Clustering and the K-means algorithm

Yihui Saw

18.304 Seminar Talk 1

March 6, 2013

Clustering examples

- Customer purchase patterns
- Language family models
- Data compression

Original Image

2 colors

4 colors

8 colors

The clustering problem

Input: Training set $S_n = \{x^{(i)}, i = 1, ..., n\}$, where $x^{(i)} \in R^d$, integer k clusters

Output: A set of clusters $C_1, C_2, ..., C_k$

Distance metric

Squared Euclidean Distance

$$dist(x^{(i)}, x^{(j)}) = \sum_{l=1}^{d} (x_l^{(i)} - x_l^{(j)})^2$$

Cosine Similarity

$$cos(x^{(i)}, x^{(j)}) = \frac{x^{(i)} \cdot x^{(j)}}{\| x^{(i)} \| \| x^{(j)} \|} = \frac{\sum_{l=1}^{d} x_l^{(i)} x_l^{(j)}}{\sqrt{\sum_{l=1}^{d} (x_l^{(i)})^2} \sqrt{\sum_{l=1}^{d} (x_l^{(j)})^2}}$$

The cost of clustering

Clusters based on representatives

$C_j = \{i \in \{1, ..., n\}\ s.t.\ the\ closest\ representative\ of\ x^{(i)}\ is\ z^{(j)}\}$

Cost function based on representatives

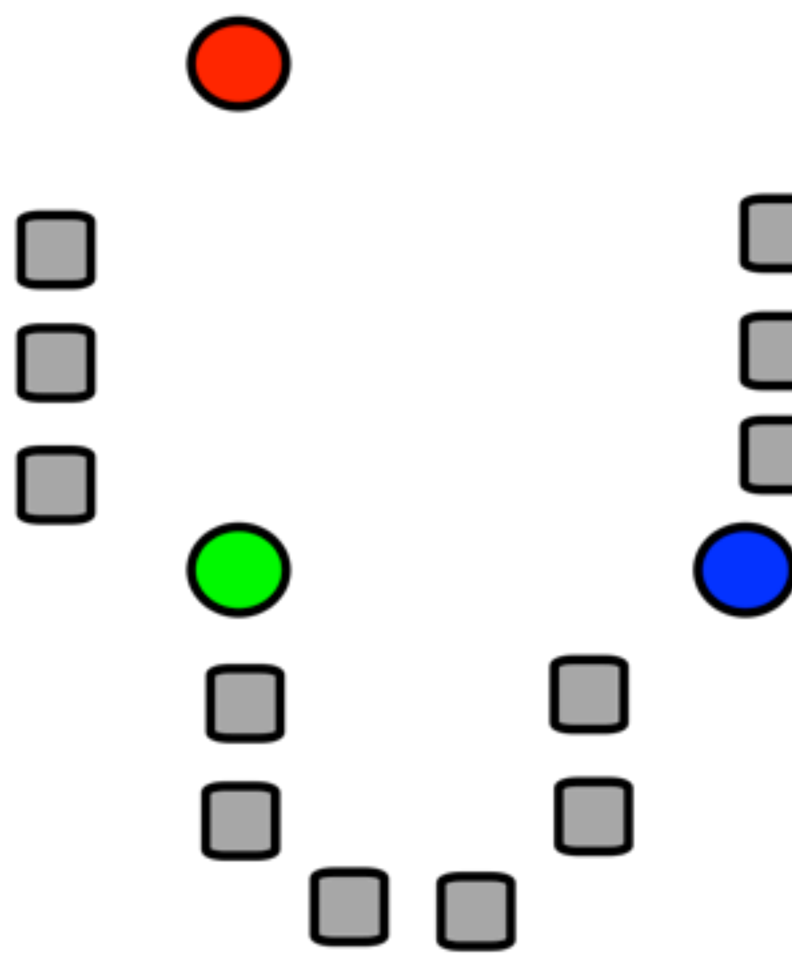
$$
\begin{aligned}
cost(z^{(1)}, ..., z^{(k)}) &= \min_{C_1,...,C_k} cost(C_1, ..., C_k, z^{(1),...,z^{(k)}}) \\
&= \min_{C_1,...,C_k} \sum_{j=1...k} \sum_{i \in C_j} \| x^{(i)} - z^{(j)} \|^2 \\
&= \sum_{i=1,...,n} \min_{j=1...k} \| x^{(i)} - z^{(j)} \|^2
\end{aligned}
$$

K-means algorithm

1. Initialize centroids $z^{(1)}, ..., z^{(k)}$

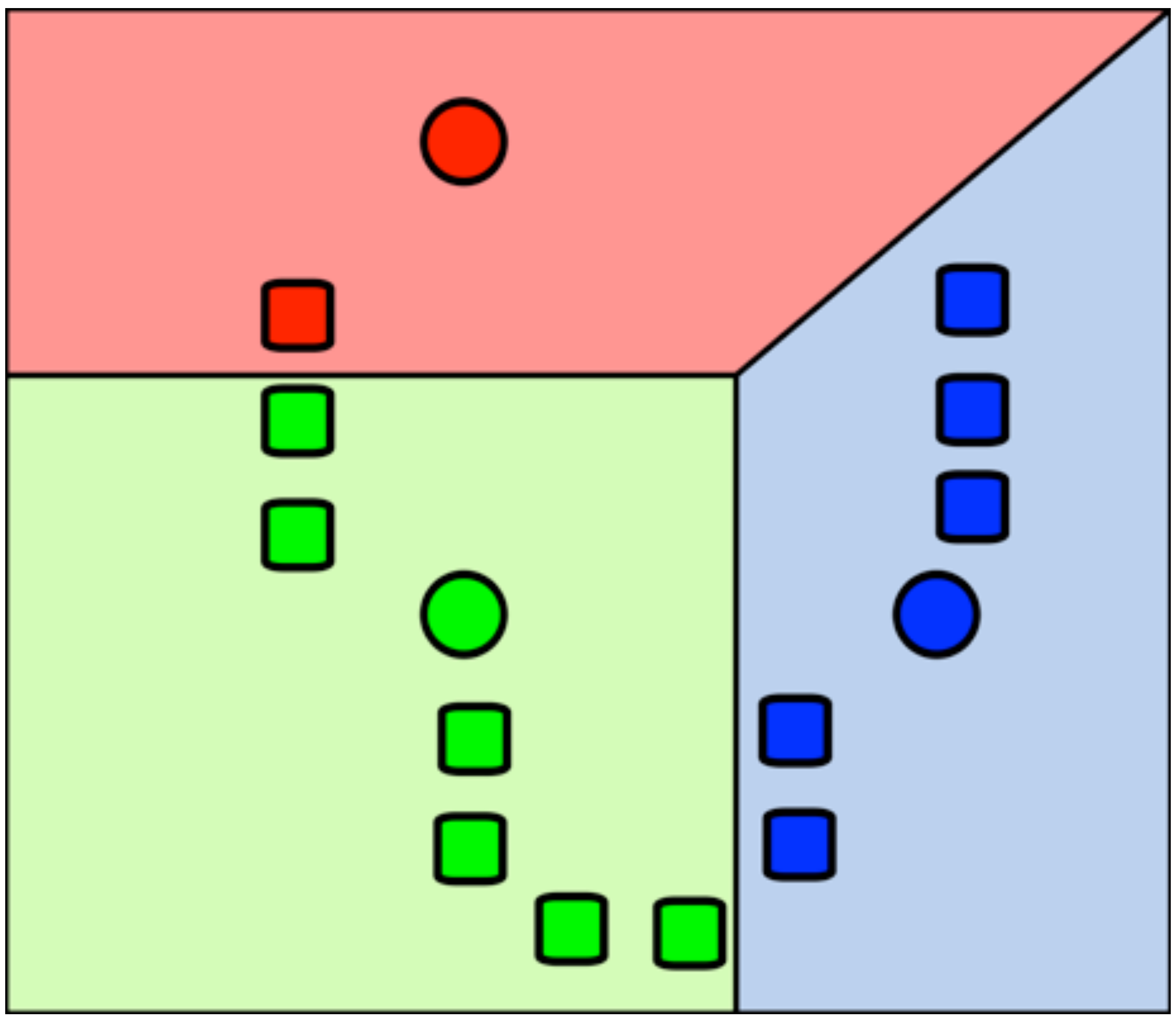
K-means algorithm

1. Initialize centroids $z^{(1)}, \ldots, z^{(k)}$



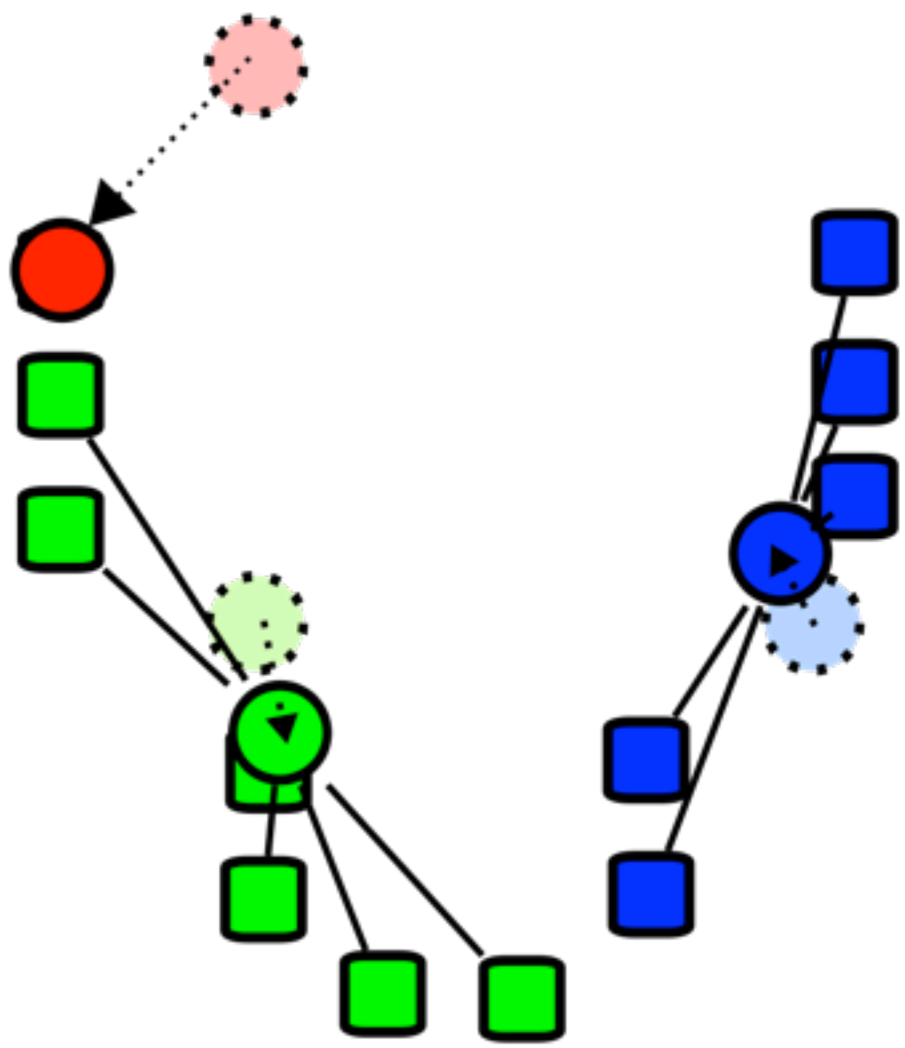
K-means algorithm

for each $j = 1, ..., k : C_j = \{i \text{ s.t. } x^{(i)} \text{ is closest to } z^{(j)}\}$



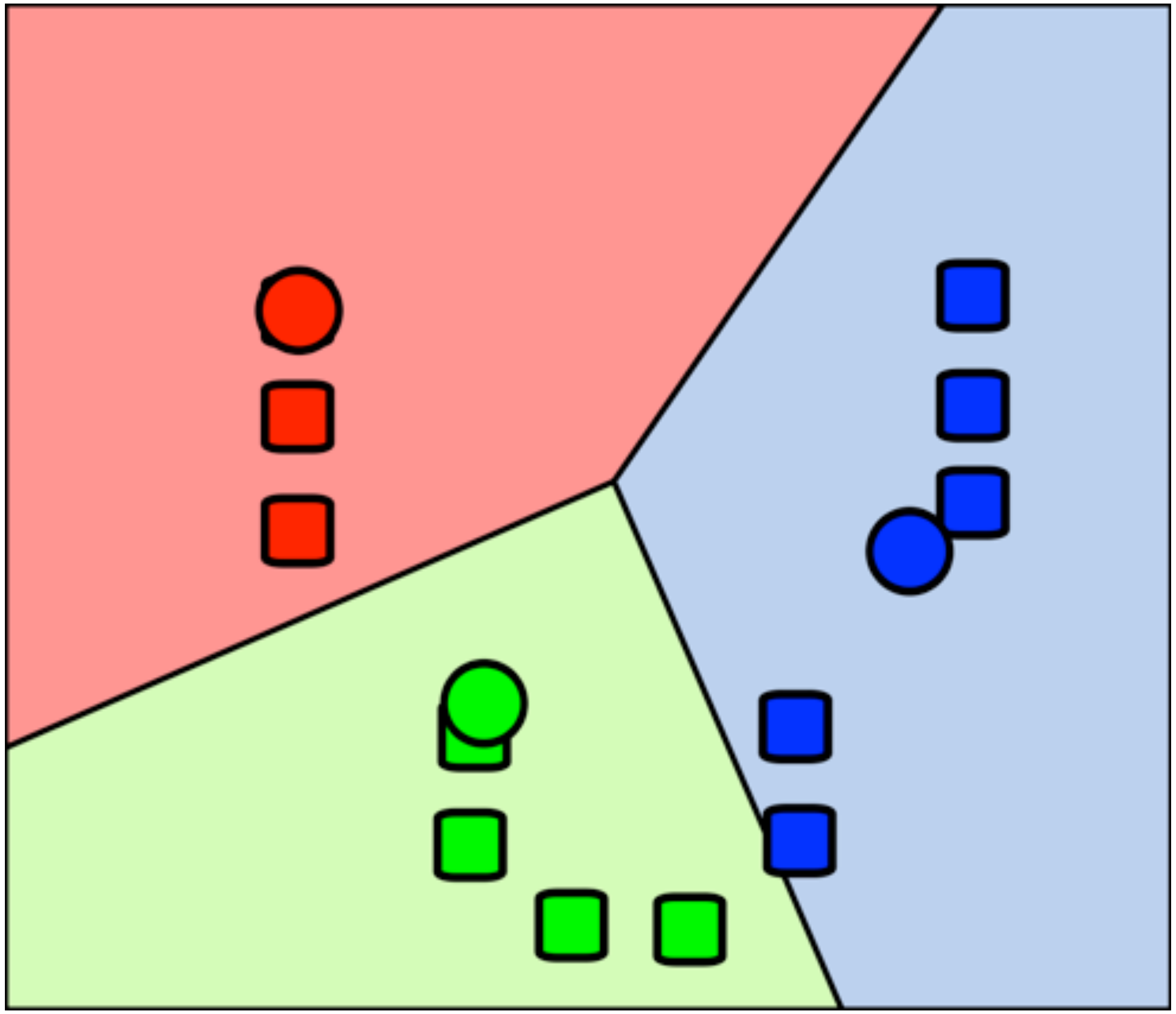
K-means algorithm

for each $j = 1, ..., k : z^{(j)} = \frac{1}{|C_j|} \sum_{i \in C_j} x^{(i)}$ (cluster mean)



K-means algorithm

Repeat until there is no further change in cost



K-means algorithm

An approximate method:

1. Initialize centroids $z^{(1)}, ..., z^{(k)}$

2. Repeat until there is no further change in cost

(a) for each $j = 1, ..., k : C_j = \{i \text{ s.t. } x^{(i)} \text{ is closest to } z^{(j)}\}$

(b) for each $j = 1, ..., k : z^{(j)} = \frac{1}{|C_j|} \sum_{i \in C_j} x^{(i)}$ (cluster mean)

Each iteration requires $O(kn)$ operations.

Proof of convergence

- Each iterative step necessarily lowers the cost - the cost monotonically decrease

Step 1 : reassign clusters based on distance

Old clusters : $C_1, C_2, ..., C_k$

New clusters : $C'_1, C'_2, ..., C'_k$

$$\begin{aligned} cost(C_1, C_2, \ldots, C_k, z^{(1)}, \ldots, z^{(k)}) &\overset{(a)}{\geq} \min_{C_1,\ldots,C_k} cost(C_1, C_2, \ldots, C_k, z^{(1)}, \ldots, z^{(k)}) \quad (10) \\ &= cost(C'_1, C'_2, \ldots, C'_k, z^{(1)}, \ldots, z^{(k)}) \quad (11) \end{aligned}$$

Proof of convergence

- Each iterative step necessarily lowers the cost - the cost monotonically decrease

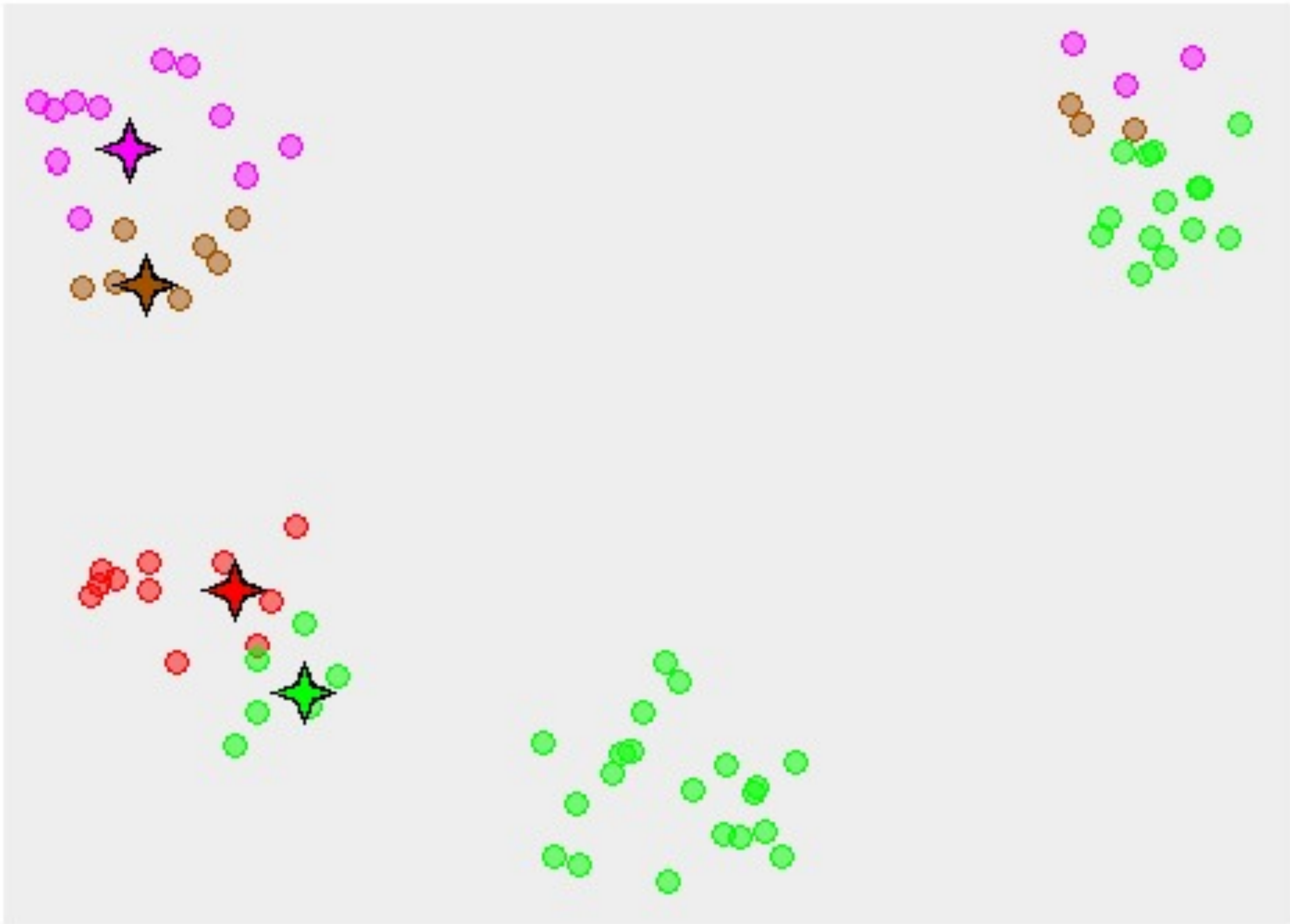
Step 2 : reassign centroids based on clusters

Old centroids : $z^{(1)}, ..., z^{(k)}$

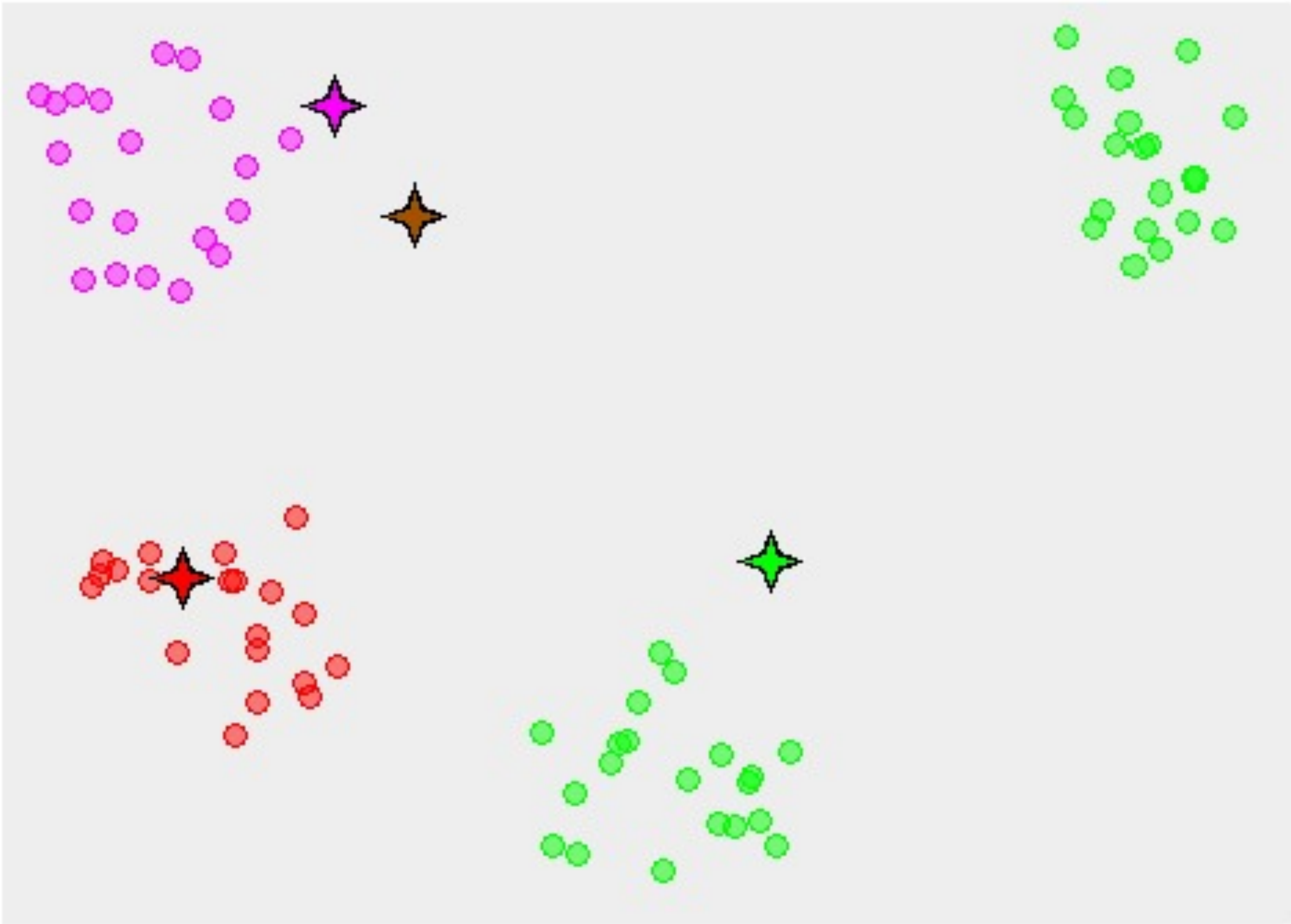
New centroids : $z'^{(1)}, ..., z'^{(k)}$

$$\begin{aligned} cost(C_1', C_2', \ldots, C_k', z^{(1)}, \ldots, z^{(k)}) &\overset{(b)}{\geq} \min_{z^{(1)},\ldots,z^{(2)}} cost(C_1', C_2', \ldots, C_k', z^{(1)}, \ldots, z^{(k)}) && (12) \\ &= cost(C_1', C_2', \ldots, C_k', z'^{(1)}, \ldots, z'^{(k)}) && (13) \end{aligned}$$

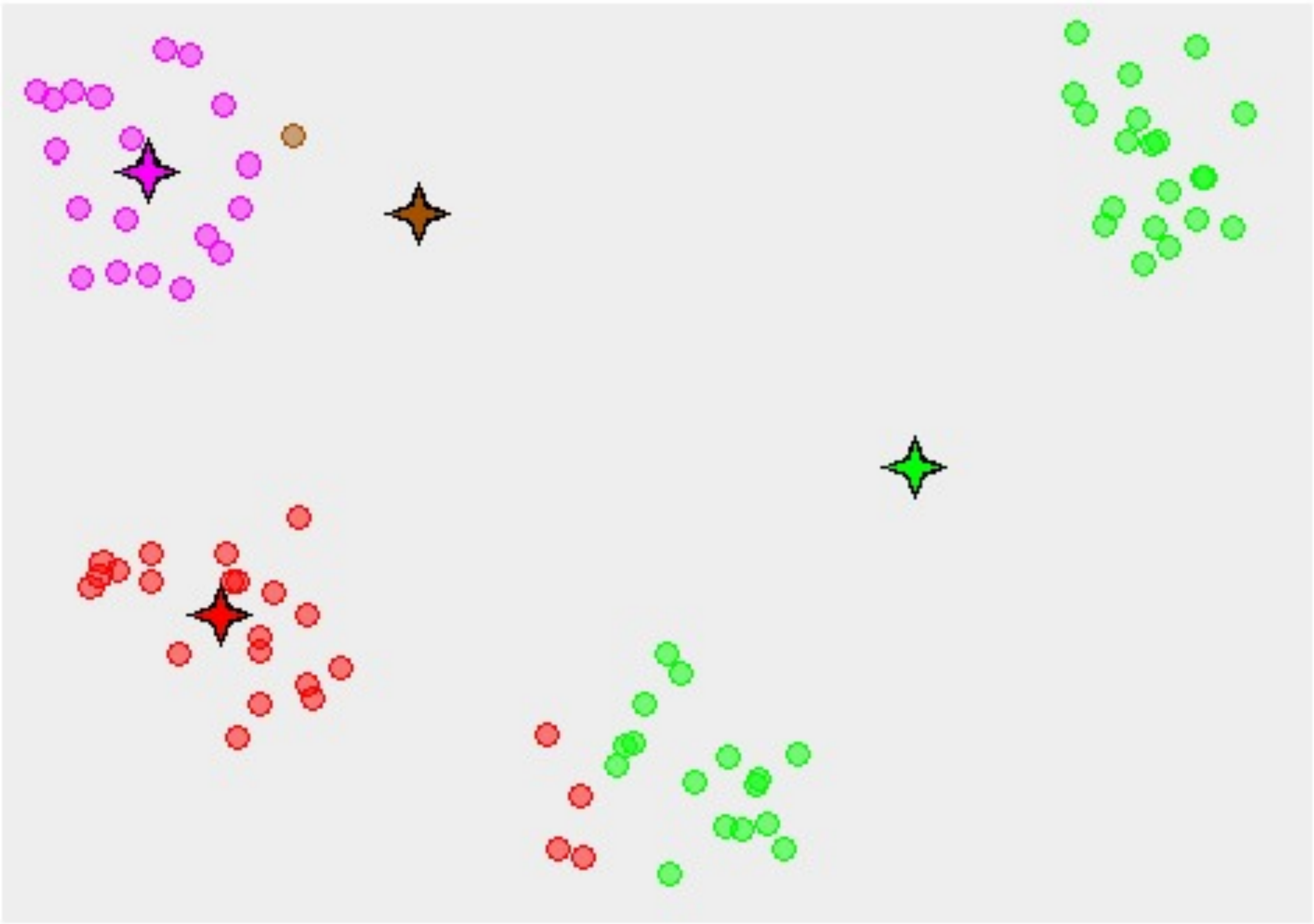
Convergence to local minimum



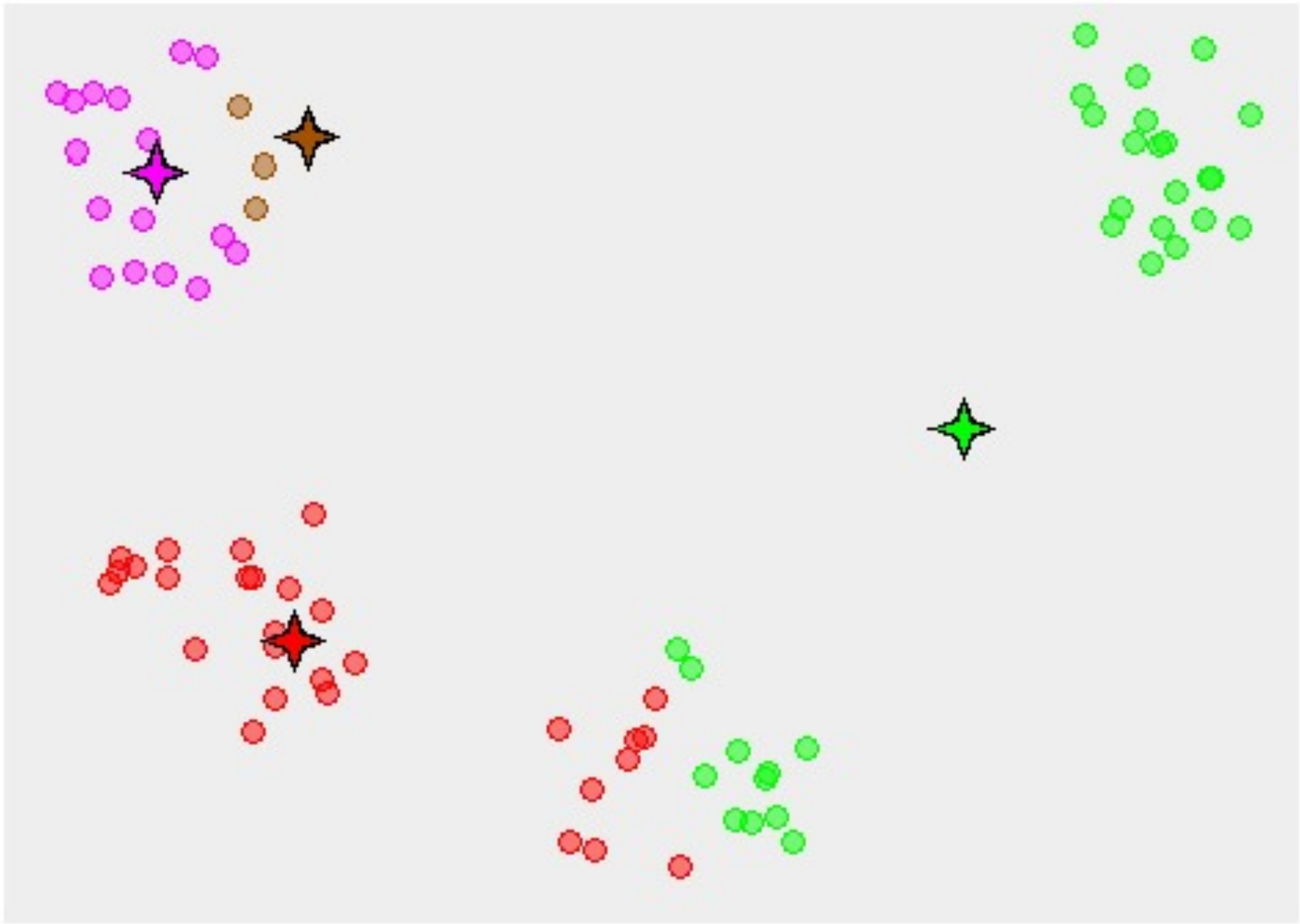
Convergence to local minimum



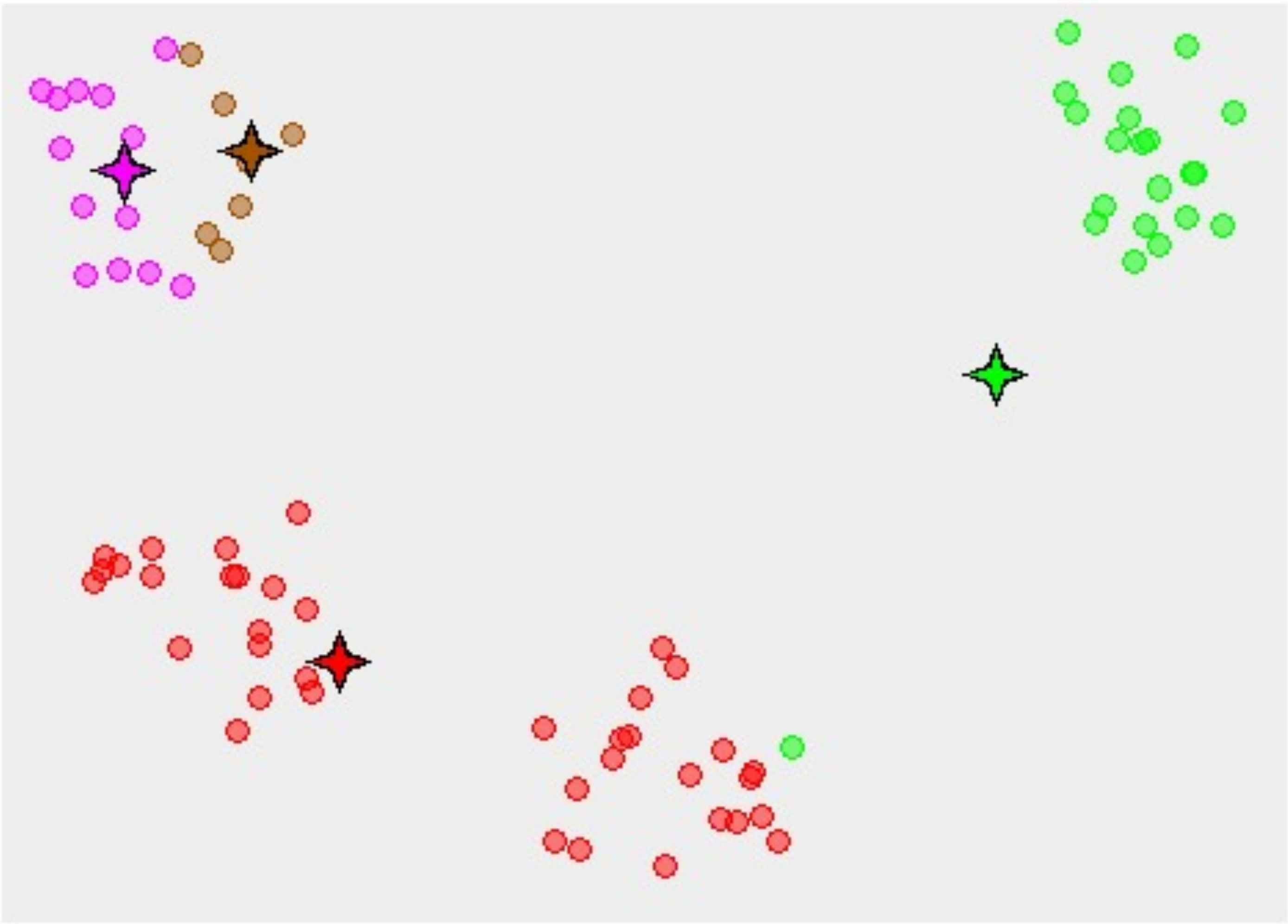
Convergence to local minimum



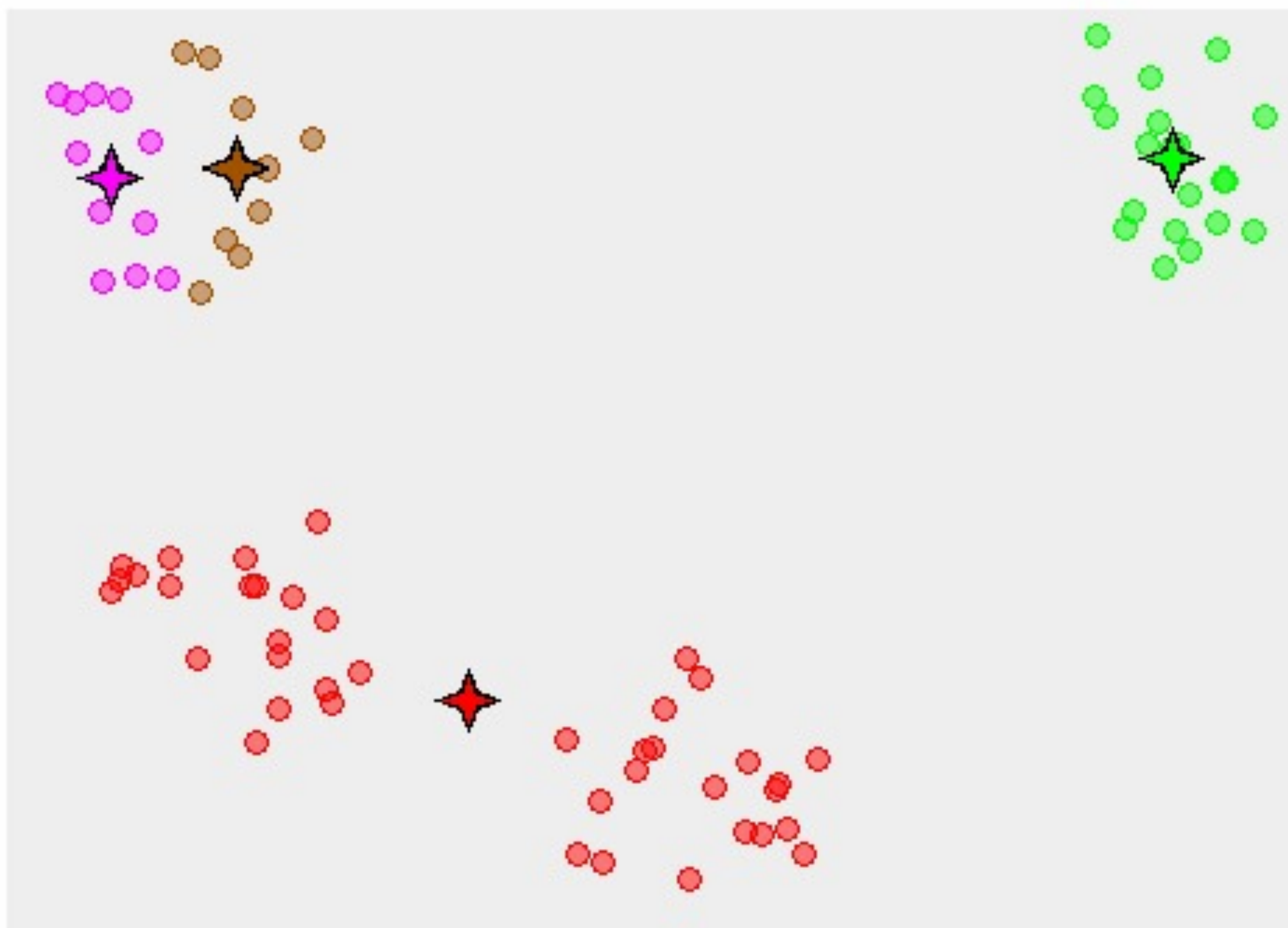
Convergence to local minimum



Convergence to local minimum



Convergence to local minimum



Convergence to local minimum