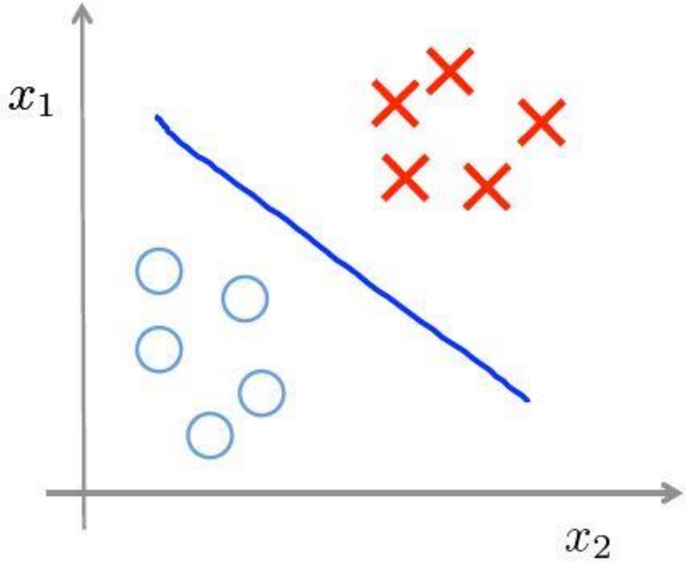
Clustering

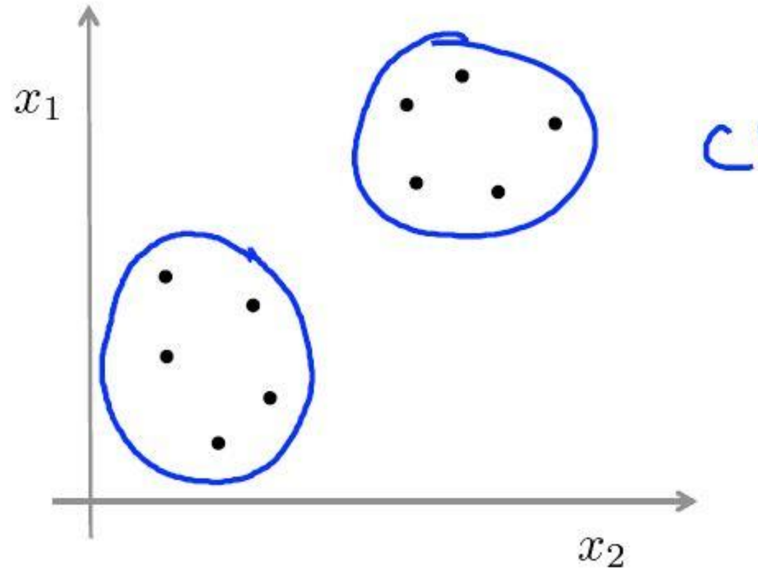
Unsupervised learning introduction

Supervised learning



Training set: $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), (x^{(3)}, y^{(3)}), \ldots, (x^{(m)}, y^{(m)})\}$

Unsupervised learning

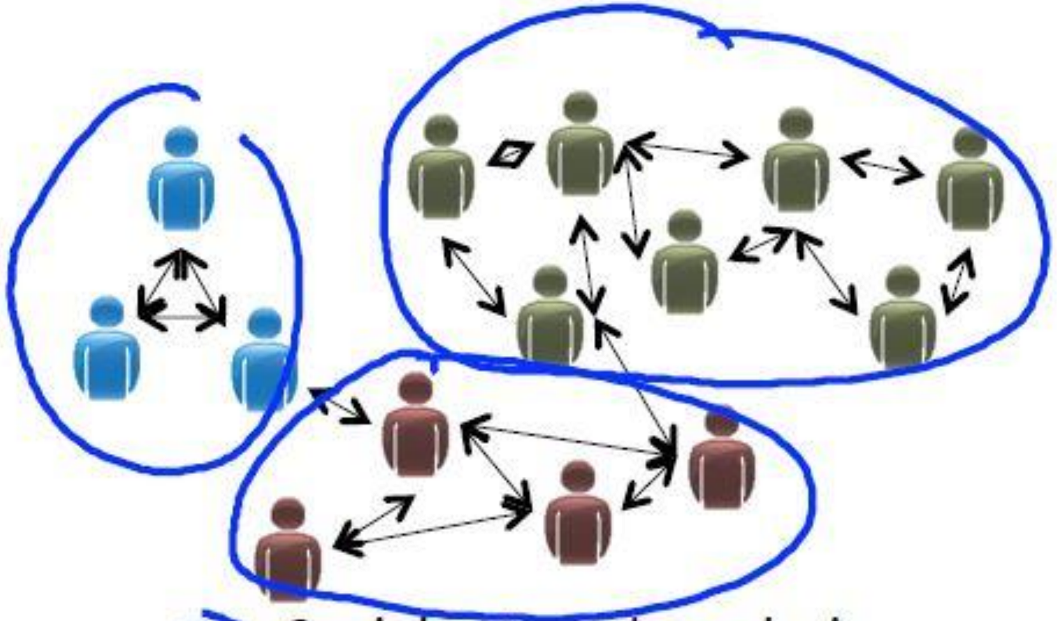


Clustering algorithm

Training set: $\{\underline{x^{(1)}}, \underline{x^{(2)}}, x^{(3)}, \dots, \underline{x^{(m)}}\} \leftarrow$

Applications of clustering

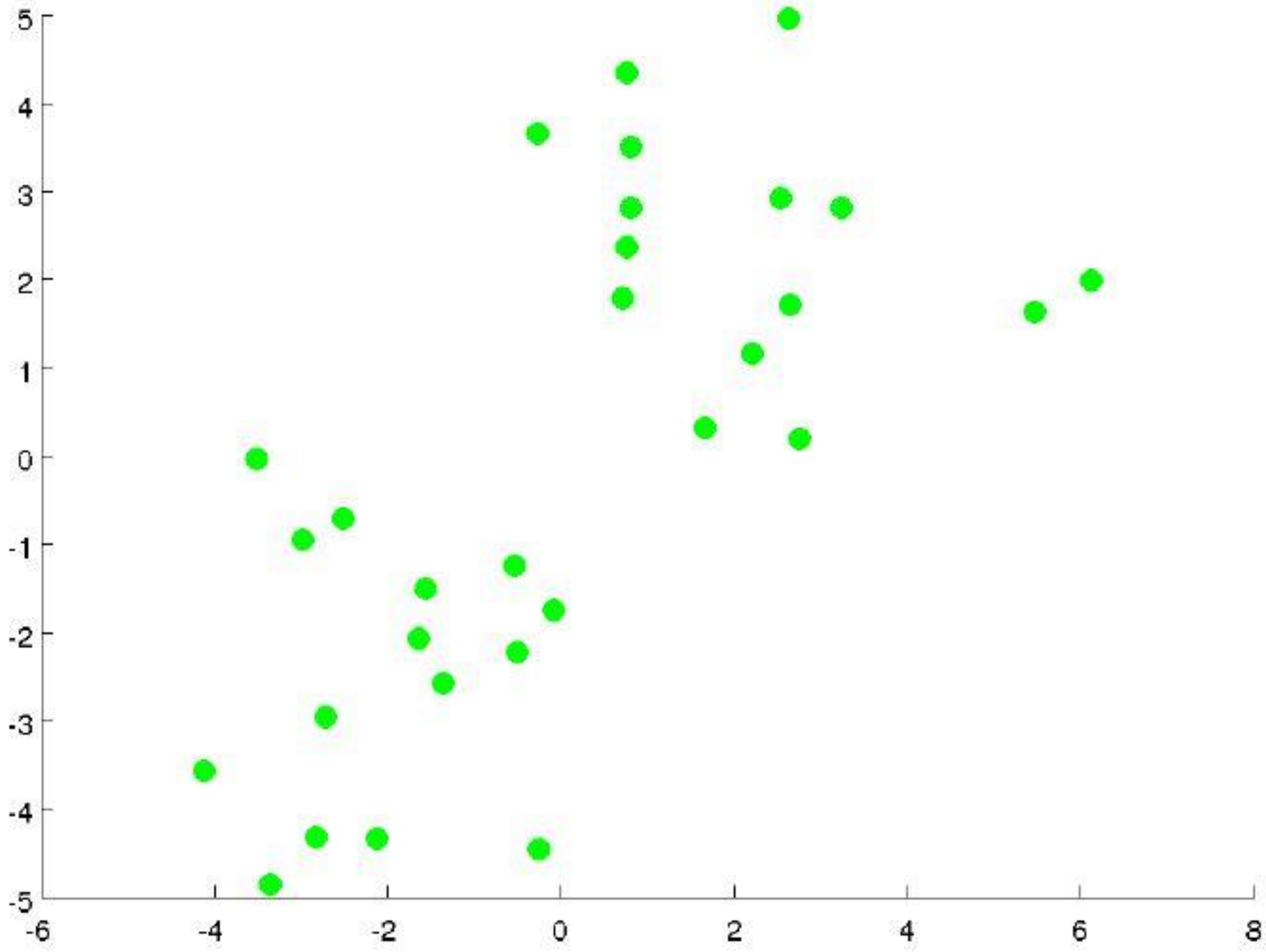
→ Market segmentation

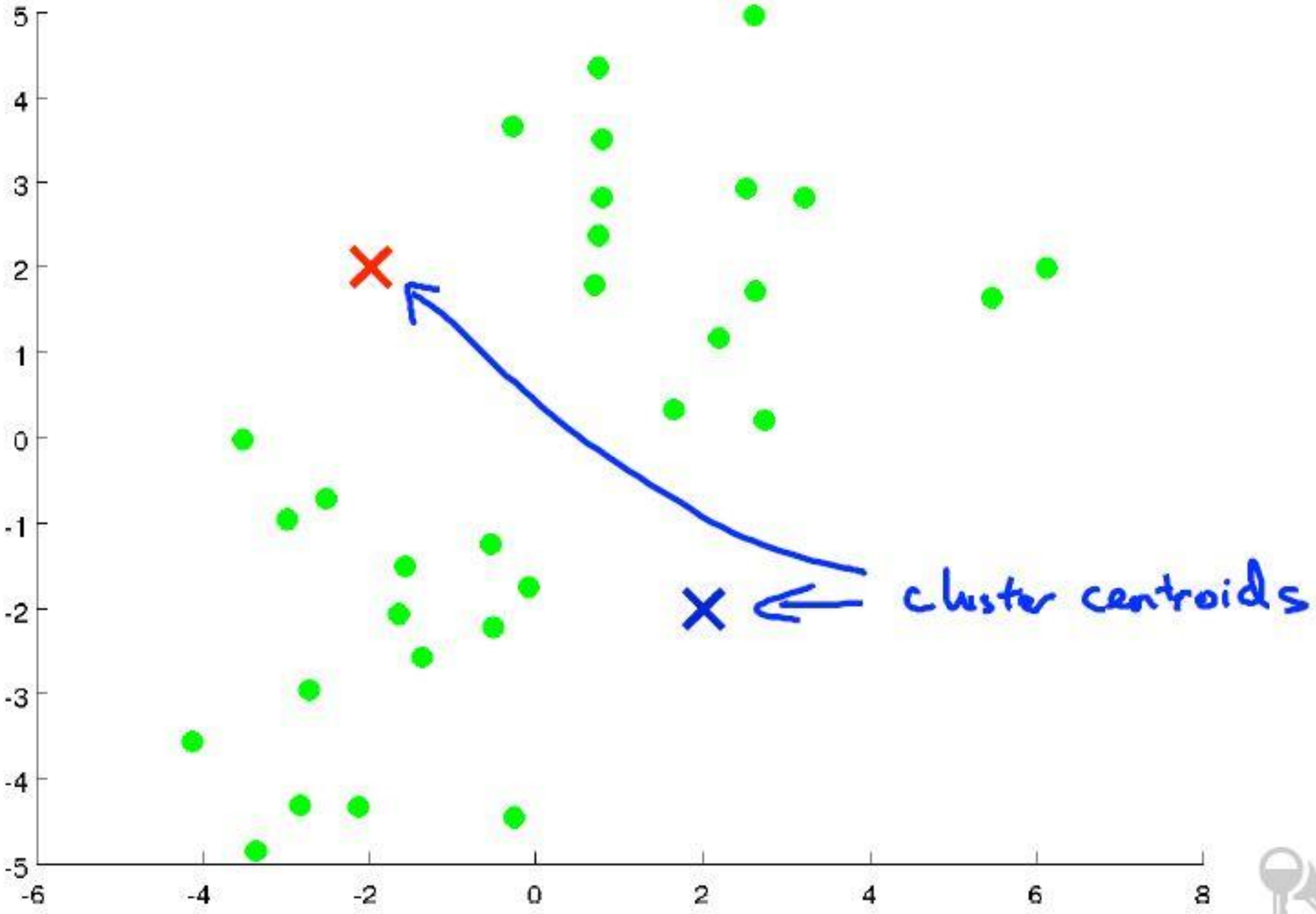


→ Social network analysis

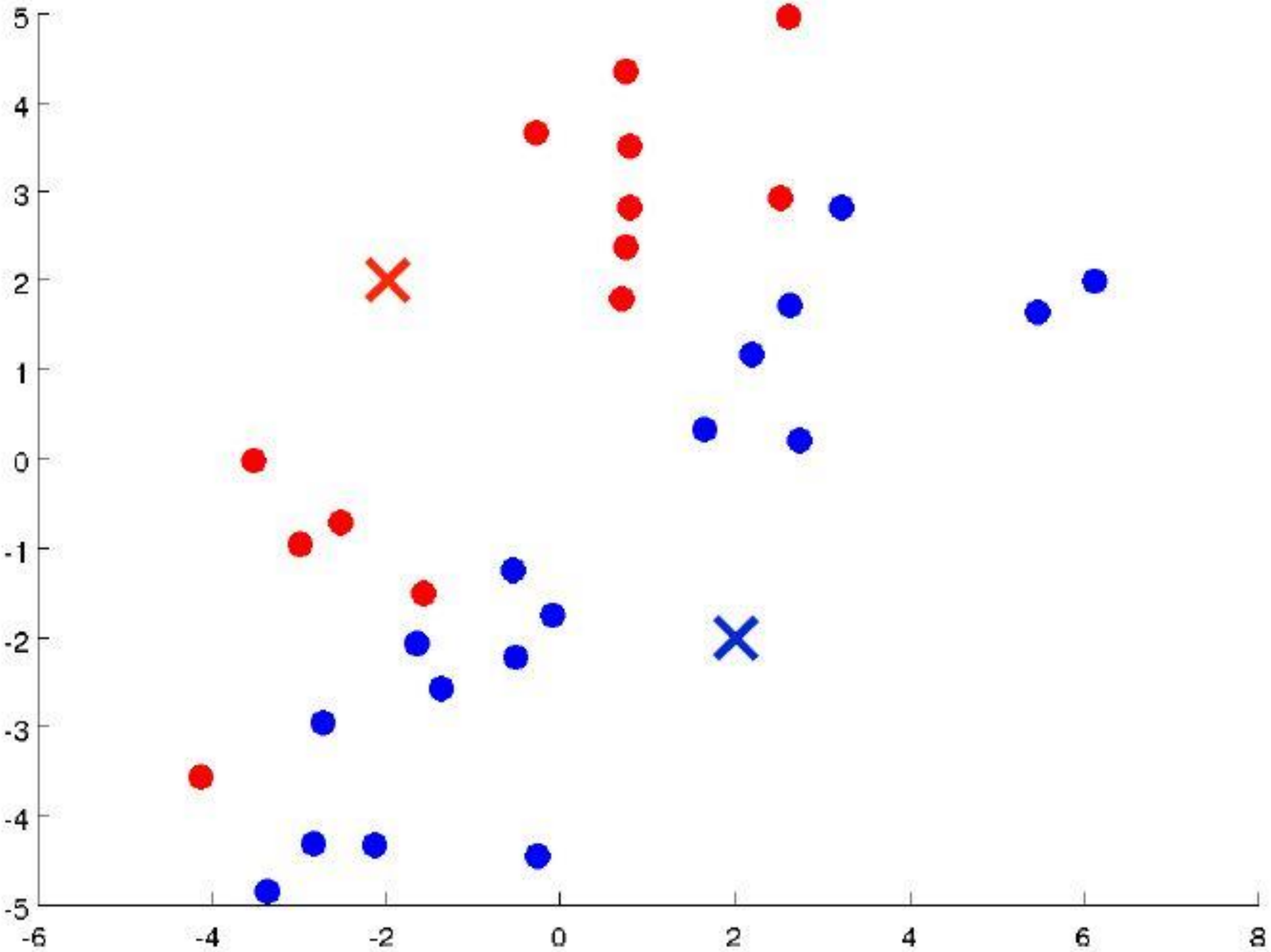
Clustering

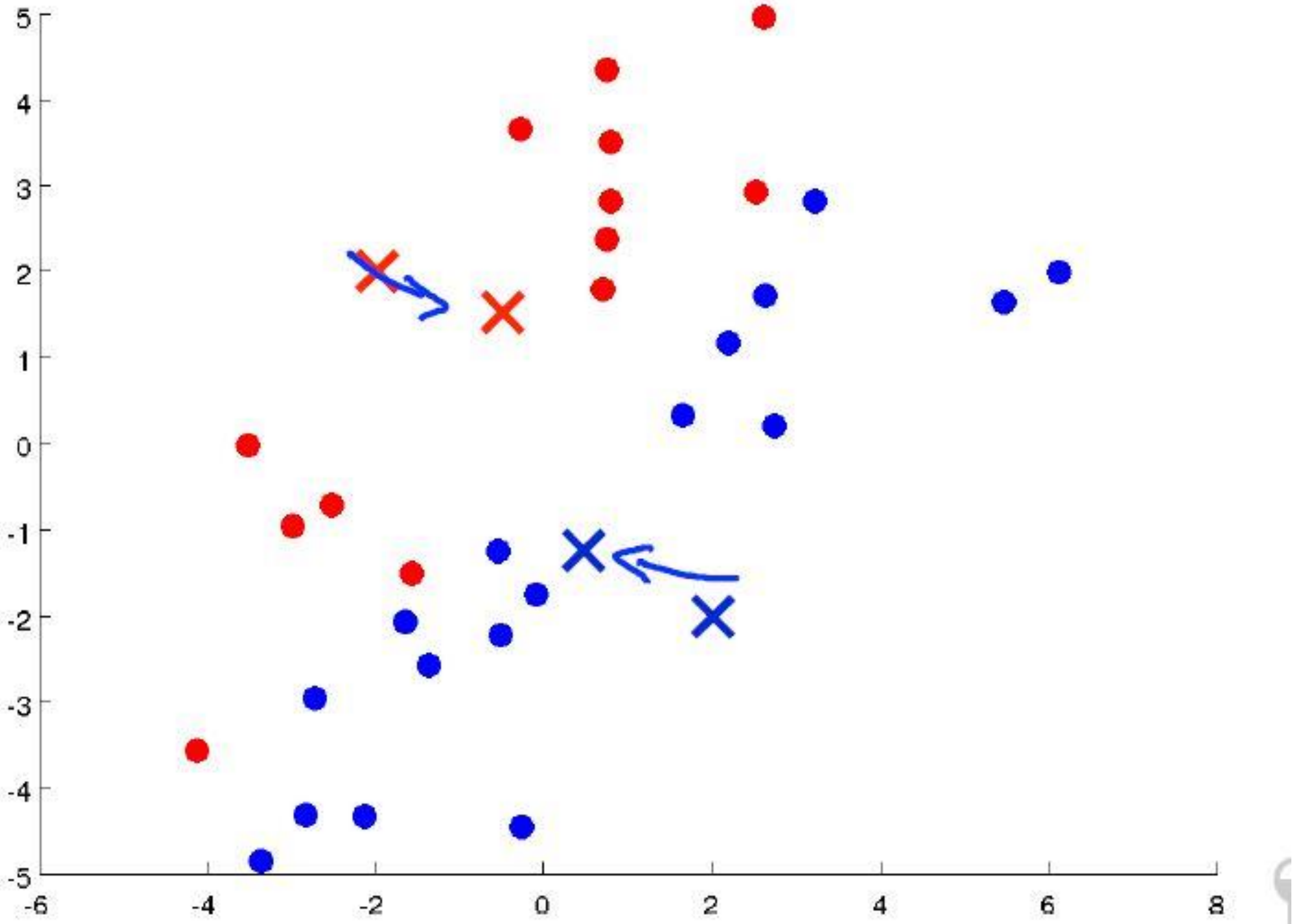
K-means algorithm


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8

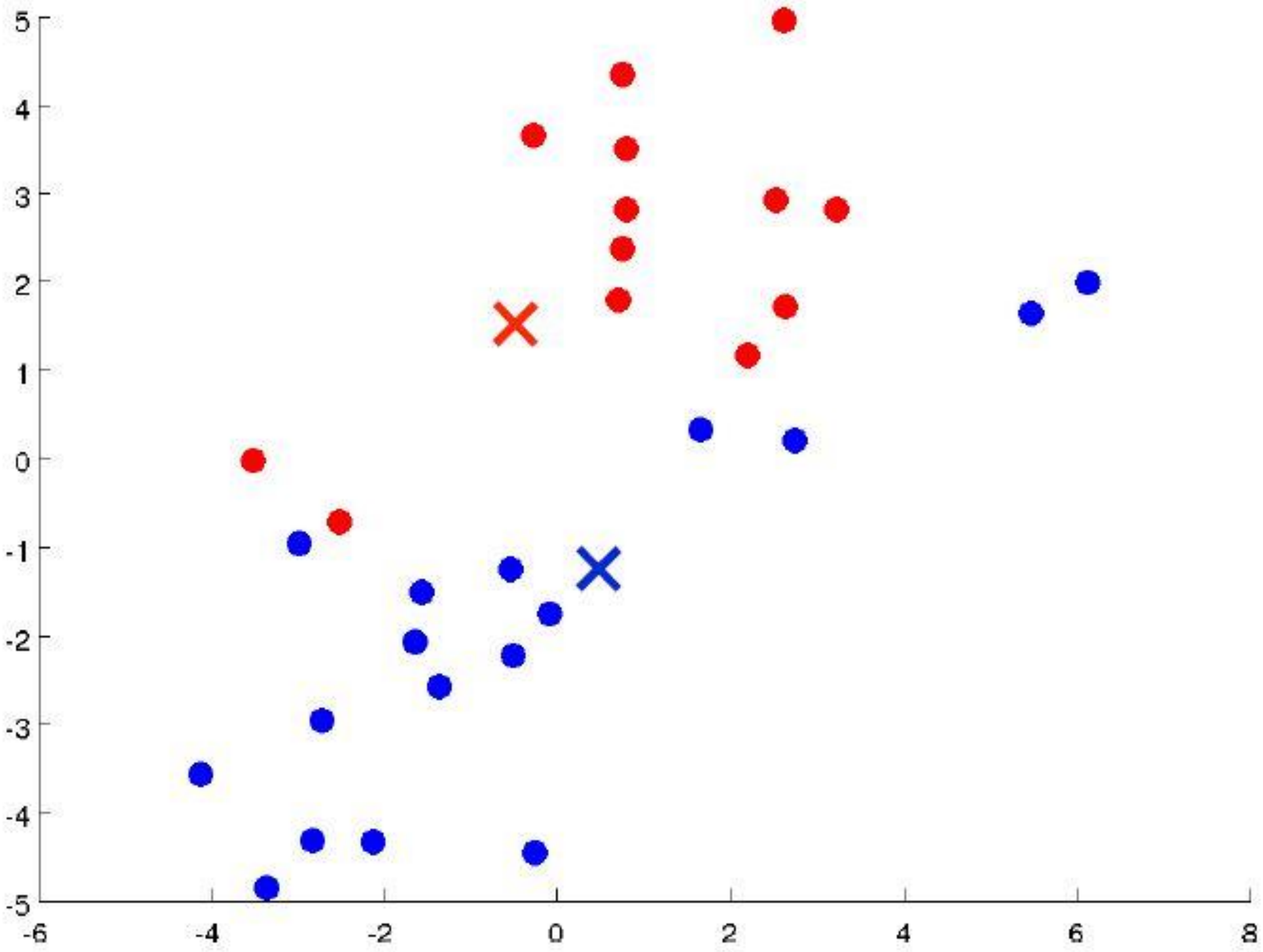


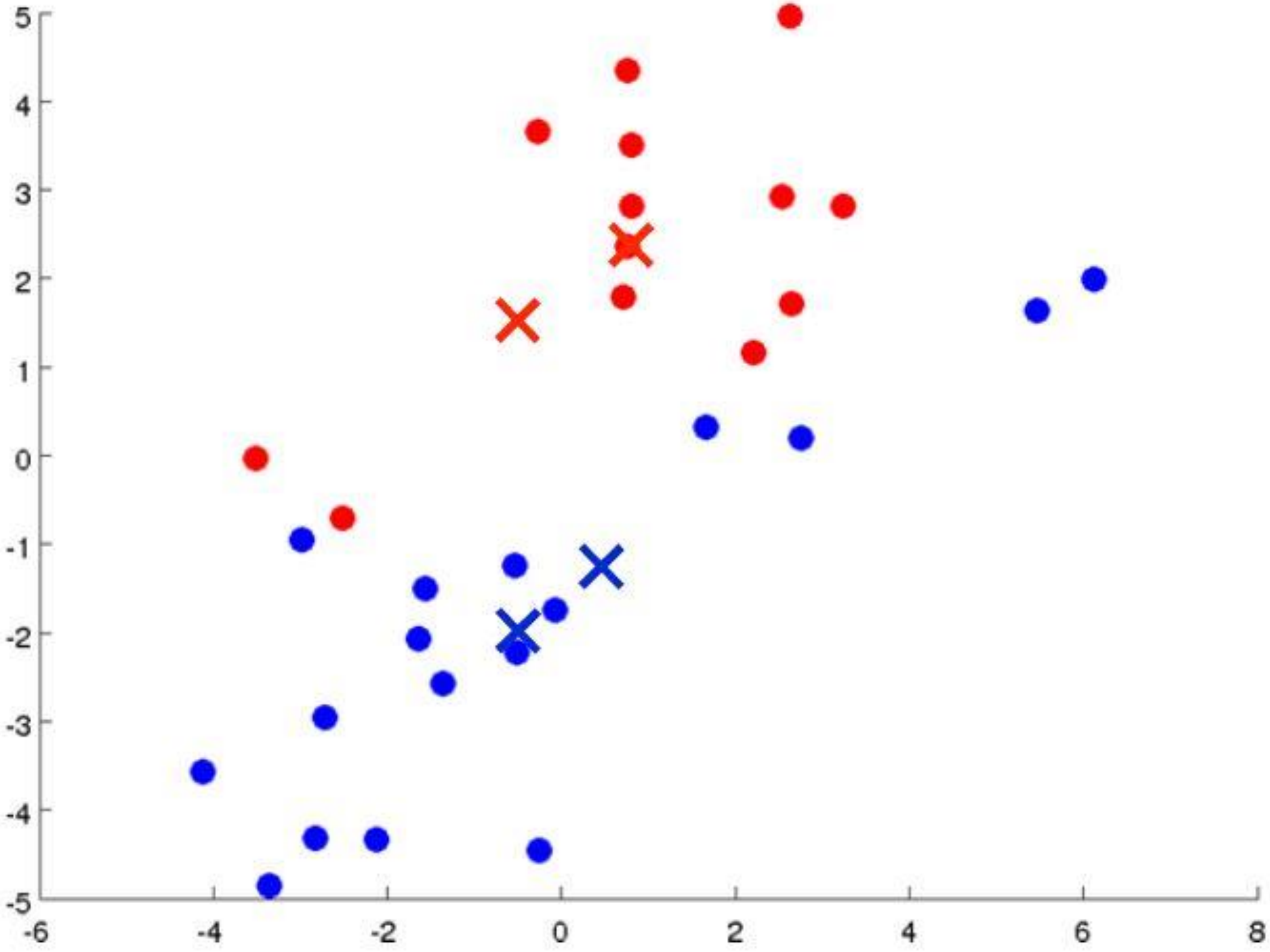
cluster centroids

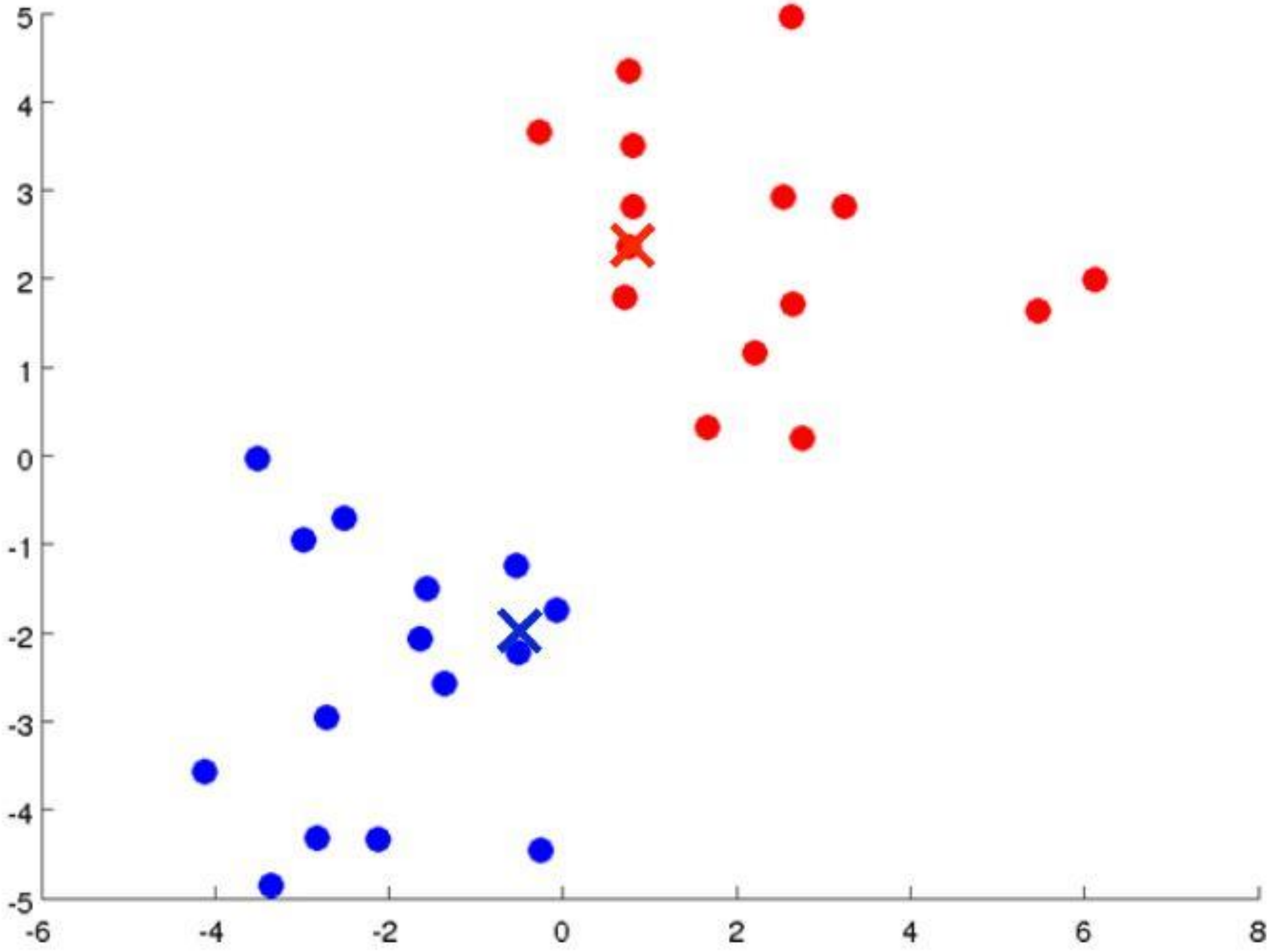

5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8

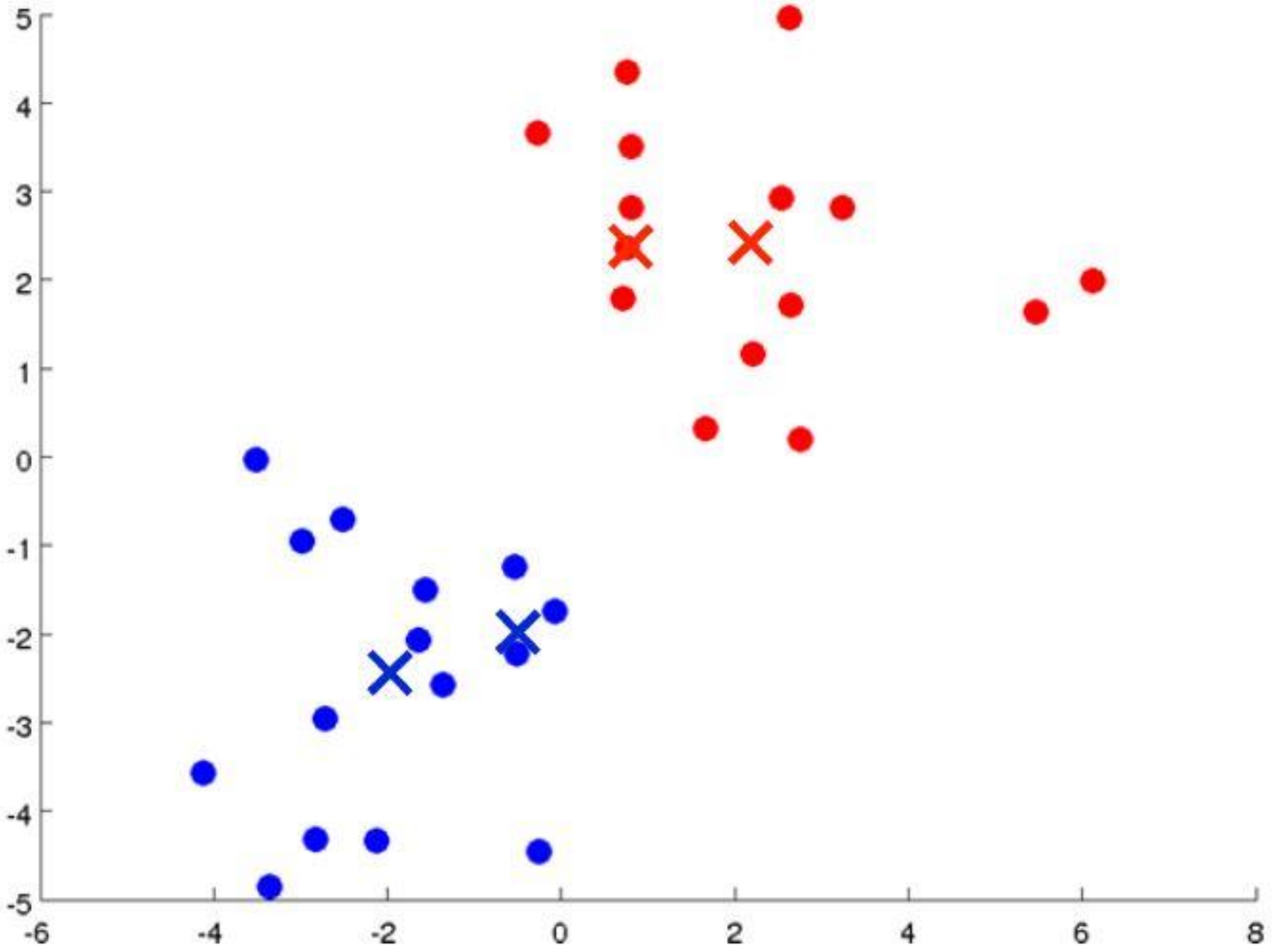


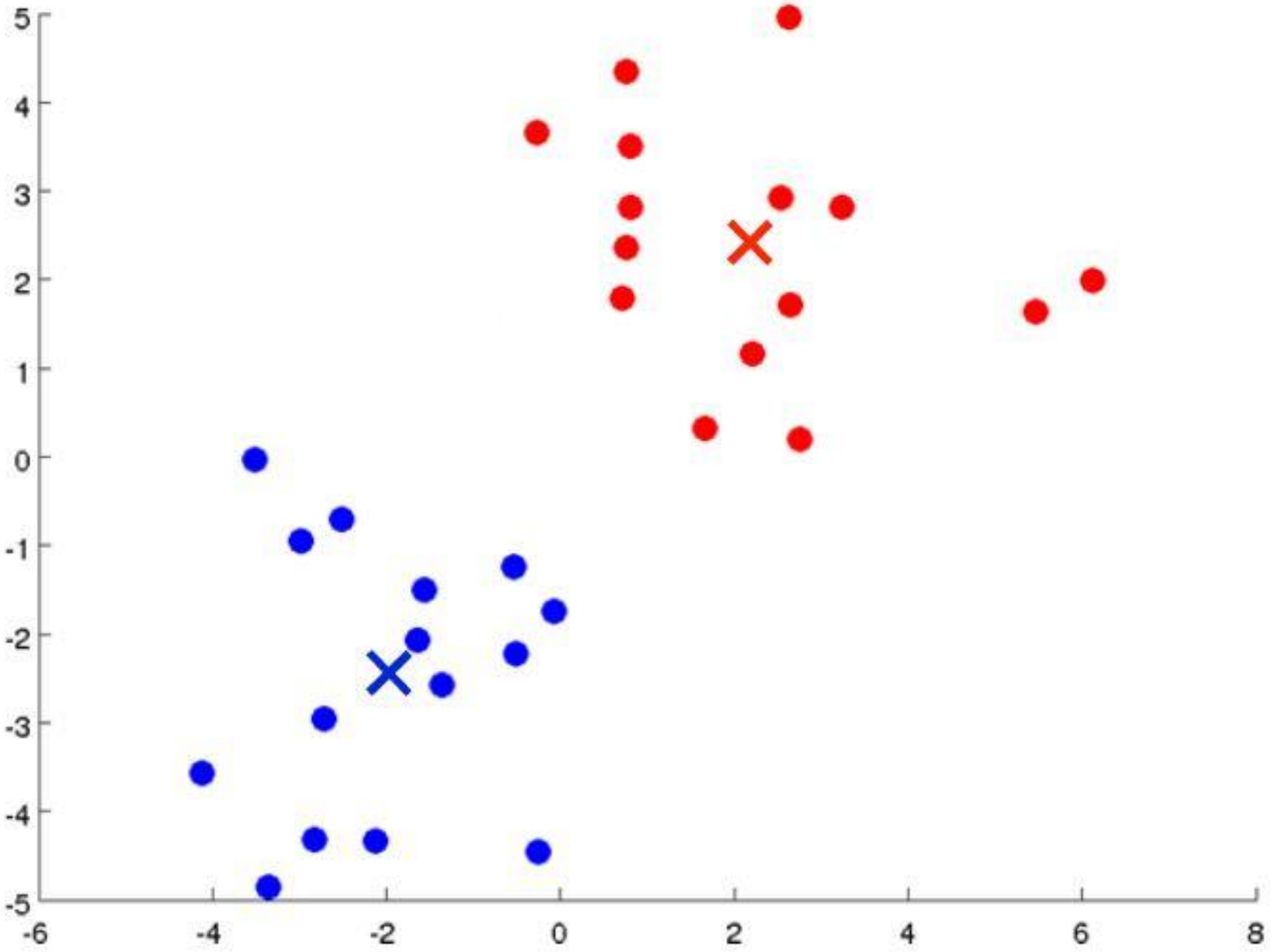
5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8

K-means algorithm

Input:

- K (number of clusters) ←
- Training set $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$ ←

K-means algorithm

μ_1 × μ_2 ×

Randomly initialize K cluster centroids $\mu_1, \mu_2, \dots, \mu_K \in \mathbb{R}^n$

```
Repeat {
    Cluster assignment step
    for $i$ = 1 to $m$
        $c^{(i)}$ := index (from 1 to $K$) of cluster centroid
                closest to $x^{(i)}$
    Move centroid
    for $k$ = 1 to $K$
        → $\mu_k$ := average (mean) of points assigned to cluster $k$
}
```

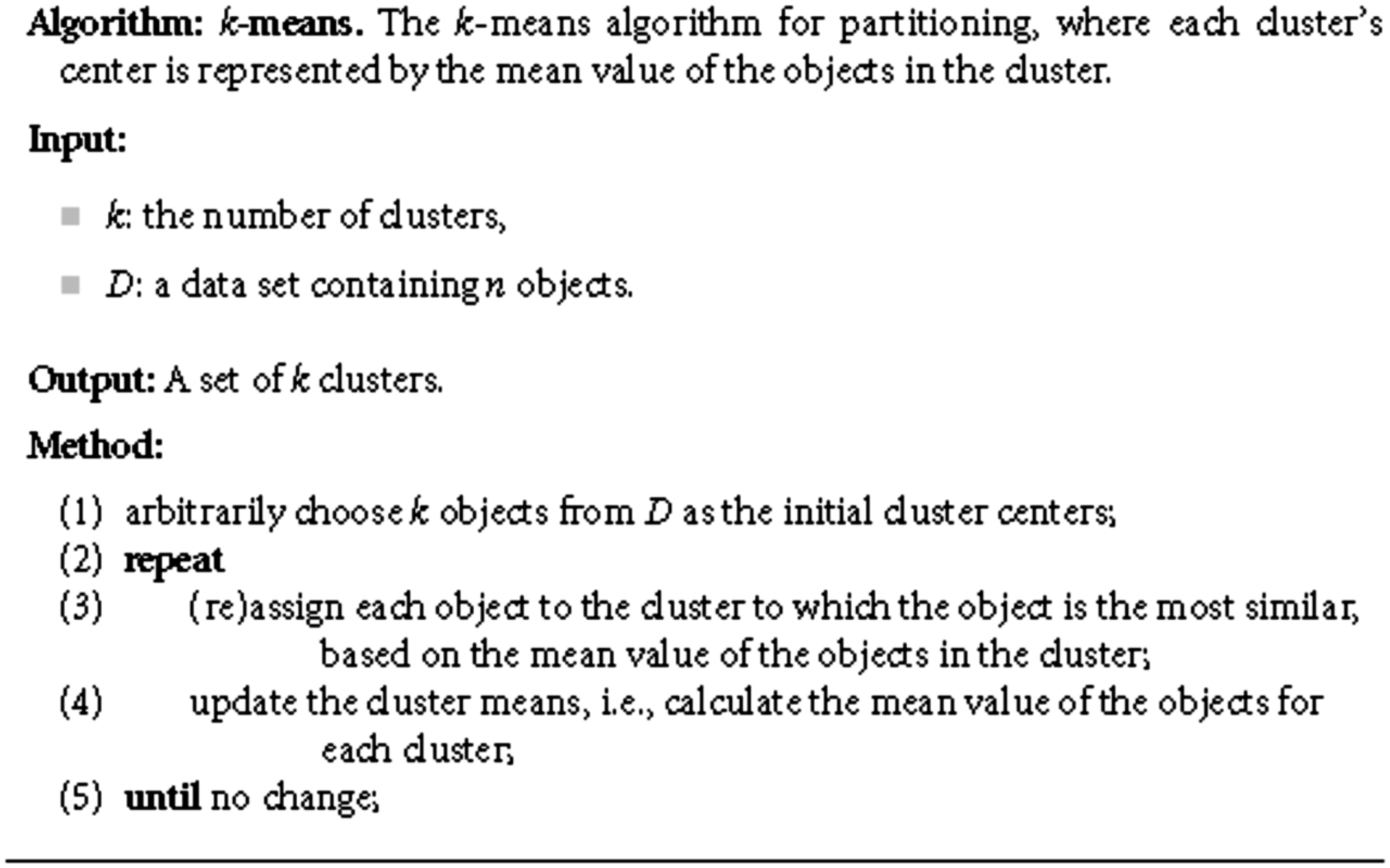
$\min_k \|x^{(i)} - \mu_k\|^2 \rightarrow c^{(i)}$

$x^{(1)}, x^{(5)}, x^{(6)}, x^{(10)}$ → $c^{(1)}=2,\ c^{(5)}=2,\ c^{(6)}=2,\ c^{(10)}=2$

$$\mu_2 = \frac{1}{4}\left[x^{(1)} + x^{(5)} + x^{(6)} + x^{(10)}\right] \in \mathbb{R}^n$$

Unsupervised Learning

- K-means Algorithm [1]



Algorithm: ***k*-means.** The k-means algorithm for partitioning, where each duster's center is represented by the mean value of the objects in the duster.

Input:

- k: the number of dusters,
- D: a data set containing n objects.

Output: A set of k clusters.

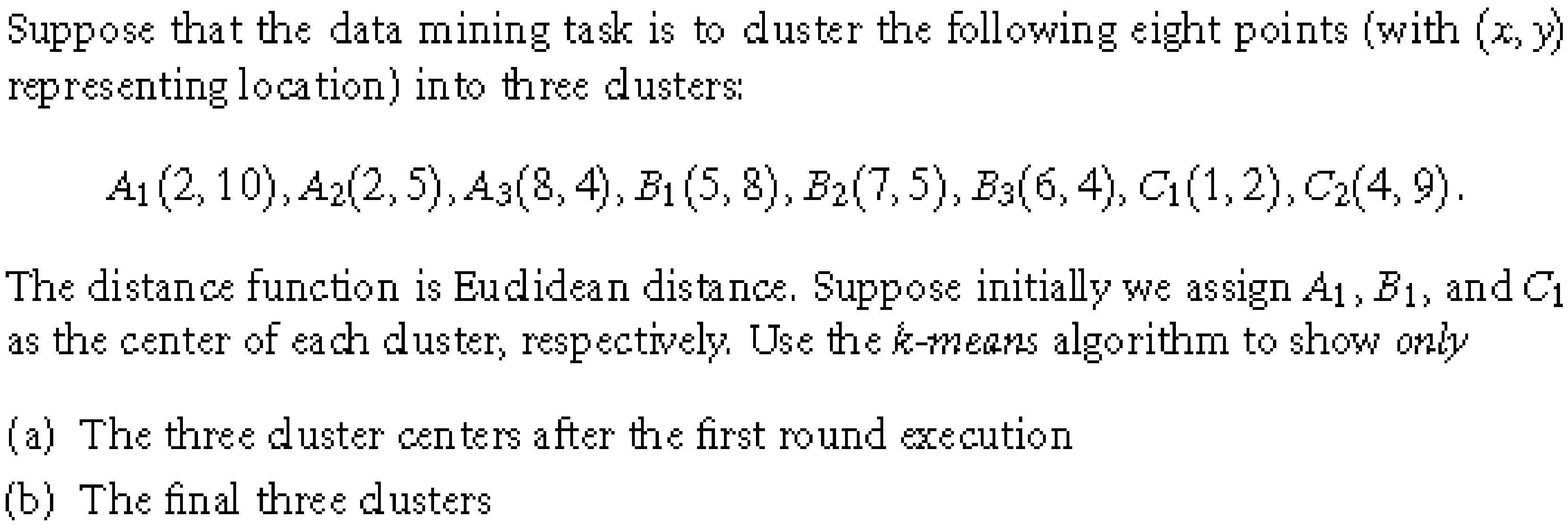
Method:

```
arbitrarily choose $k$ objects from $D$ as the initial duster centers;
repeat
      (re)assign each object to the duster to which the object is the most similar,
            based on the mean value of the objects in the duster;
      update the duster means, i.e., calculate the mean value of the objects for
            each duster;
until no change;
```

Figure 7.2 The k-means partitioning algorithm.

Unsupervised Learning

- K-means Algorithm [1]



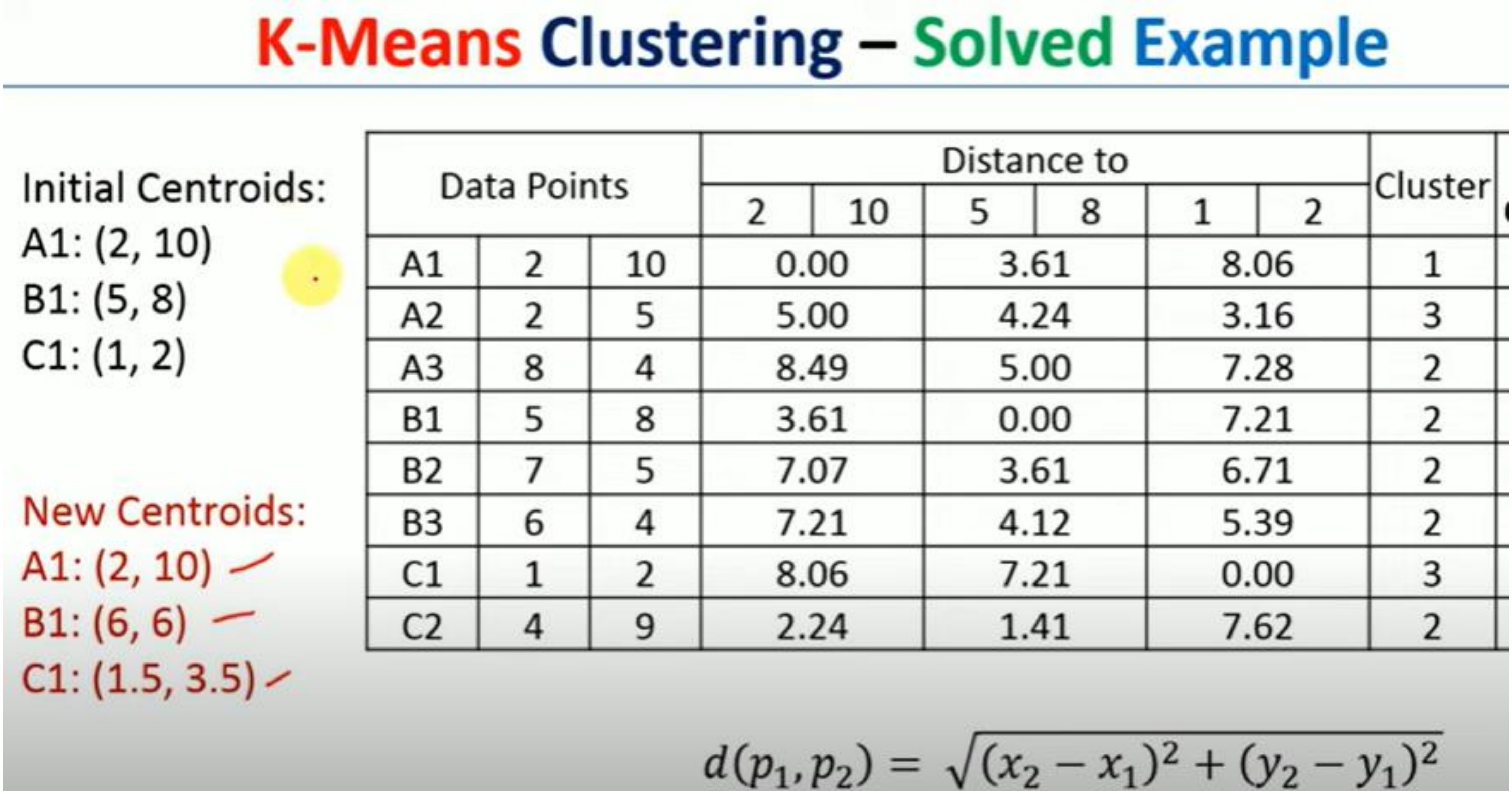
Suppose that the data mining task is to cluster the following eight points (with (x, y) representing location) into three clusters:

$$A_1(2,10), A_2(2,5), A_3(8,4), B_1(5,8), B_2(7,5), B_3(6,4), C_1(1,2), C_2(4,9).$$

The distance function is Euclidean distance. Suppose initially we assign A_1, B_1, and C_1 as the center of each cluster, respectively. Use the *k-means* algorithm to show *only*

(a) The three cluster centers after the first round execution

(b) The final three clusters



K-Means Clustering – Solved Example

Initial Centroids:

A1: (2, 10)

B1: (5, 8)

C1: (1, 2)

New Centroids:

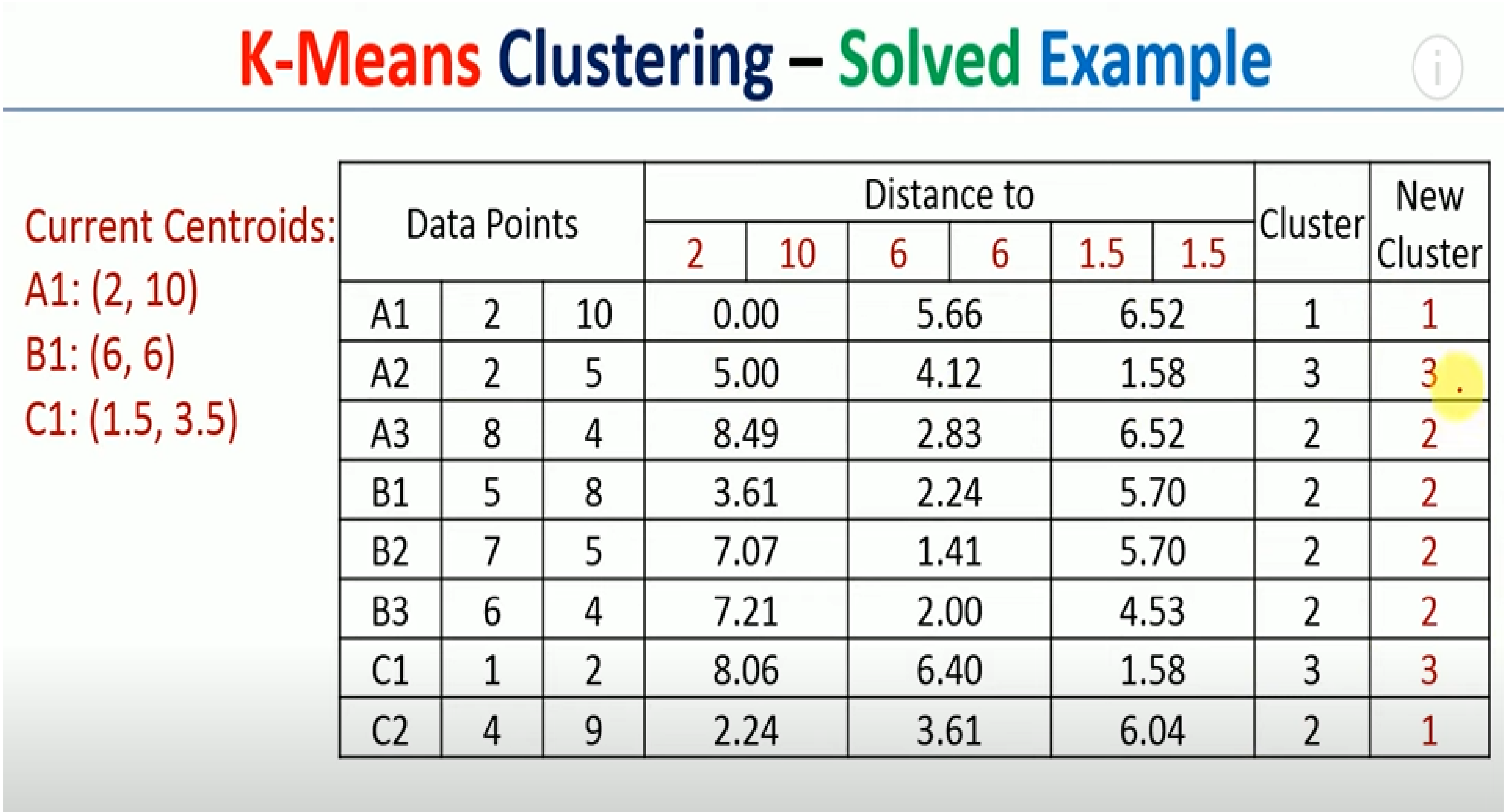
A1: (2, 10)

B1: (6, 6)

C1: (1.5, 3.5)

<table>
<tr><th colspan="3" rowspan="2">Data Points</th><th colspan="6">Distance to</th><th rowspan="2">Cluster</th></tr>
<tr><th>2</th><th>10</th><th>5</th><th>8</th><th>1</th><th>2</th></tr>
<tr><td>A1</td><td>2</td><td>10</td><td colspan="2">0.00</td><td colspan="2">3.61</td><td colspan="2">8.06</td><td>1</td></tr>
<tr><td>A2</td><td>2</td><td>5</td><td colspan="2">5.00</td><td colspan="2">4.24</td><td colspan="2">3.16</td><td>3</td></tr>
<tr><td>A3</td><td>8</td><td>4</td><td colspan="2">8.49</td><td colspan="2">5.00</td><td colspan="2">7.28</td><td>2</td></tr>
<tr><td>B1</td><td>5</td><td>8</td><td colspan="2">3.61</td><td colspan="2">0.00</td><td colspan="2">7.21</td><td>2</td></tr>
<tr><td>B2</td><td>7</td><td>5</td><td colspan="2">7.07</td><td colspan="2">3.61</td><td colspan="2">6.71</td><td>2</td></tr>
<tr><td>B3</td><td>6</td><td>4</td><td colspan="2">7.21</td><td colspan="2">4.12</td><td colspan="2">5.39</td><td>2</td></tr>
<tr><td>C1</td><td>1</td><td>2</td><td colspan="2">8.06</td><td colspan="2">7.21</td><td colspan="2">0.00</td><td>3</td></tr>
<tr><td>C2</td><td>4</td><td>9</td><td colspan="2">2.24</td><td colspan="2">1.41</td><td colspan="2">7.62</td><td>2</td></tr>
</table>

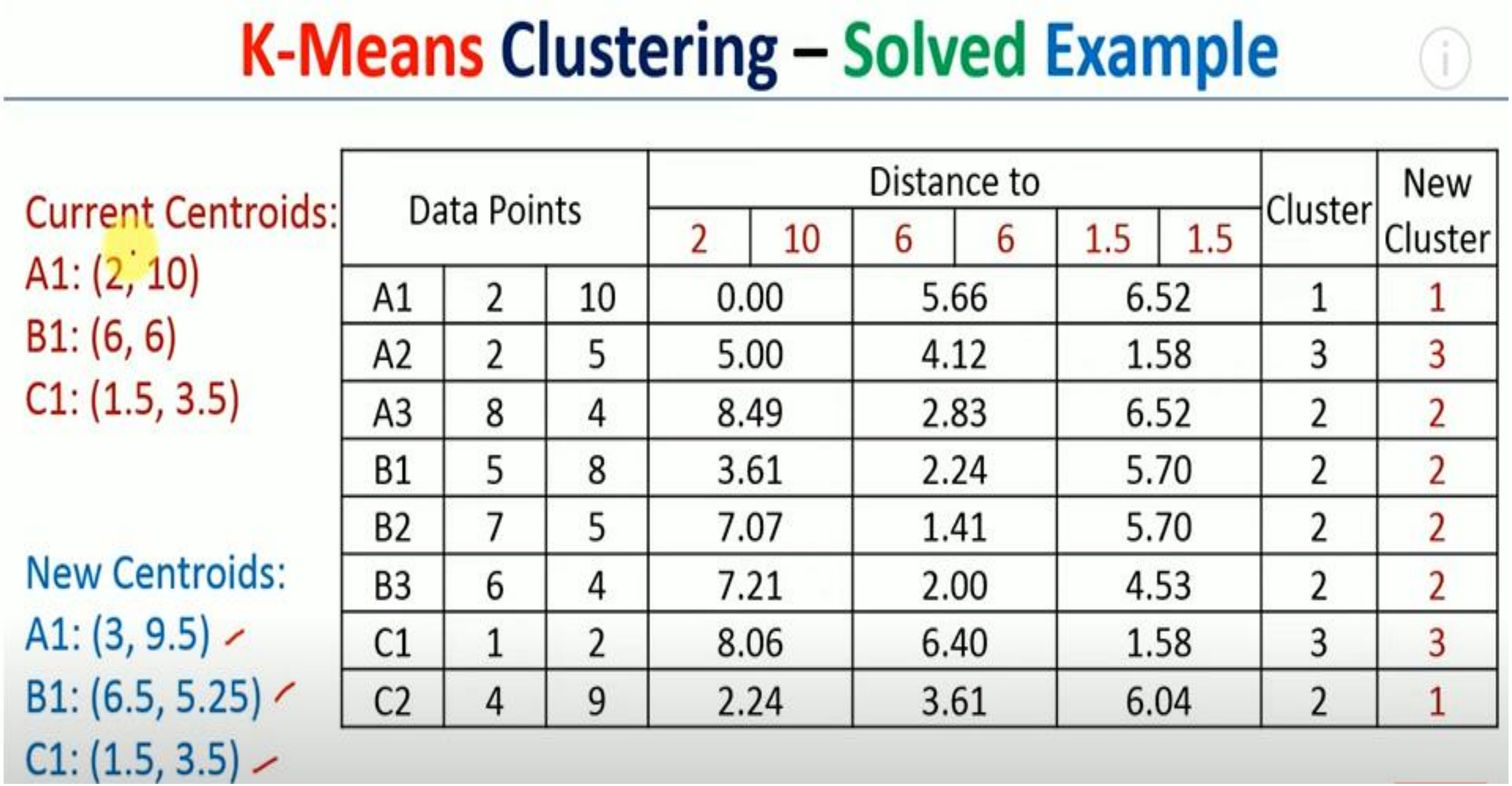
$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$



K-Means Clustering – Solved Example

Current Centroids:
A1: (2, 10)
B1: (6, 6)
C1: (1.5, 3.5)

Data Points			Distance to						Cluster	New Cluster
			2	10	6	6	1.5	1.5		
A1	2	10	0.00		5.66		6.52		1	1
A2	2	5	5.00		4.12		1.58		3	3
A3	8	4	8.49		2.83		6.52		2	2
B1	5	8	3.61		2.24		5.70		2	2
B2	7	5	7.07		1.41		5.70		2	2
B3	6	4	7.21		2.00		4.53		2	2
C1	1	2	8.06		6.40		1.58		3	3
C2	4	9	2.24		3.61		6.04		2	1



K-Means Clustering – Solved Example

Current Centroids:

A1: (2, 10)

B1: (6, 6)

C1: (1.5, 3.5)

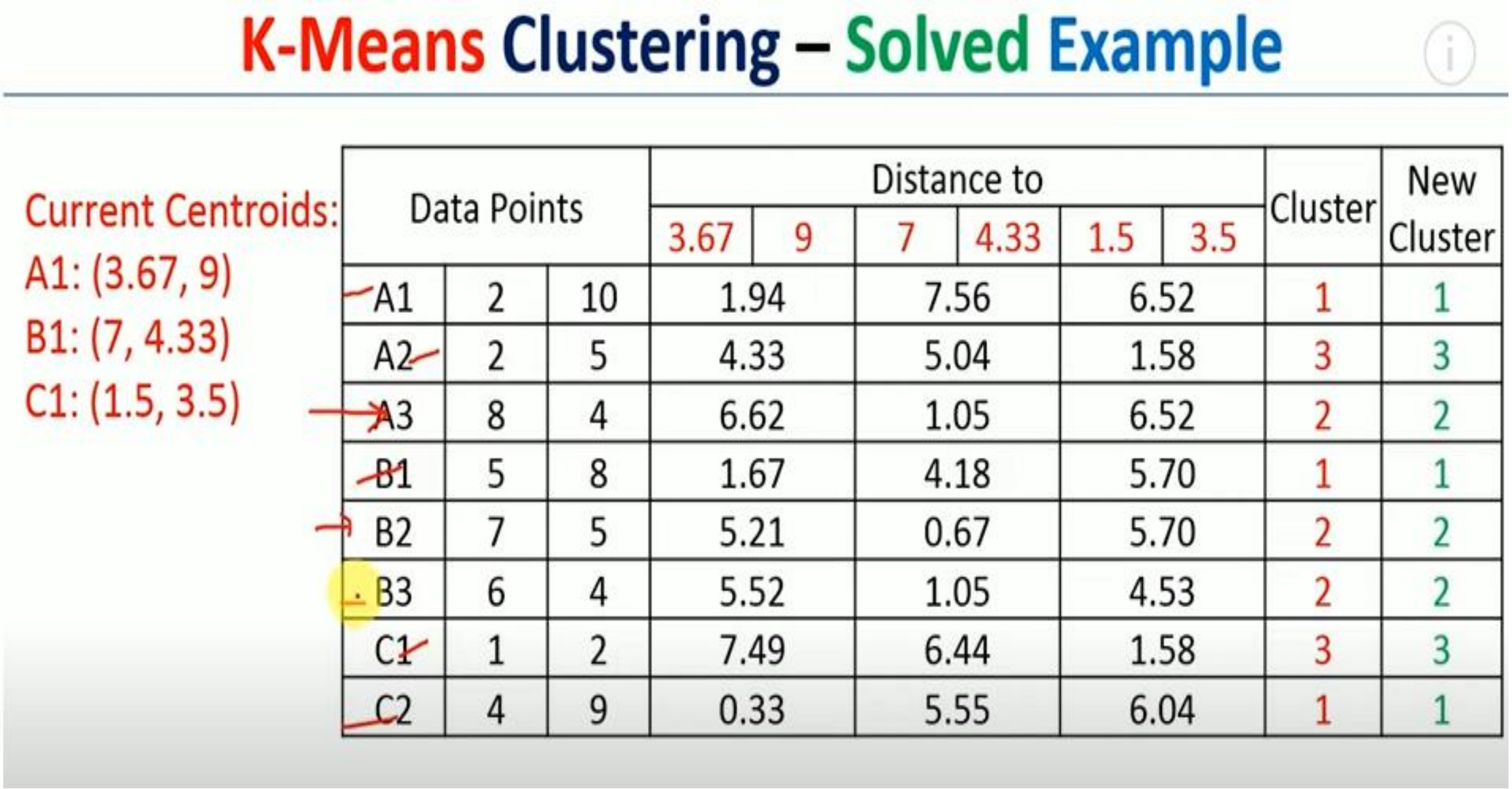
New Centroids:

A1: (3, 9.5)

B1: (6.5, 5.25)

C1: (1.5, 3.5)

Data Points			Distance to						Cluster	New Cluster
			2	10	6	6	1.5	1.5		
A1	2	10	0.00		5.66		6.52		1	1
A2	2	5	5.00		4.12		1.58		3	3
A3	8	4	8.49		2.83		6.52		2	2
B1	5	8	3.61		2.24		5.70		2	2
B2	7	5	7.07		1.41		5.70		2	2
B3	6	4	7.21		2.00		4.53		2	2
C1	1	2	8.06		6.40		1.58		3	3
C2	4	9	2.24		3.61		6.04		2	1



K-Means Clustering – Solved Example

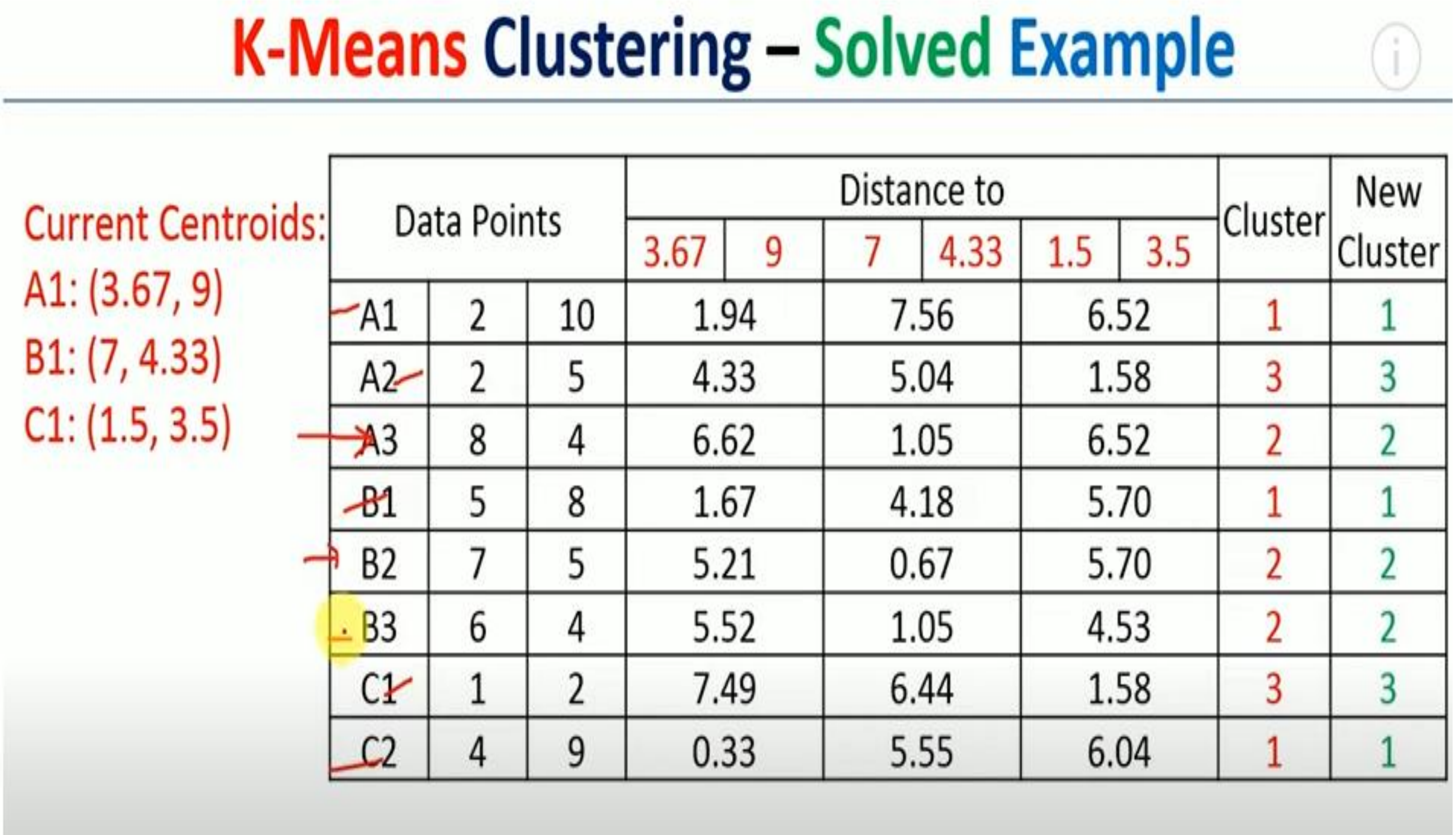
Current Centroids:

A1: (3.67, 9)

B1: (7, 4.33)

C1: (1.5, 3.5)

Data Points			Distance to			Cluster	New Cluster
			3.67 9	7 4.33	1.5 3.5		
A1	2	10	1.94	7.56	6.52	1	1
A2	2	5	4.33	5.04	1.58	3	3
A3	8	4	6.62	1.05	6.52	2	2
B1	5	8	1.67	4.18	5.70	1	1
B2	7	5	5.21	0.67	5.70	2	2
B3	6	4	5.52	1.05	4.53	2	2
C1	1	2	7.49	6.44	1.58	3	3
C2	4	9	0.33	5.55	6.04	1	1



K-Means Clustering – Solved Example

Current Centroids:

A1: (3.67, 9)

B1: (7, 4.33)

C1: (1.5, 3.5)

Data Points			Distance to						Cluster	New Cluster
			3.67	9	7	4.33	1.5	3.5		
A1	2	10	1.94		7.56		6.52		1	1
A2	2	5	4.33		5.04		1.58		3	3
A3	8	4	6.62		1.05		6.52		2	2
B1	5	8	1.67		4.18		5.70		1	1
B2	7	5	5.21		0.67		5.70		2	2
B3	6	4	5.52		1.05		4.53		2	2
C1	1	2	7.49		6.44		1.58		3	3
C2	4	9	0.33		5.55		6.04		1	1