جامعة الطائف
TAIF UNIVERSITY

DATA MINING

Assoc. Prof. Dr. Salha Alzahrani

College of Computers and Information Technology
Taif University
Saudi Arabia

s.zahrani@tu.edu.sa

Prediction | Association Rule Mining- Part: II

Recap of Lecture 8

- Association Rule Mining- Part: II: Introduction
- Transactions and Itemsets
- Support for an Itemset
- Association Rules
- Generating Association Rules: Apriori Algorithm
- Generating Association Rules: Apriori-gen Algorithm
- Generating association rules using Apriori algorithm: Example
- Handout (7): Generating Association Rules for Market Basked Using Apriori Algorithm

Recap of Previous Lecture
Content of This Lecture
Summary & Checklist

Content of Lecture 9

- Introduction: What is clustering?
- Distance Measure between Objects
- Centroid
- K-means Clustering Algorithm
- K-means Clustering Algorithm: Example
- Handout (8) : Example of using the k-means Clustering Algorithm
- Summary & Checklist.

Clustering | What is clustering?

Clustering: Grouping together objects (e.g. instances in a dataset) that are similar to each other and (relatively) dissimilar to the objects belonging to other clusters.

- In many fields there are obvious benefits to be had from grouping together similar objects. For example
  - In an economics application, we might be interested in finding countries whose economies are similar.
  - In a marketing application, we might wish to find clusters of customers with similar buying behaviour.
  - In a medical application, we might wish to find clusters of patients with similar symptoms.
  - In a document retrieval application, we might wish to find clusters of documents with related content.

Clustering | Introduction

- Object can be described by the values of just two attributes.
- So, we can represent them as points in a two-dimensional space (a plane).

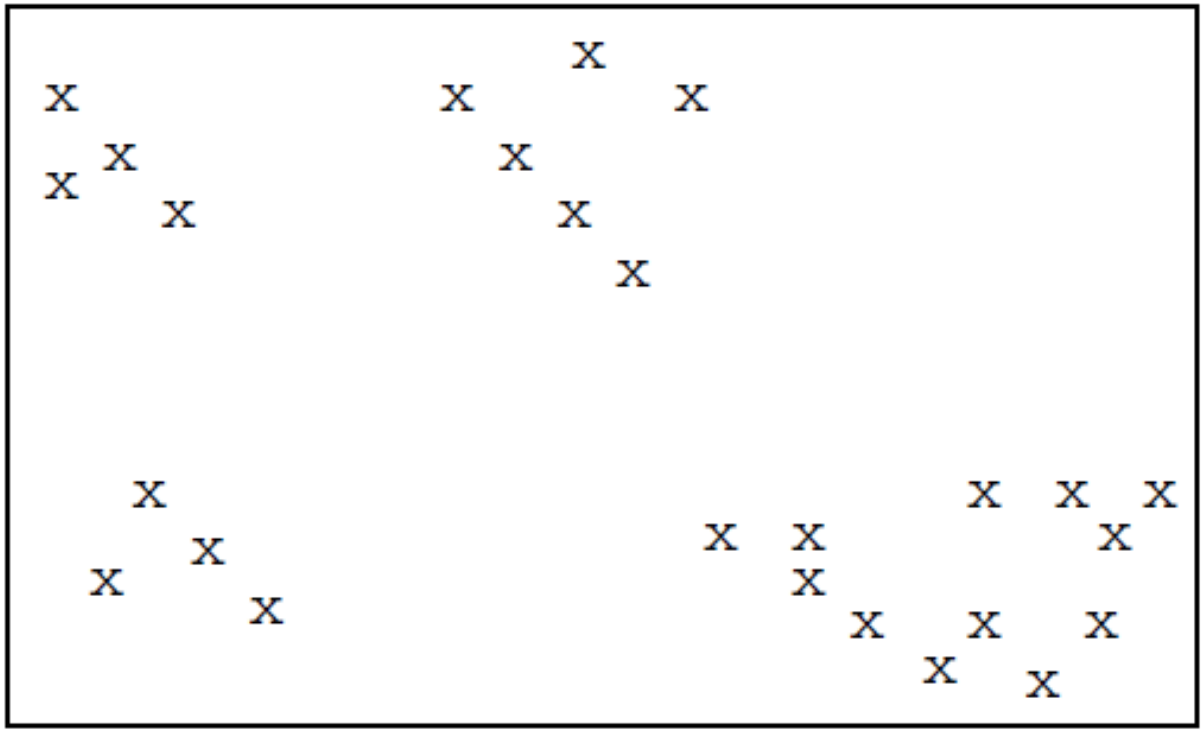


Figure 14.1 Objects for Clustering

Clustering | Introduction

- It is usually easy to visualise clusters in two dimensions.
- The points seem to fall naturally into four groups as shown by the curves drawn surrounding sets of points.

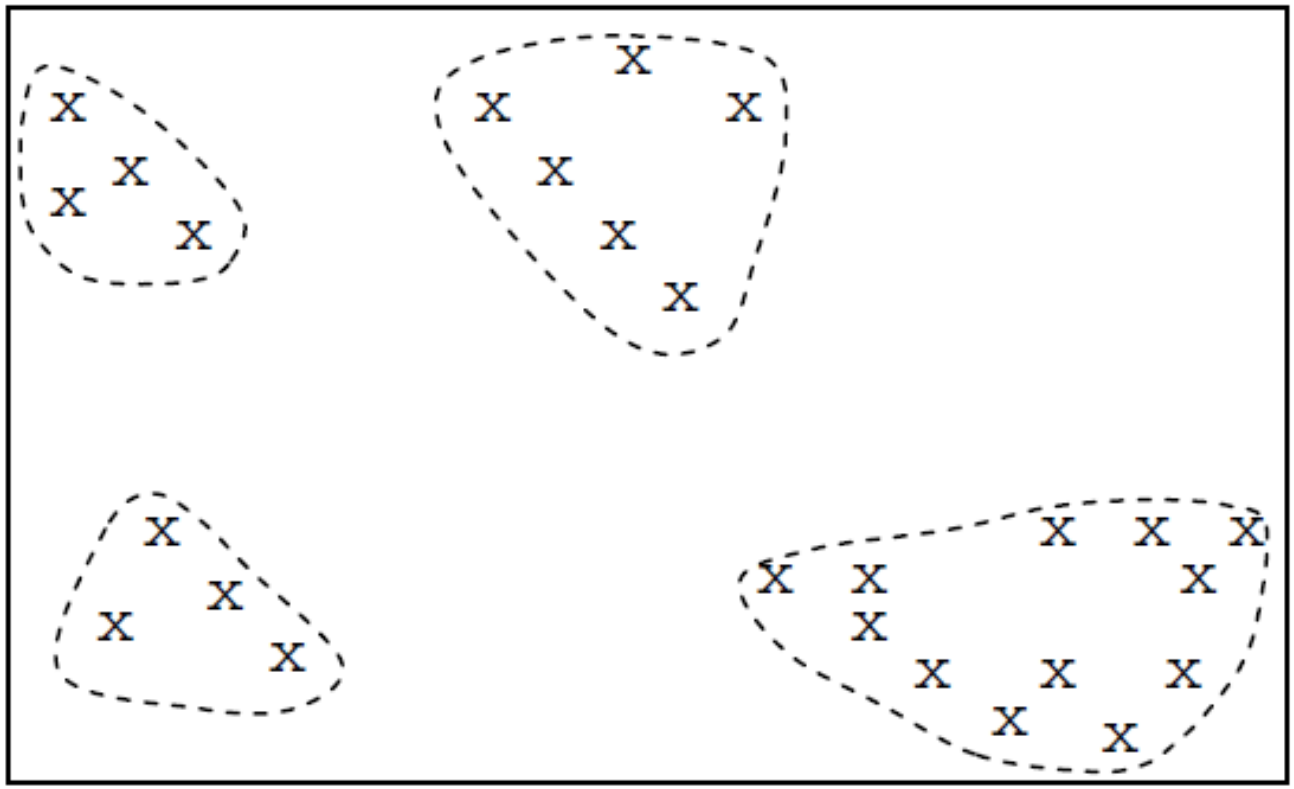


Figure 14.2 Clustering of Objects in Figure 14.1(a)

Clustering | Introduction

- However, there is frequently more than one possibility.

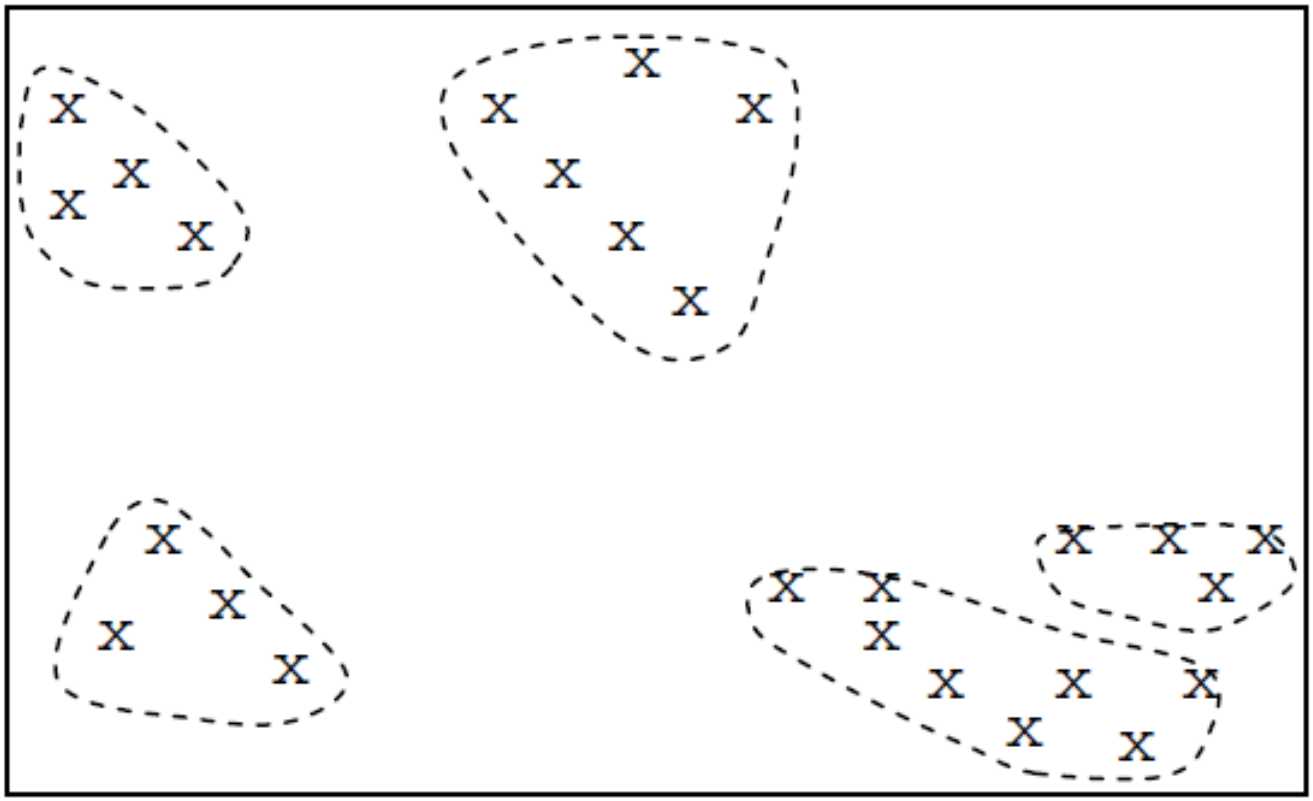
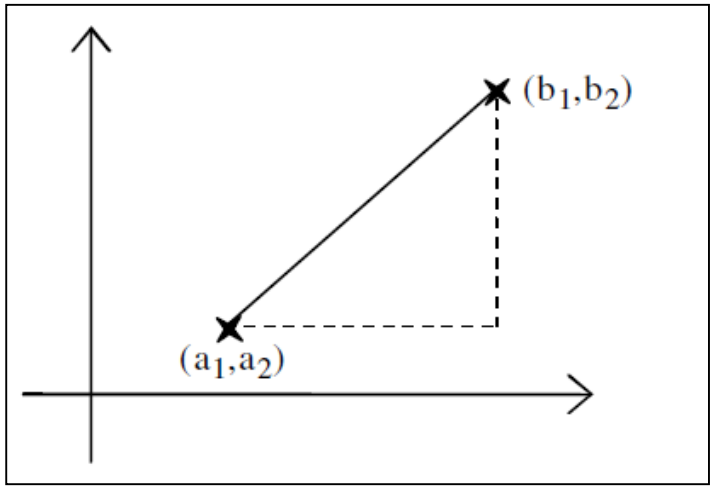


Figure 14.3 Clustering of Objects in Figure 14.1(b)

Clustering | Distance Measure between Objects

- It is first necessary to decide on a way of measuring the distance between two points.
- As for nearest neighbour classification, a measure commonly used when clustering is the Euclidean distance.
- We will assume that all attribute values are continuous.

$$\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$$



$$\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + (a_3 - b_3)^2}$$

$$\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + ... + (a_n - b_n)^2}$$

Clustering | Centroid

- We need to introduce the notion of **the 'centre' of a cluster**, generally called its ***centroid***.
- Assuming that we are using **Euclidean distance** or something similar as a measure, we define the centroid of a cluster to be the point for which each attribute value is the average of the values of the corresponding attribute for all the points in the cluster.
- Example: the centroid of the four points (with 6 attributes):

8.0	7.2	0.3	23.1	11.1	−6.1
2.0	−3.4	0.8	24.2	18.3	−5.2
−3.5	8.1	0.9	20.6	10.2	−7.3
−6.0	6.7	0.5	12.5	9.2	−8.4

would be

0.125	4.65	0.625	20.1	12.2	−6.75

Clustering | K-means Clustering Algorithm

There are many methods of clustering. The most commonly used algorithms are: *k-means clustering* and *hierarchical clustering*.

- *k*-means clustering is an ***exclusive clustering algorithm***. Each object is assigned to precisely one of a set of clusters.
- For this method of clustering, we start by deciding how many clusters we would like to form from the data.

  We call this value ***k***.
- The value of *k* is generally a small integer, such as 2, 3, 4 or 5, but may be larger.

Clustering | K-means Clustering Algorithm

1. Choose a value of k.
2. Select k objects in an arbitrary fashion. Use these as the initial set of k centroids.
3. Assign each of the objects to the cluster for which it is nearest to the centroid.
4. Recalculate the centroids of the k clusters.
5. Repeat steps 3 and 4 until the centroids no longer move.

Figure 14.4 The k-Means Clustering Algorithm

Clustering | K-means Clustering Algorithm: Example

- We will use the *k*-means algorithm to cluster **16 objects**.
- Each object with **two attributes** *x* and *y*.

x	y
6.8	12.6
0.8	9.8
1.2	11.6
2.8	9.6
3.8	9.9
4.4	6.5
4.8	1.1
6.0	19.9
6.2	18.5
7.6	17.4
7.8	12.2
6.6	7.7
8.2	4.5
8.4	6.9
9.0	3.4
9.6	11.1

Figure 14.5 Objects For Clustering (Attribute Values)

Clustering | K-means Clustering Algorithm: Example

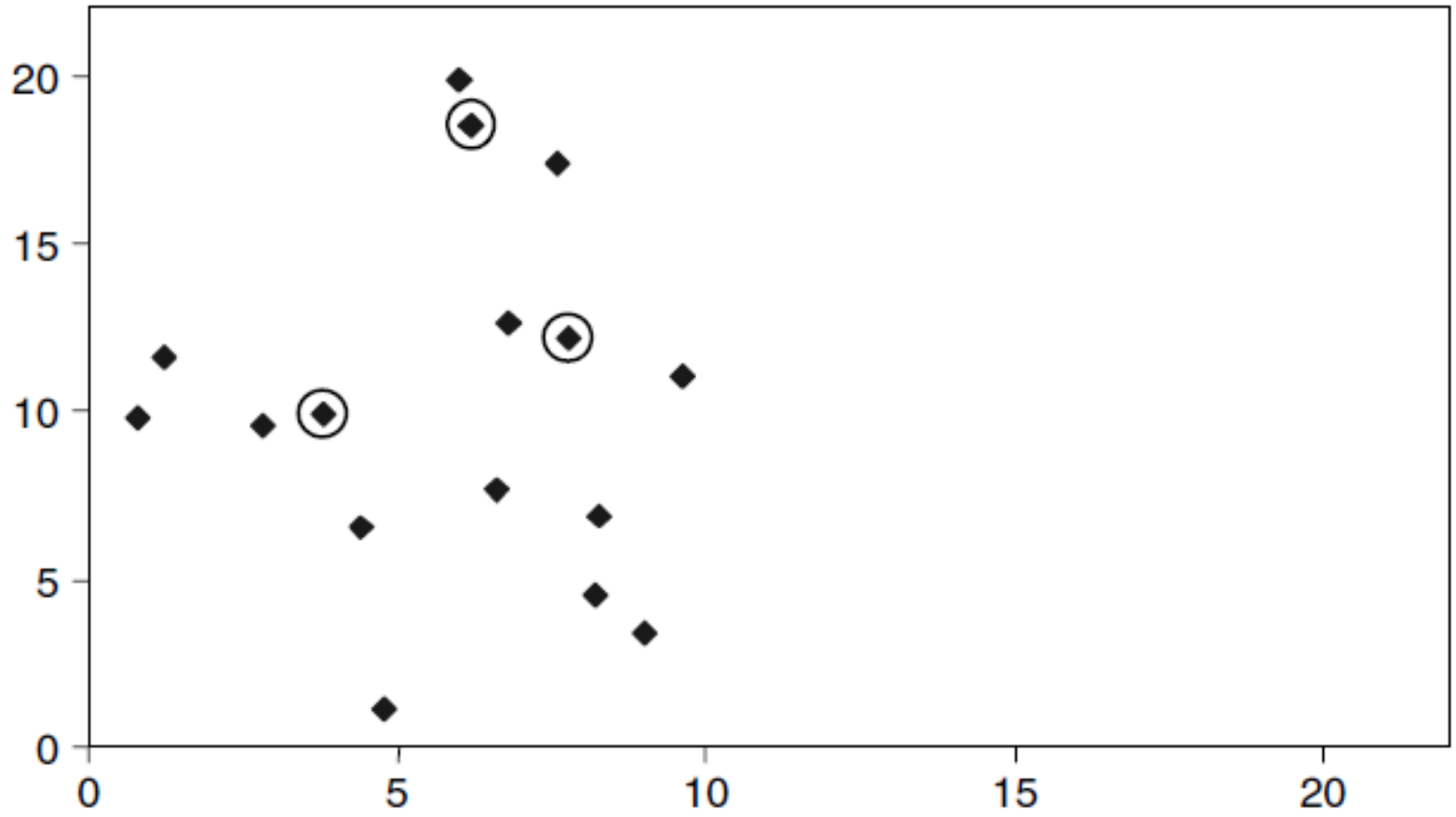


Figure 14.6 Objects For Clustering

Step 1: Set initial k and initial centroids

- We will assume that we have chosen ***k* = 3** and that three points have been selected to be the locations of the initial three **centroids**.
- This initial (arbitrary) centroids are shown below.

	Initial	
	x	y
Centroid 1	3.8	9.9
Centroid 2	7.8	12.2
Centroid 3	6.2	18.5

Figure 14.7 Initial Choice of Centroids

Step 2: Calculate the Euclidean distance

- Calculate the Euclidean distance of each of the 16 points from the three centroids.
- For example, the distance of the first point (6.8, 12.6) from the first centroid (3.8, 9.9) is simply

$$\sqrt{(6.8-3.8)^2+(12.6-9.9)^2} = 4.0 \text{ (to one decimal place)}$$

Step 3: Assign objects to clusters

- The column ‘cluster’ indicates the centroid closest to each point and thus the cluster to which it should be assigned.

Clustering | K-means Clustering Algorithm: Example

x	y	$d1$	$d2$	$d3$	cluster
6.8	12.6	4.0	1.1	5.9	2
0.8	9.8	3.0	7.4	10.2	1
1.2	11.6	3.1	6.6	8.5	1
2.8	9.6	1.0	5.6	9.5	1
3.8	9.9	0.0	4.6	8.9	1
4.4	6.5	3.5	6.6	12.1	1
4.8	1.1	8.9	11.5	17.5	1
6.0	19.9	10.2	7.9	1.4	3
6.2	18.5	8.9	6.5	0.0	3
7.6	17.4	8.4	5.2	1.8	3
7.8	12.2	4.6	0.0	6.5	2
6.6	7.7	3.6	4.7	10.8	1
8.2	4.5	7.0	7.7	14.1	1
8.4	6.9	5.5	5.3	11.8	2
9.0	3.4	8.3	8.9	15.4	1
9.6	11.1	5.9	2.1	8.1	2

Figure 14.8 Objects For Clustering (Augmented)

Clustering | K-means Clustering Algorithm: Example

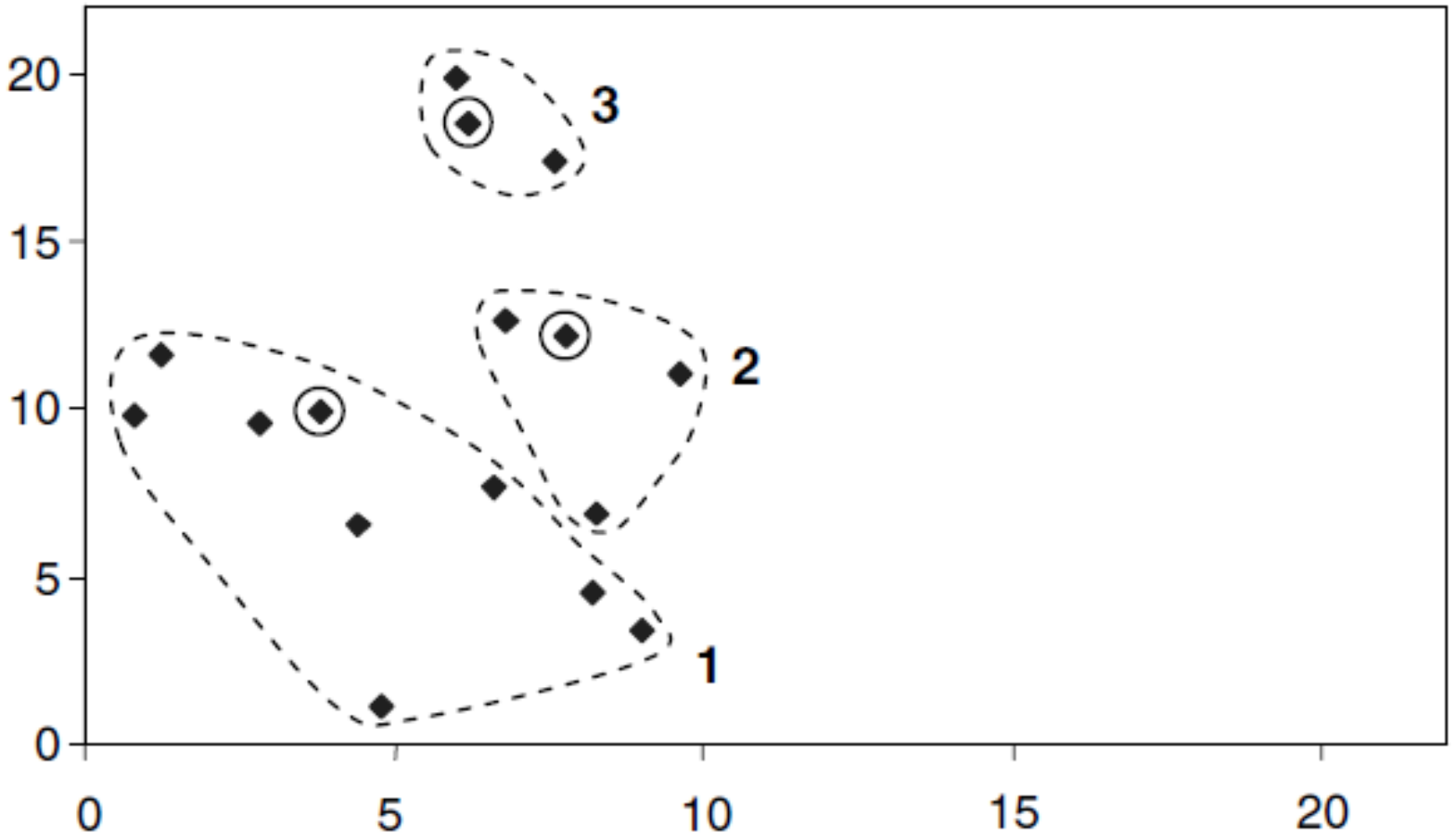


Figure 14.9 Initial Clusters

Step 4: Recalcaute the centroids of each cluster

We next calculate the centroids of the three clusters using the *x* and *y* values of the objects currently assigned to each one.

	Initial		After first iteration	
	x	y	x	y
Centroid 1	3.8	9.9	4.6	7.1
Centroid 2	7.8	12.2	8.2	10.7
Centroid 3	6.2	18.5	6.6	18.6

Figure 14.10 Centroids After First Iteration

Clustering | K-means Clustering Algorithm: Example

- The three centroids have all been moved by the assignment process, but the movement of the third one is less than for the other two.

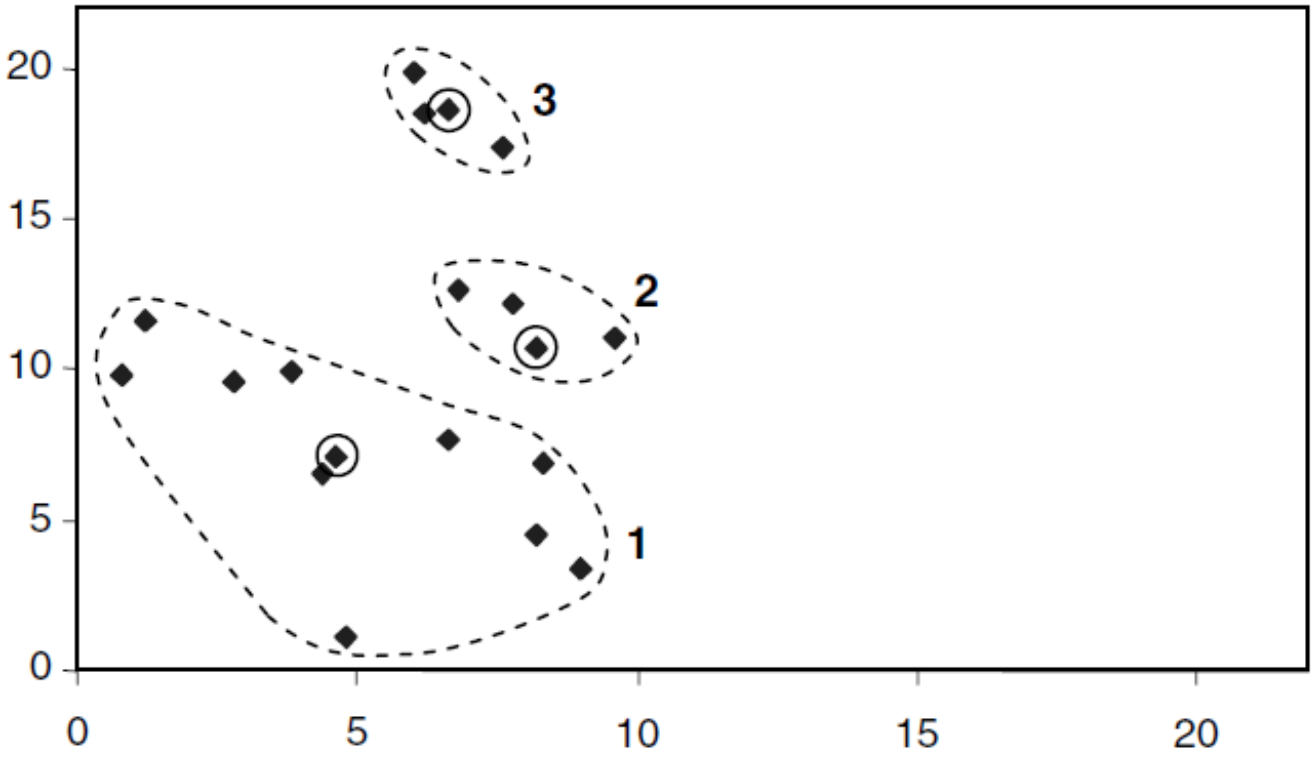


Figure 14.11 Revised Clusters

Clustering | K-means Clustering Algorithm: Example

Repeat Step 2 & Step 3: Calculate the Euclidean distance from new centroids, then assign objects to clusters

- This gives the revised set of clusters shown in Figure 14.11.
- The centroids from now on the
- centroids are 'imaginary points' corresponding to the 'centre' of each cluster, not actual points within the clusters.
- In fact only one point has moved. The object at (8.3, 6.9) has moved from cluster 2 to cluster 1.

x	y	*d1*	*d2*	*d3*	*Cluster*
6.8	12.6				
0.8	9.8				
1.2	11.6				
2.8	9.6				
3.8	9.9				
4.4	6.5				
4.8	1.1				
6.0	19.9				
6.2	18.5				
7.6	17.4				
7.8	12.2				
6.6	7.7				
8.2	4.5				
8.4	6.9				
9.0	3.4				
9.6	11.1				

Repeat Step 4: recalculate the positions of the three centroids.

	Initial		After first iteration		After second iteration	
	x	y	x	y	x	y
Centroid 1	3.8	9.9	4.6	7.1	5.0	7.1
Centroid 2	7.8	12.2	8.2	10.7	8.1	12.0
Centroid 3	6.2	18.5	6.6	18.6	6.6	18.6

Figure 14.12 Centroids After First Two Iterations

Clustering | K-means Clustering Algorithm: Example

- The first two centroids have moved a little, but the third has not moved at all.
- We assign the 16 objects to clusters once again, as below.

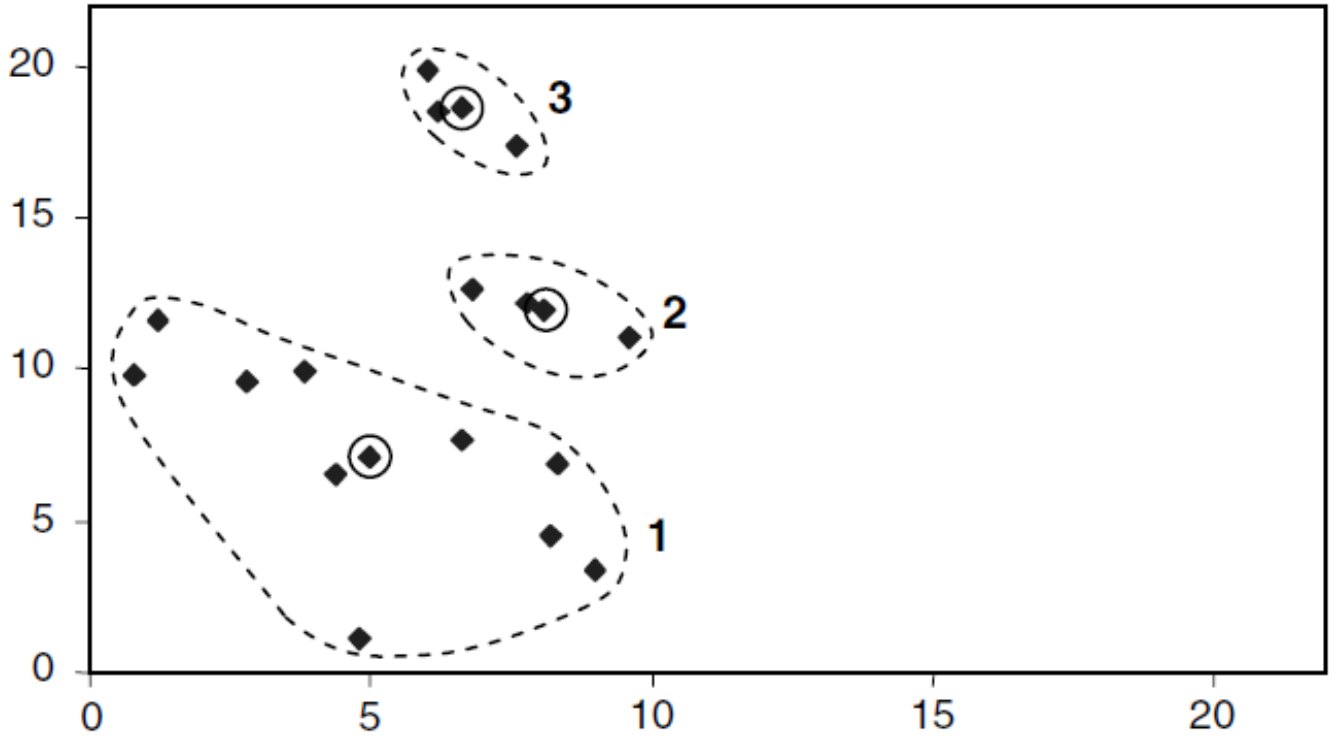


Figure 14.13 Third Set of Clusters

Clustering | K-means Clustering Algorithm: Example

- These are **the same clusters** as before.
- Their **centroids will be the same** as those from which the clusters were generated.
- Hence the termination condition of the *k*-means algorithm **'repeat ... until the centroids no longer move'** has been met and these are the final clusters produced by the algorithm.

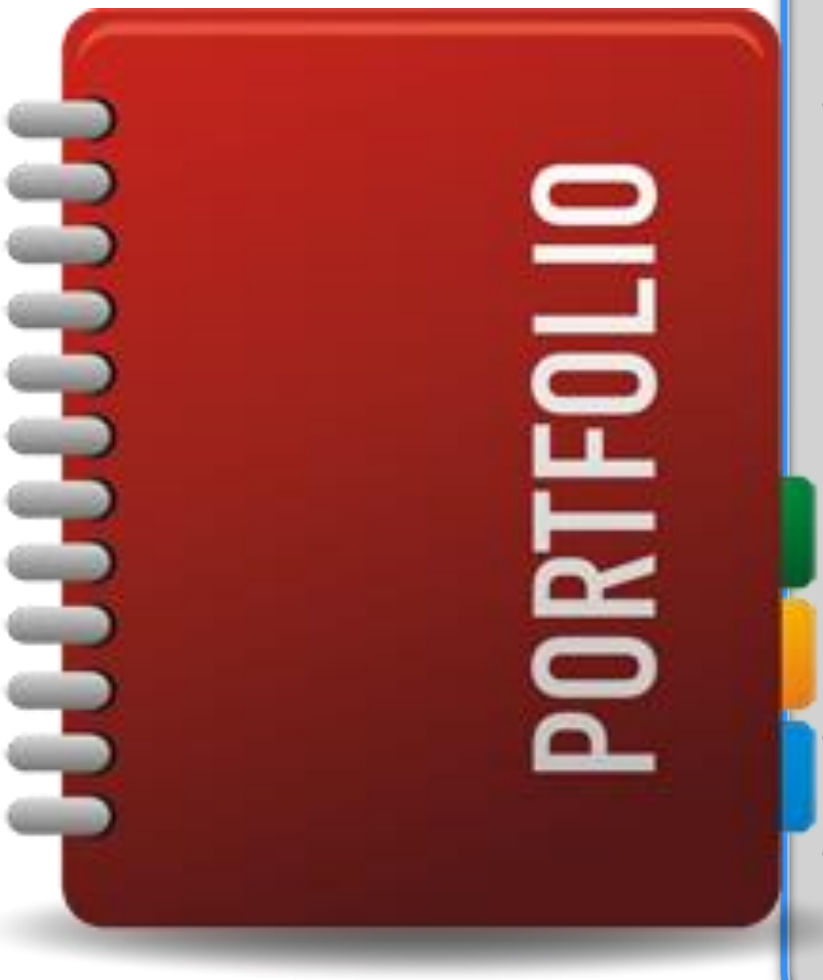
Clustering | Handout (8) : Example of using the k-means Clustering Algorithm

Recap of Previous Lecture
Content of This Lecture
Summary & Checklist

Summary & Checklist

- ☑ Introduction: What is clustering?
- ☑ Distance Measure between Objects
- ☑ Centroid
- ☑ K-means Clustering Algorithm
- ☑ K-means Clustering Algorithm: Example
- ☑ Handout (8) : Example of using the k-means Clustering Algorithm.

Reminder | Student Portfolio



- **Each student should prepare her own course portfolio!**
- **Portfolios should include the following parts:**
  1) *Course Syllabus*
  2) *Lecture notes (slides)*
  3) *Assignments*
  4) *Quizzes*
  5) *Mid-term exam and answer sheet.*
  6) *Research articles and other supporting materials.*
  7) *Lab lecture notes, exercises, and MATLAB codes.*
  8) *Glossary*
- **Portfolios will be checked regularly by the instructor.**
- **Students who prepare good course portfolios may be given a BONUS +2/+5 on their examinations, if needed.**

Final Exam | Be Ready!

Final Exam [40]

Covers everything: Lecture 1 - Lecture 9

Thank You !

✉ s.zahrani@tu.edu.sa