



$$\begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 7 & 8 & 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 & 17 & 18 \end{bmatrix}$$

Related command: reshape

```
X = imread('annie19980405.jpg');
image(X);
[m, n, p]=size(X);
index=reshape(1:m*n*p, m*n, 3)';
data=double(X(index));
```

Code

```
X = imread('annie19980405.jpg');
image(X)
[m, n, p]=size(X);
index=reshape(1:m*n*p, m*n, 3)';
data=double(X(index));
maxI=6;
for i=1:maxI
	centerNum=2^i;
	fprintf('i=%d/%d: no. of centers=%d\n', i, maxI, centerNum);
	center=kMeansClustering(data, centerNum);
	distMat=distPairwise(center, data);
	[minValue, minIndex]=min(distMat);
	X2=reshape(minIndex, m, n);
	map=center'/255;
	figure; image(X2); colormap(map); colorbar; axis image;
end
```

Extensions to Image Compression

Extensions to image data compression via clustering

1. **Use blocks as the unit for VQ (see exercise)**
   - Smart indexing by creating the indices of the blocks of page 1 first.
   - True-color image display (No way to display the compressed image as an index-color image)
2. **Use separate code books for RGB**

What are the corresponding compression ratios?

Extension: K-medians Clustering

- **Difference from k-means clustering**
  - **Use L1 norm instead of L2 in the objective function**
- **Optimization strategy**
  - **Same as k-means clustering, except that the centers are found by the median operator**

Quiz!

- **Advantage**
  - **Less susceptible to outliners**

Quiz!

Quiz

25

Extension to Circle Finding

How to find circles via a k-means like algorithm?