



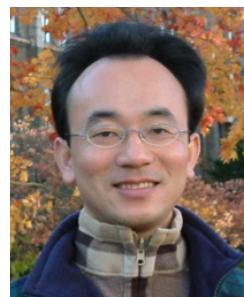
Data Mining 資料探勘

分群分析 (Cluster Analysis)

1022DM04

MI4

Wed, 6,7 (13:10-15:00) (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/>

2014-03-12



課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
1	103/02/19	資料探勘導論 (Introduction to Data Mining)
2	103/02/26	關連分析 (Association Analysis)
3	103/03/05	分類與預測 (Classification and Prediction)
4	103/03/12	分群分析 (Cluster Analysis)
5	103/03/19	個案分析與實作一 (SAS EM 分群分析)： Case Study 1 (Cluster Analysis – K-Means using SAS EM)
6	103/03/26	個案分析與實作二 (SAS EM 關連分析)： Case Study 2 (Association Analysis using SAS EM)
7	103/04/02	教學行政觀摩日 (Off-campus study)
8	103/04/09	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)

課程大綱 (Syllabus)

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

9 103/04/16 期中報告 (Midterm Project Presentation)

10 103/04/23 期中考試週 (Midterm Exam)

11 103/04/30 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)

12 103/05/07 文字探勘與網頁探勘 (Text and Web Mining)

13 103/05/14 海量資料分析 (Big Data Analytics)

14 103/05/21 期末報告 (Final Project Presentation)

15 103/05/28 畢業考試週 (Final Exam)

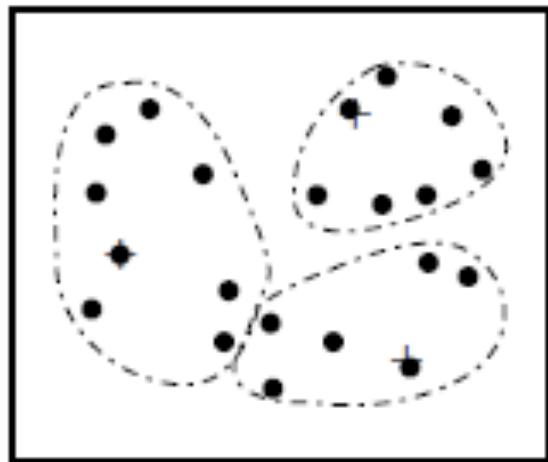
Outline

- Cluster Analysis
- *K-Means* Clustering

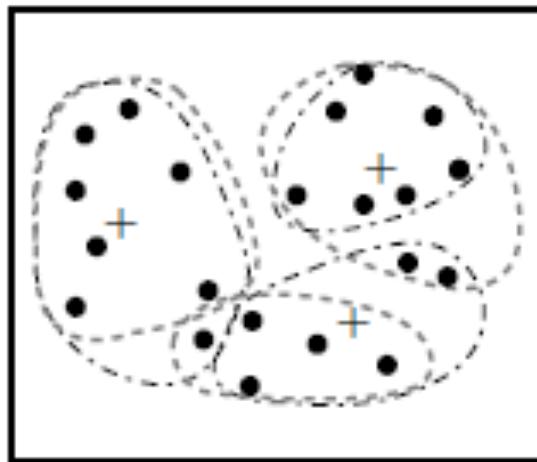
Cluster Analysis

- 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**

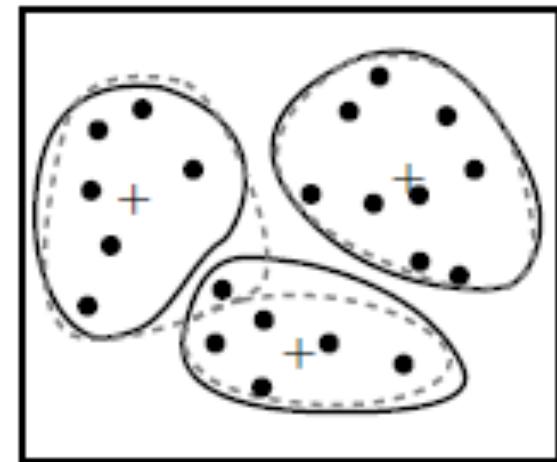
Cluster Analysis



(a)



(b)



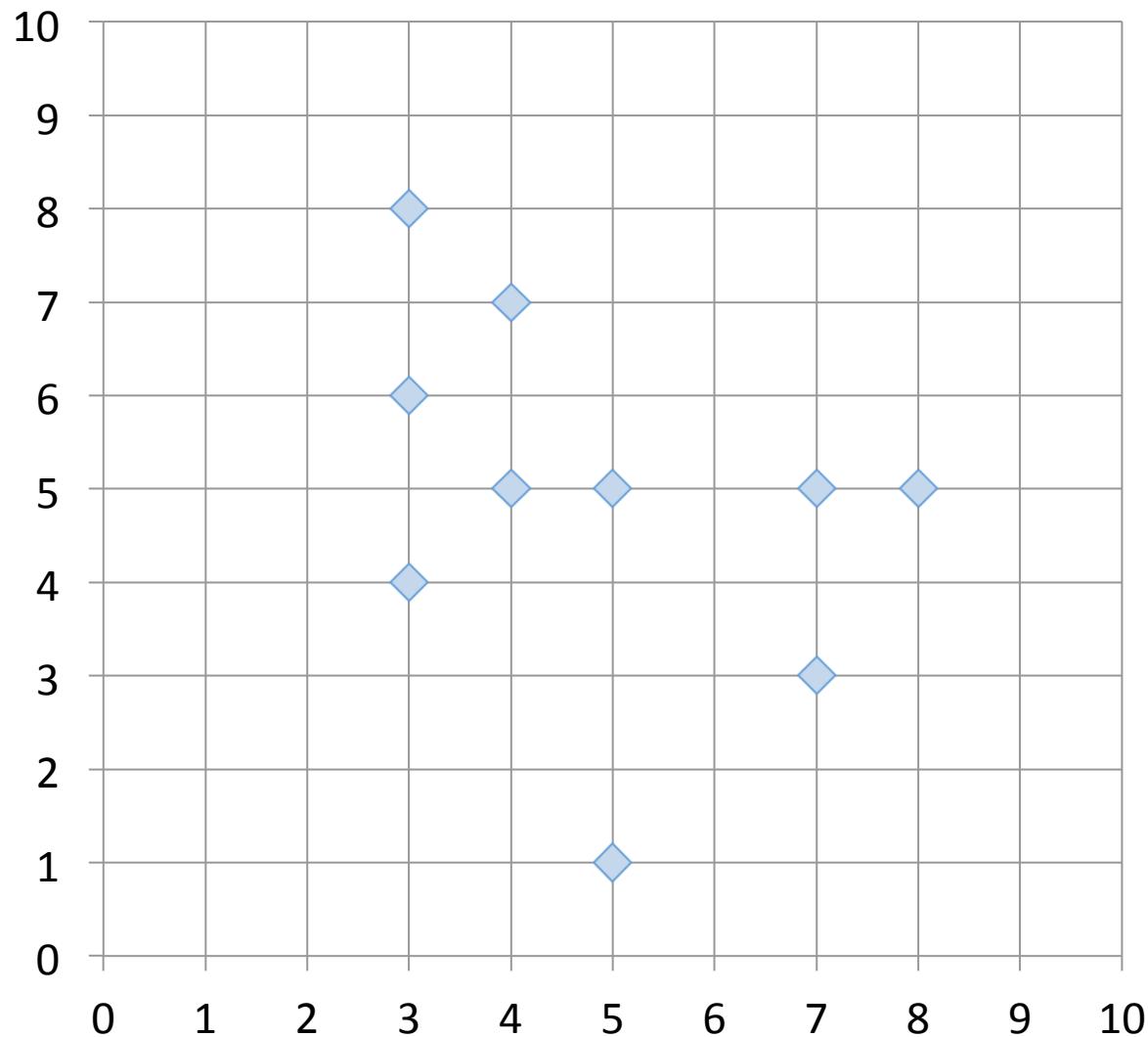
(c)

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)

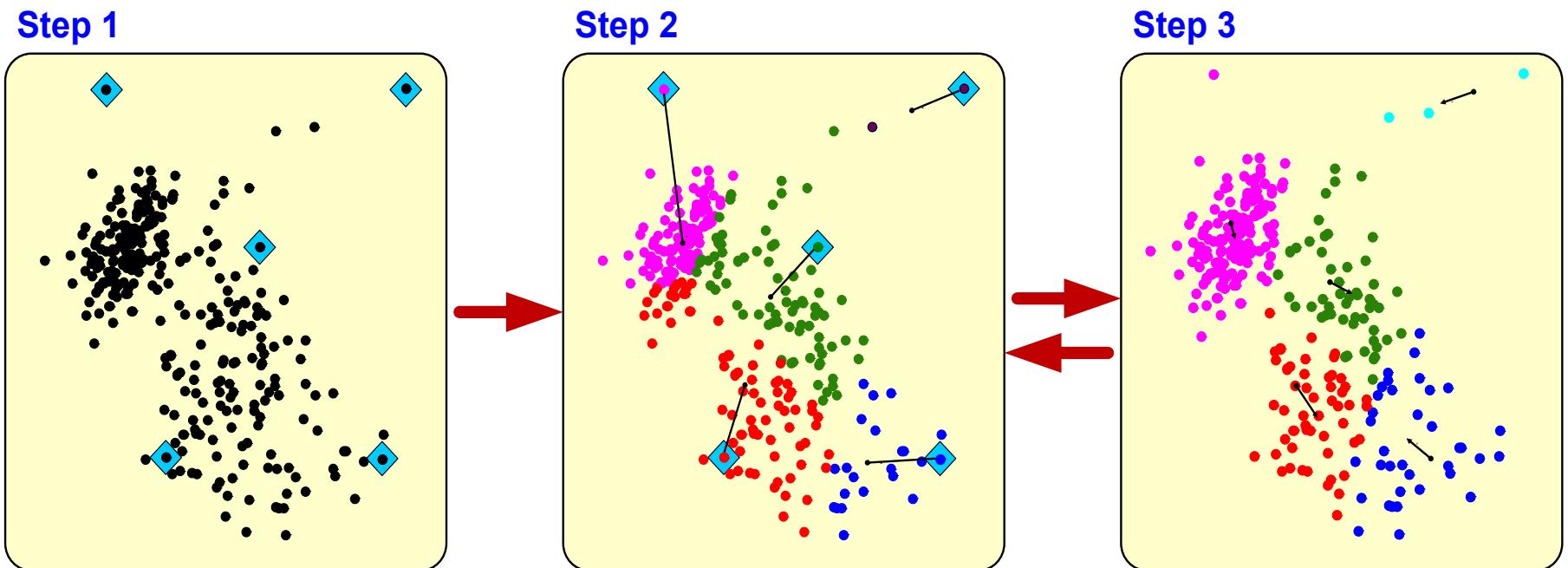
Step 1: Randomly generate k random points as initial cluster centers

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 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:

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

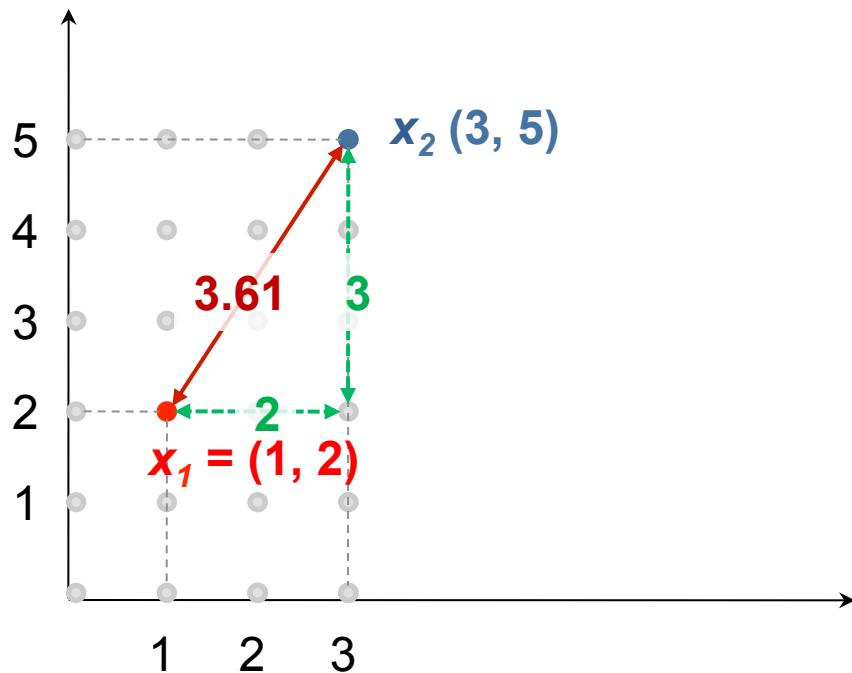
- Properties

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

- Also, one can use weighted distance, parametric Pearson product moment correlation, or other dissimilarity measures

Euclidean distance vs Manhattan distance

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

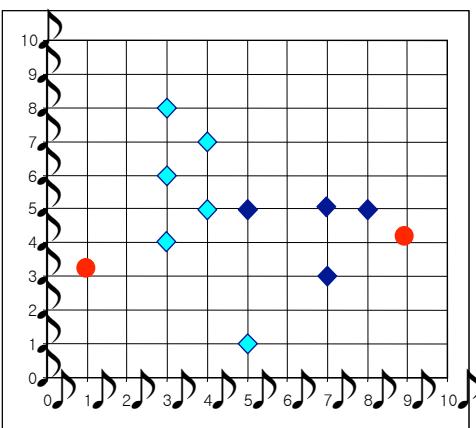


Euclidean distance:
 $= ((3-1)^2 + (5-2)^2)^{1/2}$
 $= (2^2 + 3^2)^{1/2}$
 $= (4 + 9)^{1/2}$
 $= (13)^{1/2}$
 $= 3.61$

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

The *K*-Means Clustering Method

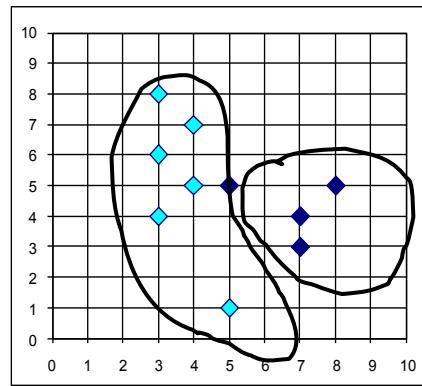
- Example



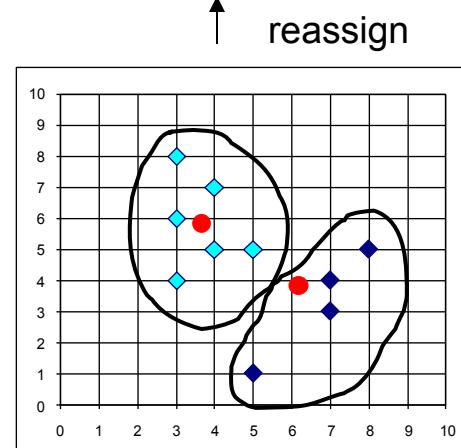
$K=2$

Arbitrarily choose K object as initial cluster center

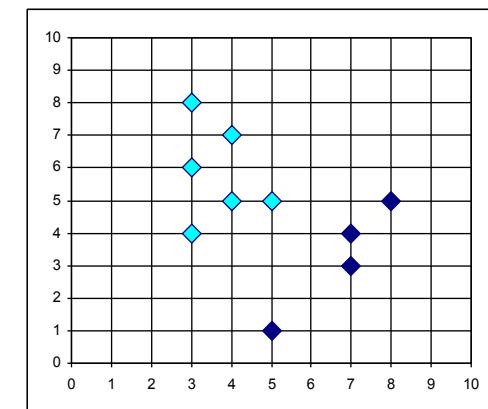
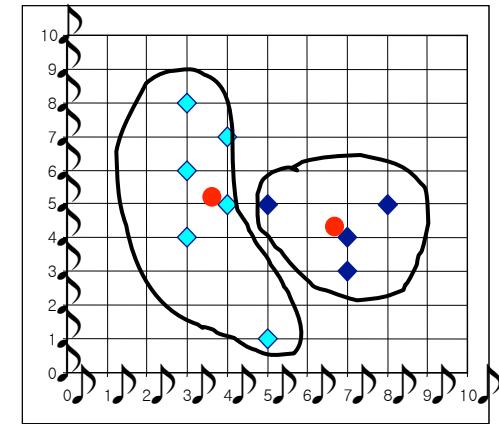
Assign each objects to most similar center



Update the cluster means

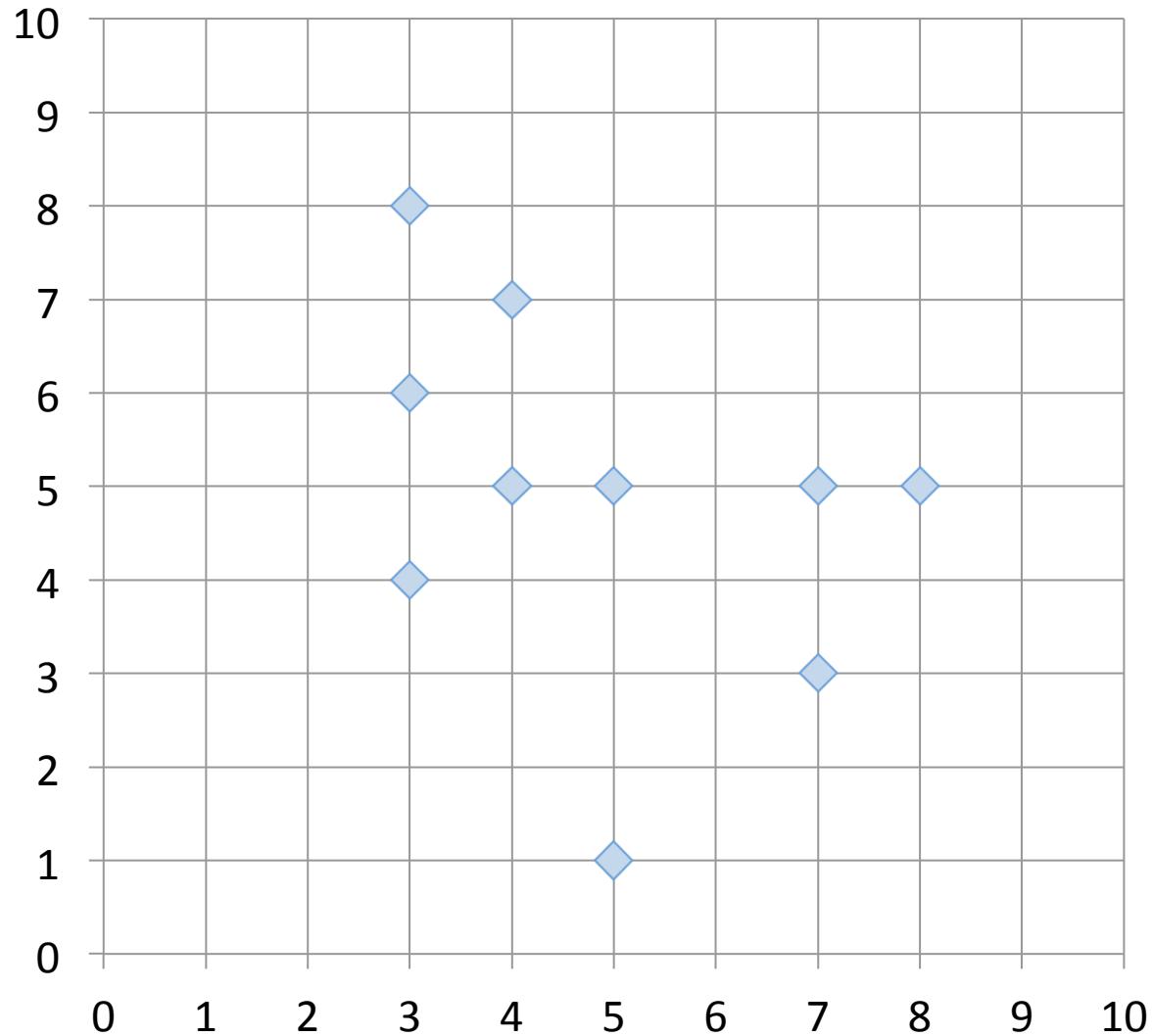


Update the cluster means



K-Means Clustering

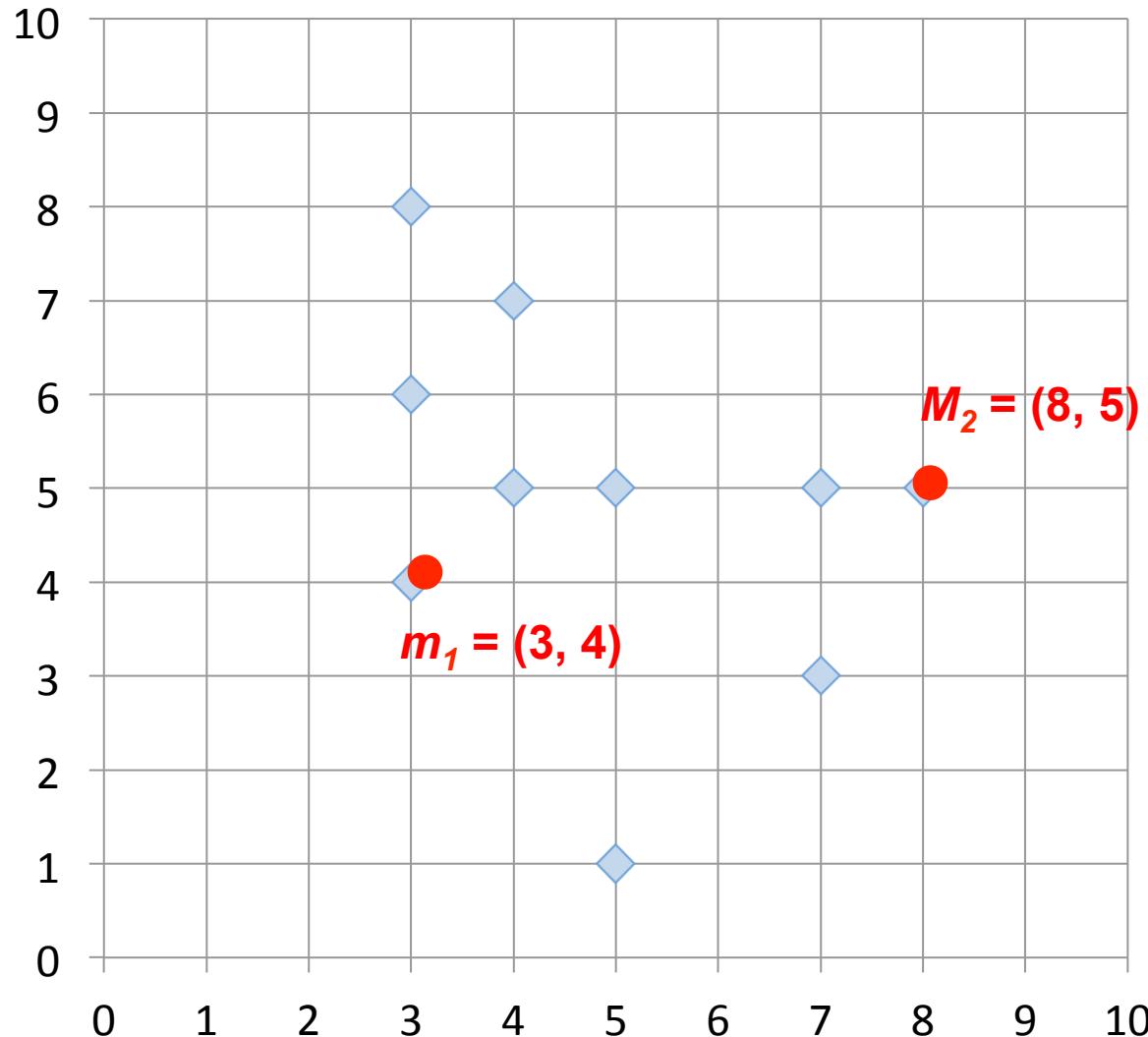
Step by Step



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

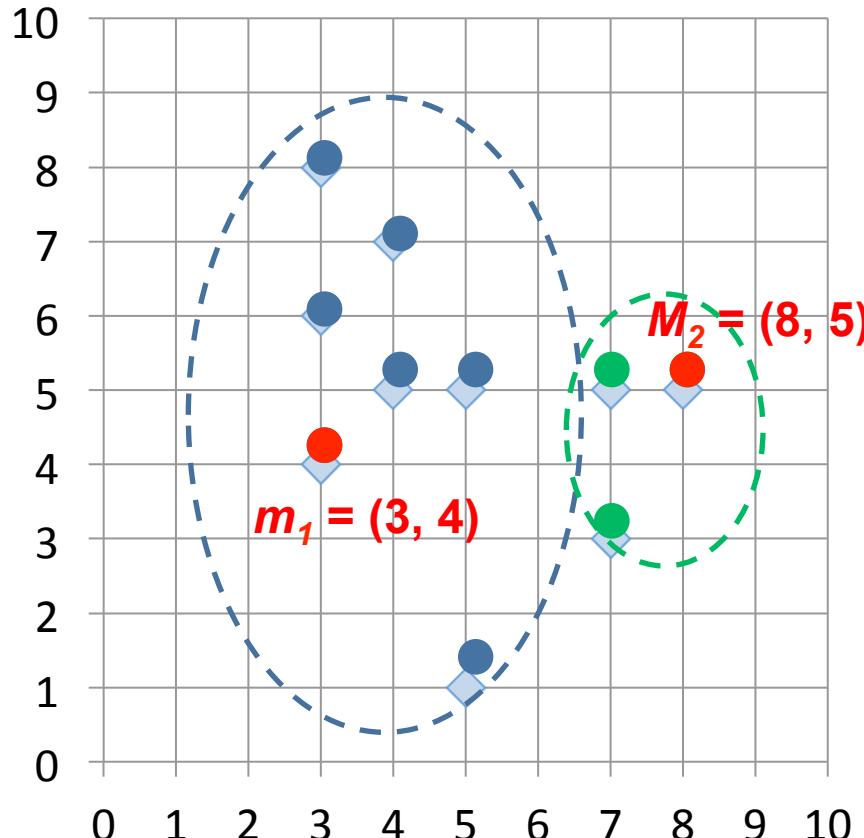
Step 1: K=2, Arbitrarily choose K object as initial cluster center



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



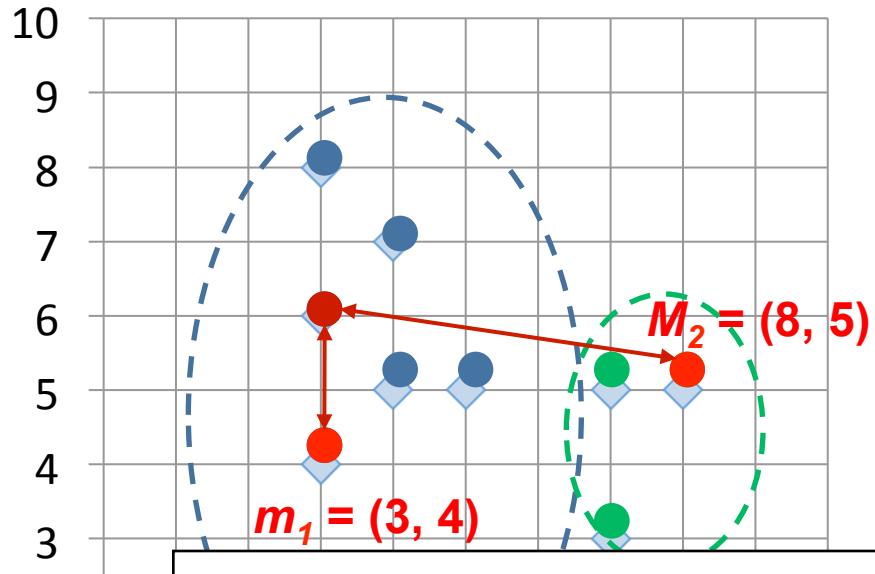
K-Means Clustering

Initial m_1 (3, 4)

Initial m_2 (8, 5)

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

Step 3: Assign each objects to most similar center



$$\begin{aligned}
 & \text{Euclidean distance} \\
 & b(3,6) \leftrightarrow m1(3,4) \\
 & = ((3-3)^2 + (6-4)^2)^{1/2} \\
 & = (0^2 + (-2)^2)^{1/2} \\
 & = (0 + 4)^{1/2} \\
 & = (4)^{1/2} \\
 & = 2.00
 \end{aligned}$$

K-1

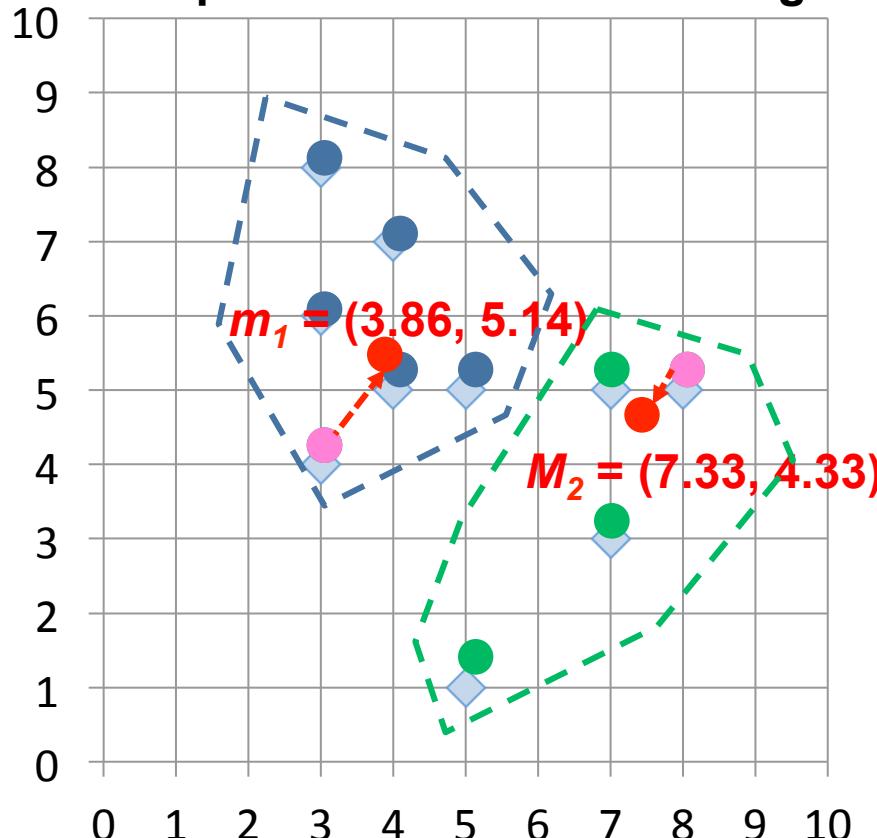
Point	P	P(x,y)	m1 distance	m2 distance	Cluster
p01	a	(3, 4)	0.00	5.10	Cluster1
p02	b	(3, 6)	2.00	5.10	Cluster1
p03	c	(3, 8)	4.00	5.83	Cluster1
p04	d	(4, 5)	1.41	4.00	Cluster1
p05					ster1
p06					ster1
p07					ster1
p08					ster2
p09					ster2
p10					ster2

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

Initial m1 (3, 4)

Initial m2 (8, 5)

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

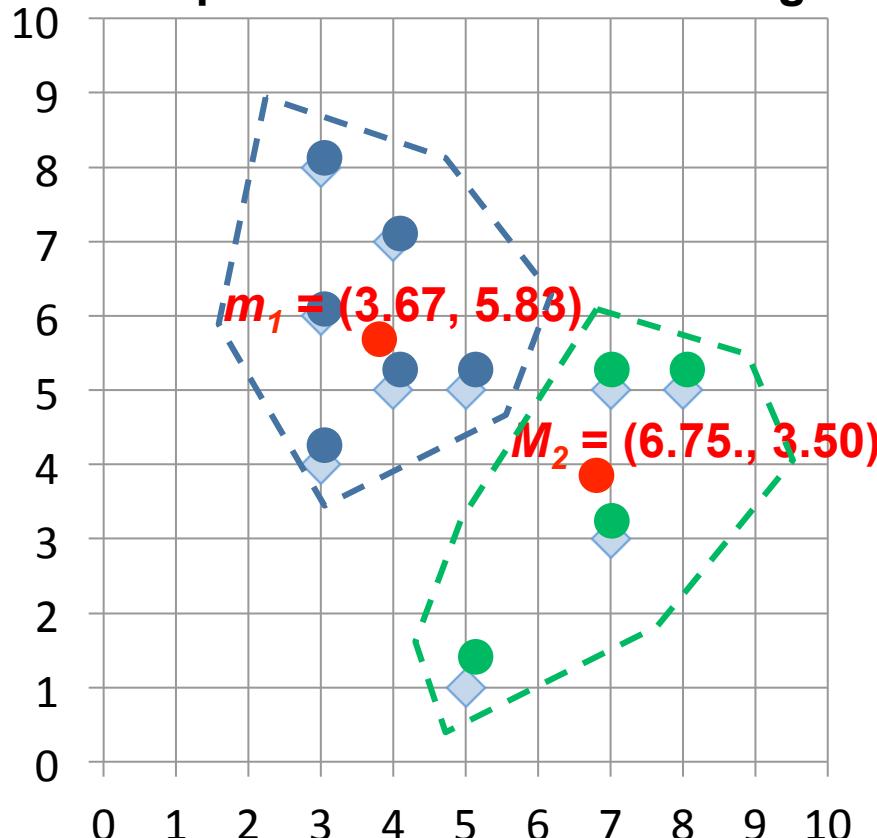


Point	P	P(x,y)	m1 distance	m2 distance	Cluster
p01	a	(3, 4)	1.43	4.34	Cluster1
p02	b	(3, 6)	1.22	4.64	Cluster1
p03	c	(3, 8)	2.99	5.68	Cluster1
p04	d	(4, 5)	0.20	3.40	Cluster1
p05	e	(4, 7)	1.87	4.27	Cluster1
p06	f	(5, 1)	4.29	4.06	Cluster2
p07	g	(5, 5)	1.15	2.42	Cluster1
p08	h	(7, 3)	3.80	1.37	Cluster2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

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

K-Means Clustering

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



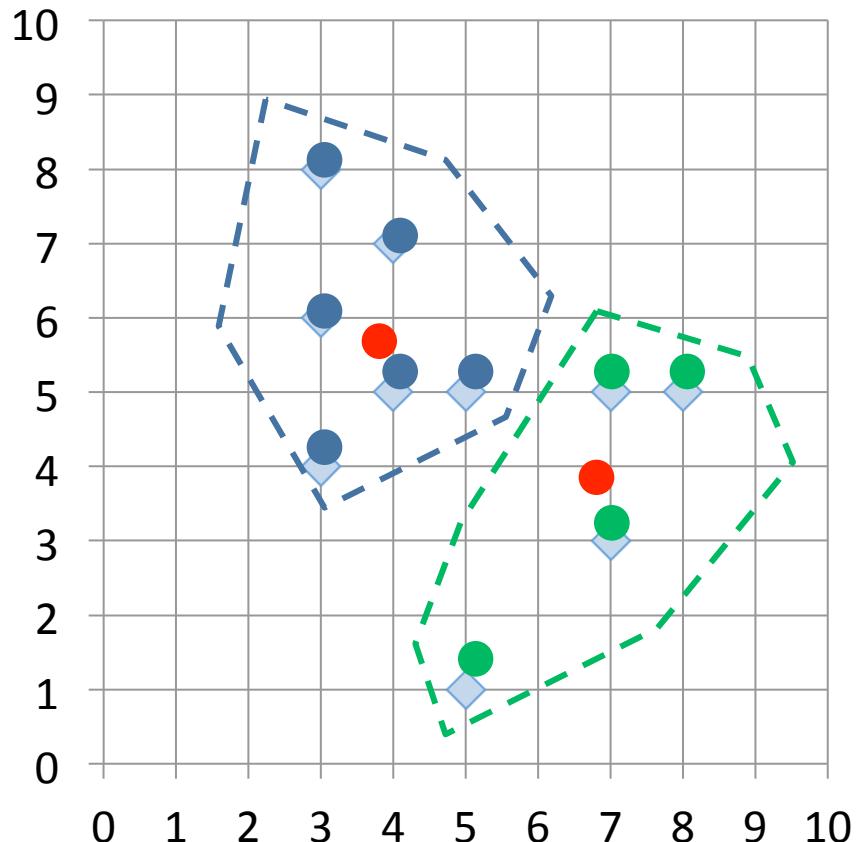
Point	P	P(x,y)	m1 distance	m2 distance	Cluster
p01	a	(3, 4)	1.95	3.78	Cluster1
p02	b	(3, 6)	0.69	4.51	Cluster1
p03	c	(3, 8)	2.27	5.86	Cluster1
p04	d	(4, 5)	0.89	3.13	Cluster1
p05	e	(4, 7)	1.22	4.45	Cluster1
p06	f	(5, 1)	5.01	3.05	Cluster2
p07	g	(5, 5)	1.57	2.30	Cluster1
p08	h	(7, 3)	4.37	0.56	Cluster2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

$$m_1 \quad (3.67, 5.83)$$

$$m_2 \quad (6.75, 3.50)$$

K-Means Clustering

stop when no more new assignment



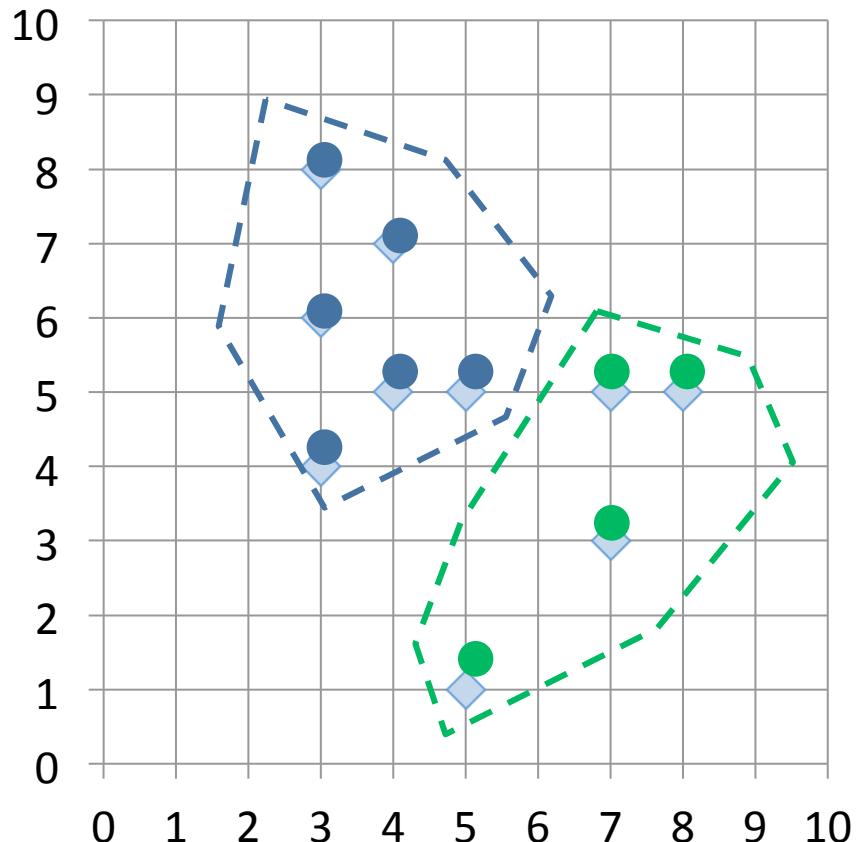
K-Means Clustering

$m_1 (3.67, 5.83)$

$m_2 (6.75, 3.50)$

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

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.