

# KHAI THÁC DỮ LIỆU & ỨNG DỤNG (*DATA MINING*)

GV : NGUYỄN HOÀNG TÚ ANH



1

## BÀI 5 – Phần 1 GOM NHÓM DỮ LIỆU



2

# NỘI DUNG



## 1. Giới thiệu

2. Phương pháp phân hoạch
3. Phương pháp phân cấp

3

# GIỚI THIỆU



## 1. Gom nhóm là gì ? :

- Nhóm/cụm/lớp : tập các đối tượng DL
- *Gom nhóm là quá trình nhóm các đối tượng thành những nhóm/cụm/lớp có ý nghĩa. Các đối tượng trong cùng một nhóm có nhiều tính chất chung và có những tính chất khác với các đối tượng ở nhóm khác.*
- Cho CSDL  $D=\{t_1,t_2,\dots,t_n\}$  và số nguyên  $k$ , **gom nhóm** là bài toán xác định ánh xạ  $f : D \rightarrow \{1,\dots,k\}$  sao cho mỗi  $t_i$  được gán vào một nhóm (lớp)  $K_j$ ,  $1 \leq j \leq k$ .
- *Không giống bài toán phân lớp, các nhóm/cụm/lớp không được biết trước.*

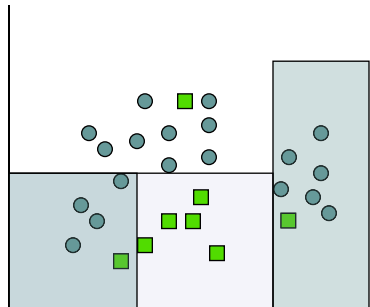
4

# PHÂN LỚP <> GOM NHÓM



**Phân lớp** : học có giám sát (Supervised learning)

*Tìm phương pháp để dự đoán lớp của mẫu mới từ các mẫu đã gán nhãn lớp (phân lớp) trước*



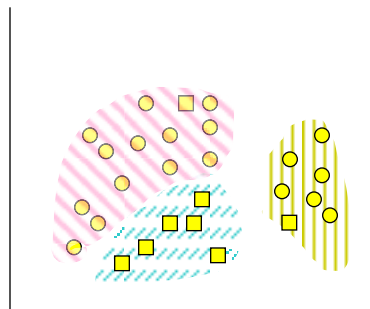
5

# PHÂN LỚP <> GOM NHÓM



**Gom nhóm** : học không giám sát (Unsupervised learning)

*Tìm các nhóm/cụm/lớp “tự nhiên” của các mẫu chưa được gán nhãn*

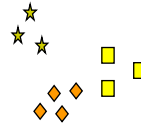
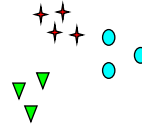
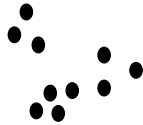
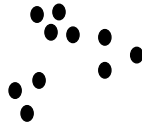


6

# GIỚI THIỆU

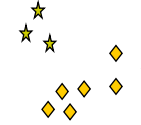
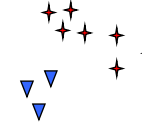
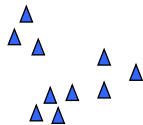
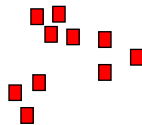


## Khái niệm nhóm/cụm – nhập nhằng



Có bao nhiêu nhóm /cụm?

6 nhóm/cụm



2 nhóm/cụm

4 nhóm/cụm

# GIỚI THIỆU



## ● Ứng dụng

- Nhận dạng
- Phân tích dữ liệu không gian
- Xử lý ảnh
- Khoa học kinh tế ( đặc biệt nghiên cứu tiếp thị)
- W W W
  - Gom nhóm tài liệu liên quan để dễ tìm kiếm
  - Gom dữ liệu Weblog thành nhóm để tìm các nhóm có cùng kiểu truy cập
- Giảm kích thước dữ liệu lớn

# GIỚI THIỆU



## ● Ví dụ

- ✚ Gom gen và protein có cùng chức năng
- ✚ Nhóm các cổ phiếu có xu hướng giá dao động giống nhau
- ✚ Nhóm các vùng theo lượng mưa ở Úc

	Discovered Clusters	Industry Group
1	Applied-Matl-DOWN, Bay-Network-DOWN, 3-COM-DOWN, Cabletron-Sys-DOWN, CISCO-DOWN, HP-DOWN, DSC-Comm-DOWN, INTEL-DOWN, LSI-Logic-DOWN, Micron-Tech-DOWN, Texas-Inst-DOWN, Tellabs-Inc-DOWN, Natl-Semiconduct-DOWN, Oracl-DOWN, SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN, Autodesk-DOWN, DEC-DOWN, ADV-Micro-Device-DOWN, Andren-Corp-DOWN, Computer-Asoc-DOWN, Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN, Microsoft-DOWN, Scientific-Ail-DOWN	Technology2-DOWN
3	Fannie-Mac-DOWN, Fed-Home-Loan-DOWN, MBNA-Corp-DOWN, Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP, Dresser-Inds-UP, Halliburton-HLD-UP, Louisiana-Land-UP, Phillips-Petro-UP, Unocal-UP, Schlumberger-UP	Oil-UP



9

# GIỚI THIỆU

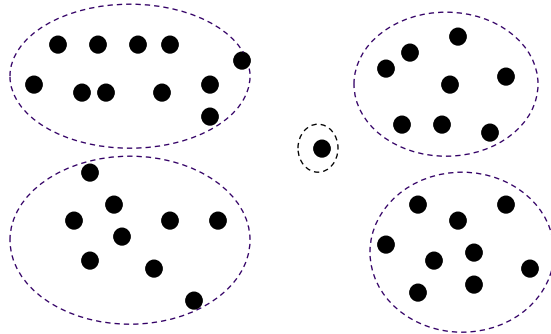


## ● Ví dụ :

- ✚ **Tiếp thị** : phát hiện các nhóm khách hàng trong CSDL khách hàng để xây dựng chương trình tiếp thị có mục tiêu
- ✚ **Đất đai** : xác định các vùng đất trồng trọt giống nhau trong CSDL quan sát trái đất
- ✚ **Bảo hiểm** : tìm nhóm khách hàng có khả năng hay gặp tai nạn
- ✚ **Nghiên cứu động đất** : gom nhóm các tâm chấn động đất quan sát được theo vết nứt lục địa

10

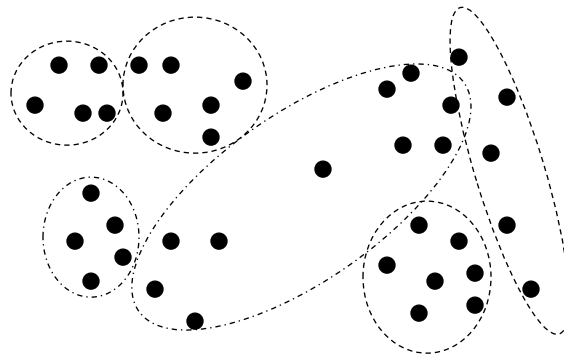
## VÍ DỤ : Gom nhóm các ngôi nhà



Dựa trên khoảng cách địa lý

11

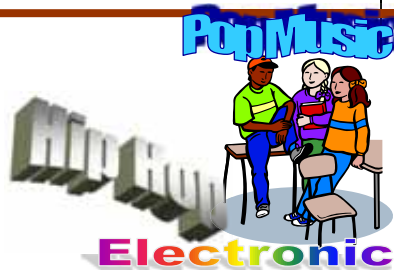
## VÍ DỤ : Gom nhóm các ngôi nhà



Dựa trên kích thước

12

# VÍ DỤ : Gom nhóm



Classical

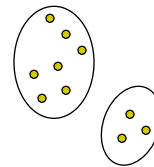
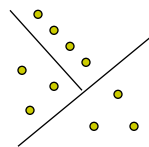


13

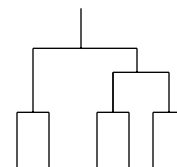
# GIỚI THIỆU

## Cách biểu diễn các nhóm/cụm

- + Phân chia bằng các đường ranh giới
- + Các khối cầu
- + Theo xác suất
- + Sơ đồ hình cây
- + ...



	1	2	3
I1	0.5	0.2	0.3
I2			
...			
In			



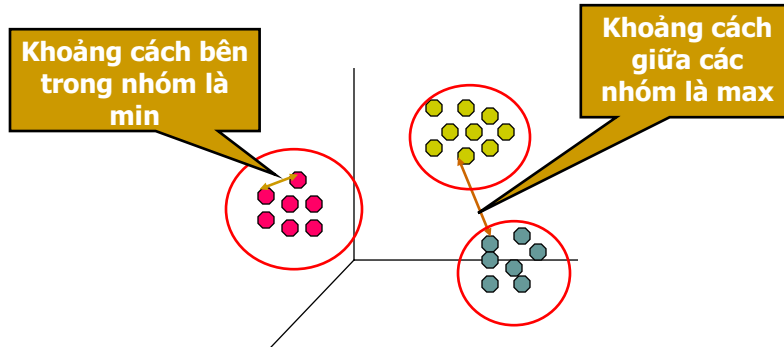
14

# GIỚI THIỆU



## 2. Tiêu chuẩn gom nhóm :

- Phương pháp gom nhóm tốt là phương pháp sẽ tạo các nhóm có chất lượng :
  - Sự giống nhau giữa đối tượng trong cùng một nhóm cao.*
  - Giữa các nhóm thì sự giống nhau thấp.*



15

# GIỚI THIỆU



## 2. Tiêu chuẩn gom nhóm (tt):

- Chất lượng của kết quả gom nhóm dựa trên 2 yếu tố :
  - Độ đo sự giống nhau dùng trong phương pháp gom nhóm và*
  - Sự thi hành nó*
- Một số độ đo chất lượng :
  - Bình phương sai (Sum of Squared Error - SSE)*
  - Entropy*

16



# GIỚI THIỆU



## 3. Độ đo khoảng cách :

- Độ đo khoảng cách thường dùng để xác định sự khác nhau hay giống nhau giữa hai đối tượng .
- Khoảng cách Minkowski :

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

với  $i = (x_{i1}, x_{i2}, \dots, x_{ip})$  và  $j = (x_{j1}, x_{j2}, \dots, x_{jp})$  : hai đối tượng  $p$ -chiều và  $q$  là số nguyên dương

- Nếu  $q=1$ ,  $d$  là khoảng cách Manhattan :

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

17

# GIỚI THIỆU



## 3. Độ đo khoảng cách (tt)

- Nếu  $q=2$ ,  $d$  là khoảng cách Euclide :

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

- Tính chất của độ đo khoảng cách

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

18

# GIỚI THIỆU



## 4. Các kiểu dữ liệu

- ✦ Các kiểu dữ liệu khác nhau yêu cầu độ đo sự khác nhau cũng khác nhau.
  - Các biến tỷ lệ theo khoảng : Khoảng cách Euclide
  - Các biến nhị phân : hệ số so khớp, hệ số Jaccard
  - Các biến tên, thứ tự, tỷ lệ : khoảng cách Minkowski
  - Các biến dạng hỗn hợp : công thức trọng lượng



19

# GIỚI THIỆU



## 5. Một số phương pháp gom nhóm :

- ✦ Phương pháp phân hoạch
- ✦ Phương pháp phân cấp
- ✦ Phương pháp dựa trên mật độ
- ✦ Phương pháp dựa trên lưới
- ✦ Phương pháp dựa trên mô hình



20

# NỘI DUNG



1. Giới thiệu
2. Phương pháp phân hoạch
3. Phương pháp phân cấp

21

# PHƯƠNG PHÁP PHÂN HOẠCH



## 1. Khái niệm cơ bản :

- Phương pháp phân hoạch : xây dựng  $k$  ( $k < n$ ) phân hoạch của CSDL  $D$  gồm  $n$  đối tượng. Mỗi phân hoạch – 1 nhóm/cụm
- Cho số  $k$ , cần tìm  $k$  nhóm thỏa mãn tiêu chuẩn phân hoạch đã chọn ( ví dụ độ đo bình phương sai - SSE nhỏ nhất).
- Biểu diễn mỗi nhóm bằng giá trị trung bình của dữ liệu trong nhóm đó : *thuật toán K-means (1967)*
- Biểu diễn nhóm bằng một đối tượng nằm gần trung tâm của nhóm : *thuật toán k-medoids, PAM (1987)*

22

# PHƯƠNG PHÁP PHÂN HOẠCH



## 1. Khái niệm cơ bản (tt):

- Công thức tính **Bình phương sai** ( **Sum of Squared Error - SSE**)

$$SSE = \sum_{i=1}^K \sum_{x \in C_i} dist^2(m_i, x)$$

Với  $x$  là một điểm DL trong nhóm  $C_i$  và  $m_i$  là điểm đại diện cho nhóm (điểm TB nhóm hoặc điểm trung tâm nhóm),  $K$ -số nhóm.  $dist()$ : khoảng cách Euclide

- Ví dụ : ta có 2 nhóm/cụm với các trung tâm tương ứng  $m_1=3, m_2=4$ 
  - $K_1=\{2,3\}, K_2=\{4,10,12,20,30,11,25\}$
  - **SSE =  $1^2+0+0+6^2+8^2+16^2+26^2+7^2+21^2=1523$**  23

# PHƯƠNG PHÁP PHÂN HOẠCH



## 2. Thuật toán k-means :

Cho số  $k$ , mỗi nhóm được biểu diễn bằng giá trị TB của DL trong nhóm

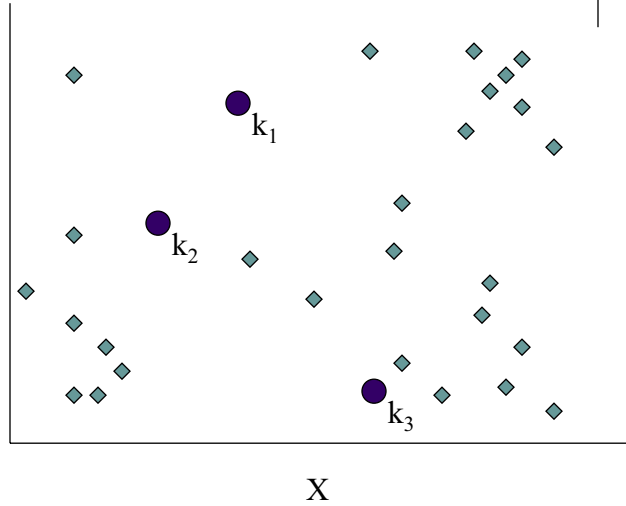
- **B1:** Chọn ngẫu nhiên  $k$  đối tượng như là những trung tâm của các nhóm .
- **B2 :** Gán từng đối tượng còn lại vào nhóm có trung tâm nhóm gần nó nhất (dựa trên độ đo khoảng cách Euclide)
- **B3 :** Tính lại giá trị trung tâm của từng nhóm
  - Di chuyển trung tâm nhóm về = giá trị TB mới của nhóm
  - Cho nhóm  $K_i=\{t_{i1}, t_{i2}, \dots, t_{im}\}$ , giá trị trung bình của nhóm là  $m_i = (1/m)(t_{i1} + \dots + t_{im})$
- **B4 :** Nếu các trung tâm nhóm không có gì thay đổi thì dừng, ngược lại quay lại B2.

24

## Ví dụ : K-means, Bước 1



Chọn 3  
trung tâm  
nhóm bất  
kỳ :  $k_1, k_2,$   
 $k_3$

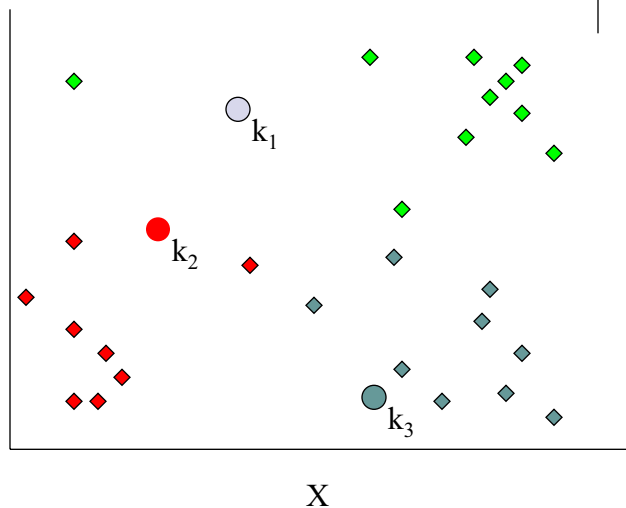


25

## Ví dụ : K-means, Bước 2



Gán từng  
điểm vào  
nhóm có  
trung tâm  
nhóm gần  
nhất

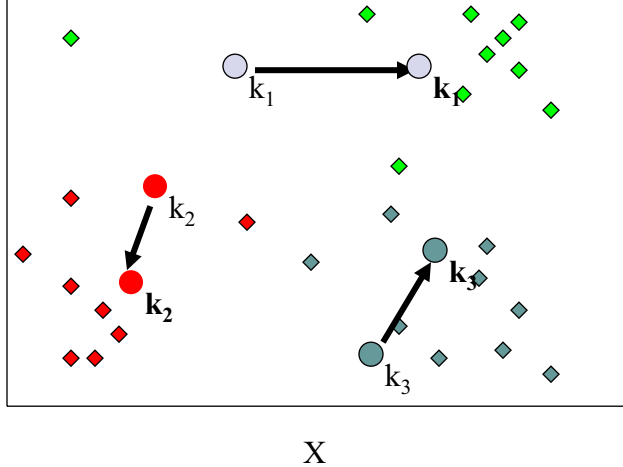


26

## Ví dụ : K-means, Bước 3



Di chuyển  
trung tâm  
từng nhóm  
về điểm  
trung bình  
mới của  
nhóm



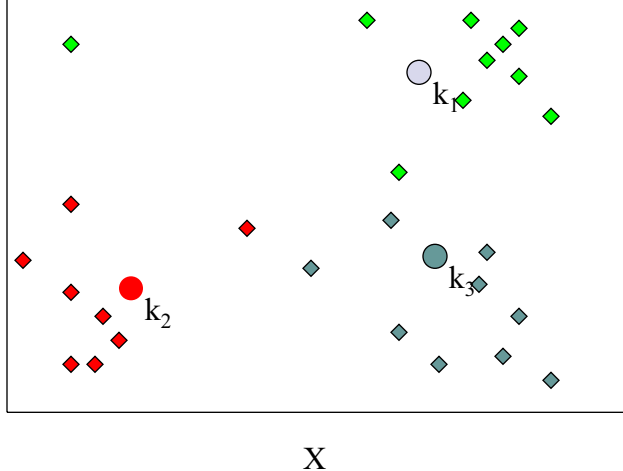
27

## Ví dụ : K-means, Bước 4



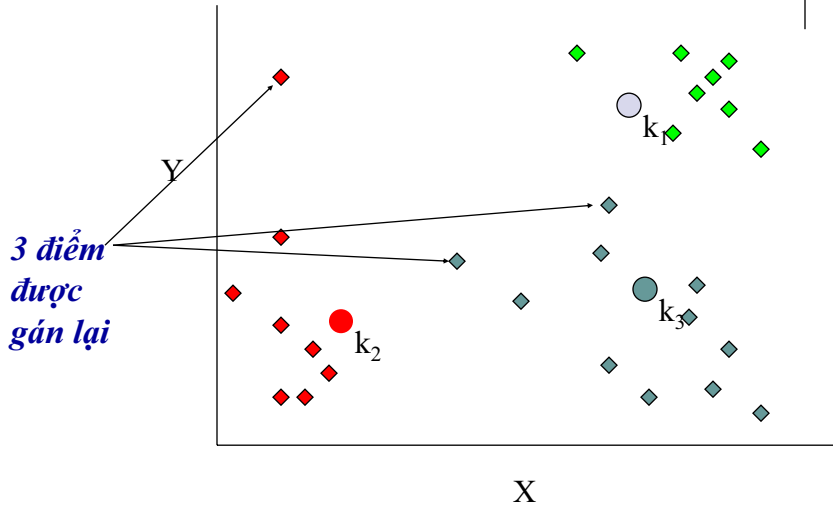
Gán lại các  
điểm cho gần  
với các trung  
tâm nhóm mới

*Q: Các điểm  
nào được gán  
lại ?*



28

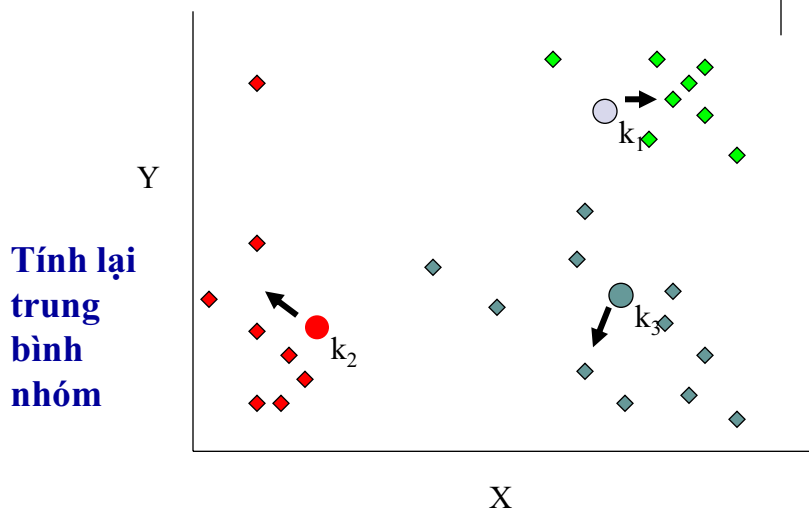
## Ví dụ : K-means, Bước 4 ...



3 điểm  
được  
gán lại

29

## Ví dụ : K-means, Bước 4b



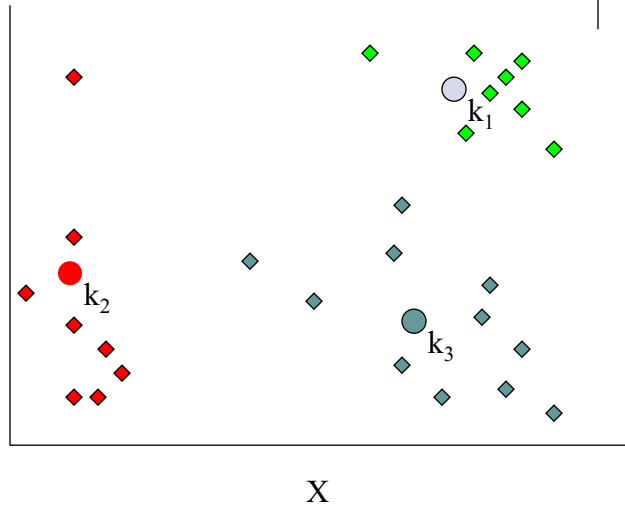
Tính lại  
trung  
bình  
nhóm

30

## Ví dụ : K-means, Bước 5









Di chuyển trung tâm nhóm về giá trị TB nhóm mới, ...

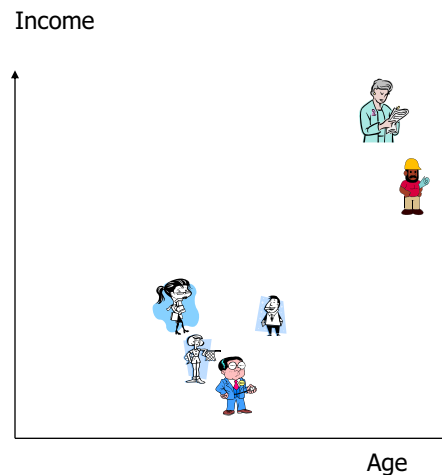


31

## Ví dụ : k-mean



Customer	Age	Income (K)
John 	0.55	0.175
Rachel 	0.34	0.25
Hannah 	1	1
Tom 	0.93	0.85
nellie 	0.39	0.2
David 	0.58	0.25



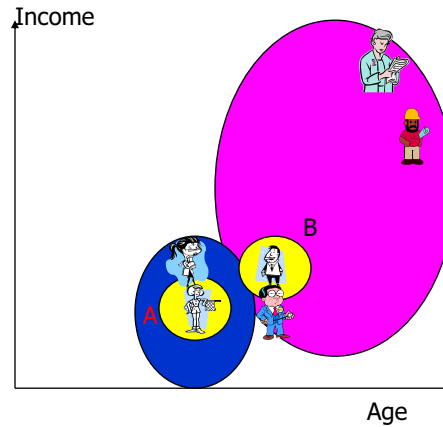


# Ví dụ : k-mean



**BƯỚC 1:** Chọn Nellie và David là trung tâm nhóm/cụm A và B

Customer	Distance from David	Distance from Nellie
John	0.08	0.161
Rachel	0.24	0.07
Hannah	0.859	1.006
Tom	0.694	0.845
Nellie		
David		



# Ví dụ: k-mean



**B2 :** Tính các trung tâm mới của nhóm/cụm A và B :

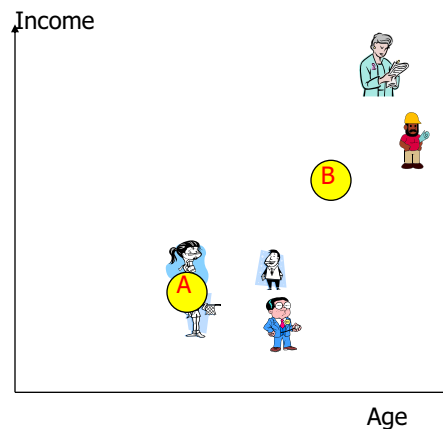
**Trung tâm của Cluster A :**

- Age 0.37, Income=0.23

**Trung tâm của Cluster B :**







- Age 0.77, Income=0.57

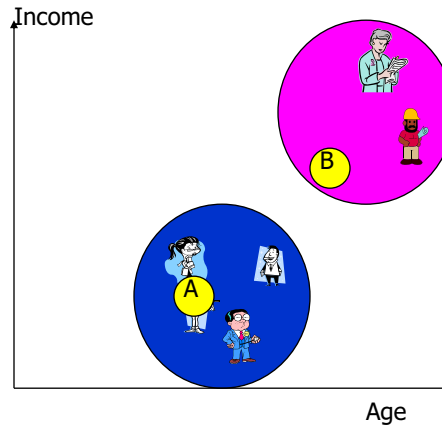
Dựa trên các trung tâm cụm mới, gán các khách hàng vào các nhóm/cụm.



# Ví dụ: k-mean



Customer	Distance A	Distance B
John 	0.19	0.45
Rachel 	0.04	0.53
Hannah 	1.00	0.49
Tom 	0.84	0.33
Nellie 	0.04	0.53
David 	0.22	0.37



# Ví dụ: k-mean



**B3: Tính các trung tâm mới của nhóm/cụm A và B :**

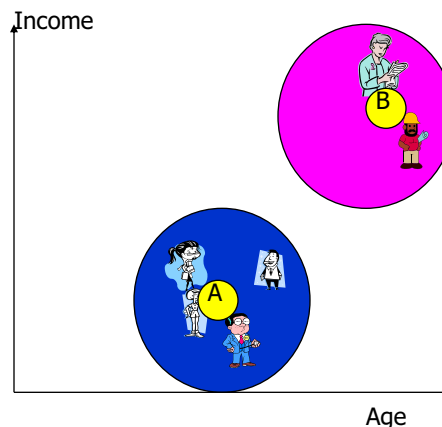
**Trung tâm của Cluster A :**

- Age 0.47, Income=0.22

**Trung tâm của Cluster B :**

- Age 0.97, Income= 0.93

- **Với các trung tâm nhóm mới này, thành phần của các nhóm không thay đổi.**
- **Thuật toán dừng.**



# Thuật toán K-means



## Ưu điểm :

- Đơn giản, dễ hiểu, tương đối hiệu quả.
- *Các đối tượng tự động gán vào các nhóm.*
- Thường đạt được tối ưu cục bộ.

37

# Thuật toán K-means



## Nhược điểm :

- Thuộc tính phi số ?
- *Cần xác định số nhóm (k) trước*
- Tất cả các đối tượng phải gán vào các nhóm
- Phụ thuộc vào việc chọn các nhóm đầu tiên
- Gặp vấn đề khi các nhóm có kích thước, mật độ khác nhau hoặc hình dáng không phải là hình cầu
- *Nhạy cảm với DL nhiễu, cá biệt*

38

# PHƯƠNG PHÁP PHÂN HOẠCH



## 3. Thuật toán k-medoids : PAM

Cho số  $k$ , mỗi nhóm được biểu diễn bằng một trong các đối tượng gần trung tâm nhóm nhất

- B1: Chọn ngẫu nhiên  $k$  đối tượng như là những trọng tâm của các nhóm .
- B2 : gán từng đối tượng còn lại vào nhóm có trọng tâm cụm gần nó nhất .
- B3 : Chọn một đối tượng bất kỳ. Hoán đổi nó với trọng tâm của nhóm. Nếu chất lượng của các nhóm tăng lên thì quay lại B2. Ngược lại tiếp tục thực hiện B3 cho đến khi không còn có thay đổi.

39

# PHƯƠNG PHÁP PHÂN HOẠCH



## 3. Thuật toán k-medoids (tt):

**Nhận xét :**

- Thuật toán PAM hiệu quả hơn so với  $k$ -means khi có mặt DL nhiễu, cá biệt.
- PAM hiệu quả với tập DL nhỏ nhưng không có dẫn tốt với tập DL lớn
- **Phát triển :**
  - CLARA (Clustering LARge Applications) : dựa trên phương pháp lấy mẫu (1990)
  - CLARANS(Clustering LARge Application based upon RANdomized Search) : lấy mẫu động (1994)

40

# NỘI DUNG



1. Giới thiệu
2. Phương pháp phân hoạch
3. Phương pháp phân cấp

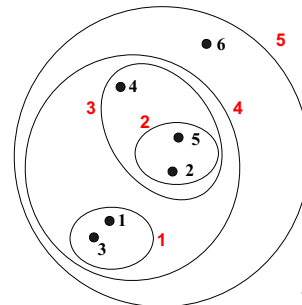
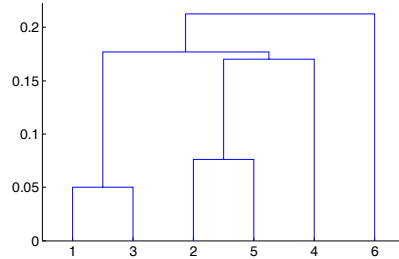
41

# PHƯƠNG PHÁP PHÂN CẤP



## 1. Giới thiệu :

- Phương pháp phân cấp : xây dựng các nhóm và tổ chức như cây phân cấp.
- **Biểu diễn dưới dạng sơ đồ hình cây (dendrogram):** lưu lại quá trình gom lại / phân chia nhóm



42

# PHƯƠNG PHÁP PHÂN CẤP



## 1. Giới thiệu (tt):

- Không cần xác định trước số nhóm  $k$ .
- Xác định số nhóm cần thiết bằng việc cắt ngang sơ đồ hình cây tại mức thích hợp.
- Hai loại phân cấp chính :
  - Tích tụ : từ dưới lên trên, mỗi đối tượng là một nhóm
  - Chia nhỏ : từ trên xuống, tất cả các đối tượng là 1 nhóm
- Thuật toán :
  - AGNES, DIANA
  - BIRCH (Balance Iterative Reducing & Clustering using Hierachies)
  - CURE (Clustering Using Representative)
  - ROCK (Robust Clustering using linKs)
  - CHAMELEON

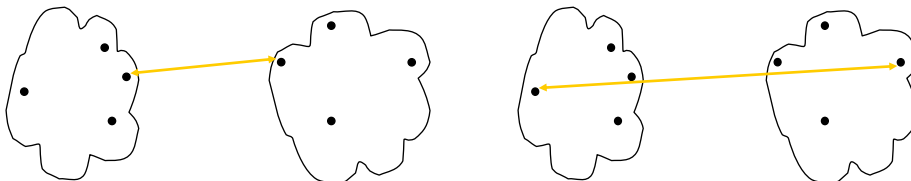
43

# PHƯƠNG PHÁP PHÂN CẤP



## 1. Giới thiệu (tt):

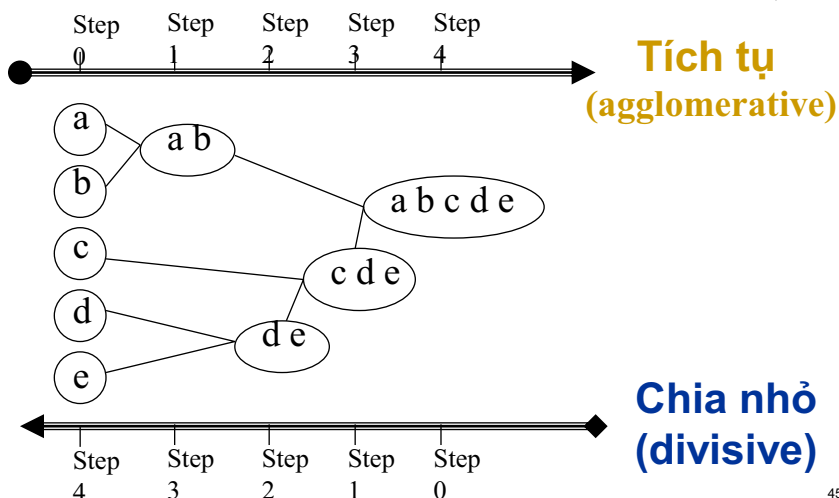
- Cách xác định khoảng cách giữa các nhóm :
  - Single Link : khoảng cách gần nhất giữa hai đối tượng thuộc hai nhóm



- Complete Link : khoảng cách xa nhất giữa hai đối tượng thuộc hai nhóm

44

# PHƯƠNG PHÁP PHÂN CẤP



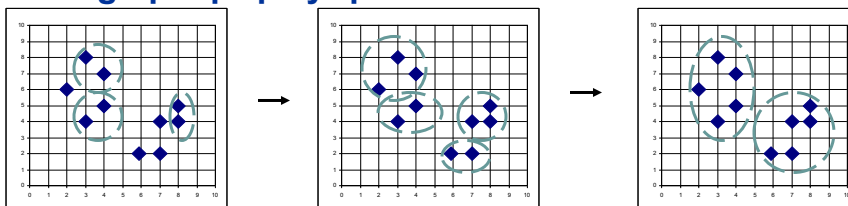
45

# PHƯƠNG PHÁP PHÂN CẤP



## 2. Thuật toán AGNES (Agglomerative Nesting):

- 🔥 **B1:** Mỗi đối tượng là một nhóm.
- 🔥 **B2:** Hợp nhất các nhóm có khoảng cách giữa các nhóm là nhỏ nhất (Single Link/ Complete Link)
- 🔥 **B3:** Nếu thu được nhóm “toàn bộ” thì dừng, ngược lại quay lại B2.

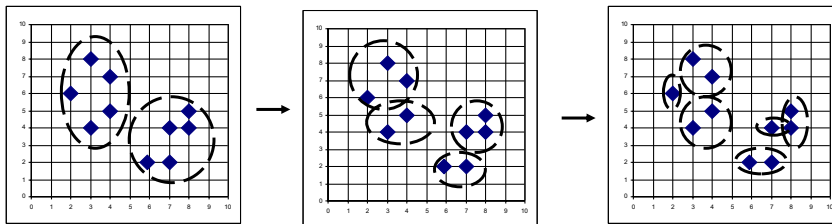


# PHƯƠNG PHÁP PHÂN CẤP



## 3. Thuật toán DIANA (Divisive Analysis):

- **B1:** Tất cả các đối tượng là một nhóm.
- **B2:** Chia nhỏ nhóm có khoảng cách giữa những đối tượng trong nhóm là lớn nhất.
- **B3:** Nếu mỗi nhóm chỉ chứa 1 đối tượng thì dừng, ngược lại quay lại B2.



## VÍ DỤ : THUẬT TOÁN AGNES



- Cho tập DL gồm 6 điểm trong không gian 2 chiều. Sử dụng thuật toán AGNES với Single link (khoảng cách gần nhất giữa 2 điểm của 2 nhóm khác nhau) để gom nhóm

Điểm	Tọa độ x	Tọa độ y
P1	0.40	0.53
P2	0.22	0.38
P3	0.353	0.32
P4	0.26	0.19
P5	0.08	0.41
P6	0.45	0.30



## VÍ DỤ : THUẬT TOÁN AGNES



- Xây dựng ma trận khoảng cách (độ đo Euclide) giữa các điểm

	P1	P2	P3	P4	P5	P6
P1	0.00	0.23	0.22	0.37	0.34	0.24
P2	0.23	0.00	0.15	0.19	0.14	0.24
P3	0.22	0.15	0.00	0.16	0.29	0.10
P4	0.37	0.19	0.16	0.00	0.28	0.22
P5	0.34	0.14	0.29	0.28	0.00	0.39
P6	0.24	0.24	0.10	0.22	0.39	0.00

49

## VÍ DỤ : THUẬT TOÁN AGNES



### Sử dụng Single Link :

1. Bước 1 : mỗi điểm là một nhóm

2. Bước 2 :

- Trong số các nhóm gồm một điểm thì  $\text{dist}(3,6)$  - min nên gộp điểm P3 và P6 với nhau thành một nhóm
- Thu được các nhóm : {1}, {4}, {2}, {5}, {3,6},

3. Quay lại bước 2 do chưa thu được nhóm “toàn bộ”

4. Tính khoảng cách giữa các nhóm . Ví dụ :

- $\text{Dist}(\{3,6\},\{1\}) = \min(\text{dist}(3,1), \text{dist}(6,1))$   
 $= \min(0.22, 0.24) = 0.22$

50

## VÍ DỤ : THUẬT TOÁN AGNES



### Sử dụng Single Link :

5.  $\text{dist}(2,5)$  là nhỏ nhất nên gộp P2 và P5. Ta có các nhóm sau :  $\{1\}$ ,  $\{4\}$ ,  $\{3,6\}$ ,  $\{2,5\}$

6. Tính khoảng cách giữa các nhóm. Ví dụ :

- $\text{dist}(\{3,6\},\{2,5\})$   
 $= \min(\text{dist}(3,2), \text{dist}(6,2), \text{dist}(3,5), \text{dist}(6,5))$   
 $= \min(0.15, 0.24, 0.28, 0.39) = 0.15$

....

- $\text{dist}(\{3,6\},\{2,5\})$  nhỏ nhất nên gộp các nhóm  $\{3,6\}$ ,  $\{2,5\}$  thành một nhóm.

- Ta thu được các nhóm :  $\{1\}, \{4\}, \{2,3,5,6\}$

51

## VÍ DỤ : THUẬT TOÁN AGNES



### Sử dụng Single Link :

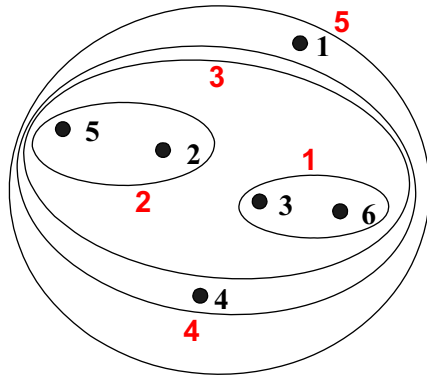
7. Tiếp tục :

- Tính khoảng cách giữa các nhóm.
- Gộp  $\{4\}$  với  $\{2,3,5,6\}$  thu được các nhóm  $\{1\}$ ,  $\{2,3,4,5,6\}$

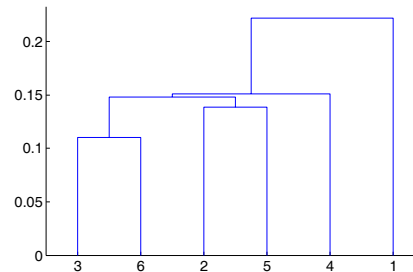
8. Gộp 2 nhóm này ta thu được nhóm “toàn bộ” và thuật toán dừng

52

## VÍ DỤ : THUẬT TOÁN AGNES



Các nhóm  
(Single Link)



Sơ đồ hình cây

53

## PHƯƠNG PHÁP PHÂN CẤP

### 4. Nhược điểm :

- ✦ Tính cơ bản thấp : Độ phức tạp là  $O(n^2)$  với  $n$  - số đối tượng
- ✦ Không thể quay lui về bước trước .
- ✦ Khó xác định phương pháp tích tụ hay chia nhỏ
- ✦ Nhạy cảm với nhiễu, cá biệt
- ✦ *Gặp vấn đề khi các nhóm có kích thước khác nhau và có hình dáng lồi*
- ✦ *Có xu hướng phân chia các nhóm DL lớn*
- *Tích hợp phương pháp phân cấp với phương pháp phân hoạch (dựa trên khoảng cách) : BIRCH, CURE, CHAMELEON*

54

# CÁC CÔNG VIỆC CẦN LÀM



1. Thảo luận và tự thực hiện các bài tập của chương 5 – Phần 1 (không nộp)
2. Chuẩn bị bài 5 – Phần 2 : Gom nhóm dữ liệu
  - Xem nội dung bài 5 – P.2.
  - **Cách thực hiện :**
    - **Đọc slide, xem các ví dụ**
    - **Tham khảo trên Internet và tài liệu tham khảo**
3. Chuẩn bị báo cáo xemina.

55

## Bài tập cá nhân



- **Thời gian : 20'**
- Cho tập DL gồm 6 điểm trong không gian 2 chiều với ma trận khoảng cách như sau.
- Sử dụng thuật toán AGNES với Complete link (khoảng cách giữa 2 nhóm là khoảng cách xa nhất giữa 2 điểm của 2 nhóm) để gom nhóm.
- **Nộp bài theo nhóm.**

56

## Bài tập cá nhân



- Cho ma trận khoảng cách giữa các điểm như sau :

	P1	P2	P3	P4	P5	P6
P1	0.00	0.23	0.22	0.37	0.34	0.24
P2	0.23	0.00	0.15	0.19	0.14	0.24
P3	0.22	0.15	0.00	0.16	0.29	0.10
P4	0.37	0.19	0.16	0.00	0.28	0.22
P5	0.34	0.14	0.29	0.28	0.00	0.39
P6	0.24	0.24	0.10	0.22	0.39	0.00

57

## Qui định trình bày bài nộp



### Bài tập nộp cá nhân

- Ngày nộp :
- Tên nhóm :
  - Họ và tên:
  - Mã số SV :
- Nội dung :
- *Lưu ý : nộp bài làm theo từng nhóm*

58

## TÀI LIỆU THAM KHẢO



1. J.Han, M.Kamber, Chương 8 – Data mining : Concepts and Techniques  
<http://www.cs.sfu.ca/~han/dmbook>  
<http://www-faculty.cs.uiuc.edu/~hanj/bk2/slidesindex.html> : 2<sup>nd</sup>
3. P.-N. Tan, M. Steinbach, V. Kumar, Chương 8 - Introduction to Data Mining  
<http://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf>

59

## BÀI TẬP



1. Cho 2 đối tượng :  $(22, 1, 42, 10)$  và  $(20, 0, 36, 8)$ 
  - a) Tính khoảng cách Euclide giữa 2 đối tượng
  - b) Tính khoảng cách Mahanttan giữa 2 đối tượng
  - c) Tính khoảng cách Minkowski giữa 2 đối tượng với  $q=3$
2. Thế nào là gom nhóm?. Trình bày chi tiết phương pháp phân hoạch, phân cấp. Cho ví dụ cụ thể từng phương pháp. So sánh ưu, khuyết điểm của 2 phương pháp.

60

# BÀI TẬP



3. Ta có 8 đối tượng sau (biểu diễn thông qua tọa độ  $(x,y)$ ) :  $A_1(2,10)$ ,  $A_2(2,5)$ ,  $A_3(8,4)$ ,  $B_1(5,8)$ ,  $B_2(7,5)$ ,  $B_3(6,4)$ ,  $C_1(1,2)$ ,  $C_2(4,9)$ .
- Sử dụng khoảng cách Euclide và giả sử gán  $A_1, B_1, C_1$  là các trung tâm của các nhóm tương ứng. Sử dụng thuật toán k-means ( với  $k=3$ ) để xác định:
    - Ba trung tâm của nhóm sau vòng lặp thi hành đầu tiên. Tính độ đo SSE cho các nhóm.
    - Ba nhóm kết quả cuối cùng. Tính độ đo SSE cho các nhóm.
  - Tự chọn 3 trung tâm nhóm bất kỳ không trùng với 8 đối tượng đã cho. Sử dụng k-mean ( $k=3$ ) để xác định các nhóm. So sánh kết quả với câu a)**
  - Sử dụng thuật toán phân cấp lần lượt với Single link và Complete link để xác định 3 nhóm từ DL trên. Vẽ sơ đồ hình cây tương ứng**

61

# BÀI TẬP



Customer	Age	Income (K)	No. cards
Thảo	35	37	3
Hưng	25	51	3
Gia	29	44	1
Thành	45	100	3
Thủy	20	30	4
Đức	33	57	2
Minh	65	200	1
Nhung	54	142	2
Nhật	58	175	1
Tùng	25	40	5

4. Cho CSDL bên :

- Sử dụng k-mean để gom cụm với  $k=3$ . Tính độ đo SSE và so sánh kết quả.
- Chuẩn hóa CSDL và gom cụm với  $k=3$ . So sánh kết quả với câu a).**

62

## Bài tập



5. Cho tập DL gồm 5 điểm trong không gian 2 chiều với ma trận khoảng cách đã cho. Sử dụng thuật toán AGNES lần lượt với Single Link và Complete link để gom nhóm. Vẽ sơ đồ hình cây.
- *Xác định 3 nhóm thu được từ sơ đồ hình cây theo cả 2 cách.*

	P1	P2	P3	P4	P5
P1	0.00	0.10	0.41	0.55	0.35
P2	0.10	0.00	0.64	0.47	0.98
P3	0.41	0.64	0.00	0.44	0.85
P4	0.55	0.47	0.44	0.00	0.76
P5	0.35	0.98	0.85	0.76	0.00

63

## Q & A



64