# Data Mining
# 資料探勘

## 分群分析
## (Cluster Analysis)

1012DM04
MI4
Thu 9, 10 (16:10-18:00) B216

Min-Yuh Day
戴敏育
Assistant Professor
專任助理教授
Dept. of Information Management, Tamkang University
淡江大學 資訊管理學系
http://mail. tku.edu.tw/myday/
2013-03-21

# 課程大綱 (Syllabus)

| 週次 | 日期 | 內容 (Subject/Topics) |
|---|---|---|
| 1 | 102/02/21 | 資料探勘導論 (Introduction to Data Mining) |
| 2 | 102/02/28 | 和平紀念日 (放假一天)<br>(Peace Memorial Day) (No Classes) |
| 3 | 102/03/07 | 關連分析 (Association Analysis) |
| 4 | 102/03/14 | 分類與預測 (Classification and Prediction) |
| 5 | 102/03/21 | 分群分析 (Cluster Analysis) |
| 6 | 102/03/28 | SAS企業資料採礦實務<br>(Data Mining Using SAS Enterprise Miner) |
| 7 | 102/04/04 | 清明節、兒童節(放假一天)<br>(Children's Day, Tomb Sweeping Day)(No Classes) |
| 8 | 102/04/11 | 個案分析與實作一 (SAS EM 分群分析)：<br>Banking Segmentation (Cluster Analysis – K-Means using SAS EM) |

# 課程大綱 (Syllabus)

| 週次 | 日期 | 內容 (Subject/Topics) |
| --- | --- | --- |
| 9 | 102/04/18 | 期中報告 (Midterm Presentation) |
| 10 | 102/04/25 | 期中考試週 |
| 11 | 102/05/02 | 個案分析與實作二 (SAS EM 關連分析)：<br>Web Site Usage Associations ( Association Analysis using SAS EM) |
| 12 | 102/05/09 | 個案分析與實作三 (SAS EM 決策樹、模型評估)：<br>Enrollment Management Case Study<br>(Decision Tree, Model Evaluation using SAS EM) |
| 13 | 102/05/16 | 個案分析與實作四 (SAS EM 迴歸分析、類神經網路)：<br>Credit Risk Case Study<br>(Regression Analysis, Artificial Neural Network using SAS EM) |
| 14 | 102/05/23 | 期末專題報告 (Term Project Presentation) |
| 15 | 102/05/30 | 畢業考試週 |

# Outline

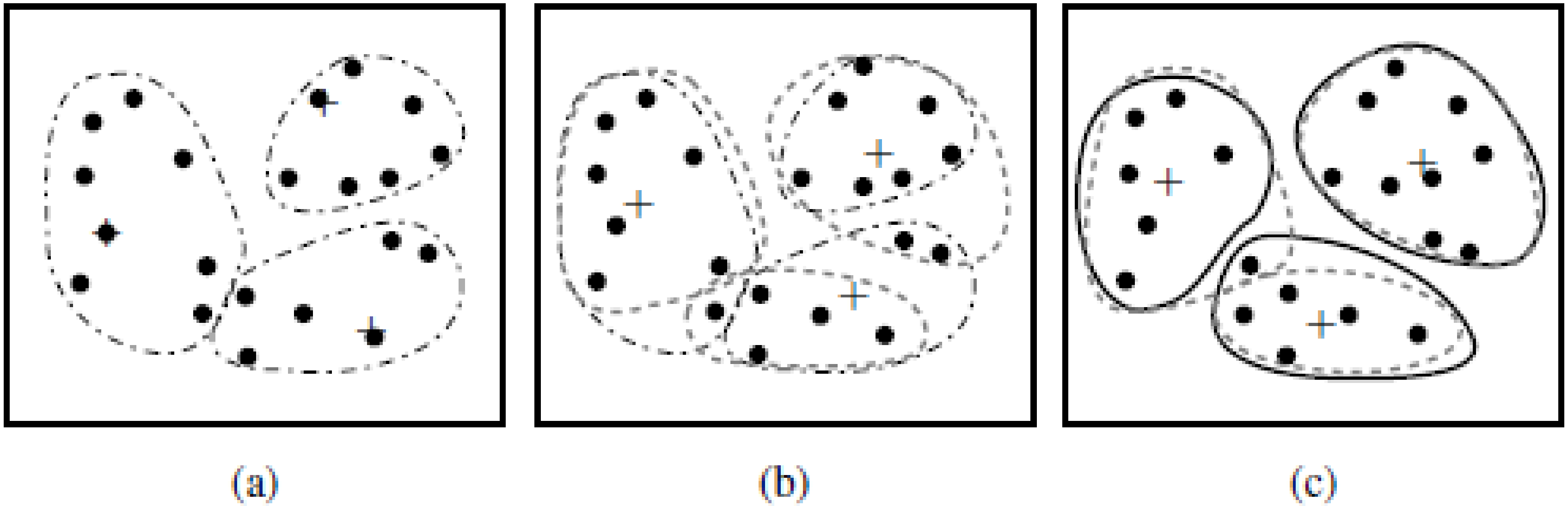
- Cluster Analysis
- *K-Means* Clustering

# Cluster Analysis

- Used for automatic identification of natural groupings of things
- Part of the machine-learning family
- Employ unsupervised learning
- Learns the clusters of things from past data, then assigns new instances
- There is not an output variable
- Also known as segmentation

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

# Cluster Analysis




Clustering of a set of objects based on the *k-means method.*
*(The mean of each cluster is* marked by a “+”.)

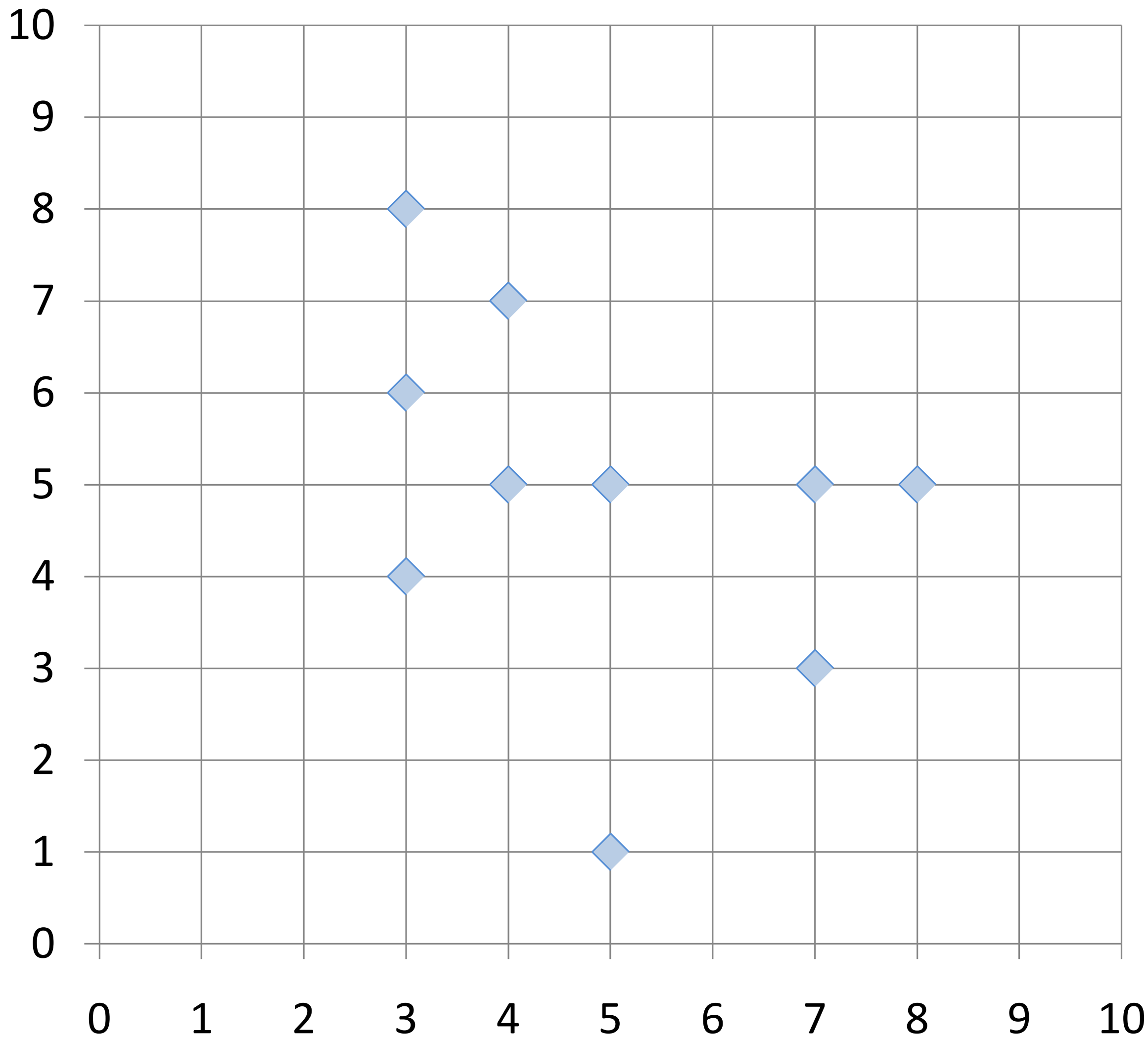
Source: Han & Kamber (2006)

# Cluster Analysis

- Clustering results may be used to
  - Identify natural groupings of customers
  - Identify rules for assigning new cases to classes for targeting/diagnostic purposes
  - Provide characterization, definition, labeling of populations
  - Decrease the size and complexity of problems for other data mining methods
  - Identify outliers in a specific domain (e.g., rare-event detection)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

# Example of Cluster Analysis




| Point | P | P(x,y) |
|---|---|---|
| p01 | a | (3, 4) |
| p02 | b | (3, 6) |
| p03 | c | (3, 8) |
| p04 | d | (4, 5) |
| p05 | e | (4, 7) |
| p06 | f | (5, 1) |
| p07 | g | (5, 5) |
| p08 | h | (7, 3) |
| p09 | i | (7, 5) |
| p10 | j | (8, 5) |

# Cluster Analysis for Data Mining

- Analysis methods
  - Statistical methods
    (including both hierarchical and nonhierarchical), such as *k*-means, *k*-modes, and so on
  - Neural networks
    (adaptive resonance theory [ART], self-organizing map [SOM])
  - Fuzzy logic (e.g., fuzzy c-means algorithm)
  - Genetic algorithms
- Divisive versus Agglomerative methods

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

# Cluster Analysis for Data Mining

- How many clusters?
  - There is not a "truly optimal" way to calculate it
  - Heuristics are often used
    1. Look at the sparseness of clusters
    2. Number of clusters = $(n/2)^{1/2}$ (n: no of data points)
    3. Use Akaike information criterion (AIC)
    4. Use Bayesian information criterion (BIC)
- Most cluster analysis methods involve the use of a distance measure to calculate the closeness between pairs of items
  - Euclidian versus Manhattan (rectilinear) distance

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

# *k*-Means Clustering Algorithm

- *k* : pre-determined number of clusters
- Algorithm (Step 0: determine value of *k*)

Step 1: Randomly generate *k* random points as initial cluster centers

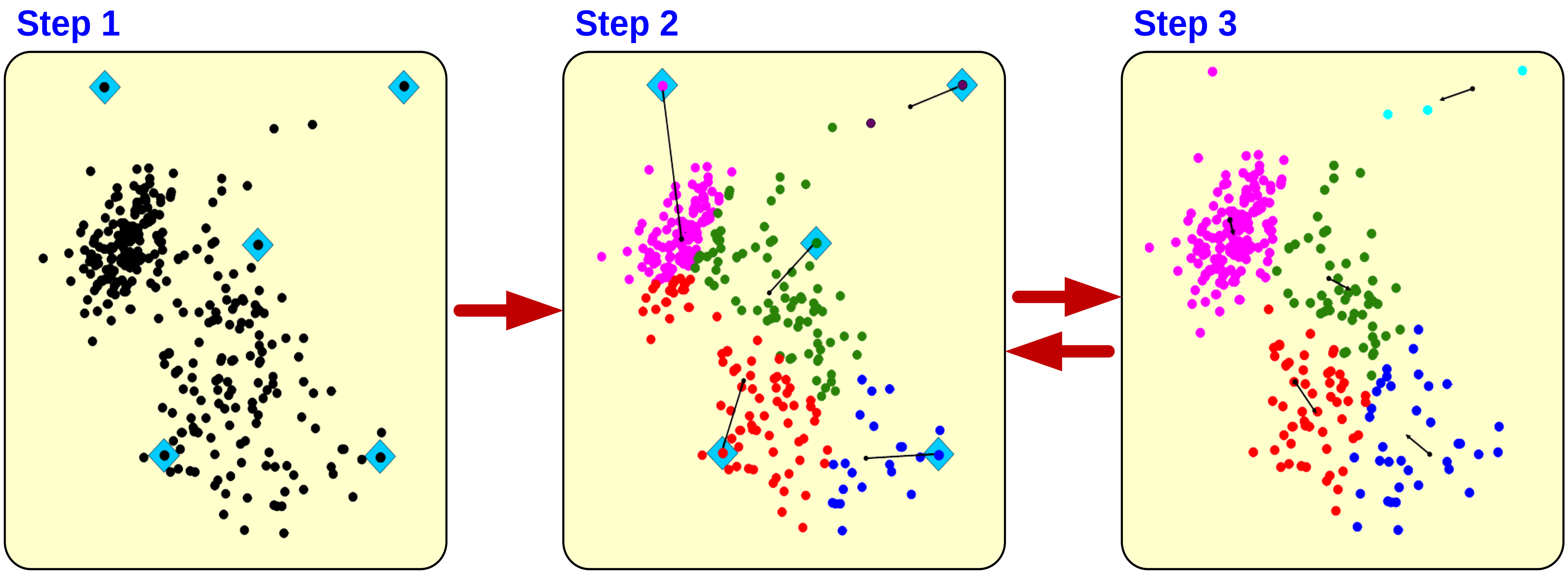
Step 2: Assign each point to the nearest cluster center

Step 3: Re-compute the new cluster centers

Repetition step: Repeat steps 2 and 3 until some convergence criterion is met (usually that the assignment of points to clusters becomes stable)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

# Cluster Analysis for Data Mining - *k*-Means Clustering Algorithm




Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

# Quality: 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

Source: Han & Kamber (2006)

# 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*:

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

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

- If $q = 1$, $d$ is Manhattan distance

$$d(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + ... + |x_{i_p} - x_{j_p}|$$

Source: Han & Kamber (2006)

# Similarity and Dissimilarity Between Objects (Cont.)

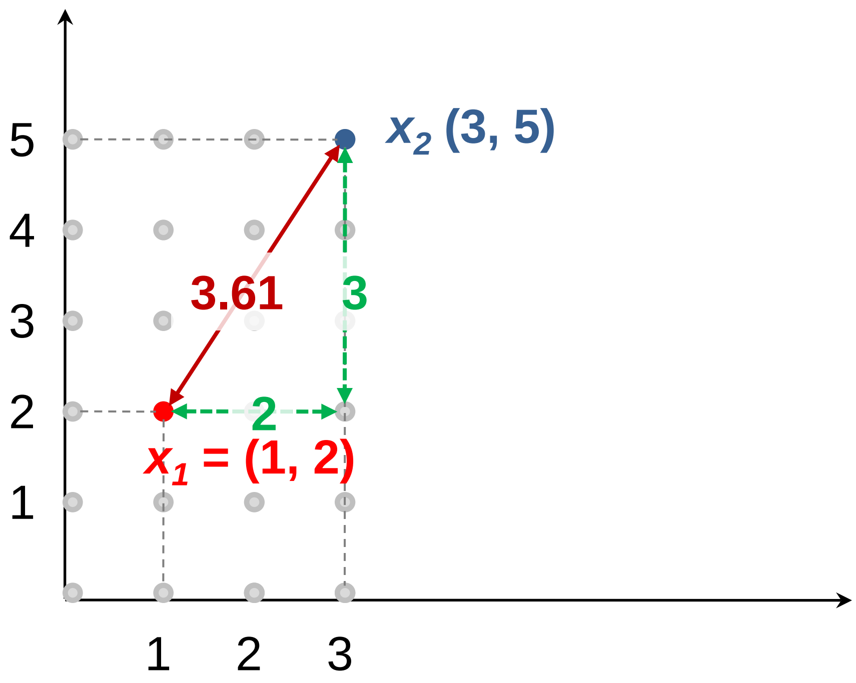
- *If q = 2, d* is Euclidean distance:

$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + ... + |x_{i_p} - x_{j_p}|^2)}$$

  - Properties
    - $d(i,j) \geq 0$
    - $d(i,i) = 0$
    - $d(i,j) = d(j,i)$
    - $d(i,j) \leq d(i,k) + d(k,j)$
- Also, one can use weighted distance, parametric Pearson product moment correlation, or other disimilarity measures

# Euclidean distance vs Manhattan distance

- Distance of two point $x_1$ = (1, 2) and $x_2$ (3, 5)




Euclidean distance:

$= ((3-1)^2 + (5-2)^2)^{1/2}$

$= (2^2 + 3^2)^{1/2}$

$= (4 + 9)^{1/2}$

$= (13)^{1/2}$

$= 3.61$

Manhattan distance:

$= (3-1) + (5-2)$

$= 2 + 3$

$= 5$

# Binary Variables

- A contingency table for binary data

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

- Distance measure for symmetric binary variables:

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

- Distance measure for asymmetric binary variables:

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

- Jaccard coefficient (*similarity* measure for *asymmetric* binary variables):

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

Source: Han & Kamber (2006)

# 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

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

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

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

Source: Han & Kamber (2006)

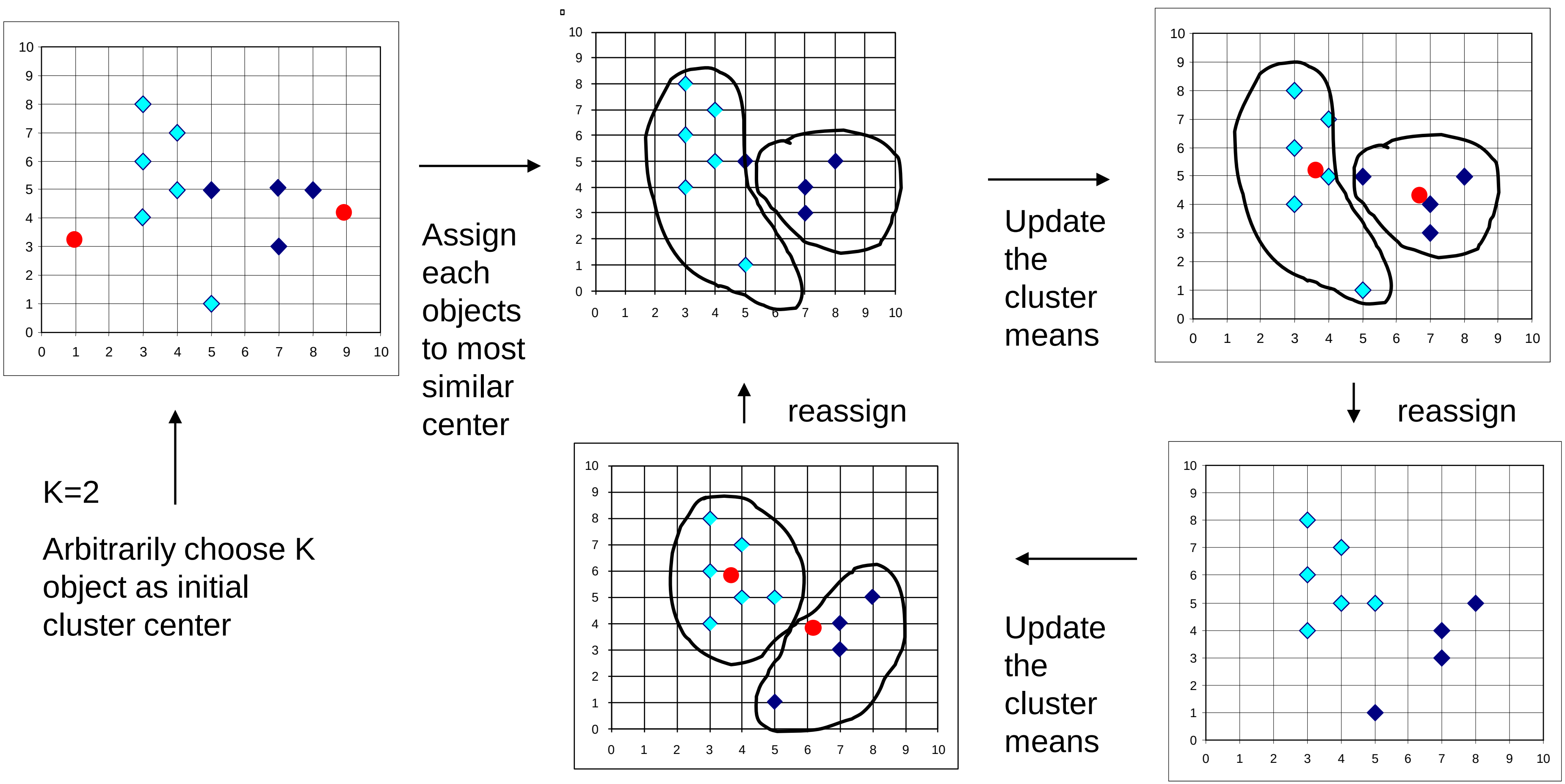
# The *K-Means* Clustering Method

- Given *k*, the *k-means* algorithm is implemented in four steps:
  1. Partition objects into *k* nonempty subsets
  2. Compute seed points as the centroids of the clusters of the current partition
     (the centroid is the center, i.e., *mean point*, of the cluster)
  3. Assign each object to the cluster with the nearest seed point
  4. Go back to Step 2, stop when no more new assignment

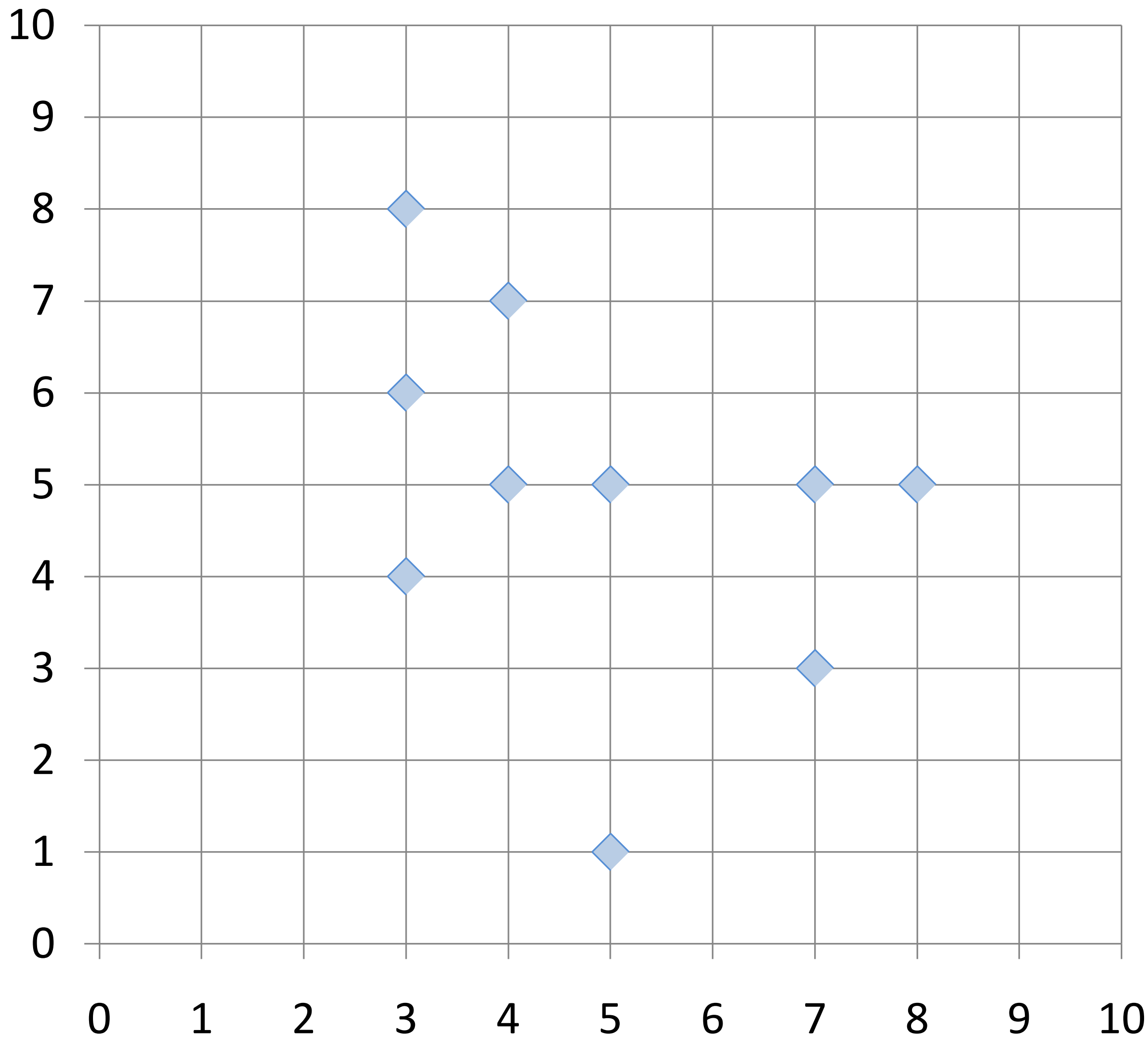
Source: Han & Kamber (2006)

# The *K-Means* Clustering Method

- Example



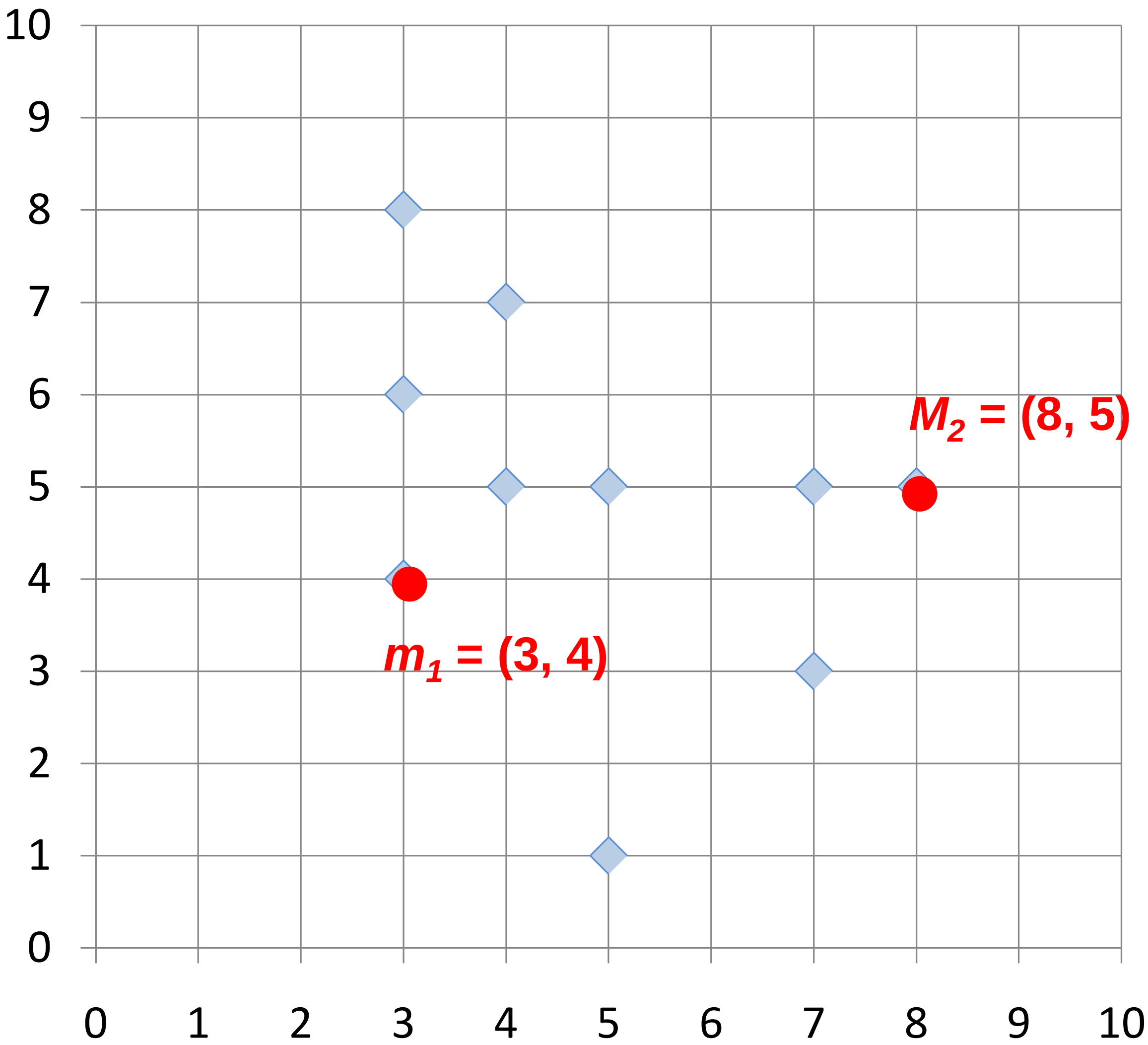
# *K-Means* Clustering Step by Step




| Point | P | P(x,y) |
|---|---|---|
| p01 | a | (3, 4) |
| p02 | b | (3, 6) |
| p03 | c | (3, 8) |
| p04 | d | (4, 5) |
| p05 | e | (4, 7) |
| p06 | f | (5, 1) |
| p07 | g | (5, 5) |
| p08 | h | (7, 3) |
| p09 | i | (7, 5) |
| p10 | j | (8, 5) |

# *K-Means* Clustering

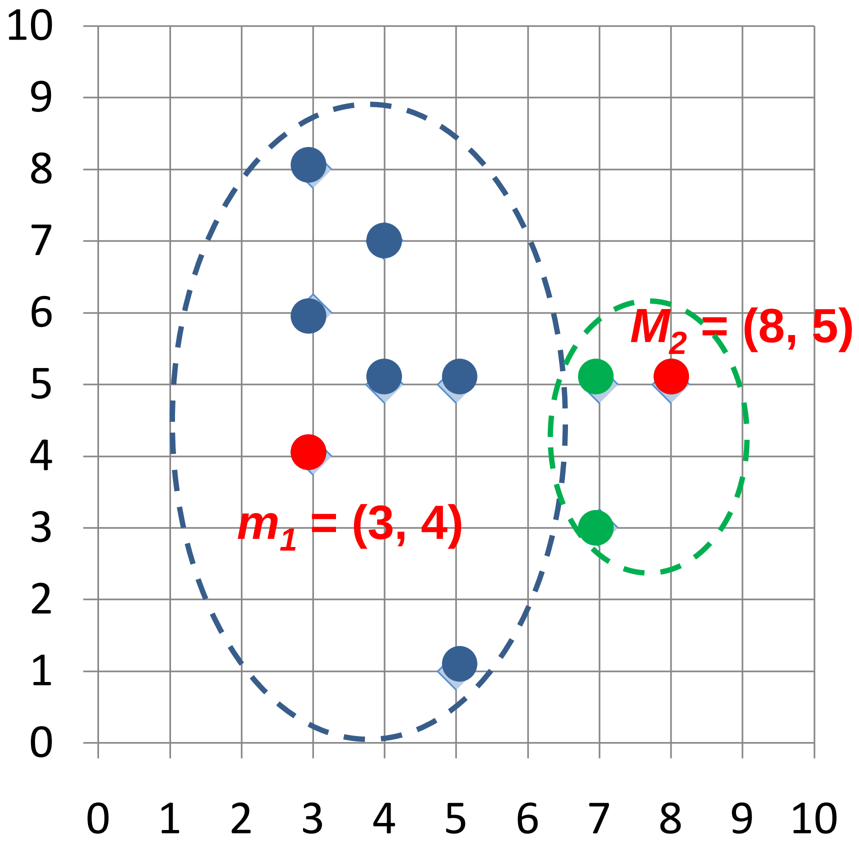
**Step 1: K=2, Arbitrarily choose K object as initial cluster center**




| Point | P | P(x,y) |
|---|---|---|
| p01 | a | (3, 4) |
| p02 | b | (3, 6) |
| p03 | c | (3, 8) |
| p04 | d | (4, 5) |
| p05 | e | (4, 7) |
| p06 | f | (5, 1) |
| p07 | g | (5, 5) |
| p08 | h | (7, 3) |
| p09 | i | (7, 5) |
| p10 | j | (8, 5) |
| | | |
| Initial | m1 | (3, 4) |
| Initial | m2 | (8, 5) |

**Step 2: Compute seed points as the centroids of the clusters of the current partition**

**Step 3: Assign each objects to most similar center**




| Point | P | P(x,y) | m1 distance | m2 distance | Cluster |
|---|---|---|---|---|---|
| p01 | a | (3, 4) | 0.00 | 5.10 | Cluster1 |
| p02 | b | (3, 6) | 2.00 | 5.10 | Cluster1 |
| p03 | c | (3, 8) | 4.00 | 5.83 | Cluster1 |
| p04 | d | (4, 5) | 1.41 | 4.00 | Cluster1 |
| p05 | e | (4, 7) | 3.16 | 4.47 | Cluster1 |
| p06 | f | (5, 1) | 3.61 | 5.00 | Cluster1 |
| p07 | g | (5, 5) | 2.24 | 3.00 | Cluster1 |
| p08 | h | (7, 3) | 4.12 | 2.24 | Cluster2 |
| p09 | i | (7, 5) | 4.12 | 1.00 | Cluster2 |
| p10 | j | (8, 5) | 5.10 | 0.00 | Cluster2 |

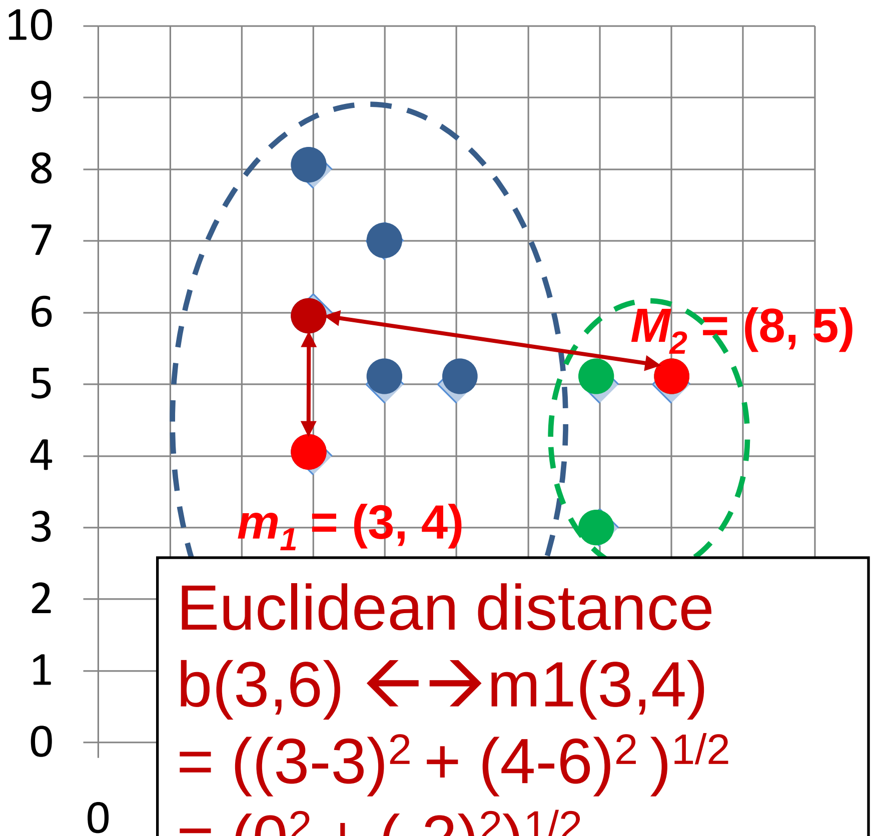
# *K-Means* Clustering

Initial m1 (3, 4)

Initial m2 (8, 5)

**Step 2: Compute seed points as the centroids of the clusters of the current partition**

**Step 3: Assign each objects to most similar center**




| Point | P | P(x,y) | m1 distance | m2 distance | Cluster |
|---|---|---|---|---|---|
| p01 | a | (3, 4) | 0.00 | 5.10 | Cluster1 |
| p02 | b | (3, 6) | 2.00 | 5.10 | Cluster1 |
| p03 | c | (3, 8) | 4.00 | 5.83 | Cluster1 |
| p0 | | | | | er1 |
| p0 | | | | | er1 |
| p0 | | | | | er1 |
| p0 | | | | | er1 |
| p0 | | | | | er2 |
| p0 | | | | | er2 |
| p1 | | | | | er2 |

Euclidean distance
b(3,6) ←→m2(8,5)
$= ((8-3)^2 + (5-6)^2)^{1/2}$
$= (5^2 + (-1)^2)^{1/2}$
$= (25 + 1)^{1/2}$
$= (26)^{1/2}$
$= 5.10$

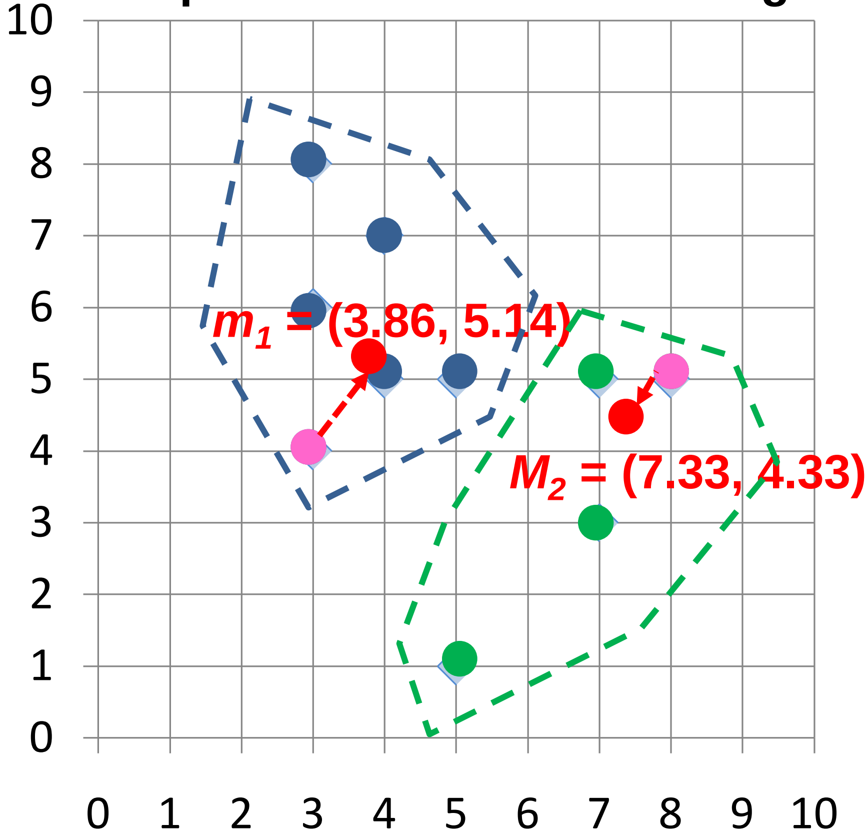
Euclidean distance
b(3,6) ←→m1(3,4)
$= ((3-3)^2 + (4-6)^2)^{1/2}$
$= (0^2 + (-2)^2)^{1/2}$
$= (0 + 4)^{1/2}$
$= (4)^{1/2}$
$= 2.00$

***K-I***

Initial m1 (3, 4)

Initial m2 (8, 5)

**Step 4: Update the cluster means,**
**Repeat Step 2, 3,**
**stop when no more new assignment**



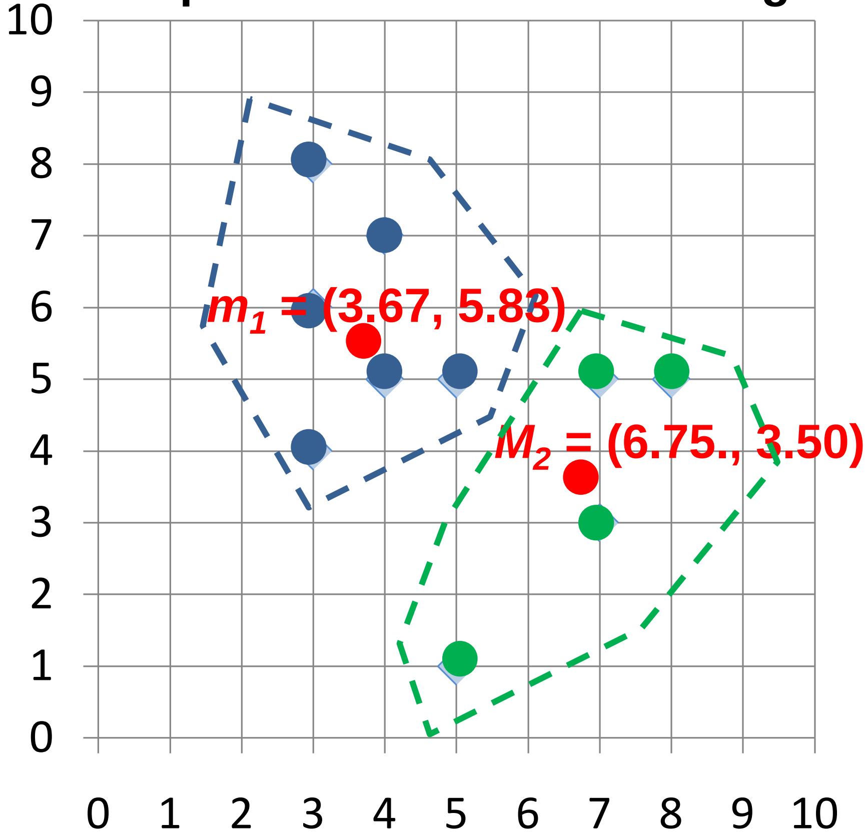

| Point | P | P(x,y) | m1 distance | m2 distance | Cluster |
|---|---|---|---|---|---|
| p01 | a | (3, 4) | 1.43 | 4.34 | Cluster1 |
| p02 | b | (3, 6) | 1.22 | 4.64 | Cluster1 |
| p03 | c | (3, 8) | 2.99 | 5.68 | Cluster1 |
| p04 | d | (4, 5) | 0.20 | 3.40 | Cluster1 |
| p05 | e | (4, 7) | 1.87 | 4.27 | Cluster1 |
| p06 | f | (5, 1) | 4.29 | 4.06 | Cluster2 |
| p07 | g | (5, 5) | 1.15 | 2.42 | Cluster1 |
| p08 | h | (7, 3) | 3.80 | 1.37 | Cluster2 |
| p09 | i | (7, 5) | 3.14 | 0.75 | Cluster2 |
| p10 | j | (8, 5) | 4.14 | 0.95 | Cluster2 |

m1 (3.86, 5.14)

m2 (7.33, 4.33)

# *K-Means* Clustering

**Step 4: Update the cluster means,**
**Repeat Step 2, 3,**
**stop when no more new assignment**



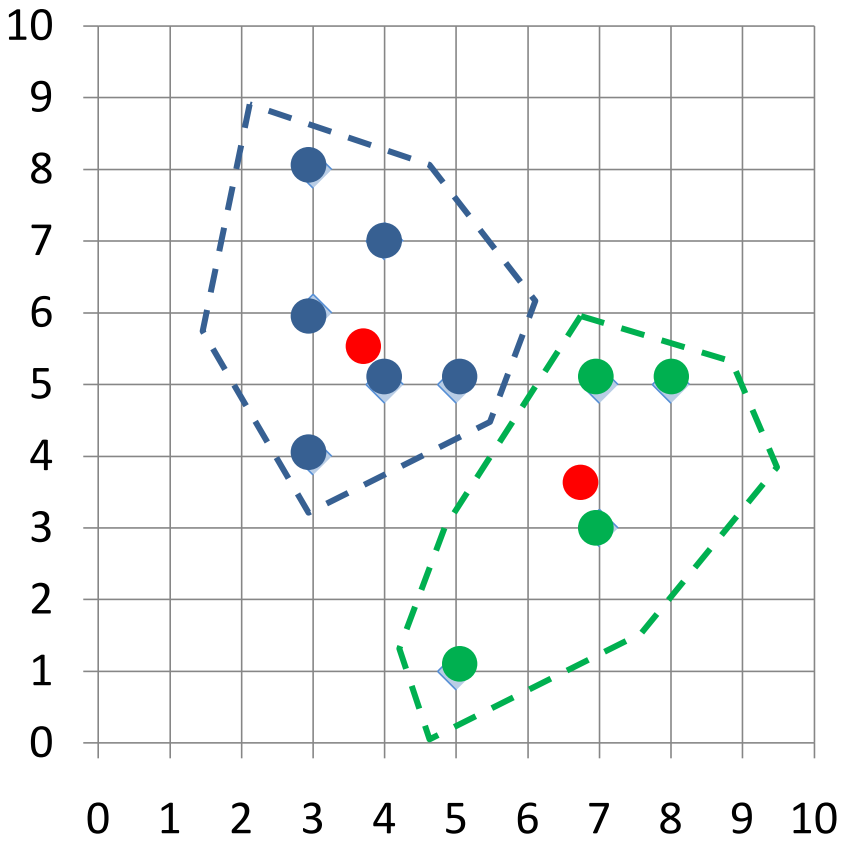

| Point | P | P(x,y) | m1 distance | m2 distance | Cluster |
|---|---|---|---|---|---|
| p01 | a | (3, 4) | 1.95 | 3.78 | Cluster1 |
| p02 | b | (3, 6) | 0.69 | 4.51 | Cluster1 |
| p03 | c | (3, 8) | 2.27 | 5.86 | Cluster1 |
| p04 | d | (4, 5) | 0.89 | 3.13 | Cluster1 |
| p05 | e | (4, 7) | 1.22 | 4.45 | Cluster1 |
| p06 | f | (5, 1) | 5.01 | 3.05 | Cluster2 |
| p07 | g | (5, 5) | 1.57 | 2.30 | Cluster1 |
| p08 | h | (7, 3) | 4.37 | 0.56 | Cluster2 |
| p09 | i | (7, 5) | 3.43 | 1.52 | Cluster2 |
| p10 | j | (8, 5) | 4.41 | 1.95 | Cluster2 |

m1 (3.67, 5.83)

m2 (6.75, 3.50)

# *K-Means* Clustering

**stop when no more new assignment**



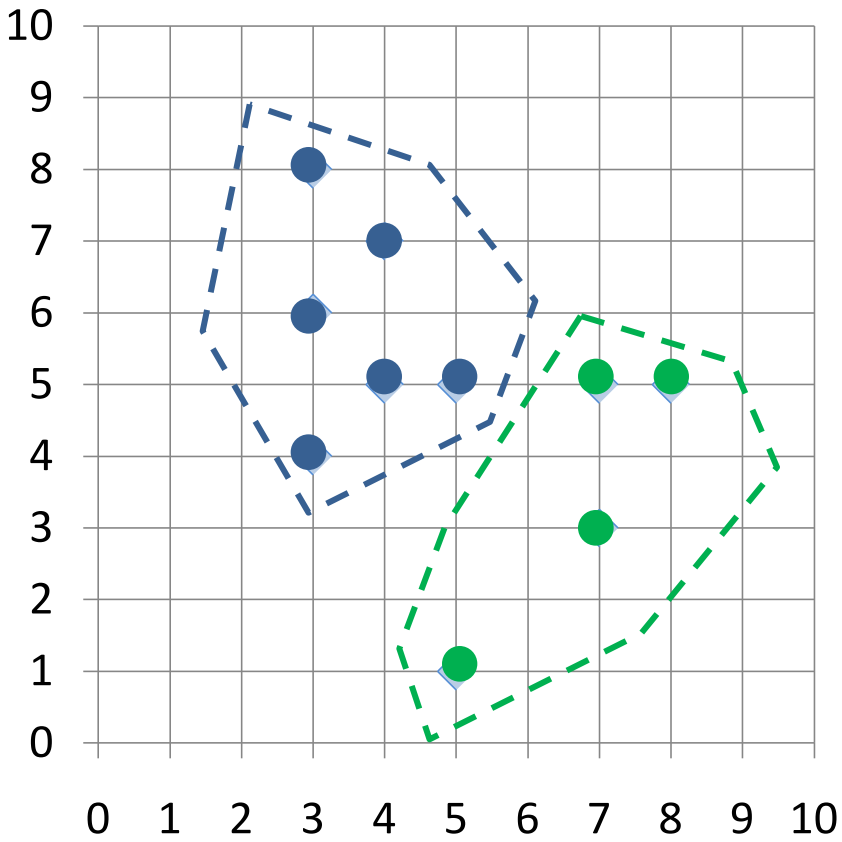

| Point | P | P(x,y) | m1 distance | m2 distance | Cluster |
|---|---|---|---|---|---|
| p01 | a | (3, 4) | 1.95 | 3.78 | Cluster1 |
| p02 | b | (3, 6) | 0.69 | 4.51 | Cluster1 |
| p03 | c | (3, 8) | 2.27 | 5.86 | Cluster1 |
| p04 | d | (4, 5) | 0.89 | 3.13 | Cluster1 |
| p05 | e | (4, 7) | 1.22 | 4.45 | Cluster1 |
| p06 | f | (5, 1) | 5.01 | 3.05 | Cluster2 |
| p07 | g | (5, 5) | 1.57 | 2.30 | Cluster1 |
| p08 | h | (7, 3) | 4.37 | 0.56 | Cluster2 |
| p09 | i | (7, 5) | 3.43 | 1.52 | Cluster2 |
| p10 | j | (8, 5) | 4.41 | 1.95 | Cluster2 |

m1 (3.67, 5.83)

m2 (6.75, 3.50)

# *K-Means* Clustering

**stop when no more new assignment**




| Point | P | P(x,y) | m1 distance | m2 distance | Cluster |
|---|---|---|---|---|---|
| p01 | a | (3, 4) | 1.95 | 3.78 | Cluster1 |
| p02 | b | (3, 6) | 0.69 | 4.51 | Cluster1 |
| p03 | c | (3, 8) | 2.27 | 5.86 | Cluster1 |
| p04 | d | (4, 5) | 0.89 | 3.13 | Cluster1 |
| p05 | e | (4, 7) | 1.22 | 4.45 | Cluster1 |
| p06 | f | (5, 1) | 5.01 | 3.05 | Cluster2 |
| p07 | g | (5, 5) | 1.57 | 2.30 | Cluster1 |
| p08 | h | (7, 3) | 4.37 | 0.56 | Cluster2 |
| p09 | i | (7, 5) | 3.43 | 1.52 | Cluster2 |
| p10 | j | (8, 5) | 4.41 | 1.95 | Cluster2 |

m1 (3.67, 5.83)

m2 (6.75, 3.50)

# *K-Means* Clustering

# Summary

- Cluster Analysis
- *K-Means* Clustering

# References


- Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, Second Edition, 2006, Elsevier
- Efraim Turban, Ramesh Sharda, Dursun Delen, Decision Support and Business Intelligence Systems, Ninth Edition, 2011, Pearson.