



# Big Data Mining

## 巨量資料探勘

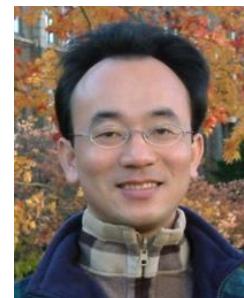
# 分群分析

## (Cluster Analysis)

1042DM05

MI4 (M2244) (3094)

Tue, 3, 4 (10:10-12:00) (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/>

2016-03-15



# 課程大綱 (Syllabus)

週次 (Week) 日期 (Date) 內容 (Subject/Topics)

- |   |            |   |
|---|------------|---|
| 1 | 2016/02/16 | 巨量資料探勘課程介紹<br>(Course Orientation for Big Data Mining)  |
| 2 | 2016/02/23 | 巨量資料基礎：MapReduce典範、Hadoop與Spark生態系統<br>(Fundamental Big Data: MapReduce Paradigm, Hadoop and Spark Ecosystem) |
| 3 | 2016/03/01 | 關連分析 (Association Analysis)   |
| 4 | 2016/03/08 | 分類與預測 (Classification and Prediction)   |
| 5 | 2016/03/15 | 分群分析 (Cluster Analysis)   |
| 6 | 2016/03/22 | 個案分析與實作一 (SAS EM 分群分析)：<br>Case Study 1 (Cluster Analysis – K-Means using SAS EM)                             |
| 7 | 2016/03/29 | 個案分析與實作二 (SAS EM 關連分析)：<br>Case Study 2 (Association Analysis using SAS EM)                                   |

# 課程大綱 (Syllabus)

週次 (Week) 日期 (Date) 內容 (Subject/Topics)

8 2016/04/05 教學行政觀摩日 (Off-campus study)

9 2016/04/12 期中報告 (Midterm Project Presentation)

10 2016/04/19 期中考試週 (Midterm Exam)

11 2016/04/26 個案分析與實作三 (SAS EM 決策樹、模型評估) :  
Case Study 3 (Decision Tree, Model Evaluation using SAS EM)

12 2016/05/03 個案分析與實作四 (SAS EM 迴歸分析、類神經網路) :  
Case Study 4 (Regression Analysis,  
Artificial Neural Network using SAS EM)

13 2016/05/10 Google TensorFlow 深度學習  
(Deep Learning with Google TensorFlow)

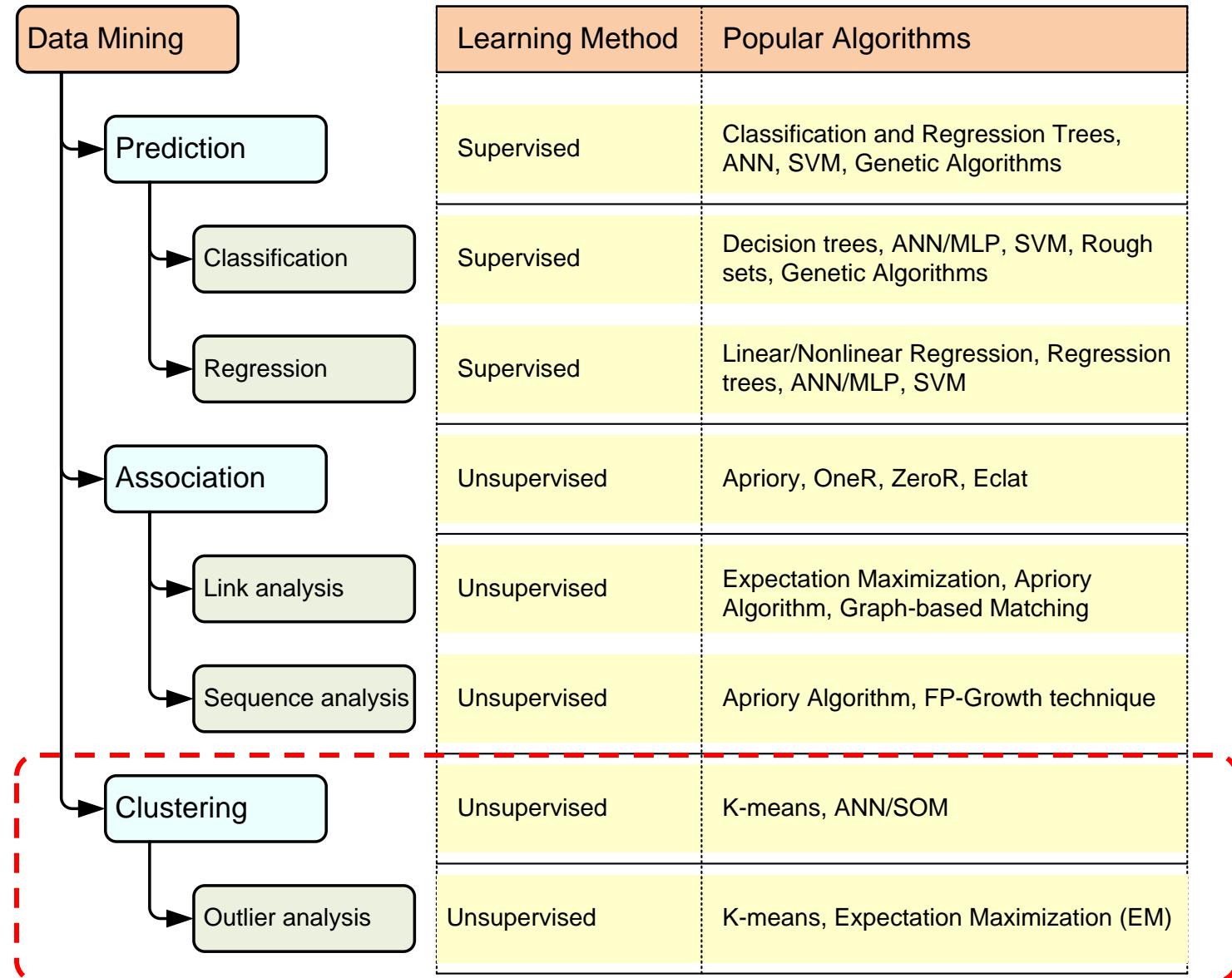
14 2016/05/17 期末報告 (Final Project Presentation)

15 2016/05/24 畢業班考試 (Final Exam)

# Outline

- Cluster Analysis
- *K-Means Clustering*

# A Taxonomy for Data Mining Tasks



# 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)

# K-Means Clustering

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)

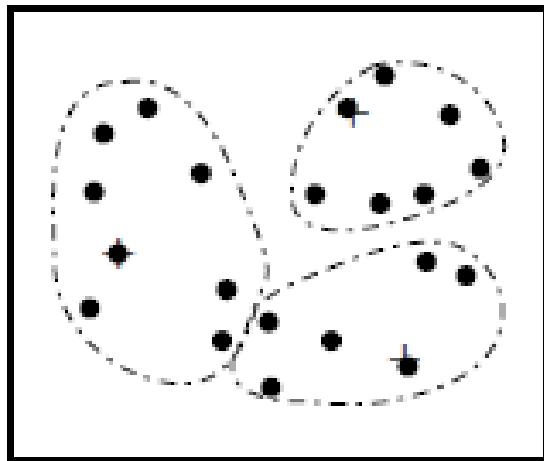
m2                   (6.75, 3.50)

# Cluster Analysis

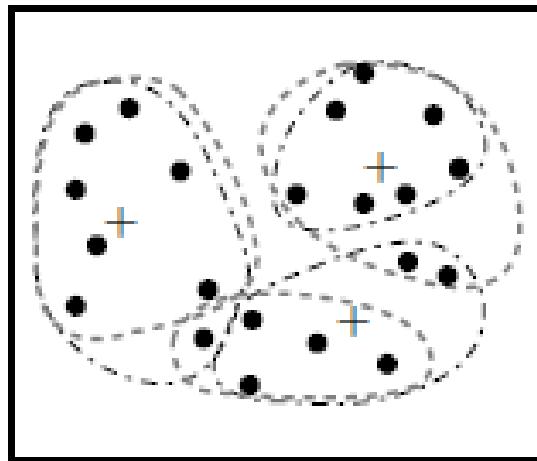
# Cluster Analysis

- 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

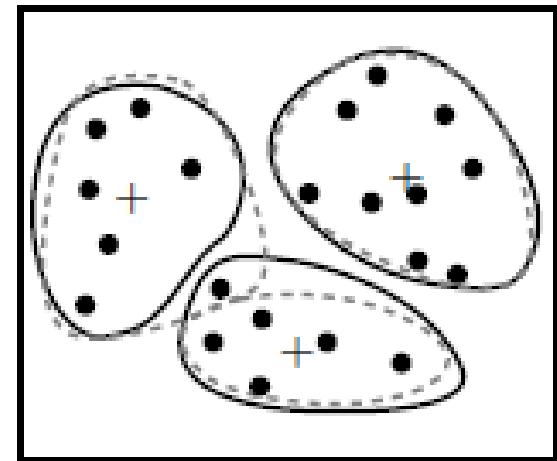
# Cluster Analysis



(a)



(b)



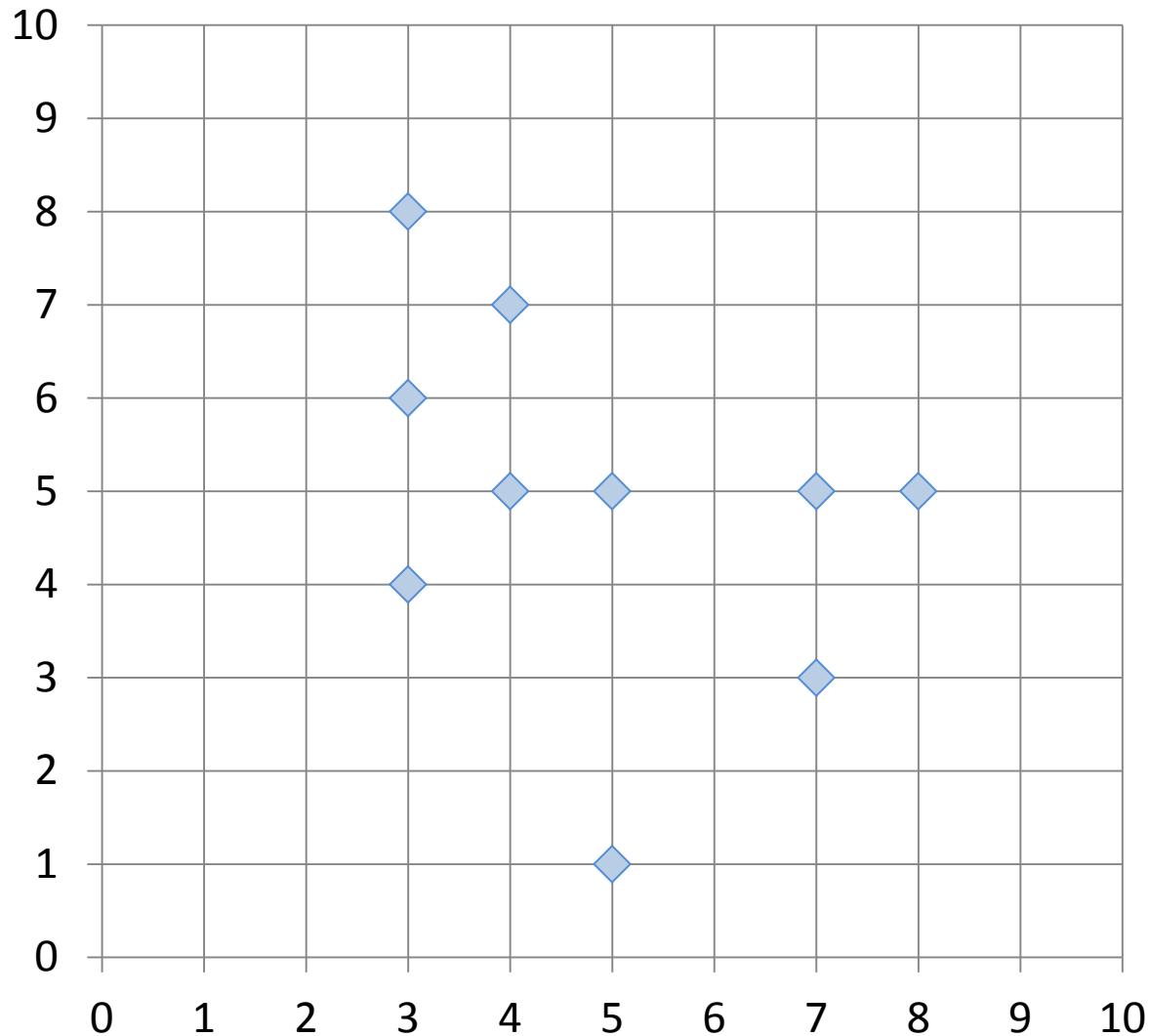
(c)

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

# 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)

# 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

# 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

# ***k*-Means Clustering Algorithm**

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

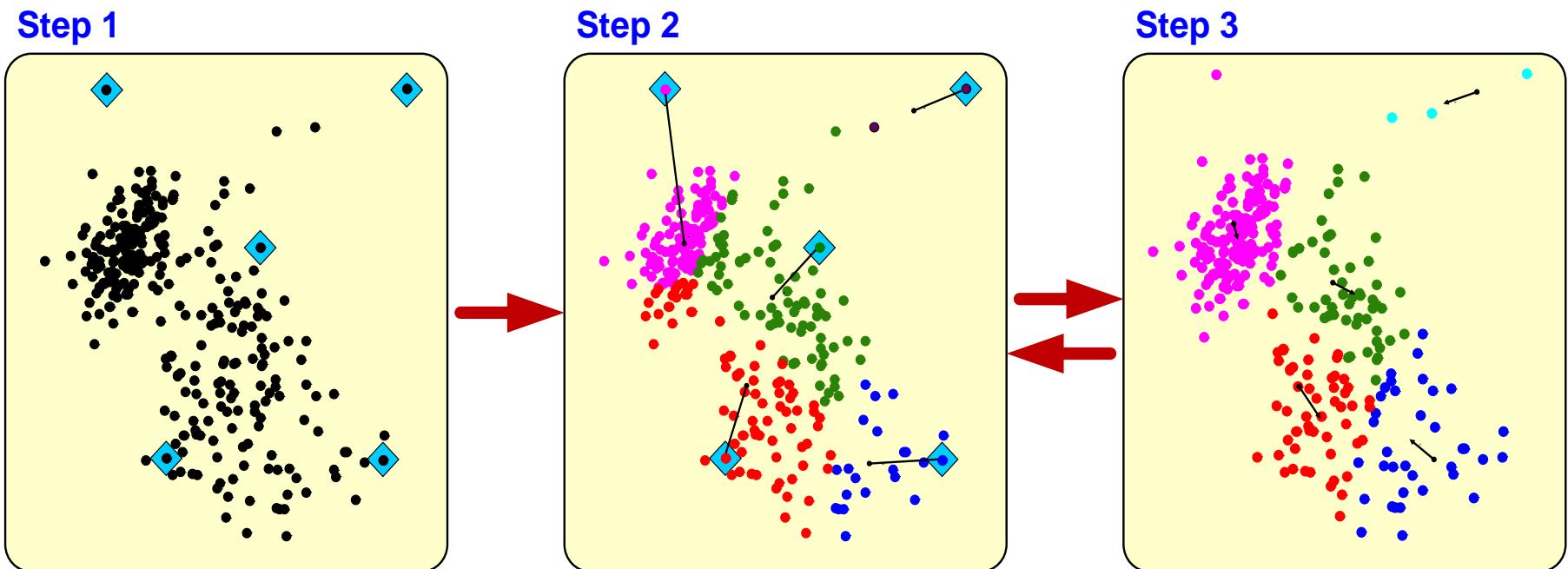
**Step 1:** Randomly generate  $k$  random points as initial cluster centers

**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)

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



# Similarity

# Distance

# 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*:

$$d(i, j) = \sqrt[q]{(|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \dots + |x_{ip} - x_{jp}|^q)}$$

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

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

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{ip} - x_{jp}|$$

# Similarity and Dissimilarity Between Objects (Cont.)

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

$$d(i, j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \dots + |x_{ip} - x_{jp}|^2)}$$

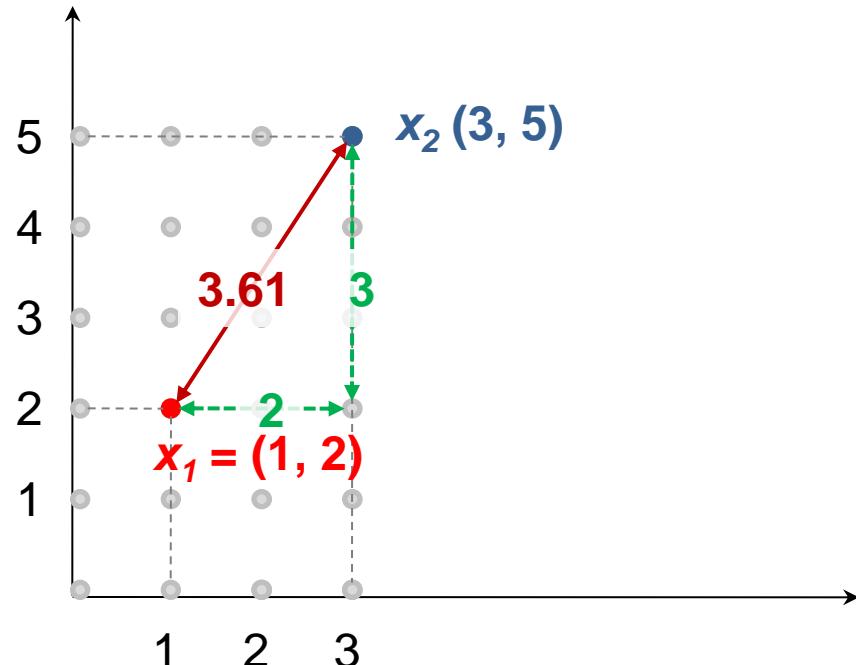
- Properties

- $d(i, j) \geq 0$
- $d(i, i) = 0$
- $d(i, j) = d(j, i)$
- $d(i, j) \leq d(i, k) + d(k, j)$

- 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  $x_2 (3, 5)$

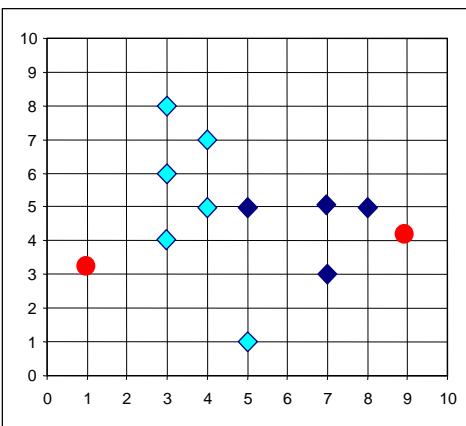


Euclidean distance:  
 $= ((3-1)^2 + (5-2)^2 )^{1/2}$   
 $= (2^2 + 3^2)^{1/2}$   
 $= (4 + 9)^{1/2}$   
 $= (13)^{1/2}$   
 $= 3.61$

Manhattan distance:  
 $= (3-1) + (5-2)$   
 $= 2 + 3$   
 $= 5$

# The *K*-Means Clustering Method

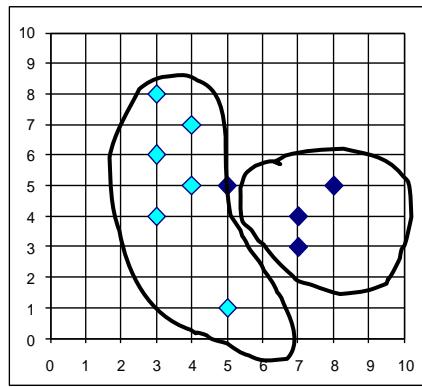
- Example



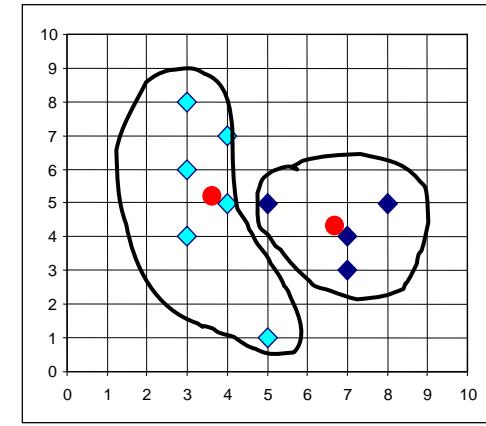
K=2

Arbitrarily choose K object as initial cluster center

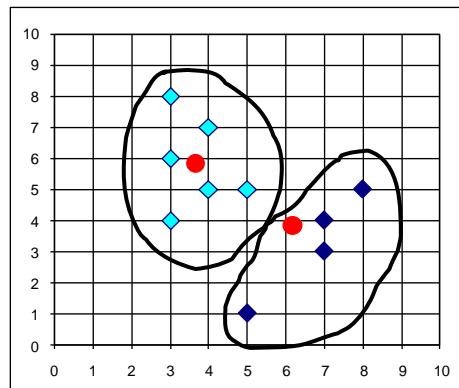
Assign each objects to most similar center



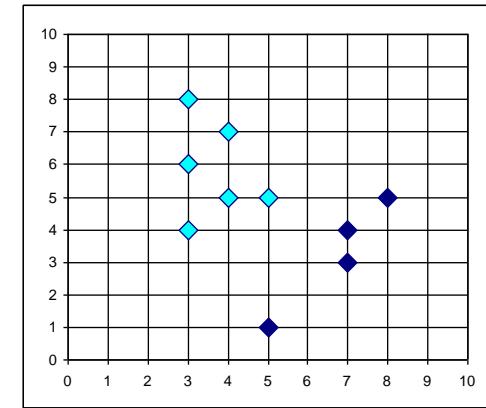
Update the cluster means



reassign



Update the cluster means



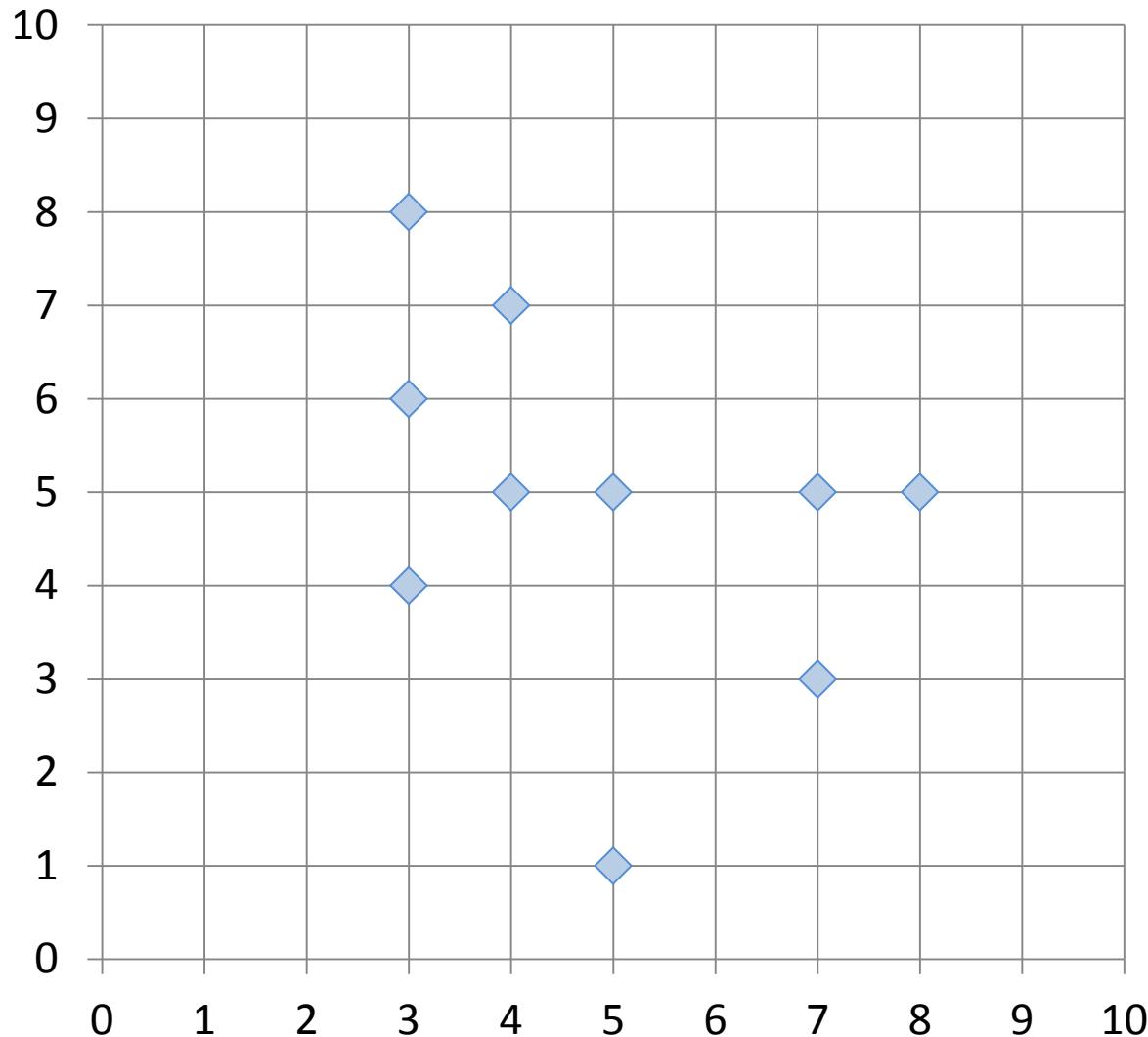
# *K-Means* Clustering

# 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)

# K-Means Clustering

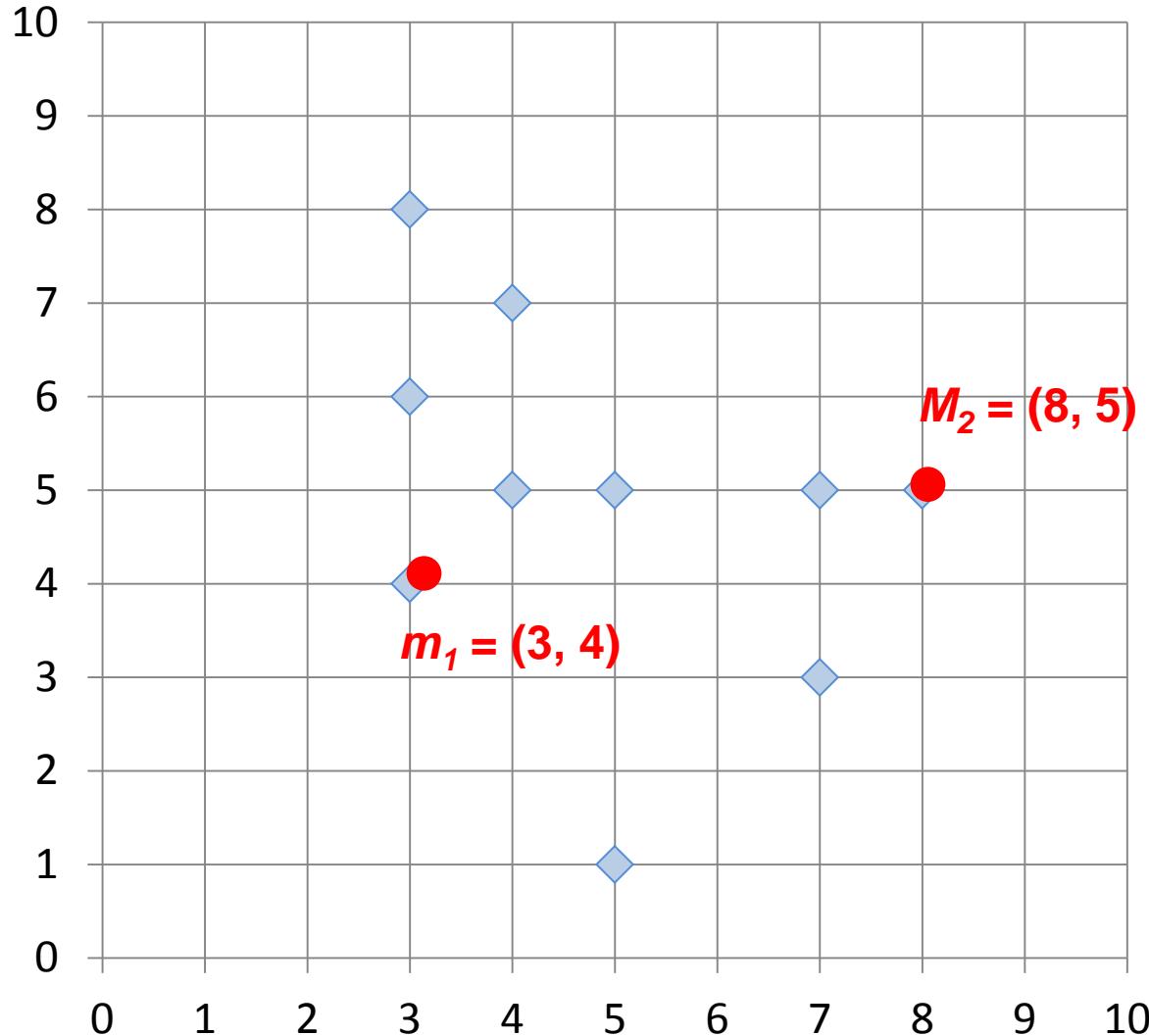
## Step by Step



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

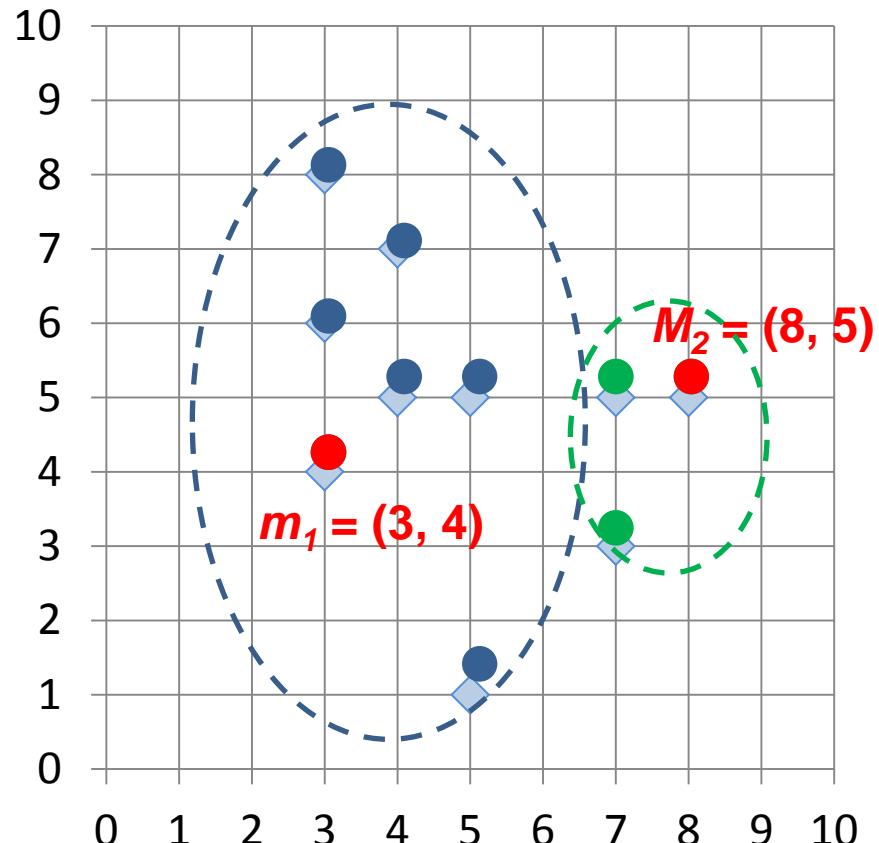


Point	P	P(x,y)
p01	a	(3, 4)
p02	b	(3, 6)
p03	c	(3, 8)
p04	d	(4, 5)
p05	e	(4, 7)
p06	f	(5, 1)
p07	g	(5, 5)
p08	h	(7, 3)
p09	i	(7, 5)
p10	j	(8, 5)

Initial  $m_1$  (3, 4)  
Initial  $m_2$  (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**



## K-Means Clustering

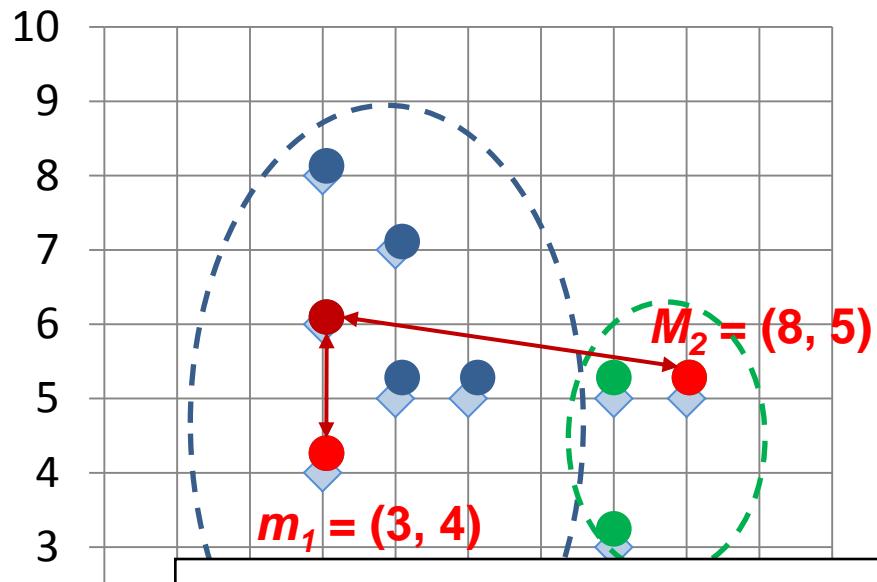
Initial  $m_1$  (3, 4)

Initial  $m_2$  (8, 5)

Point	P	P(x,y)	$m_1$ distance	$m_2$ 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

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

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



**Euclidean distance**  
 $b(3,6) \leftrightarrow m_1(3,4)$   
 $= ((3-3)^2 + (4-6)^2)^{1/2}$   
 $= (0^2 + (-2)^2)^{1/2}$   
 $= (0 + 4)^{1/2}$   
 $= (4)^{1/2}$   
 $= 2.00$

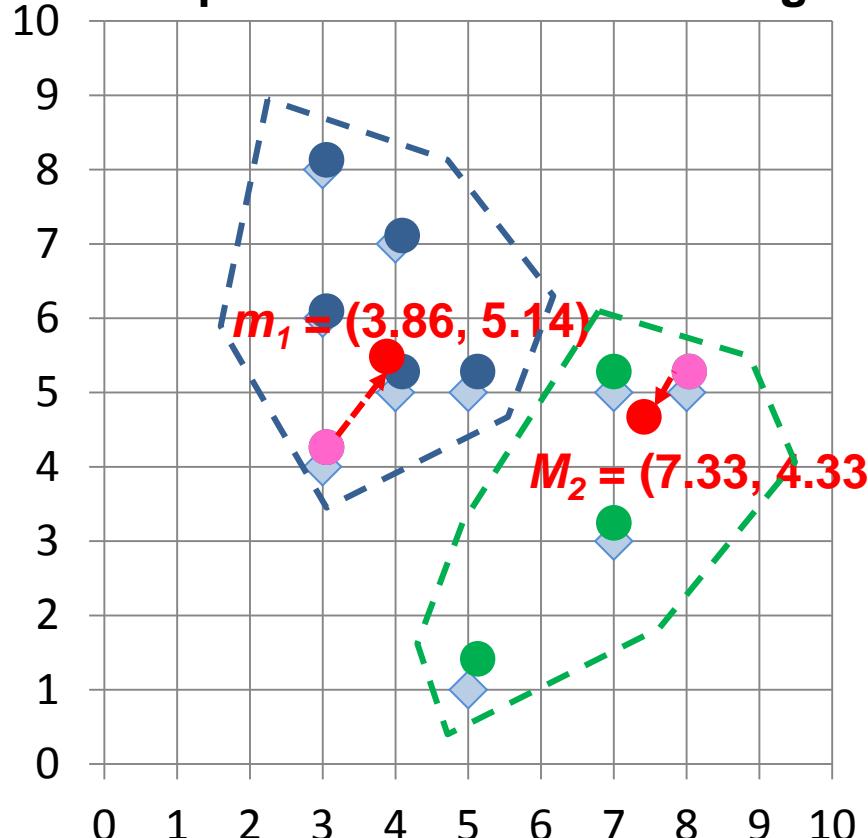
Point	P	P(x,y)	m1 distance	m2 distance	Cluster
p01	a	(3, 4)	0.00	5.10	Cluster1
p02	b	(3, 6)	2.00	5.10	Cluster1
p03	c	(3, 8)	4.00	5.83	Cluster1
p04	d	(4, 5)	1.41	4.00	Cluster1

**Euclidean distance**  
 $b(3,6) \leftrightarrow m_2(8,5)$   
 $= ((8-3)^2 + (5-6)^2)^{1/2}$   
 $= (5^2 + (-1)^2)^{1/2}$   
 $= (25 + 1)^{1/2}$   
 $= (26)^{1/2}$   
 $= 5.10$

Initial  $m_1 (3, 4)$

Initial  $m_2 (8, 5)$

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

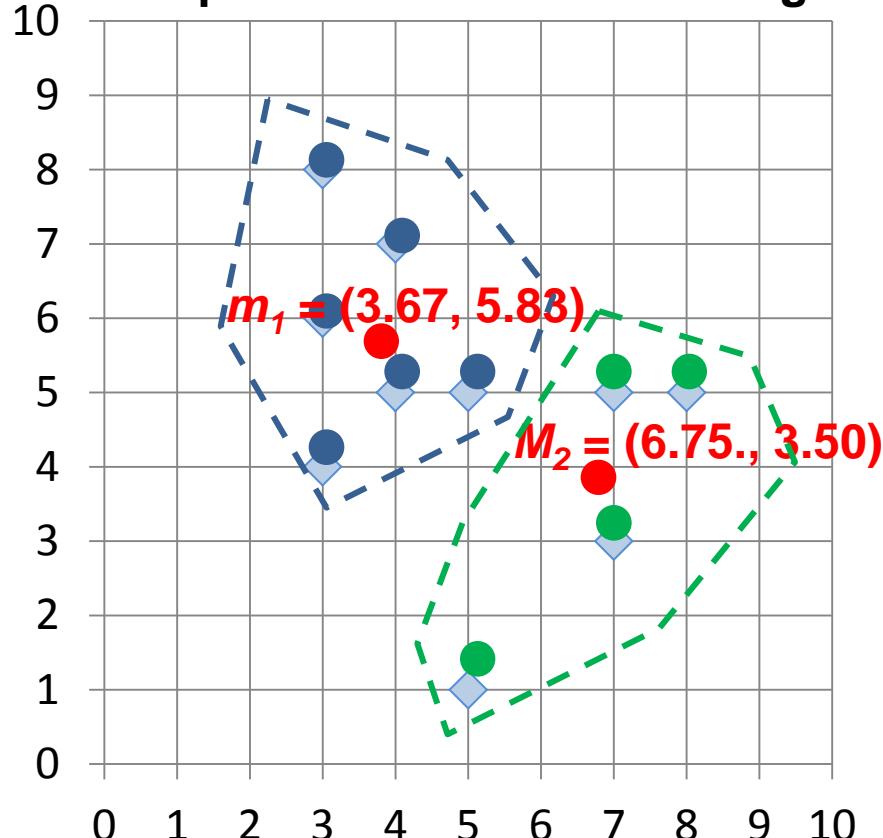


Point	P	P(x,y)	m1 distance	m2 distance	Cluster
p01	a	(3, 4)	1.43	4.34	Cluster1
p02	b	(3, 6)	1.22	4.64	Cluster1
p03	c	(3, 8)	2.99	5.68	Cluster1
p04	d	(4, 5)	0.20	3.40	Cluster1
p05	e	(4, 7)	1.87	4.27	Cluster1
p06	f	(5, 1)	4.29	4.06	Cluster2
p07	g	(5, 5)	1.15	2.42	Cluster1
p08	h	(7, 3)	3.80	1.37	Cluster2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

$$\begin{aligned}m1 &= (3.86, 5.14) \\m2 &= (7.33, 4.33)\end{aligned}$$

## K-Means Clustering

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

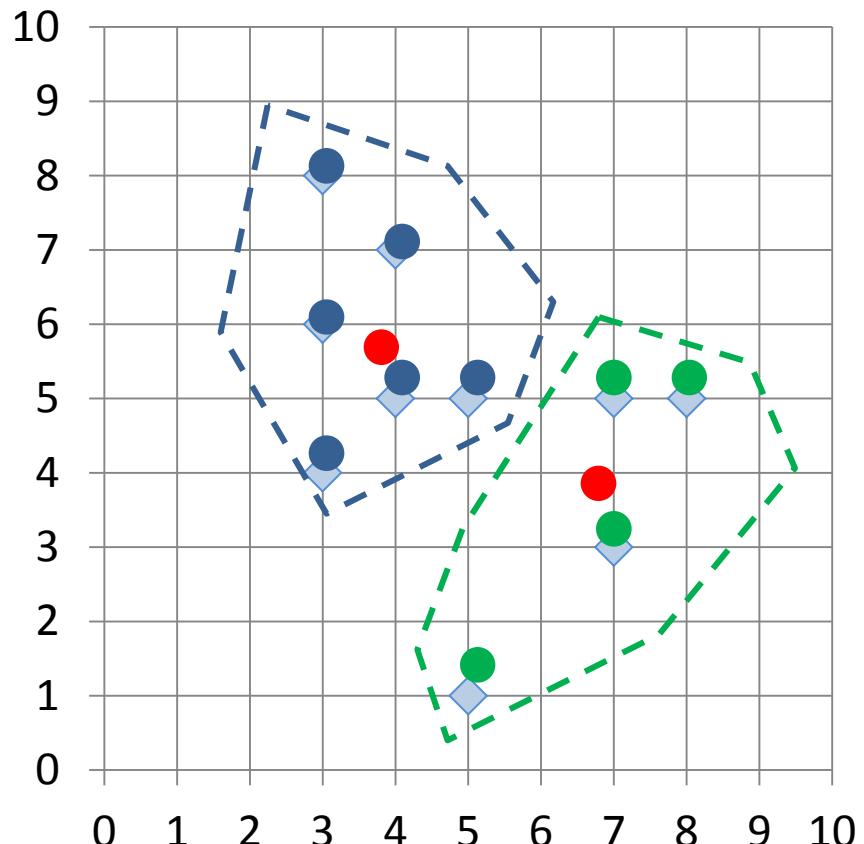


Point	P	P(x,y)	m1 distance	m2 distance	Cluster
p01	a	(3, 4)	1.95	3.78	Cluster1
p02	b	(3, 6)	0.69	4.51	Cluster1
p03	c	(3, 8)	2.27	5.86	Cluster1
p04	d	(4, 5)	0.89	3.13	Cluster1
p05	e	(4, 7)	1.22	4.45	Cluster1
p06	f	(5, 1)	5.01	3.05	Cluster2
p07	g	(5, 5)	1.57	2.30	Cluster1
p08	h	(7, 3)	4.37	0.56	Cluster2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

$$\begin{aligned}m1 & (3.67, 5.83) \\m2 & (6.75, 3.50)\end{aligned}$$

## K-Means Clustering

**stop when no more new assignment**



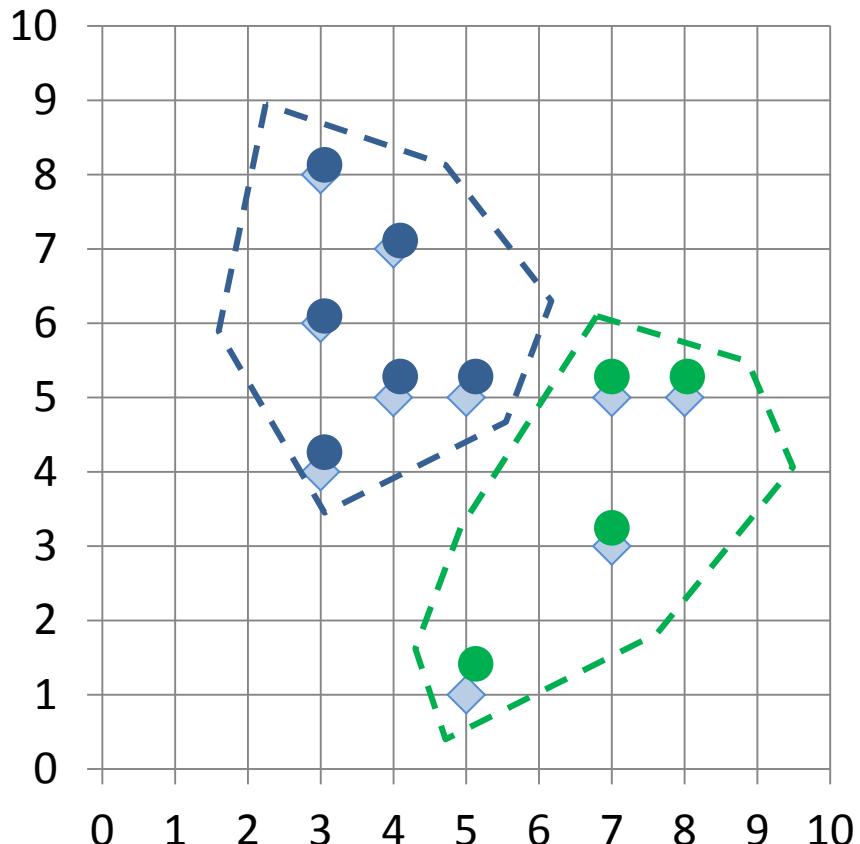
Point	P	P(x,y)	m1 distance	m2 distance	Cluster
p01	a	(3, 4)	1.95	3.78	Cluster1
p02	b	(3, 6)	0.69	4.51	Cluster1
p03	c	(3, 8)	2.27	5.86	Cluster1
p04	d	(4, 5)	0.89	3.13	Cluster1
p05	e	(4, 7)	1.22	4.45	Cluster1
p06	f	(5, 1)	5.01	3.05	Cluster2
p07	g	(5, 5)	1.57	2.30	Cluster1
p08	h	(7, 3)	4.37	0.56	Cluster2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

$$\begin{aligned} m1 & (3.67, 5.83) \\ m2 & (6.75, 3.50) \end{aligned}$$

## K-Means Clustering

# K-Means Clustering ( $K=2$ , two clusters)

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

# K-Means Clustering

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)		

# Summary

- Cluster Analysis
- *K-Means Clustering*

# References

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