

COMP37332

Introduction to clustering techniques

Goran Nenadic
School of Computer Science

1

Aims

- Understand basic principles of clustering
- Understand techniques/approaches to clustering
- Learn possible application areas

2

Plan

- Introduction to clustering
- Clustering techniques
 - k-means
 - hierarchical clustering
- Applications and problems
- Outlier discovery

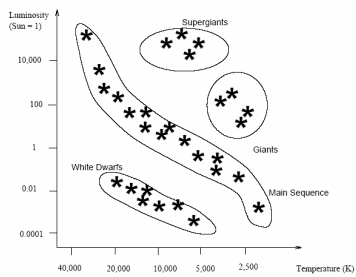
3

Clustering

- “The art of finding groups in data”
- Based on some measure of (dis)similarity among data, like **distance** or **correlation**
 - needs to be defined, depending on the problem
- Unsupervised learning
 - no predefined classes/groups

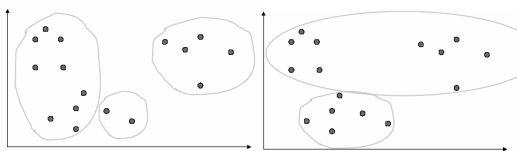
4

Example (1)



5

Example (2)



6

Clustering

- Cluster: a collection of **similar** data objects
 - minimise intra-cluster and maximise inter-cluster "distance"
 - objects similar to one another within the same cluster, and
 - dissimilar to the objects in other clusters
- Clustering approaches
 - hard clustering
 - each object in exactly one cluster
 - soft clustering
 - allow objects to be in more than one cluster

7

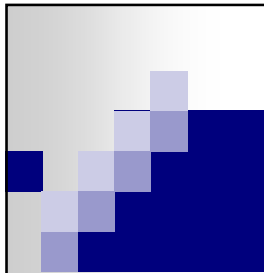
Clustering approaches

- Partitioning
 - based on elements moving between clusters
 - start with some random initial partition of data into clusters and move objects between clusters until clustering quality is optimal
- Hierarchical
 - based on merging/splitting (sub-)clusters
 - assign all objects to a different cluster and recursively merge clusters (**agglomerative** approach)
 - alternatively, start with one big cluster containing all objects and recursively divide into smaller clusters (**divisive** approach)

8



9



k-means clustering

10

k-means clustering

- Groups given data-set into **k clusters**
 - partitioning technique
- Each cluster is represented by its **centre**, and distances are calculated with regard to it
 - centre (cluster's mean element) is called **centroid**
 - calculated as the mean value (for each attribute)
$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$
 - centroid = $(\bar{x}_1, \dots, \bar{x}_n)$

11

Distance measures

- Data points represented as vectors
- Similarity/distance among objects/vectors
- Distance functions are usually different for interval-scaled, boolean, categorical, etc. data
- Various (standard) distance measures are used for numerical data
 - e.g. Euclidean, Manhattan, cosine, etc.

12

Distance measures (examples)

$$x = (x_1, \dots, x_p)$$

$$y = (y_1, \dots, y_p)$$

Euclidean distance $d(x, y) = \sqrt{\sum_{i=1}^p |x_i - y_i|^2}$

Manhattan distance $d(x, y) = \sum_{i=1}^p |x_i - y_i|$

"max" distance $d(x, y) = \max_{1 \leq i \leq p} |x_i - y_i|$

13

Distance measures (examples)

$$x = (0, 2, 4)$$

$$y = (1, 5, 4)$$

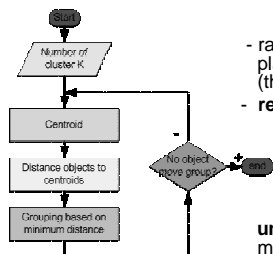
$$d_E(x, y) = \sqrt{|0-1|^2 + |2-5|^2 + |4-4|^2} = \sqrt{1+9+0} = \sqrt{10} = 3.16$$

$$d_M(x, y) = |0-1| + |2-5| + |4-4| = 1+3+0 = 4$$

$$d_{\max}(x, y) = \max\{|0-1|, |2-5|, |4-4|\} = 3$$

14

k-means algorithm



- randomly choose k elements and place them in k clusters (these are *initial* centroids)

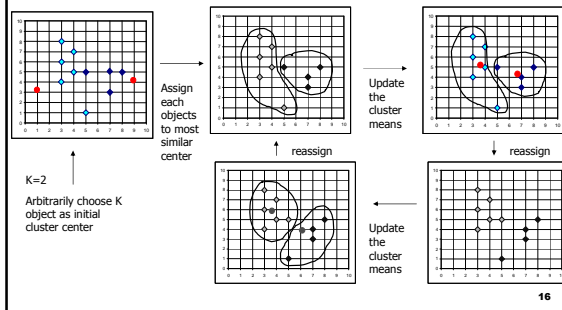
- **repeat**

- assign each element into a cluster so that its distance from cluster's centroid is minimal
- recalculate the centroid for each cluster

until no change in cluster memberships

15

k-means algorithm – example



Clustering example: two clusters

k = 2

Object	Attribute_1	Attribute_2
A	1.0	1.0
B	1.5	2.0
C	3.0	4.0
D	5.0	7.0
E	3.5	5.0
F	4.5	5.0
G	3.5	4.5

Distance between objects: Euclidean distance $d(x, y) = \sqrt{\sum_{i=1}^p |x_i - y_i|^2}$

17

continued

Clustering example: two clusters

Object	Attribute_1	Attribute_2
A	1.0	1.0
B	1.5	2.0
C	3.0	4.0
D	5.0	7.0
E	3.5	5.0
F	4.5	5.0
G	3.5	4.5

Initialisation: randomly choose two (k=2) centroids for two clusters

	Object	Centroid
Cluster 1	A	(1.0, 1.0) = m_1
Cluster 2	D	(5.0, 7.0) = m_2

18

continued

Clustering example – distances (1)

- Distances to centroids

centroid m_1 (1.0, 1.0)

centroid m_2 (5.0, 7.0)

Object	distances to	
	Centroid m_1	Centroid m_2
A (1.0, 1.0)	0	7.21
B (1.5, 2.0)	1.12	6.10
C (3.0, 4.0)	3.61	3.61
D (5.0, 7.0)	7.21	0
E (3.5, 5.5)	4.72	2.5
F (4.5, 5.0)	5.31	2.06
G (3.5, 4.5)	4.30	2.92

$$d(m_1, B) = \sqrt{|1.0 - 1.5|^2 + |1.0 - 2.0|^2} = 1.12$$

$$d(m_2, B) = \sqrt{|5.0 - 1.5|^2 + |7.0 - 2.0|^2} = 6.10$$

19

continued

Example – recalculation (1)

- Two clusters: {A, B, C} and {D, E, F, G}
- Their new centroids:

$$m_1 = \left(\frac{1}{3}(1.0 + 1.5 + 3.0), \frac{1}{3}(1.0 + 2.0 + 4.0)\right) = (1.83, 2.33)$$

$$m_2 = \left(\frac{1}{4}(5.0 + 3.5 + 4.5 + 3.5), \frac{1}{4}(7.0 + 5.0 + 5.0 + 4.5)\right) = (4.12, 5.38)$$

20

continued

Example – distances (2)

- Distances to centroids:

m_1 (1.83, 2.33)

m_2 (4.12, 5.38)

Object	distances to	
	Centroid 1	Centroid 2
A	1.57	5.38
B	0.47	4.28
C	2.04	1.78
D	5.64	1.84
E	3.15	0.73
F	3.78	0.54
G	2.74	1.08

new clusters: {A, B}, {C, D, E, F, G}

21

Example – distances (3)

Distances to centroids:

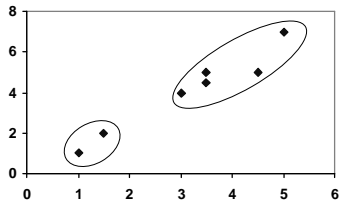
$$m_1 (1.25, 1.5)$$

$$m_2 (3.9, 5.1)$$

Object	distances to	
	Centroid 1	Centroid 2
A	0.56	5.02
B	0.56	3.92
C	3.05	1.42
D	6.66	2.20
E	4.16	0.41
F	4.78	0.61
G	3.75	0.72

no changes in cluster membership

Example – plot



Example – three clusters (1)

Object	distances to			cluster
	$m_1 = 1$	$m_2 = 2$	$m_3 = 3$	
A	0	1.11	3.61	1
B	1.12	0	2.5	2
C	3.61	2.5	0	3
D	7.21	6.10	3.61	
E	4.72	3.61	1.12	
F	5.31	4.24	1.80	
G	4.30	3.20	0.71	

clustering with the following initial centroids: A, B and C

Example – three clusters (1)

Object	distances to			cluster
	$m_1 = A$	$m_2 = B$	$m_3 = C$	
A	0	1.11	3.61	1
B	1.12	0	2.5	2
C	3.61	2.5	0	3
D	7.21	6.10	3.61	3
E	4.72	3.61	1.12	3
F	5.31	4.24	1.80	3
G	4.30	3.20	0.71	3

clustering with initial centroids (A, B, C)

25

continued

Example – three clusters (1)

Object	distances to			cluster
	m_1 (1.0, 1.0)	m_2 (1.5, 2.0)	m_3 (3.9, 5.1)	
A	0	1.11	5.02	
B	1.12	0	3.92	
C	3.61	2.5	1.42	
D	7.21	6.10	2.20	
E	4.72	3.61	0.41	
F	5.31	4.24	0.61	
G	4.30	3.20	0.72	

26

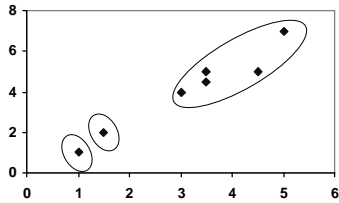
continued

Example – three clusters (1)

Object	distances to			cluster
	m_1 (1.0, 1.0)	m_2 (1.5, 2.0)	m_3 (3.9, 5.1)	
A	0	1.11	5.02	1
B	1.12	0	3.92	2
C	3.61	2.5	1.42	3
D	7.21	6.10	2.20	3
E	4.72	3.61	0.41	3
F	5.31	4.24	0.61	3
G	4.30	3.20	0.72	3

27

Example – plot (1)



28

Example – three clusters (2)

Object	distances to			cluster
	$m_1 = 1$	$m_2 = 4$	$m_3 = 7$	
A	0	7.21	4.30	
B	1.12	6.10	3.20	
C	3.61	3.61	0.71	
D	7.21	0	2.92	
E	4.72	2.50	0.5	
F	5.31	2.06	1.12	
G	4.30	2.92	0	

clustering with initial centroids (A, D, G)

29

Example – three clusters (2)

Object	distances to			cluster
	$m_1 = A$	$m_2 = D$	$m_3 = G$	
A	0	7.21	4.30	1
B	1.12	6.10	3.20	1
C	3.61	3.61	0.71	3
D	7.21	0	2.92	2
E	4.72	2.50	0.5	3
F	5.31	2.06	1.12	3
G	4.30	2.92	0	3

clustering with initial centroids (A, D, G)

30

continued

Example – three clusters (2)

Object	distances to			cluster
	m_1 (1.25, 1.5)	$m_2 = D$	m_3 (3.62, 4.62)	
A	0.56	7.21	4.47	
B	0.56	6.10	3.37	
C	3.05	3.61	0.88	
D	6.66	0	2.75	
E	4.16	2.50	0.40	
F	4.78	2.06	0.95	
G	3.75	2.92	0.17	

31

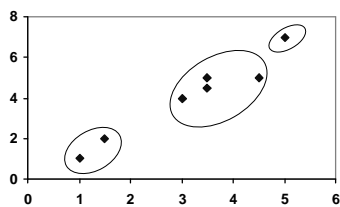
continued

Example – three clusters (2)

Object	distances to			cluster
	m_1 (1.25, 1.5)	$m_2 = D$	m_3 (3.62, 4.62)	
A	0.56	7.21	4.47	1
B	0.56	6.10	3.37	1
C	3.05	3.61	0.88	3
D	6.66	0	2.75	2
E	4.16	2.50	0.40	3
F	4.78	2.06	0.95	3
G	3.75	2.92	0.17	3

32

Example – plot (2)



33

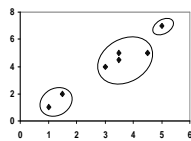
Quality of clusters

- In k-means, different selections of initial clusters may produce different clustering results
- Good clustering will produce clusters with
 - high intra-cluster similarity
 - low inter-cluster similarity
- An example metric for quality assessment
 - "**error estimation**": measure the sum of distances to the centroid: the smaller the sum of "square errors", the better the clusters

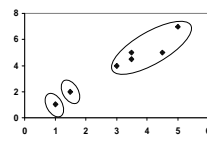
34

Quality of clusters: error estimation

$$Error = \sum_{i=1}^k \sum_{r \in C_i} d(r, m_i)^2$$



Error = 2.49
Error_{C3} = 1.87



Error = 7.92
Error_{C3} = 7.92

35

continued

Quality of clusters

- Other measures: *purity*
 - percentage of the frequent class members in the given cluster (needs to know classes!)
- The quality of a clustering method can be also measured by its ability to discover some of the hidden patterns
 - difficult to measure

36

Issues with k-means

- Problems
 - need to specify k (the number of clusters) in advance
 - unable to handle noisy data and outliers
- Choice of initial centroids may influence the final result
 - often terminates at a **local** optimum
 - we may not find the **global** optimum!!!
- Relatively efficient: $O(tkn)$, where n is # objects, k is # clusters, and t is # iterations ($k, t \ll n$)

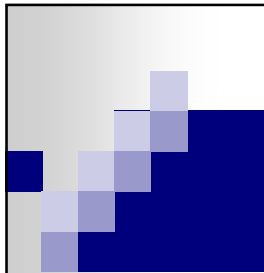
37

Centroids and medoids

- *k-means clustering*: each cluster is represented by its centroid
 - *centroid* is the "mean" vector of a cluster
 - may not be a real object
- *k-medoids clustering*: each cluster is represented by one of the objects in the cluster, the one that is **closest** to the centroid

38

The screenshot shows the Weka GUI. The main window is titled 'SimpleKMeans: N 4.5 4'. The 'Cluster mode' section has 'Use training set' checked. The 'Cluster output' section shows 'Model and evaluation on training set' with 'kMeans' selected. The 'Cluster centroid' section shows 'Number of iterations: 0' and 'Within cluster sum of squared errors: 13.95950917543607'. The 'Cluster 0' section shows 'Mean/Modes: 10 Instance_number (Num)' and 'Std Devs: Colour Cluster (Nom)'. The 'Cluster 1' section shows 'Mean/Modes: 10 Instance_number (Num)' and 'Std Devs: Colour Cluster (Nom)'. The 'Cluster 2' section shows 'Mean/Modes: 10 Instance_number (Num)' and 'Std Devs: Colour Cluster (Nom)'. The 'Cluster 3' section shows 'Mean/Modes: 10 Instance_number (Num)' and 'Std Devs: Colour Cluster (Nom)'. The 'Clustered Instance:' section shows a table with 4 columns: Instance, Cluster, Size, and Percentage. The 'Cluster colour' section shows a table with 4 columns: Cluster, Colour, and Size. The 'Visualize' window shows a scatter plot of data points colored by cluster.

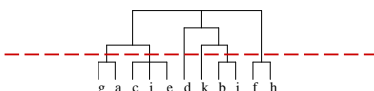


Hierarchical clustering

40

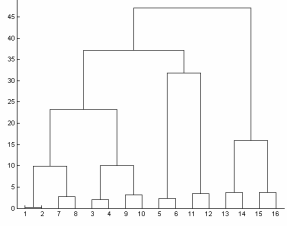
Hierarchical clustering

- Produces a hierarchy (dendrogram)
- Does not require the number of clusters in advance
- Clustering is obtained by cutting the dendrogram at the desired level
 - connected components form a cluster



41

Hierarchical clustering



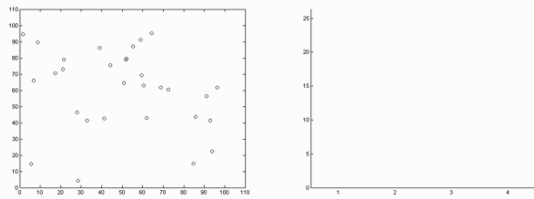
height of the bars may indicate how close the items are

42

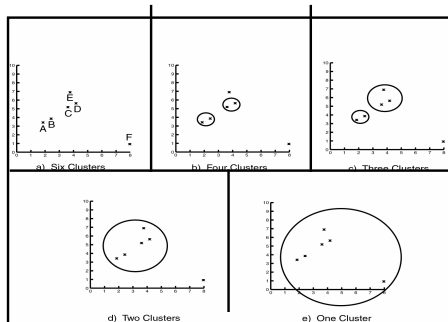
Hierarchical clustering

- Bottom up (**agglomerative**, merging)
 - start with single-object clusters
 - at each step, join the two **closest** clusters using the distance between them
- Top down (**divisive**, splitting)
 - start with one cluster containing all objects
 - split it into two clusters, and then proceed recursively on each subset

Hierarchical clustering - demo



Levels of clustering



Agglomerative clustering

- Sequentially merge pair of “closest” clusters
- Procedure
 - find two closest clusters and merge them
 - proceed until we have a single cluster
- Two prerequisites:
 - distance measure between data points (objects)
 - distance measure between clusters (“cluster linkage”)
- No notion of optimality, greedy algorithm

46

Distance matrix

- Clustering is based on distances between data objects
- Distance matrix
 - represents pre-calculated distances between objects
 - only need half the matrix, since it is symmetrical

$$\begin{bmatrix} 0 & & & & & \\ d(2,1) & 0 & & & & \\ d(3,1) & d(3,2) & 0 & & & \\ \vdots & \vdots & \vdots & & & \\ d(n,1) & d(n,2) & \dots & \dots & 0 & \end{bmatrix}$$

47

Procedure

Given a set of n items to be clustered, and an $n \times n$ distance matrix, the basic procedure is:

1. Assigning each item to its own cluster, each containing just one item
2. Find the closest pair of clusters and merge them into a single cluster
3. Compute distances between the new cluster and each of the old clusters.
4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size n .

48

Find the closest pair of clusters?

- How to calculate *inter-cluster distances*?
- **Single linkage** method (nearest neighbour)
 - distance between two clusters is the smallest distance between any pair of their elements:
 $distance(C_a, C_b) = \min_{x \in C_a, y \in C_b} distance(x, y)$
- **Complete linkage** method
 - use the maximal distance between two points:
 $distance(C_a, C_b) = \max_{x \in C_a, y \in C_b} distance(x, y)$

49

continued

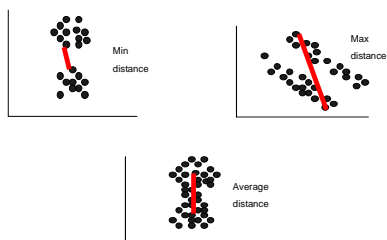
Find the closest pair of clusters?

- **Average linkage** method
 - distance between two clusters is the average distance between all pairs of their elements
 $distance(C_a, C_b) = \text{avg}_{x \in C_a, y \in C_b} distance(x, y)$
- Distance between **cluster centroids**
 - distance between two clusters is the distance between their centroids

50

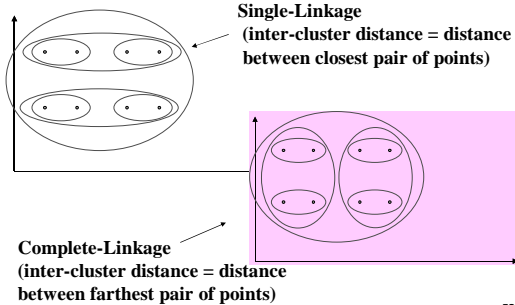
continued

Find the closest pair of clusters?



51

Inter-cluster distances



52

An example: distance matrix

e.g. pre-calculated (e.g. using the Euclidean distance)

	BA	FI	MI	NA	RM	TO
BA	0	662	877	255	412	996
FI		0	295	468	268	400
MI			0	754	564	138
NA				0	219	869
RM					0	669
TO						0

53

example

Steps 1 and 2

- Step 1: initial clusters
 $\{BA\}, \{FI\}, \{MI\}, \{NA\}, \{RM\}, \{TO\}$
- Let's use single-linkage

$$d(A, B) = \min_{a \in A, b \in B} d(a, b)$$

- Step 2: Currently, the minimal distance is 138 (between $\{MI\}$ and $\{TO\}$): this is a new cluster
 $\{BA\}, \{FI\}, \{MI, TO\}, \{NA\}, \{RM\}$

54

continued

Step 3: re-compute distances

	BA	FI	MI, TO	NA	RM
BA	0	662	877	255	412
FI		0	295	468	268
MI, TO			0	754	564
NA				0	219
RM					0

$$d(\{BA\}, \{MI, TO\}) = \min\{d(BA, MI), d(BA, TO)\} = 877$$

$$d(\{FI\}, \{MI, TO\}) = \min\{d(FI, MI), d(FI, TO)\} = 295$$

55

continued

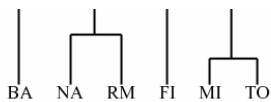
Step 2: find the closest two

	BA	FI	MI, TO	NA	RM
BA	0	662	877	255	412
FI		0	295	468	268
MI, TO			0	754	564
NA				0	219
RM					0

new clusters: {BA}, {FI}, {MI, TO}, {NA, RM}

56

Dendrogram so far



57

continued

Step 3: re-compute distances

	BA	FI	MI, TO	NA, RM
BA	0	662	877	255
FI		0	295	268
MI, TO			0	564
NA, RM				0

$d(\{BA\}, \{NA, RM\}) = \min\{d(BA, NA), d(BA, RM)\} = 255$
 $d(\{FI\}, \{NA, RM\}) = \min\{d(FI, NA), d(FI, RM)\} = 268$
 $d(\{MI, TO\}, \{NA, RM\}) = \min\{d(MI, NA), d(MI, RM), d(TO, NA), d(TO, RM)\} = 564$

58

continued

Step 2: find the closest

	BA	FI	MI, TO	NA, RM
BA	0	662	877	255
FI		0	295	268
MI, TO			0	564
NA, RM				0

59

continued

Step 3: re-compute distances

	FI	MI, TO	BA, NA, RM
FI	0	295	268
MI, TO		0	564
BA, NA, RM			0

$d(\{FI\}, \{BA, NA, RM\}) = 268$
 $d(\{MI, TO\}, \{BA, NA, RM\}) = 564$

60

continued

Step 2: find the closest

	FI	MI, TO	BA, NA, RM
FI	0	295	268
MI, TO		0	564
BA, NA, RM			0

61

continued

Step 3: re-compute distances

	MI, TO	FI, BA, NA, RM
MI, TO	0	295
FI, BA, NA, RM		0

$$d(\{MI, TO\}, \{FI, BA, NA, RM\})$$

$$= \min \{d(MI, \{FI, BA, NA, RM\}),$$

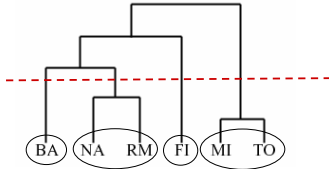
$$d(TO, \{FI, BA, NA, RM\}) = 295$$

62

Final dendrogram

63

Example clusters



64

Example 2

- The same example with average linkage

$$d(A, B) = \text{avg}_{a \in A, b \in B} d(a, b)$$

- Step 1: initial clusters (the same)
 $\{BA\}, \{FI\}, \{MI\}, \{NA\}, \{RM\}, \{TO\}$
- Step 2: the minimal avg. distance is 138 (again between $\{MI\}$ and $\{TO\}$):
 $\{BA\}, \{FI\}, \{MI, TO\}, \{NA\}, \{RM\}$

65

continued

Step 3: re-compute distances

	BA	FI	MI, TO	NA	RM
BA	0	662	936.5	255	412
FI		0	347.5	468	268
MI, TO			0	811.5	616.5
NA				0	219
RM					0

$$d(\{BA\}, \{MI, TO\}) = \text{avg} \{d(BA, MI), d(BA, TO)\} = \frac{1}{2} (877+996) = 936.5$$

$$d(\{FI\}, \{MI, TO\}) = \text{avg} \{d(FI, MI), d(FI, TO)\} = \frac{1}{2} (295+400) = 347.5$$

$$d(\{NA\}, \{MI, TO\}) = \text{avg} \{d(NA, MI), d(NA, TO)\} = \frac{1}{2} (754+869) = 811.5$$

$$d(\{RM\}, \{MI, TO\}) = \text{avg} \{d(RM, MI), d(RM, TO)\} = \frac{1}{2} (564+669) = 616.5$$

66

continued

Step 3: re-compute distances

	FI	MI, TO	BA, NA, RM
FI	0	347.5	466
MI, TO		0	788.17
BA, NA, RM			0

$d(\{FI\}, \{BA, NA, RM\}) = \text{avg}(d(FI, BA), d(FI, NA), d(FI, RM)) = 1/3 (662+468+268) = 466$
 $d(\{MI, TO\}, \{BA, NA, RM\}) = \text{avg}(d(MI, BA), d(MI, NA), d(MI, RM), d(TO, BA), d(TO, NA), d(TO, RM)) = 788.17$

70

continued

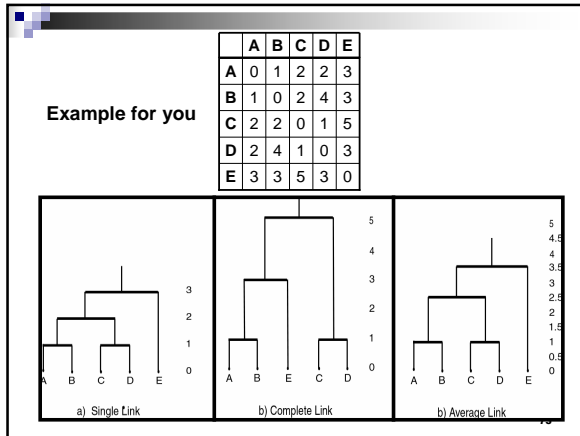
Step 2: find the closest

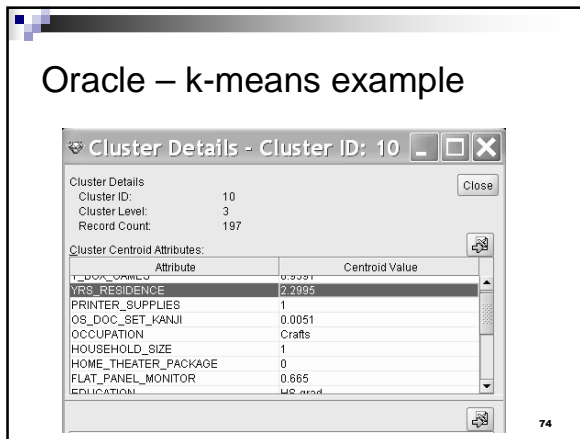
	FI	MI, TO	BA, NA, RM
FI	0	347.5	466
MI, TO		0	788.17
BA, NA, RM			0

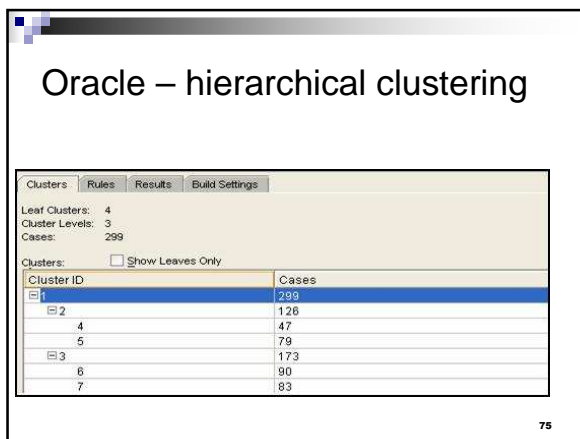
71

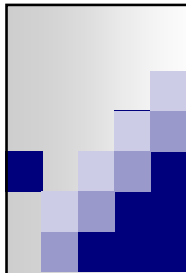
Final dendrogram (2)

72









Wrapping-up

76

Clustering applications

- Applications
 - fraud detection
 - business profiles
 - medical data
 - image processing
 - document clustering
 - web log clustering
 - ...

77

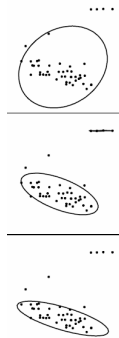
Some issues with clustering

- Selection of features
 - find representative attributes
- Number of clusters
 - selection of initial clusters (k-means)
 - depth of dendrogram cut-off (hierarchical)
- Training data may be noisy
- Efficiency
 - speed and scalability issues ($\sim O(n^2)$)
- Interpretability
 - understanding and insight provided by the model

78

Outlier discovery

- What are outliers?
 - objects that are considerably dissimilar from the remainder of the data
- Typical problem
 - find top n outlier points
- Applications:
 - credit card fraud detection
 - telecom fraud detection
 - customer segmentation
 - medical analysis



79

Oracle example

Outlier discovery

DMRCASE_ID	PREDICTION	PROBABILITY
101,501	1	0.9621
101,502	0	0.5481
101,503	1	0.51
101,504	1	0.36
101,505	0	0.732
101,506	1	0.5748
101,507	1	0.5415
101,508	1	0.5443
101,509	1	0.8447
101,510	1	0.829
101,511	1	0.5687
101,512	1	0.5508
101,513	1	0.5886

The customers with a prediction of 1 are considered typical. Customers with a prediction of 0 are considered outliers.

80

Summary

- Clustering:
 - arrange data into sets containing similar objects
- Approaches
 - partitioning (e.g. k-means)
 - hierarchical
- Agglomerative clustering
 - produces dendrogram
 - various inter-cluster distances

81

Reading for this lecture

- Chapter 27 (27.4) in [Elmasri & Navathe]

Also:

- Section 18.4.5 in A. Silberschatz, H. Korth, S. Sudarshan, Database System Concepts, 5th edition, McGraw-Hill, 2005 (ISBN 0-07-295886-3)
- Clustering lab tutorial (Weka)

<http://personalpages.manchester.ac.uk/staff/G.Nenadic/COMP37332/>

- Other on-line materials, tutorials and software

82
