Khoa Công Nghệ Thông Tin
Trường Đại Học Cần Thơ

Giải thuật gom cụm
Clustering algorithms

Đỗ Thanh Nghị
dtnghi@cit.ctu.edu.vn
Trần Nguyễn Minh Thư
tnmthu@cit.ctu.edu.vn

Cần Thơ
12/2008 – 08/2014

Nội dung

- Giới thiệu về clustering
- K-Means
- Kết luận và hướng phát triển

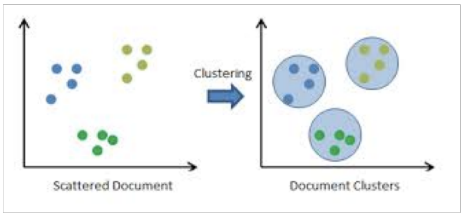
Nội dung

- **Giới thiệu về clustering**
- K-Means
- Kết luận và hướng phát triển

Clustering

- Giới thiệu về clustering
- Hierarchical clustering
- K-Means
- Kết luận và hướng phát triển

- Gom nhóm-cụm/clustering
  - Gom nhóm: mô hình gom cụm dữ liệu (không có nhãn) sao cho các dữ liệu cùng nhóm có các tính chất **tương tự nhau** và dữ liệu của 2 nhóm khác nhau sẽ có các tính chất khác nhau



  - Phương pháp học không giám sát
  - Dữ liệu thường không có nhiều thông tin sẵn có như lớp (nhãn)

Một số ứng dụng của phương pháp clustering

Phương pháp Clustering được sử dụng rộng rãi trong nhiều ứng dụng như nghiên cứu thị trường, tìm kiếm thông tin, phân tích dữ liệu, và xử lý hình ảnh

- Có thể giúp các nhà tiếp thị khám phá các nhóm khách hàng riêng biệt. Và họ có thể đặc trưng nhóm khách hàng của họ dựa trên các lịch sử mua hàng.
- Trong lĩnh vực sinh học, clustering được sử dụng để phân loại thực vật và động vật, phân loại gen có chức năng tương tự
- Clustering cũng giúp trong việc phân loại tài liệu trên web để phát hiện thông tin.

Một số ứng dụng của phương pháp clustering

- Clustering cũng được sử dụng trong các ứng dụng phát hiện outlier như phát hiện các gian lận thẻ tín dụng.
- Bảo hiểm: Xác định các nhóm chính sách bảo hiểm xe máy. Chủ sở hữu được chi phí bồi thường trung bình, cao, thấp khác nhau tuỳ đối tượng.
- Clustering cũng giúp trong việc xác định các khu vực sử dụng đất tương tự trong một cơ sở dữ liệu quan sát trái đất. Nó cũng giúp trong việc xác định các nhóm nhà ở một thành phố theo kiểu nhà, giá trị, và vị trí địa lý.

Clustering

- có nhiều nhóm giải thuật khác nhau
  - hierarchical clustering,
  - **K-Means (Partitional clustering),**
  - Dendrogram,
  - SOM, EM,…

Clustering

KDnuggets : Polls : Deployed data mi

Which data mining techniques you used i
[173 voters, 509 votes total]

- Decision Trees/Rules (88)
- Logistic regression (74)
- K-means clustering (54)
- Association rules (48)
- Neural Networks (45)
- SVM (33)
- Other methods (29)
- Self-organizing maps (24)
- Other clustering (24)
- Hybrid methods (20)
- Naive Bayes (19)
- Nearest Neighbor (17)
- Genetic algorithms (16)
- Fuzzy methods (11)
- Bayes nets (6)
- None (1)

Which methods/algorithms did you use for data analysis in 2011? [311 voters]	
Decision Trees/Rules (186)	59.8 %
Regression (180)	57.9 %
Clustering (163)	52.4 %
Statistics (descriptive) (149)	47.9 %
Visualization (119)	38.3 %
Time series/Sequence analysis (92)	29.6 %
Support Vector (SVM) (89)	28.6 %
Association rules (89)	28.6 %
Ensemble methods (88)	28.3 %
Text Mining (86)	27.7 %
Neural Nets (84)	27.0 %
Boosting (73)	23.5 %
Bayesian (68)	21.9 %
Bagging (63)	20.3 %
Factor Analysis (58)	18.7 %
Anomaly/Deviation detection (51)	16.4 %
Social Network Analysis (44)	14.2 %
Survival Analysis (29)	9.32 %
Genetic algorithms (29)	9.32 %
Uplift modeling (15)	4.82 %

Clustering

- gom nhóm
  - thường dựa trên cơ sở **khoảng cách**
  - nên chuẩn hóa dữ liệu
  - khoảng cách được tính theo từng kiểu của dữ liệu
    - Kiểu số,
    - Kiểu nhị phân
    - Kiểu rời rạc (nominal type),

Gom nhóm: mô hình gom cụm dữ liệu (không có nhãn) sao cho các dữ liệu cùng nhóm có các tính chất **tương tự nhau** và dữ liệu của 2 nhóm khác nhau sẽ có các tính chất khác nhau

9

Các độ đo khoảng cách - Kiểu số

- Khoảng cách *Minkowski*

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

$i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ và $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ là 2 phần tử dữ liệu trong p-dimensional, q là số nguyên dương

- nếu $q = 1$, d là khoảng cách Manhattan

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

- nếu $q = 2$, d là khoảng cách Euclid

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

10

Kiểu rời rạc (nominal type)

- VD: thuộc tính color có giá trị là red, green, blue, etc.
  - phương pháp matching đơn giản,
    - m là số lượng matches và
    - p là tổng số biến (thuộc tính),
    - khoảng cách được định nghĩa :

$$d(i,j)=\frac{p-m}{p}$$

Kiểu rời rạc (nominal type)

$$d(i,j)=\frac{p-m}{p}$$

- m là số lượng matches và
- p là tổng số biến (thuộc tính),

	Màu tóc	Màu mắt	Chiều cao	Cân nặng	Trình độ
Nam	Đen	Đen	Cao	Trung bình	Cao đẳng
Lan	Nâu	Đen	Thấp	Trung bình	Đại học

d(Nam, Lan) =?

Các độ đo khoảng cách - Kiểu nhị phân

		Object *j*		
		1	0	*sum*
Object *i*	1	*a*	*b*	*a+b*
	0	*c*	*d*	*c+d*
	sum	*a+c*	*b+d*	*p*

- khoảng cách đối xứng : $d(i,j) = \frac{b+c}{a+b+c+d}$
- khoảng cách bất đối xứng : $d(i,j) = \frac{b+c}{a+b+c}$
- hệ số Jaccard bất đối xứng : $sim_{Jaccard}(i,j) = \frac{a}{a+b+c}$

13

Kiểu nhị phân

- **Binary variables/attributes**
  - **Ví dụ**

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

  - **gender: symmetric**
  - **Binary attributes còn lại: asymmetric**
  - **Y, P → 1, N → 0**

$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(jack, jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(jim, mary) = \frac{1+2}{1+1+2} = 0.75$$

14

Nội dung

- Giới thiệu về clustering
- Hierarchical clustering
- **K-Means**
- Kết luận và hướng phát triển

15

Giải thuật K-Means

- Giới thiệu về clustering
- Hierarchical clustering
- K-Means
- Kết luận và hướng phát triển

- giải thuật
  1. khởi động ngẫu nhiên **K tâm** (center) của **K clusters**
  2. mỗi phần tử được gán cho tâm gần nhất với phần tử dựa vào khoảng cách (e.g. khoảng cách Euclid)
  3. **cập nhật lại các tâm của K clusters**, mỗi tâm là giá trị trung bình (mean) của các phần tử trong cluster của nó
  4. lặp lại bước 2,3 cho đến khi hội tụ

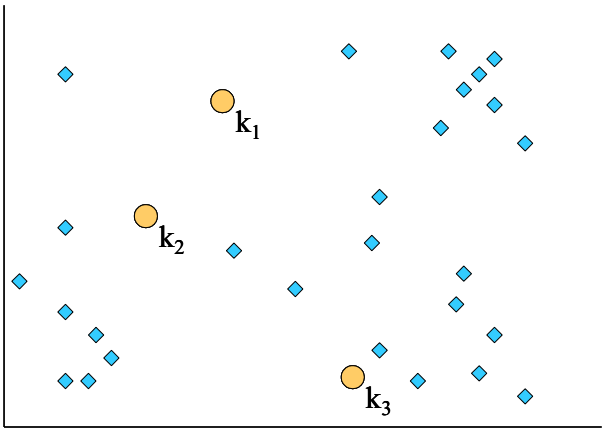
16

- Giới thiệu về clustering
- Hierarchical clustering
- **K-Means**
- Kết luận và hướng phát triển

Giải thuật K-Means

Y

khởi động ngẫu nhiên 3 tâm của 3 clusters



X

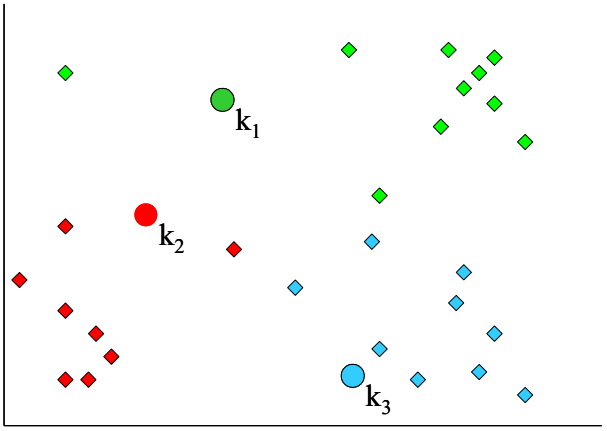
17

- Giới thiệu về clustering
- Hierarchical clustering
- **K-Means**
- Kết luận và hướng phát triển

Giải thuật K-Means

Y

mỗi phần tử được gán cho tâm cluster gần nhất của nó

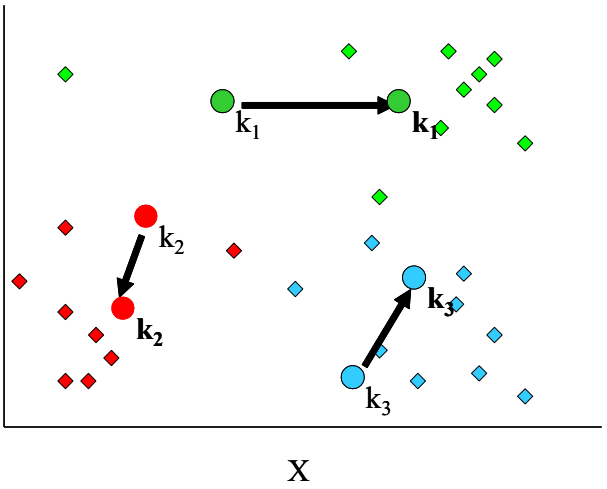


X

18

Giới thiệu về clustering
Hierarchical clustering
K-Means
Kết luận và hướng phát triển

Giải thuật K-Means

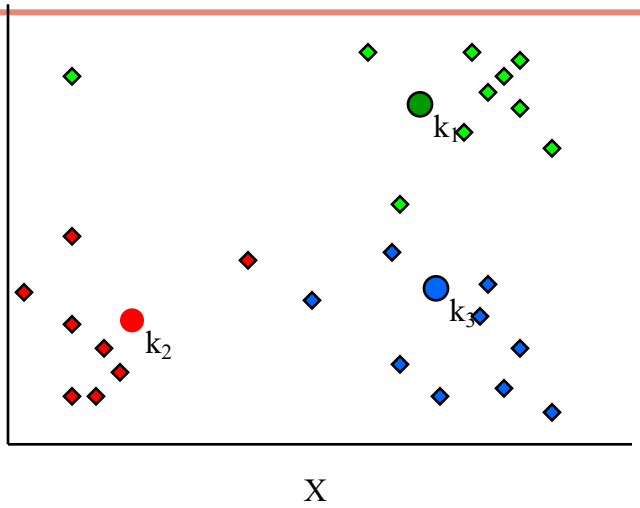


cập nhật lại tâm của các cluster (giá trị trung bình của các phần tử trong cluster)

19

Giới thiệu về clustering
Hierarchical clustering
K-Means
Kết luận và hướng phát triển

Giải thuật K-Means



cấu hình mới của lần lặp tiếp theo

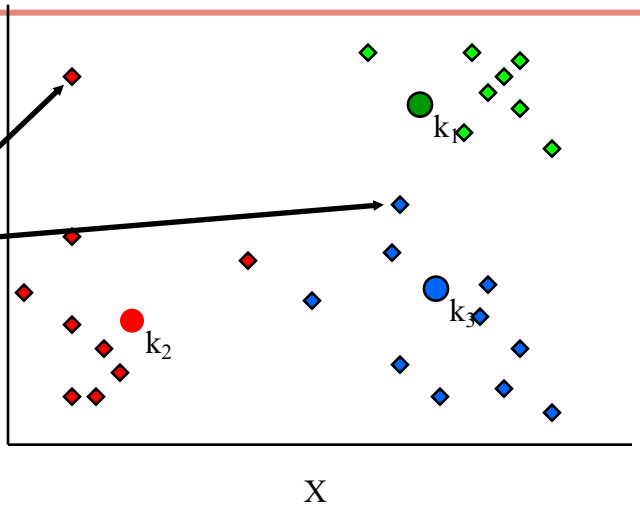
20

- Giới thiệu về clustering
- Hierarchical clustering
- **K-Means**
- Kết luận và hướng phát triển

Giải thuật K-Means

mỗi phần tử được gán lại cho tâm cluster gần nhất của nó

có 2 phần tử thay đổi nhóm

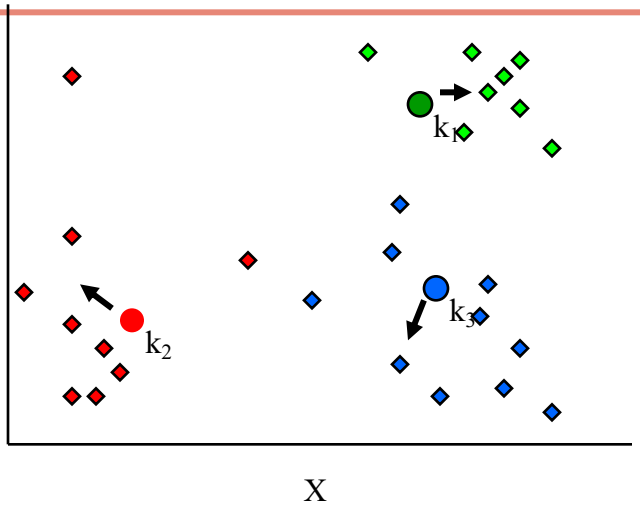


21

- Giới thiệu về clustering
- Hierarchical clustering
- **K-Means**
- Kết luận và hướng phát triển

Giải thuật K-Means

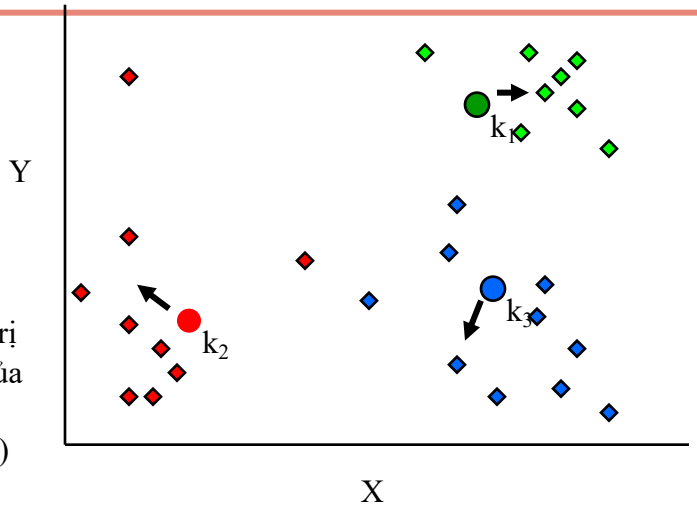
cập nhật lại tâm của các cluster (giá trị trung bình của các phần tử trong cluster)



22

- Giới thiệu về clustering
- Hierarchical clustering
- **K-Means**
- Kết luận và hướng phát triển

Giải thuật K-Means

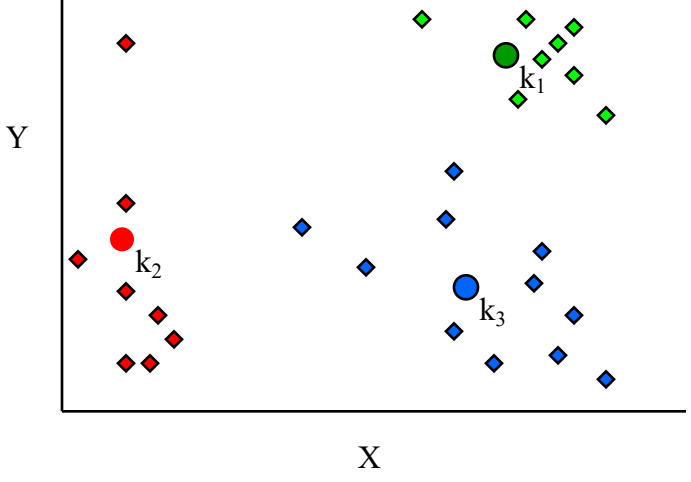


cập nhật lại tâm của các cluster (giá trị trung bình của các phần tử trong cluster)

23

- Giới thiệu về clustering
- Hierarchical clustering
- **K-Means**
- Kết luận và hướng phát triển

Giải thuật K-Means



24

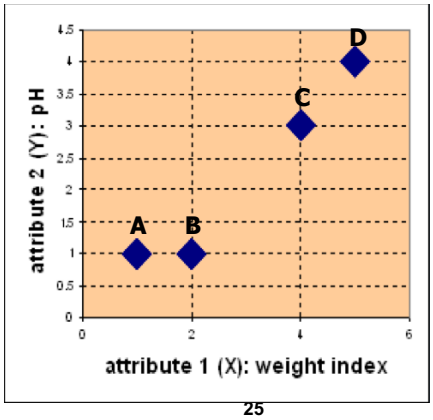
Bài tập

Bài tập 1:

Cho 4 loại thuốc mỗi loại có 2 thuộc tính pH và Weight

Yêu cầu nhóm những loại thuốc này thành 2 nhóm sử dụng khoảng cách Euclidea với 2 điểm khởi tạo là A và B

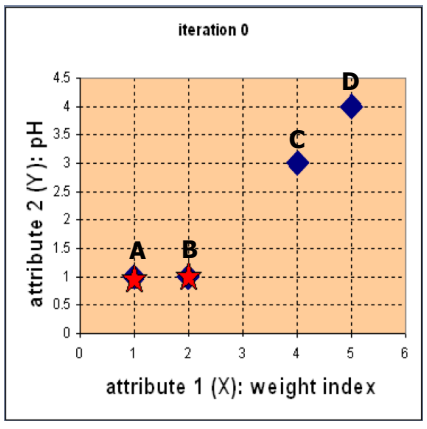
Medicine	Weight	pH-Index
A	1	1
B	2	1
C	4	3
D	5	4



25

Bài tập 1

Bước 1: khởi tạo các trọng tâm: A,B: trọng tâm nhóm 1/2 ; tính khoảng cách của các điểm còn lại đến 2 trọng tâm này



$c_1 = A, c_2 = B$

$$D^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} \quad \begin{matrix} c_1 = (1,1) & group-1 \\ c_2 = (2,1) & group-2 \end{matrix}$$

Euclidean distance

$$\begin{matrix} A & B & C & D \\ \begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} & \begin{matrix} X \\ Y \end{matrix} \end{matrix}$$

$$d(D,c_1) = \sqrt{(5-1)^2 + (4-1)^2} = 5$$

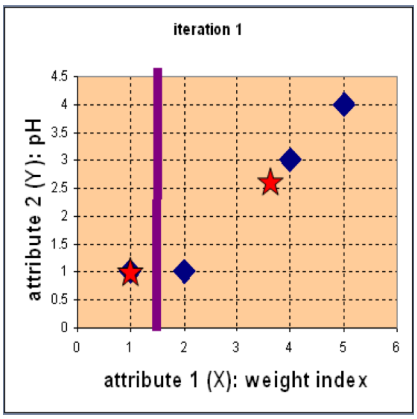
$$d(D,c_2) = \sqrt{(5-2)^2 + (4-1)^2} = 4.24$$

Assign each object to the cluste with the nearest seed point

26

Bài tập 1

Bước 2: tính lại 2 trọng tâm mới dựa vào các thành viên của nhóm vừa tạo ra ở bước 1



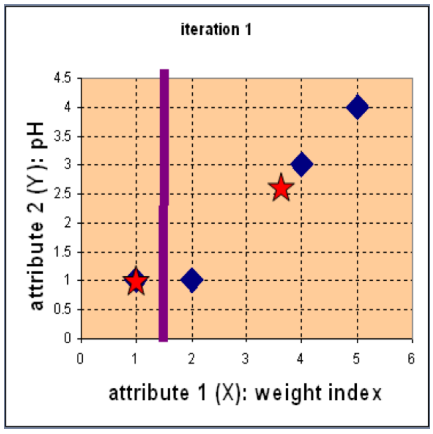
$$c_1 = (1,\ 1)$$

$$c_2 = \left(\frac{2+4+5}{3},\ \frac{1+3+4}{3}\right) = \left(\frac{11}{3},\ \frac{8}{3}\right)$$

27

Bài tập 1

- **Bước 2: Tính lại các thành viên theo 2 trọng tâm mới**



Compute the distance of all objects to the new centroids

$$D^1 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{bmatrix} \begin{matrix} c_1 = (1,1) & group-1 \\ c_2 = (\frac{11}{3}, \frac{8}{3}) & group-2 \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

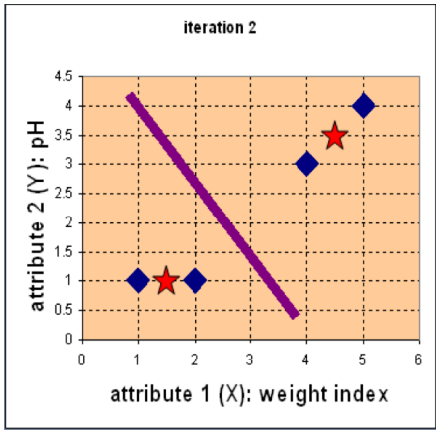
$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \begin{matrix} X \\ Y \end{matrix}$$

Assign the membership to objects

28

Bài tập 1

- **Bước 3: Lặp lai 2 bước đầu tiên cho đến khi hội tụ**



Knowing the members of each cluster, now we compute the new centroid of each group based on these new memberships.

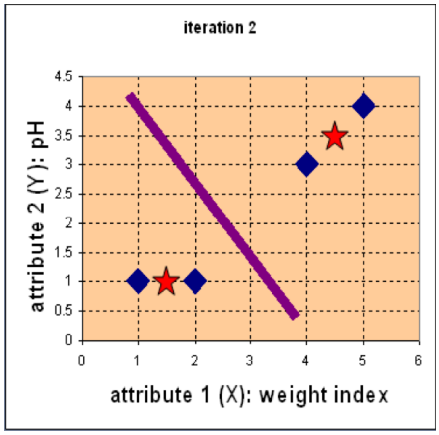
$$c_1 = \left(\frac{1+2}{2}, \frac{1+1}{2}\right) = (1\frac{1}{2}, 1)$$

$$c_2 = \left(\frac{4+5}{2}, \frac{3+4}{2}\right) = (4\frac{1}{2}, 3\frac{1}{2})$$

29

Bài tập 1

Bước 3: Lặp lai 2 bước đầu tiên cho đến khi hội tụ



Compute the distance of all objects to the new centroids

$$D^2 = \begin{bmatrix} 0.5 & 0.5 & 3.20 & 4.61 \\ 4.30 & 3.54 & 0.71 & 0.71 \end{bmatrix} \quad \begin{matrix} c_1 = (1\frac{1}{2}, 1) & group-1 \\ c_2 = (4\frac{1}{2}, 3\frac{1}{2}) & group-2 \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \begin{matrix} X \\ Y \end{matrix}$$

Stop due to no new assignment
Membership in each cluster no longer change

30

Bài tập 2: k=2

Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Bước 1:
Khởi tạo k=2 trọng tâm: m1=(1.0,1.0) và m2=(5.0,7.0).

Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

	Individual	Mean Vector
Group 1	1	(1.0, 1.0)
Group 2	4	(5.0, 7.0)

Bước 2:

- Sau bước 1 ta được 2 nhóm: {1,2,3} và {4,5,6,7}.
- Their new centroids are:

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

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

Individual	Centroid 1	Centroid 2
1	0	7.21
2 (1.5, 2.0)	1.12	6.10
3	3.61	3.61
4	7.21	0
5	4.72	2.5
6	5.31	2.06
7	4.30	2.92

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

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

Step 3:

Nhóm mới: {1,2} and {3,4,5,6,7}

- Trọng tâm mới: m1=(1.25,1.5) và m2 = (3.9,5.1)

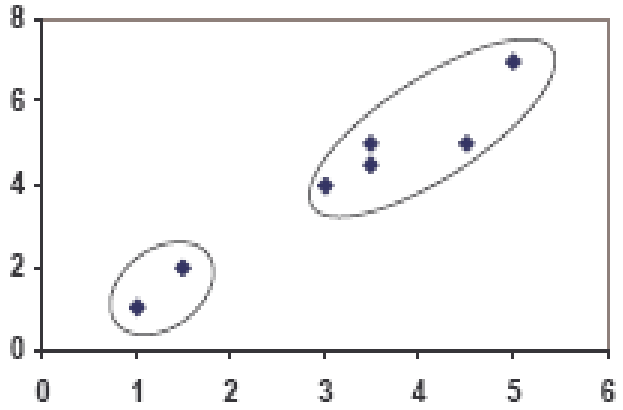
Individual	Centroid 1	Centroid 2
1	1.57	5.38
2	0.47	4.28
3	2.04	1.78
4	5.64	1.84
5	3.15	0.73
6	3.78	0.54
7	2.74	1.08

- **Bước 4:**
  Nhóm:
  {1,2} và {3,4,5,6,7}

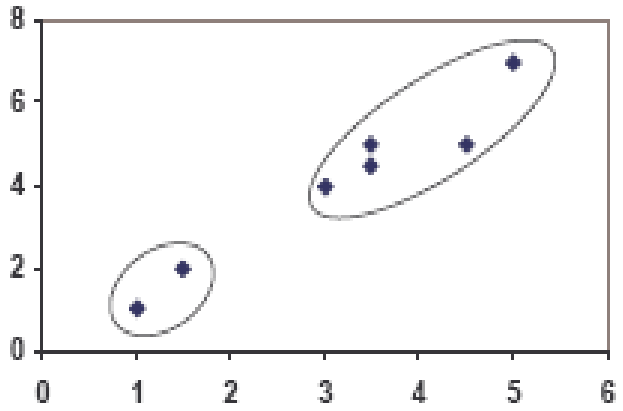
- **=> các thành viên trong nhóm không thay đổi => giải thuật dừng, ta có 2 nhóm {1,2} và {3,4,5,6,7}.**

Individual	Centroid 1	Centroid 2
1	0.56	5.02
2	0. 56	3.92
3	3.05	1.42
4	6.66	2.20
5	4.16	0.41
6	4.78	0.61
7	3.75	0.72

PLOT



PLOT



Giải thuật K-Means

- nhận xét
  1. giải thuật đơn giản
  2. cho kết quả dễ hiểu
  3. cần cho tham số K (số lượng clusters)
  4. kết quả phụ thuộc vào việc khởi động ngẫu nhiên K tâm (center) của K clusters : có thể khắc phục bằng cách khởi động lại nhiều lần.
  5. khả năng chịu đựng nhiễu không tốt (ảnh hưởng bởi các phần tử outliers) : có thể khắc phục bằng K-Medoids, không sử dụng giá trị trung bình, nhưng sử dụng phần tử ngay giữa

38

Nội dung

- Giới thiệu về clustering
- Hierarchical clustering
- K-Means
- **Kết luận và hướng phát triển**

39

Giải thuật clustering

- Giới thiệu về clustering
- Hierarchical clustering
- K-Means
- Kết luận và hướng phát triển

- còn nhiều phương pháp khác
  - density-based : DBSCAN (Ester et al., 1996), OPTICS (Ankerst et al., 1999), DENCLUE (Hinneburg & Keim, 1998)
  - model-based : EM (Expected maximization), SOM (Kohonen, 1995)

40

- Giới thiệu về clustering
- Hierarchical clustering
- K-Means
- **Kết luận và hướng phát triển**

Hướng phát triển

- các kiểu dữ liệu phức tạp
- tăng tốc độ xử lý
- các tham số đầu vào của giải thuật
- diễn dịch kết quả sinh ra
- phương pháp kiểm chứng chất lượng mô hình