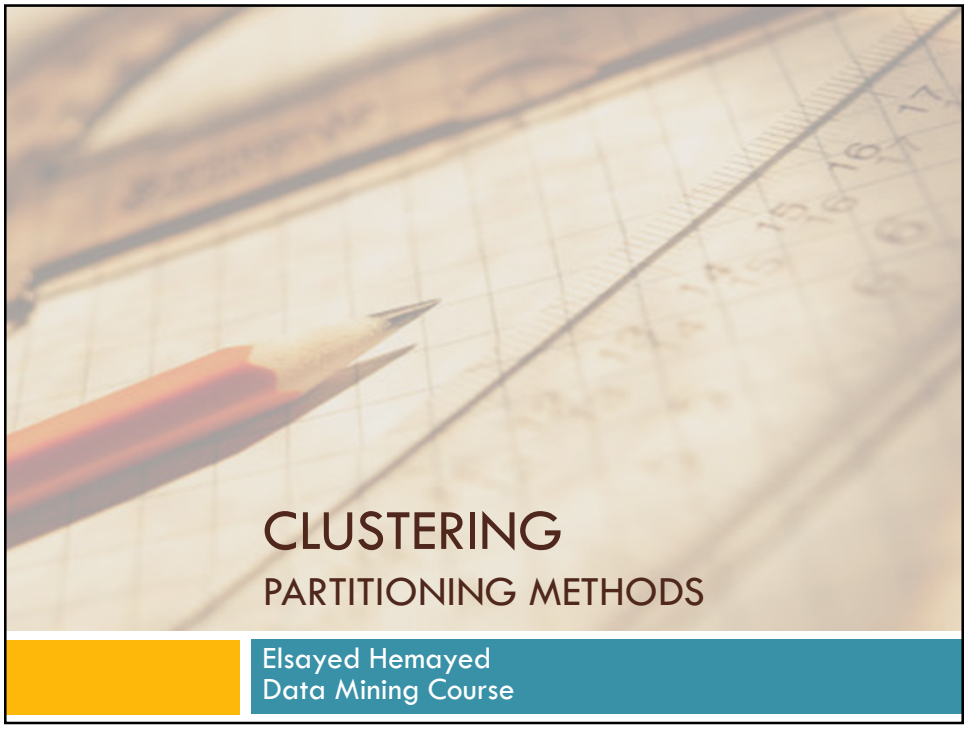


CLUSTERING

PARTITIONING METHODS

Elsayed Hemayed
Data Mining Course

Outline

- Introduction
- Clustering Methods
- K-Means Clustering
- Selecting K

- Acknowledgement: some of the material in these slides are from [Max Bramer, "Principles of Data Mining", Springer-Verlag London Limited 2007]

Introduction

3

- Extracting information from unlabelled data.
- *Clustering is* concerned with grouping together objects that are similar to each other and dissimilar to the objects belonging to other clusters.
- Examples:
  - In an economics application we might be interested in finding countries whose economies are similar.
  - In a financial application we might wish to find clusters of companies that have similar financial performance.
  - In a marketing application we might wish to find clusters of customers with similar buying behaviour.

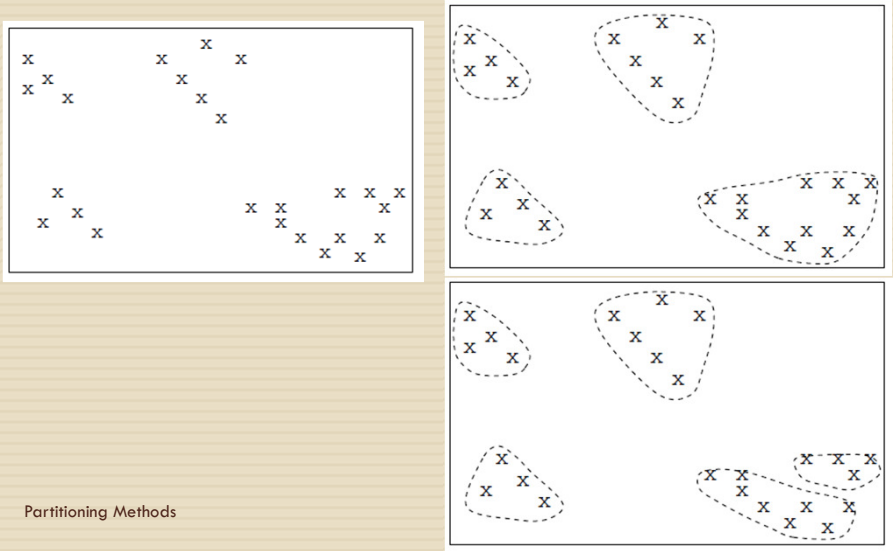
Clustering Examples

4

- 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.
- In a crime analysis application we might look for clusters of high volume crimes such as burglaries or try to cluster together much rarer (but possibly related) crimes such as murders.

Clustering Example

5



Partitioning Methods

Clustering: Application 1

6

- Market Segmentation:
  - Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
  - Approach:
    - Collect different attributes of customers based on their geographical and lifestyle related information.
    - Find clusters of similar customers.
    - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Partitioning Methods

Clustering: Application 2

7

- Document Clustering:
  - Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
  - Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
  - Gain: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.

Clustering Methods

8

- Partitioning methods
  - K-Means
- Hierarchical methods
  - Agglomerative Hierarchical Clustering
  - Divisive hierarchical clustering
- Density-based methods
  - DBSCAN: a Density-Based Spatial Clustering of Applications with Noise
- Grid-based methods
  - STING: A Statistical Information Grid Approach to Spatial Data Mining
- High Dimensional Data Clustering
  - CLIQUE: A Dimension-Growth Subspace Clustering Method

9

Partitioning Methods

K-Means Clustering

K-Means Clustering

10

- *k-means clustering is an* exclusive *clustering algorithm. Each object is assigned* to precisely one of a set of clusters. (There are other methods that allow objects to be in more than one cluster.)
- For this method of clustering we start by deciding how many clusters k we would like to form from our data.
- *The value of k is generally* a small integer, such as 2, 3, 4 or 5, but may be larger.

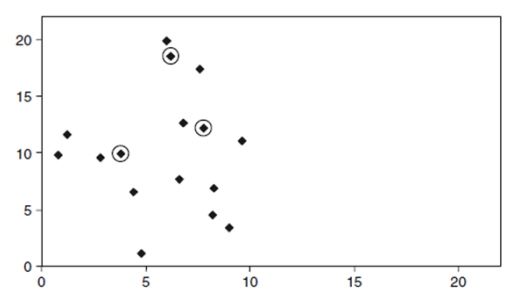
The *k*-Means Clustering Algorithm

11

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.

Example (k=3)

12

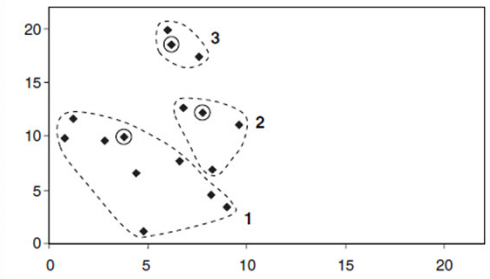


	Initial	
	x	y
Centroid 1	3.8	9.9
Centroid 2	7.8	12.2
Centroid 3	6.2	18.5

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

Initial Clusters

13

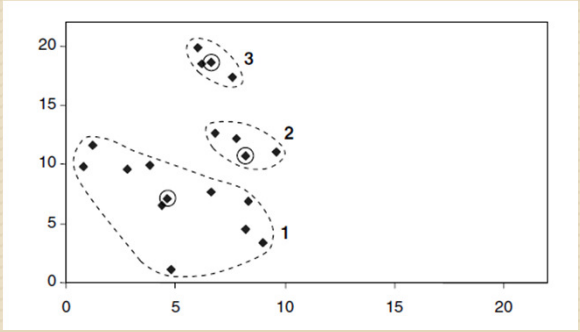


	Initial		After first iteration	
	x	y	x	y
Centroid 1	3.8	9.9	4.6	7.1
Centroid 2	7.8	12.2	8.2	10.7
Centroid 3	6.2	18.5	6.6	18.6

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

Revised Cluster

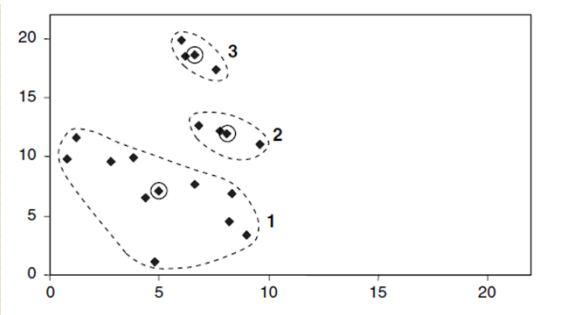
14



	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

Third Set of Clusters

15



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 has* been met and these are the final clusters produced by the algorithm for the initial choice of centroids made.

Finding the Best Set of Clusters

16

- It can be proved that the *k-means algorithm will always terminate, but it does not necessarily find the best set of clusters,* corresponding to minimising the value of the objective function.
- The initial selection of centroids can significantly affect the result.
  - Solution: Try different initial selection and take the best
  - But what should be k
    - Try different k
      - But which one to choose

Selecting K

17

- The table shown value of Objective Function For Different Values of *k*
- These results suggest that the best value of *k is probably* 3.
- *The value of* the function for *k = 3 is much less than for k = 2, but only a little better than* for *k = 4.*

Value of k	Value of objective function
1	62.8
2	12.3
3	9.4
4	9.3
5	9.2
6	9.1
7	9.05

We normally prefer to find a fairly small number of clusters as far as possible.

K-Selection Strategy

18

- We are *not trying to find the value of k with the smallest value of* the objective function.
- That will occur when the value of *k is the same as the* number of objects, i.e. each object forms its own cluster of one. The objective function will then be zero, but the clusters will be worthless.
- We usually want a fairly small number of clusters and accept that the objects in a cluster will be spread around the centroid (but ideally not too far away).

Summary

19

- Introduction
- Clustering Methods
- K-Means Clustering
- Selecting K