Data Mining: Clustering

Cluster Analysis

- What is Cluster Analysis?
- Types of Data in Cluster Analysis
- A Categorization of Major Clustering Methods
- Partitioning Methods
- Hierarchical Methods
- Grid-Based Methods
- Model-Based Clustering Methods
- Outlier Analysis

What is Cluster Analysis?

- Cluster: a collection of data objects
  - Similar to one another within the same cluster
  - Dissimilar to the objects in other clusters
- Cluster analysis
  - Grouping a set of data objects into clusters
- Clustering is **unsupervised classification**: no predefined classes
- Typical applications
  - As a **stand-alone tool** to get insight into data distribution
  - As a **preprocessing step** for other algorithms

General Applications of Clustering

- Pattern Recognition
- Spatial Data Analysis
  - create thematic maps in GIS by clustering feature spaces
  - detect spatial clusters and explain them in spatial data mining
- Image Processing
- Economic Science (especially market research)
- WWW
  - Document classification
  - Cluster Weblog data to discover groups of similar access patterns

Examples of Clustering Applications

- **Marketing**: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- **Land use:** Identification of areas of similar land use in an earth observation database
- **Insurance:** Identifying groups of motor insurance policy holders with a high average claim cost
- **City-planning:** Identifying groups of houses according to their house type, value, and geographical location
- **Earth-quake studies:** Observed earth quake epicenters should be clustered along continent faults

What is not Cluster Analysis?

- Supervised classification
  - Have class label information
- Simple segmentation
  - Dividing students into different registration groups alphabetically, by last name
- Results of a query
  - Groupings are a result of an external specification
- Graph partitioning
  - Some mutual relevance and synergy, but areas are not identical

What Is Good Clustering?

- A good clustering method will produce high quality clusters with
  - **high intra-class** similarity
  - **low inter-class** similarity
- The quality of a clustering result depends on both the similarity measure used by the method and its implementation.
- The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns.

Requirements of Clustering in Data Mining

- Scalability
- Ability to deal with different types of attributes
- Discovery of clusters with arbitrary shape
- Minimal requirements for domain knowledge to determine input parameters
- Able to deal with noise and outliers
- Insensitive to order of input records
- High dimensionality
- Incorporation of user-specified constraints
- Interpretability and usability

Data Structures

- Data matrix
  - (two modes)

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

- Dissimilarity matrix
  - (one mode)

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ : & : & : & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

Measure the Quality of Clustering

- Dissimilarity/Similarity metric: Similarity is expressed in terms of a distance function, which is typically metric: ***d*(*i, j*)**
- There is a separate "quality" function that measures the "goodness" of a cluster.
- The definitions of distance functions are usually very different for interval-scaled, boolean, categorical, ordinal and ratio variables.
- Weights should be associated with different variables based on applications and data semantics.
- It is hard to define "similar enough" or "good enough"
  - the answer is typically highly subjective.

Type of data in clustering analysis

- **Interval-scaled variables:**
- **Binary variables:**
- **Nominal, ordinal, and ratio variables:**
- **Variables of mixed types:**

Interval-valued variables

- Standardize data
  - Calculate the mean absolute deviation:

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \ldots + |x_{nf} - m_f|)$$

where $m_f = \frac{1}{n}(x_{1f} + x_{2f} + \ldots + x_{nf})$.

  - Calculate the standardized measurement (*z-score*)

$$z_{if} = \frac{x_{if} - m_f}{s_f}$$

- Using mean absolute deviation is more robust than using standard deviation

Similarity and Dissimilarity Between Objects

- Distances are normally used to measure the similarity or dissimilarity between two data objects

- Some popular ones include: *Minkowski distance*:

$$d(i,j) = \sqrt[q]{(|x_{i_1} - x_{j_1}|^q + |x_{i_2} - x_{j_2}|^q + \ldots + |x_{i_p} - x_{j_p}|^q)}$$

  where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ are two p-dimensional data objects, and q is a positive integer

- If $q = 1$, d is Manhattan distance $d(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \ldots + |x_{i_p} - x_{j_p}|$
- *If* $q = 2$, d is Euclidean distance: $d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \ldots + |x_{i_p} - x_{j_p}|^2)}$

Binary Variables

- A contingency table for binary data

		Object j		
		1	0	*sum*
	1	a	b	$a+b$
Object i	0	c	d	$c+d$
	sum	$a+c$	$b+d$	p

- Simple matching coefficient (invariant, if the binary variable is *symmetric*):

$$d(i,j) = \frac{b+c}{a+b+c+d}$$

- Jaccard coefficient (noninvariant if the binary variable is *asymmetric*):

$$d(i,j) = \frac{b+c}{a+b+c}$$

Dissimilarity between Binary Variables

- Example

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

  - Gender is a symmetric attribute
  - The remaining attributes are asymmetric binary
  - Let the values Y and P be set to 1, and the value N be set to 0

$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(jack, jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(jim, mary) = \frac{1+2}{1+1+2} = 0.75$$

Nominal Variables

- A generalization of the binary variable in that it can take more than 2 states, e.g., red, yellow, blue, green
- Method 1: Simple matching
  - *m*: # of matches, *p*: total # of variables

$$d(i, j) = \frac{p - m}{p}$$

- Method 2: use a large number of binary variables
  - creating a new binary variable for each of the *M* nominal states

Major Clustering Approaches

- **Partitioning algorithms:** Construct various partitions and then evaluate them by some criterion
- **Hierarchy algorithms:** Create a hierarchical decomposition of the set of data (or objects) using some criterion
- **Density-based:** based on connectivity and density functions
- **Grid-based:** based on a multiple-level granularity structure
- **Model-based:** A model is hypothesized for each of the clusters and the idea is to find the best fit of that model to each other

Major Clustering Approaches

- Important distinction between **partitional** and **hierarchical** sets of clusters
- **Partitional Clustering**
    - A division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset
- **Hierarchical clustering**
    - A set of nested clusters organized as a hierarchical tree

Partitioning Algorithms: Basic Concept

- <u>Partitioning method:</u> Construct a partition of a database ***D*** of ***n*** objects into a set of ***k*** clusters
- Given a *k*, find a partition of *k clusters* that optimizes the chosen partitioning criterion
  - Global optimal: exhaustively enumerate all partitions
  - Heuristic methods: *k-means* and *k-medoids* algorithms
  - <u>*k-means*</u> (MacQueen’67): Each cluster is represented by the center of the cluster
  - <u>*k-medoids*</u> or PAM (Partition around medoids) (Kaufman & Rousseeuw’87): Each cluster is represented by one of the objects in the cluster

K-Means Clustering

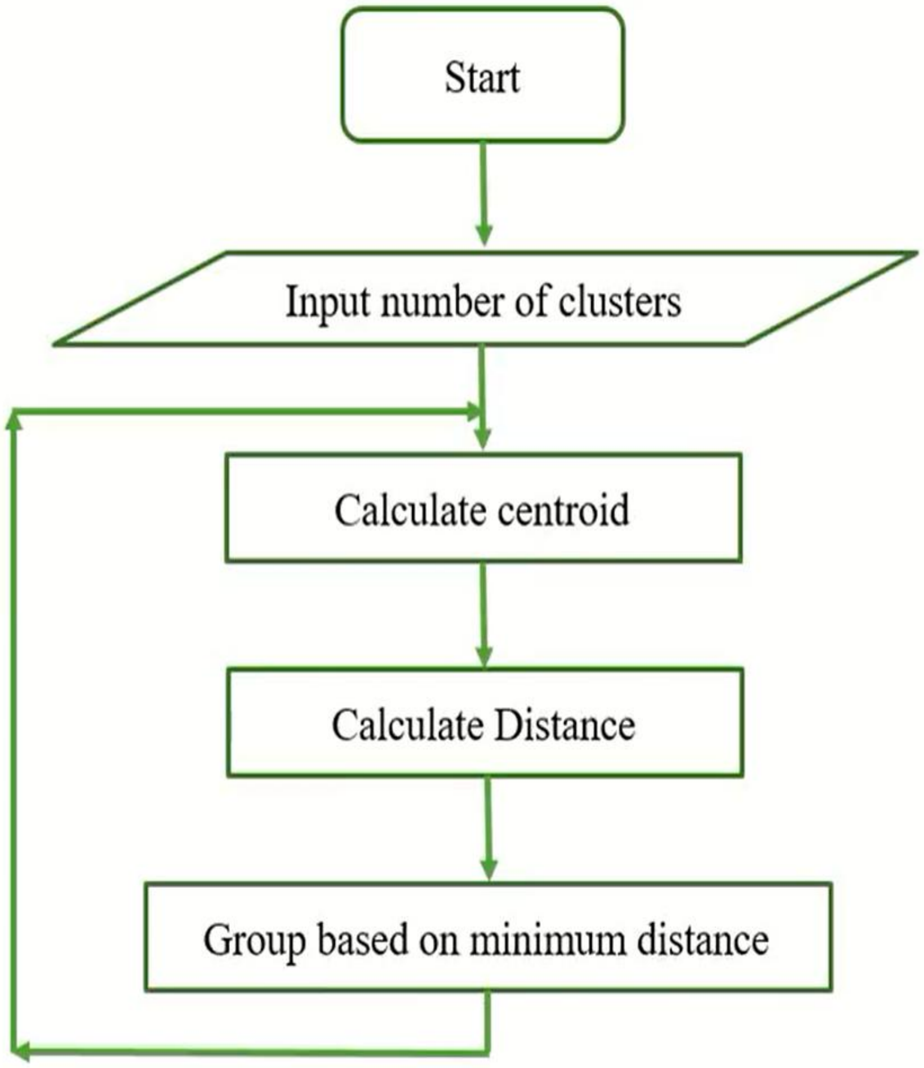
- Simple Clustering: K-means
- Given *k*, the *k-means* algorithm consists of four steps:

  (Basic version works with numeric data only)

  1) Select initial centroids at random - Pick a number (K) of cluster centers - centroids (at random)
  2) Assign every item to its nearest cluster center (e.g. using Euclidean distance)
  3) Move each cluster center to the mean of its assigned items
  4) Repeat steps 2,3 until convergence (change in cluster assignments less than a threshold)

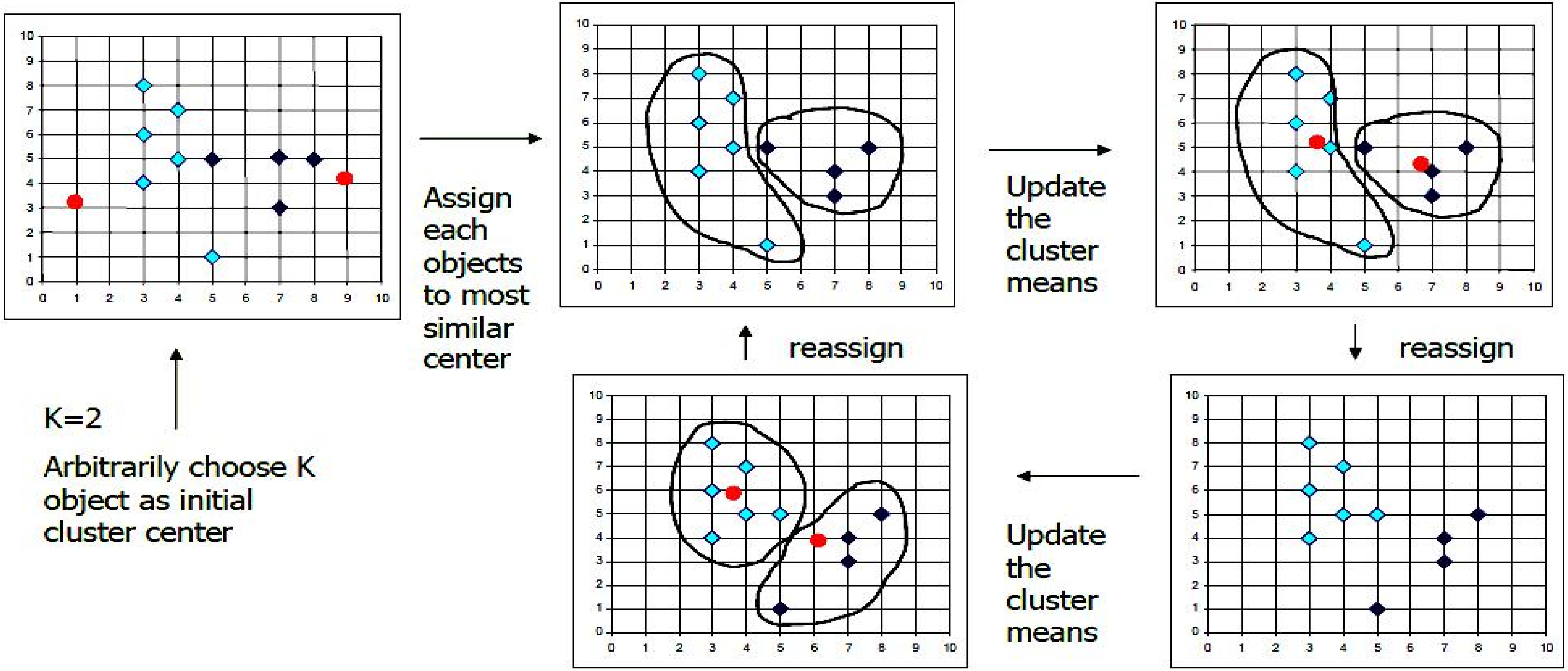
K-means Algoritms

- Initialization
  - Arbitrarily choose k objects as the initial cluster centers (centroids)
- Iteration until no change
  - For each object O_i
    - Calculate the distances between O_i and the k centroids
    - (Re)assign O_i to the cluster whose centroid is the closest to O_i
  - Update the cluster centroids based on current assignment

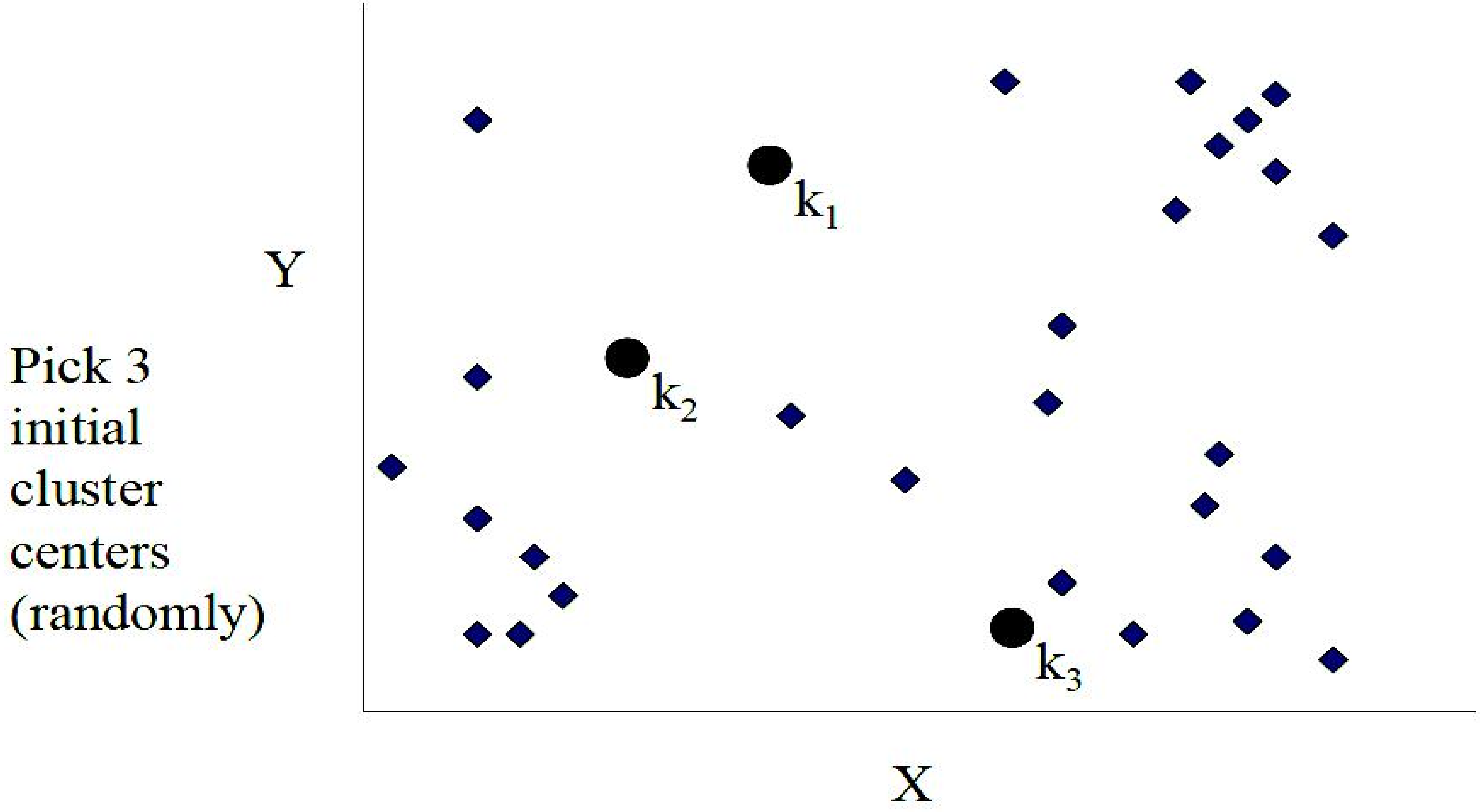


Illustrating K-Means

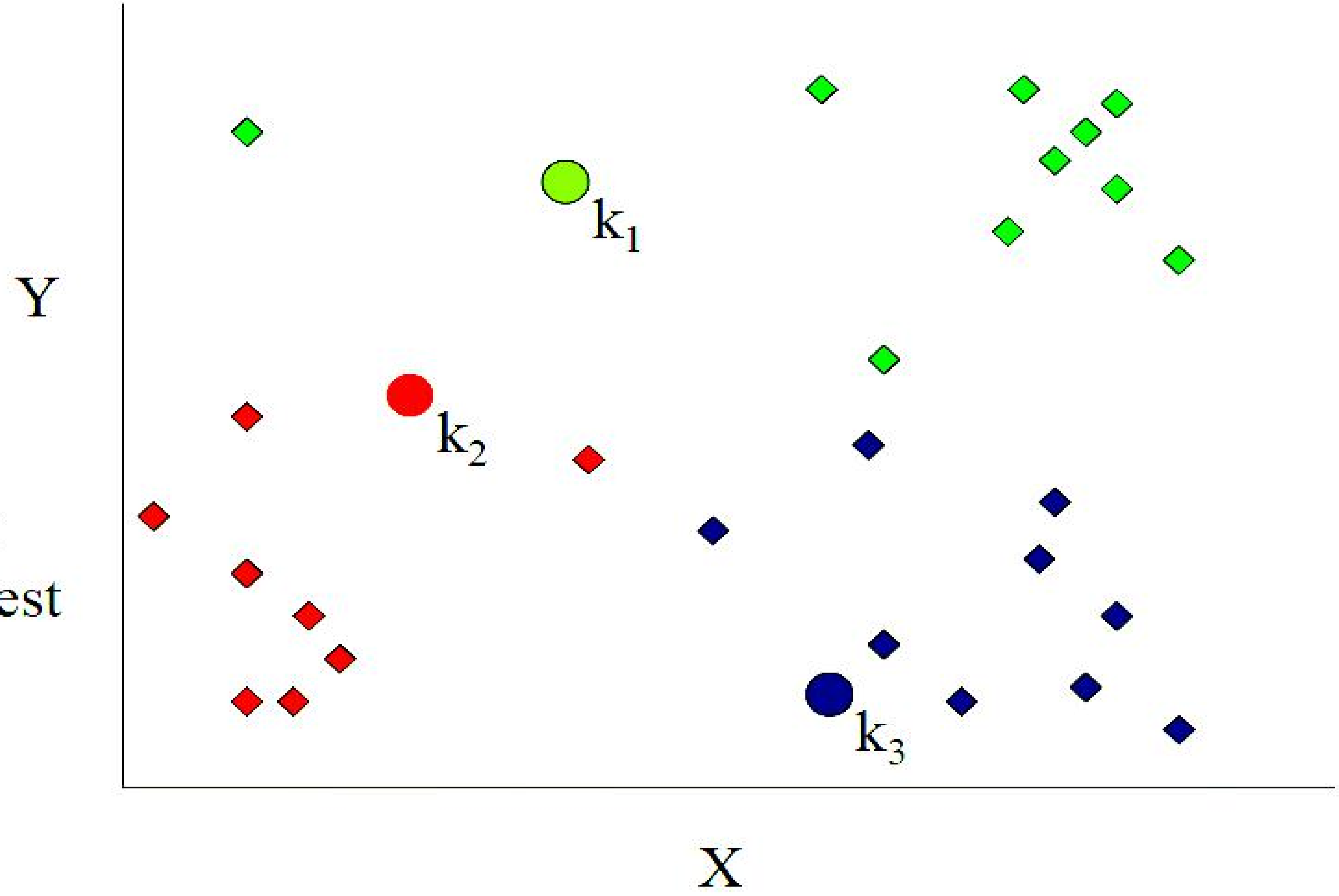
- Example



K-means example, step 1

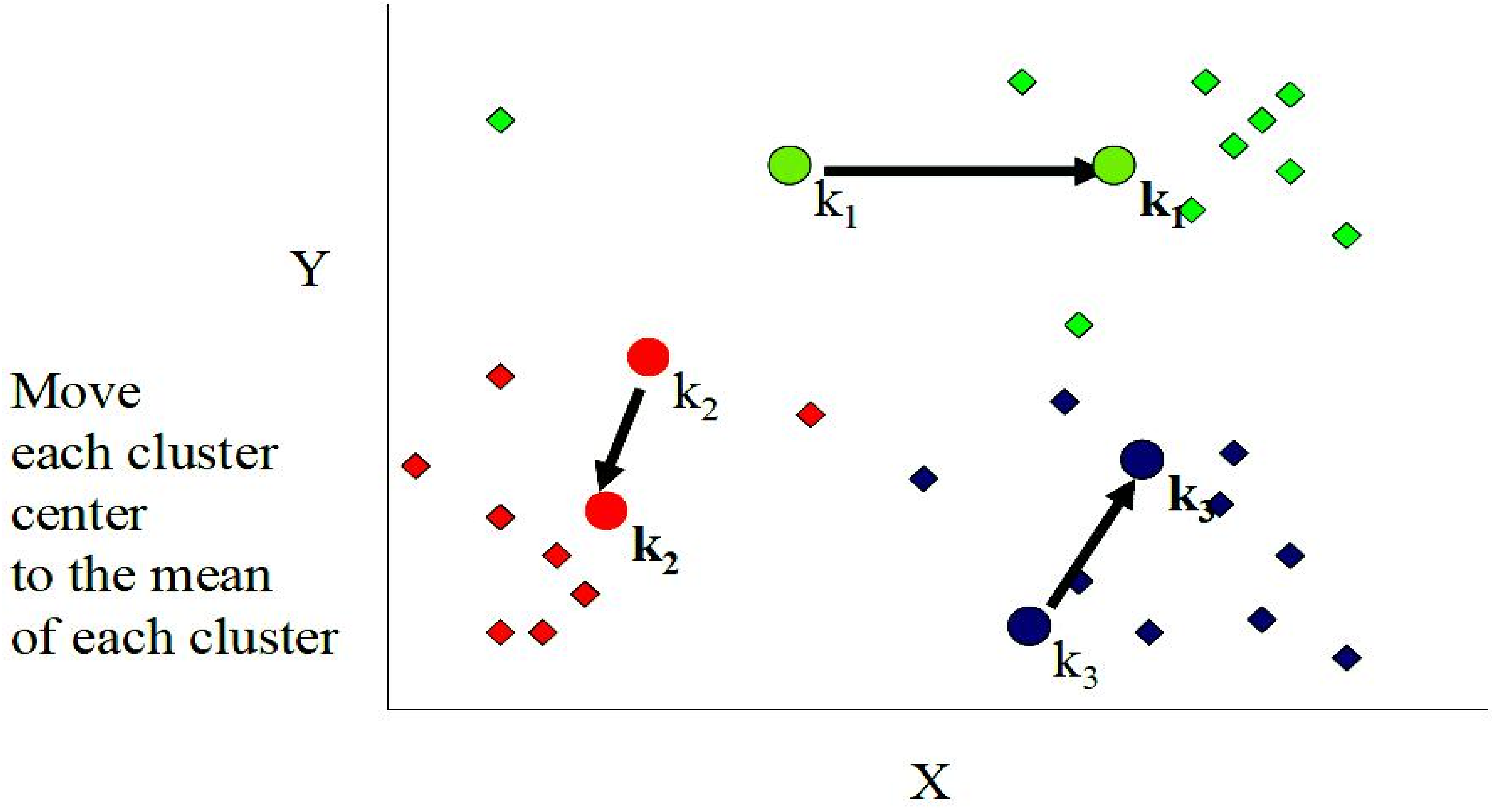


K-means example, step 2



Assign each point to the closest cluster center

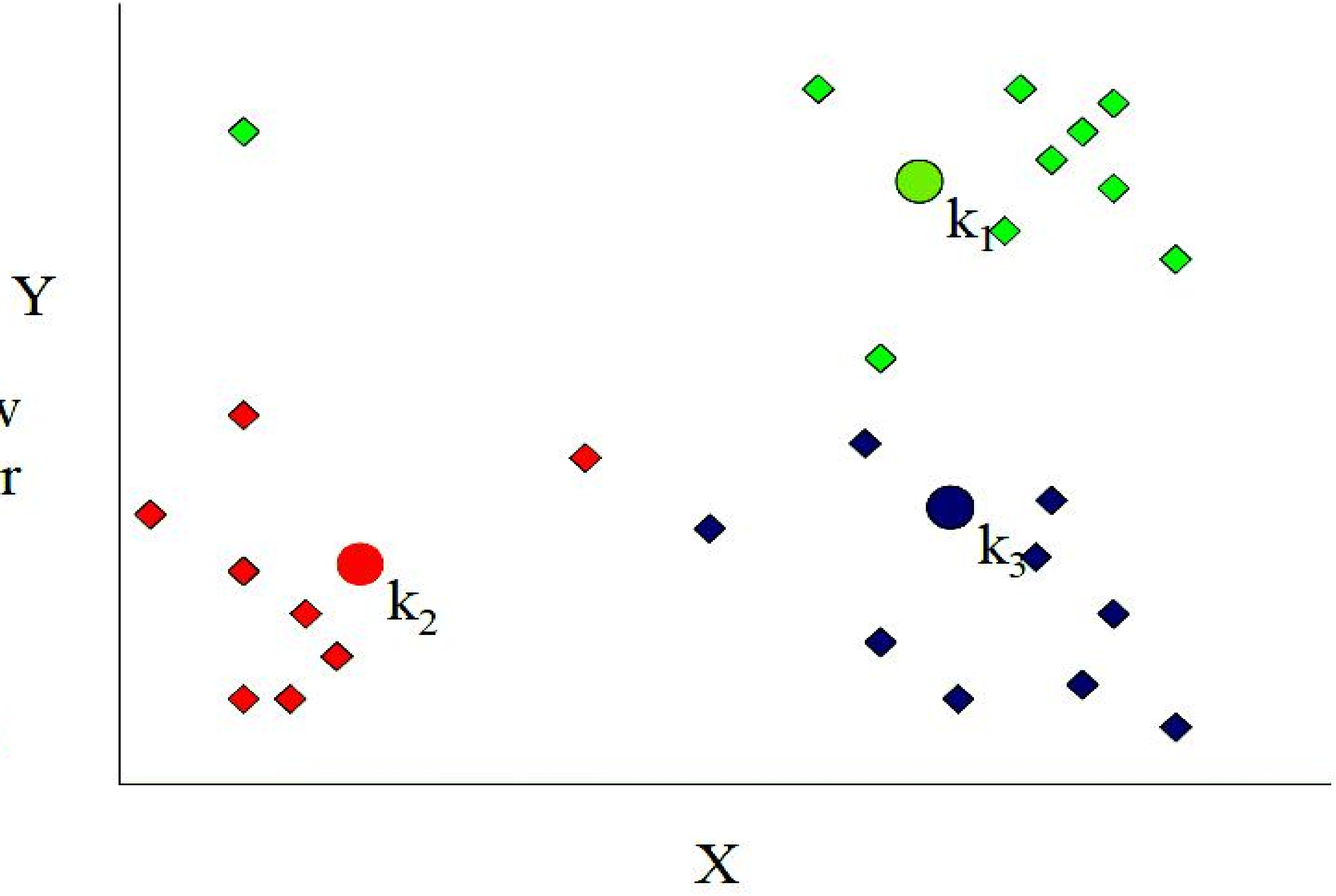
K-means example, step 3



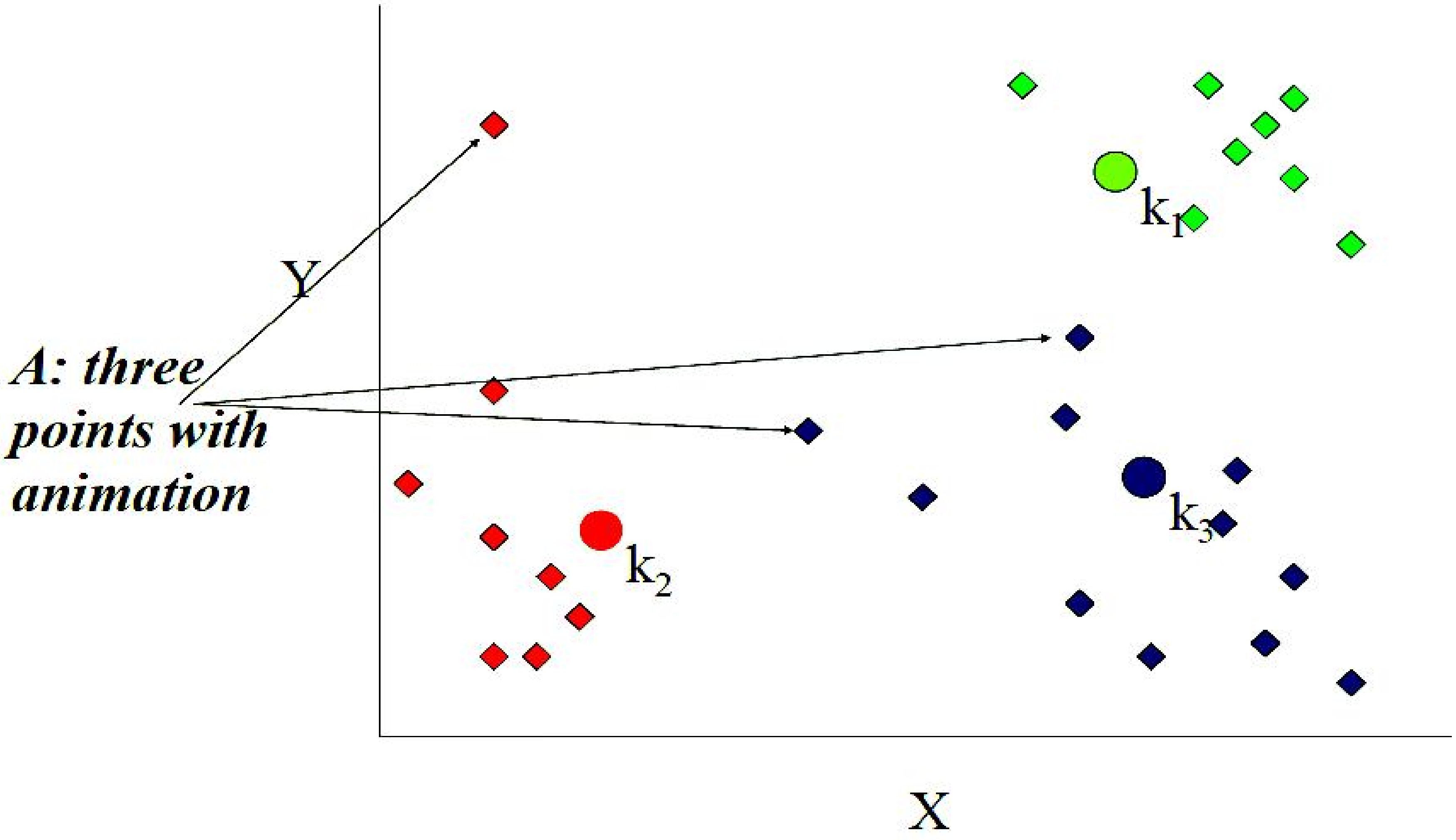
K-means example, step 4

Reassign points closest to a different new cluster center

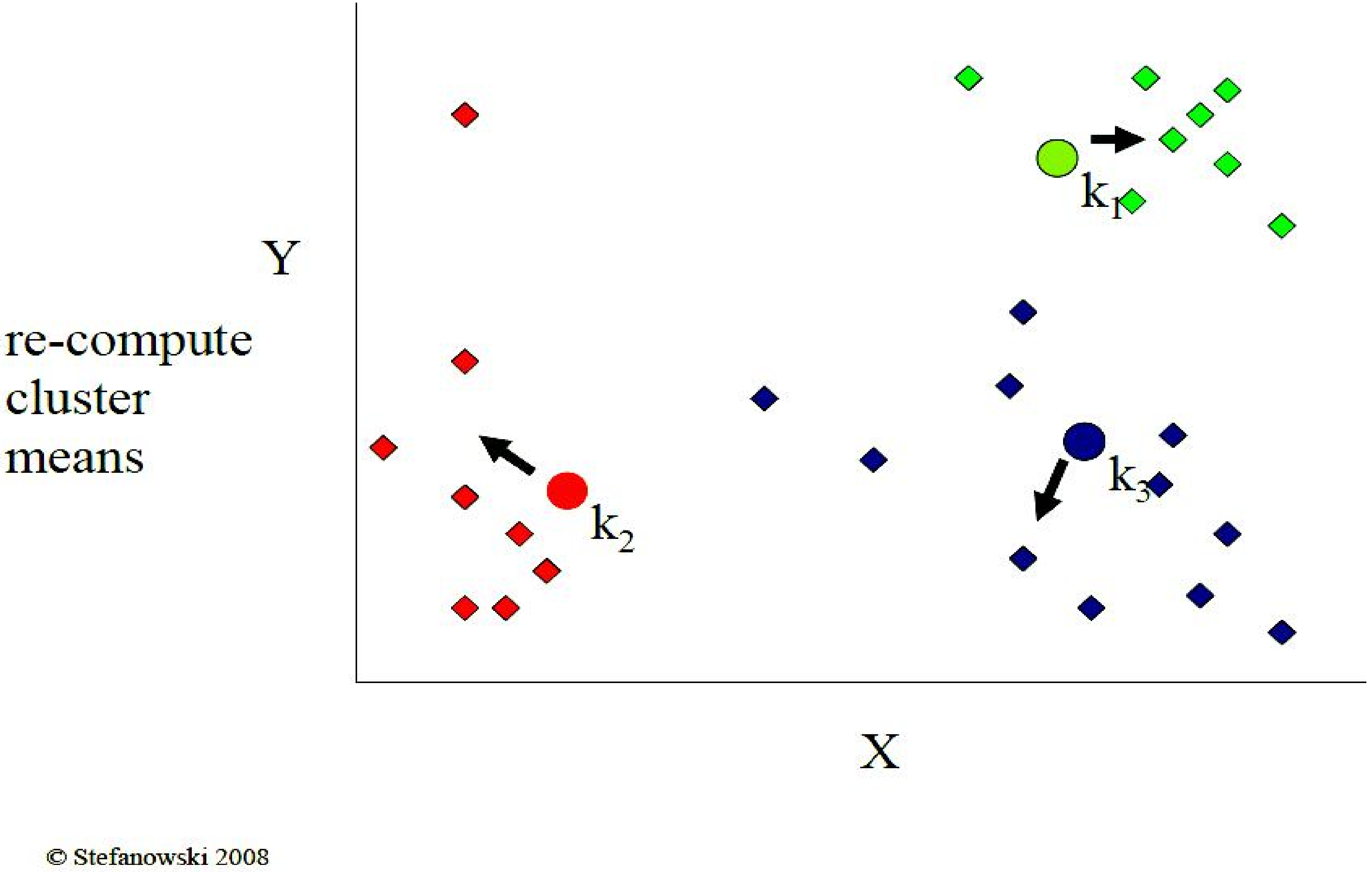
Q: Which points are reassigned?



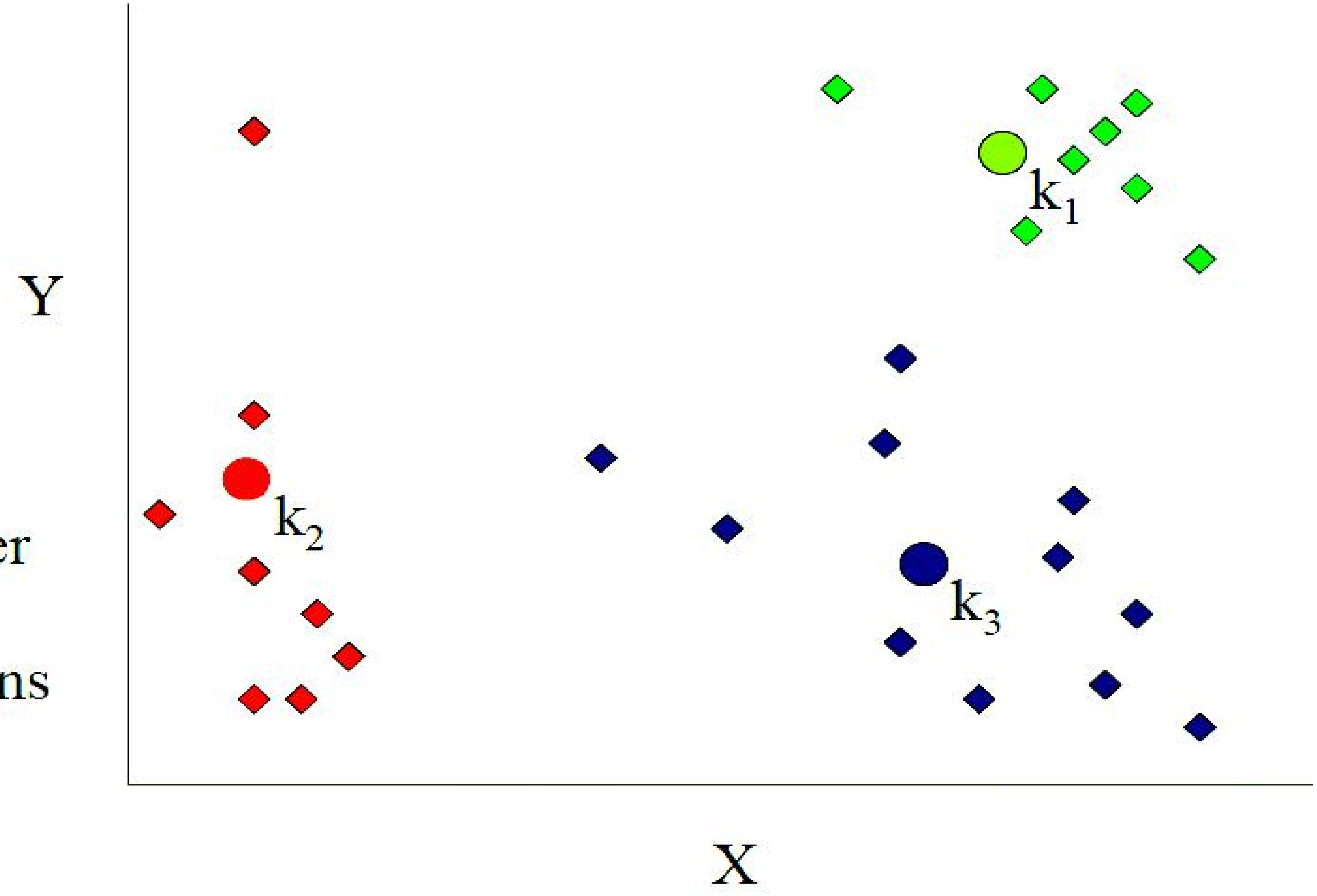
K-means example, step 4a



K-means example, step 4b



K-means example, step 5



move cluster
centers to
cluster means

Example:

- Apply K-mean clustering for the following data sets for two clusters. Tabulate all the assignments.

Sample No	X	Y
1	185	72
2	170	56
3	168	60
4	179	68
5	182	72
6	188	77

Step-1:
Given k=2, Initial Centroid

Initial Centroid

Cluster	X	Y
k1	185	72
k2	170	56

Calculate Euclidean distance using the given equation.

$$\text{Distance } [(\text{x,y}), (\text{a,b})] = \sqrt{(x-a)^2+(x-b)^2}$$

$$\text{Cluster 1 } (185{,}72) = \sqrt{(185-185)^2+(72-72)^2} = 0$$

$$\text{Distance from Cluster 2} = \sqrt{(170-185)^2+(56-72)^2}$$

$$(170{,}56) = \sqrt{(-15)^2+(-16)^2}$$

$$= \sqrt{255+256}$$

$$= \sqrt{481}$$

$$= 21.93$$

$$\text{Cluster 2 } (170{,}56) = \sqrt{(170-170)^2+(56-56)^2} = 0$$

Step-2: New Centroid

Cluster	Centroid		
	X	Y	ASSIGNMENT
k1	0	21.93	1
k2	21.93	0	2

Step-2: Distance calculation

Sample No	X	Y
1	185	72
2	170	56
3	168	60
4	179	68
5	182	72
6	188	77

Calculate Euclidean distance for the next dataset (168,60)

Distance [(x,y), (a,b)] = $\sqrt{(x-a)^2+(x-b)^2}$

Distance from Cluster 1 = $\sqrt{(168-185)^2+(60-72)^2}$

(185,72) $= \sqrt{(-17)^2+(-12)^2}$

$= \sqrt{283+144}$

$= \sqrt{433}$

$= 20.808$

Distance from Cluster 2 = $\sqrt{(168-170)^2+(60-56)^2}$

(170,56) $= \sqrt{(-2)^2+(-4)^2}$

$= \sqrt{4+16}$

$= \sqrt{20}$

$= 4.472$

Dataset	Euclidean Distance		
	Cluster 1	Cluster 2	ASSIGNMENT
(168,60)	20.808	4.472	2

Step-3: Update the cluster centroid

Cluster	X	Y
k1	185	72
k2	= (170 + 168)/ 2 = 169	= (60+56)/ 2 = 58

Step-4: Similarly process for next data set

Sample No	X	Y
1	185	72
2	170	56
3	168	60
4	179	68
5	182	72
6	188	77

Calculate Euclidean distance for the next dataset (179,68)

Distance from Cluster 1 $= \sqrt{(179-185)^2+(68-72)^2}$

(185,72)

$$= \sqrt{(-6)^2+(-4)^2}$$
$$= \sqrt{36+16}$$
$$= \sqrt{52}$$
$$= 7.211103$$

Distance from Cluster 2 $= \sqrt{(179-169)^2+(68-58)^2}$

(169,58)

$$= \sqrt{(10)^2+(10)^2}$$
$$= \sqrt{100+100}$$
$$= \sqrt{200}$$
$$= 14.14214$$

Dataset	Euclidean Distance		
	Cluster 1	Cluster 2	ASSIGNMENT
(179,68)	7.211103	14.14214	1

Step-5: Update the cluster centroid

Cluster	X	Y
k1	= 185+179/2 =182	= 72+68/2 =70
k2	169	58

Sample No	X	Y
1	185	72
2	170	56
3	168	60
4	179	68
5	182	72
6	188	77

Calculate Euclidean distance for the next dataset (182,72)

Distance from Cluster 1 $= \sqrt{(182-182)^2+(72-70)^2}$

(182,70)

$$= \sqrt{(0)^2+(2)^2}$$
$$= \sqrt{0+4}$$
$$= \sqrt{4}$$
$$= 2$$

Distance from Cluster 2 $= \sqrt{(182-169)^2+(72-58)^2}$

(169,58)

$$= \sqrt{(13)^2+(14)^2}$$
$$= \sqrt{169+196}$$
$$= \sqrt{365}$$
$$= 19.10$$

Dataset	Euclidean Distance		
	Cluster 1	Cluster 2	ASSIGNMENT
(182,72)	2	19.10	1

Step-6: Update the cluster centroid

Cluster	X	Y
k1	= 182+182/2 =182	= 70+72/2 = 71
k2	169	58

Sample No	X	Y
1	185	72
2	170	56
3	168	60
4	179	68
5	182	72
6	188	77

Calculate Euclidean distance for the next dataset (188,77)

Distance from Cluster 1 $= \sqrt{(188-182)^2+(77-71)^2}$

(182,71) $= \sqrt{(6)^2+(6)^2}$

$= \sqrt{36+36}$

$= \sqrt{72}$

$= 8.4852$

Distance from Cluster 2 $= \sqrt{(188-169)^2+(77-58)^2}$

(169,58) $= \sqrt{(19)^2+(19)^2}$

$= \sqrt{361+361}$

$= \sqrt{722}$

$= 26.87$

Dataset	Euclidean Distance		
	Cluster 1	Cluster 2	ASSIGNMENT
(188,77)	8.4852	26.87	1

Step-7: Update the cluster centroid

Cluster	X	Y
k1	= 182+188/2 = 185	= 71+77/2 = 74
k2	169	58

Final Assignment

Dataset No	X	Y	Assignment
1	185	72	1
2	170	56	2
3	168	60	2
4	179	68	1
5	182	72	1
6	188	77	1