

問題解決技法入門

クラスタ分析

堀田 敬介

クラスタ分析

● Contents

- クラスタ分析

1. クラスタ分析概要
2. 類似度の測定
3. クラスタ化の方法の決定(類似度更新法)

- クラスタ分析〔階層的方法〕の実施

4. Excelで計算したクラスタ分析, Rによるクラスタ分析
5. クラスタ分析実施上の注意点

- クラスタ分析〔非階層的方法〕

6. 非階層的クラスタ分析〔K-means法〕
7. Rによるクラスタ分析〔K-means法〕

1. クラスタ分析概要

- クラスタ分析とは？

- 複数の対象(もの, 変数など)を, その**属性**によって**類似度 (similarity)**をはかり, 均質な**集団 (cluster)**に分類する方法の総称

どれとどれが似てる？
(同じクラスター？)

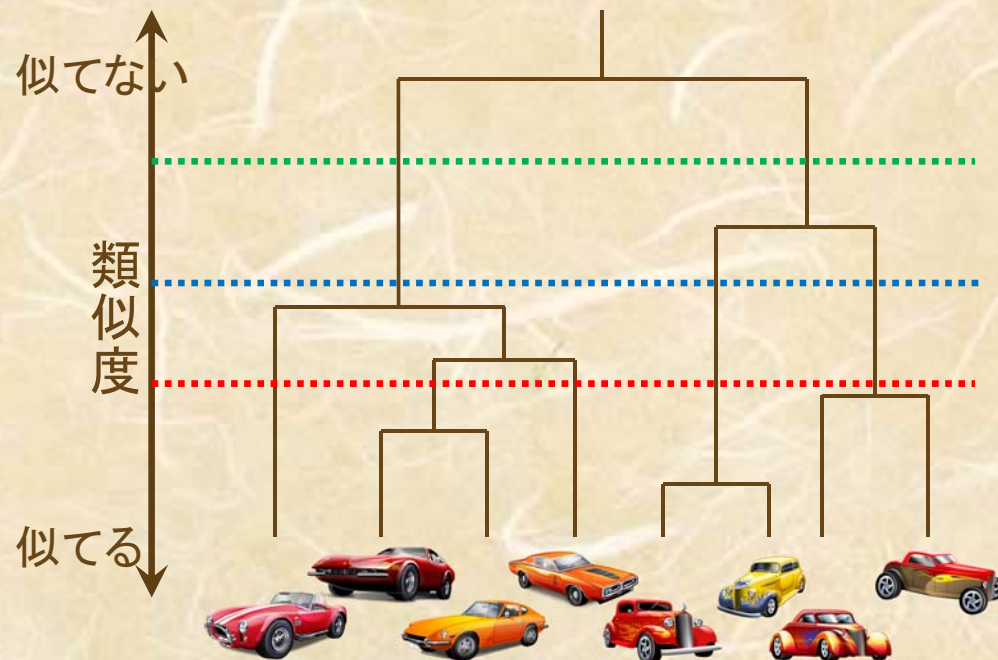


1. クラスタ分析概要

● クラスタ分析の種類

● 階層的方法

- 樹形図(デンドログラム)を作成
- 目的により高さを決めてクラスタリング



● 非階層的方法

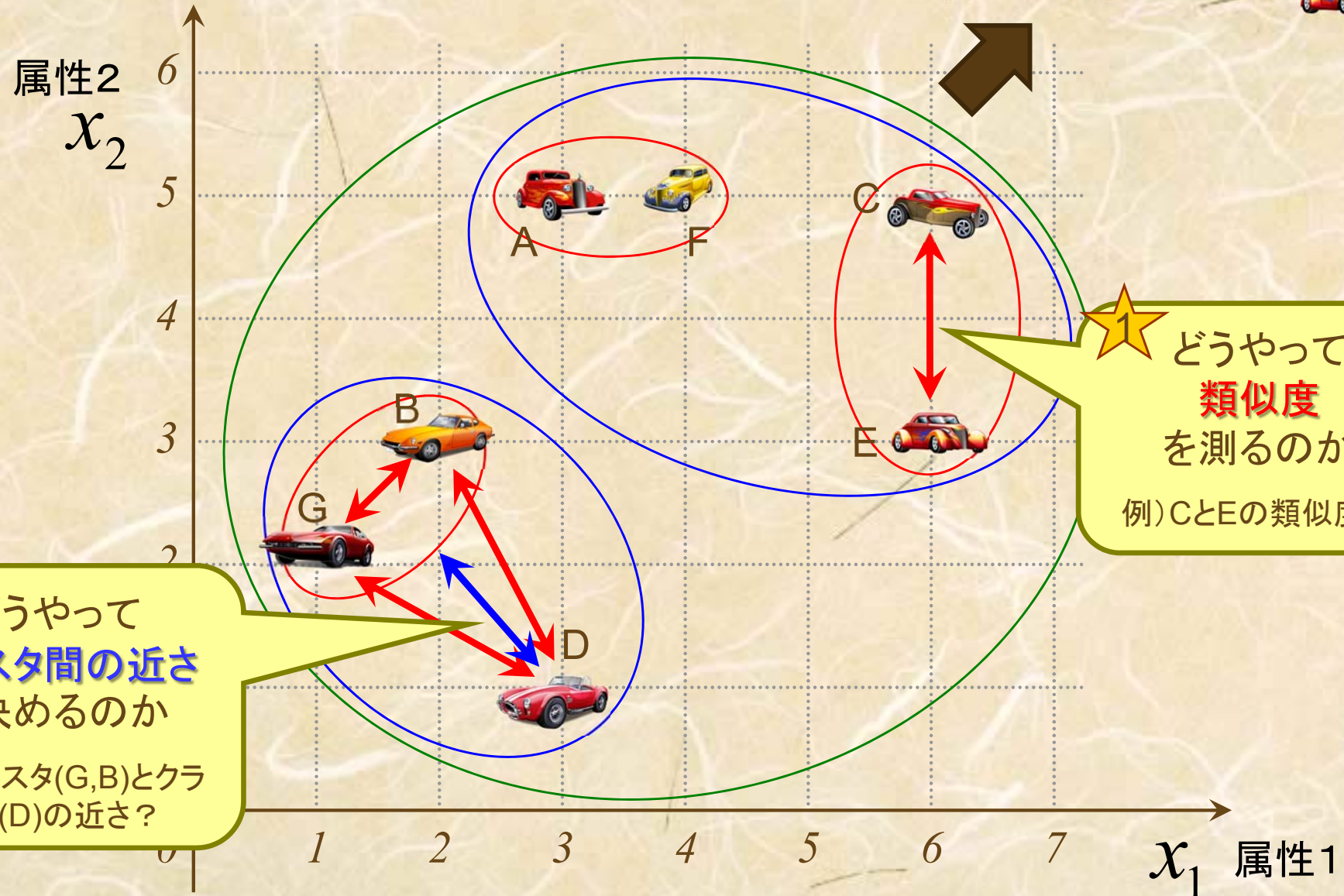
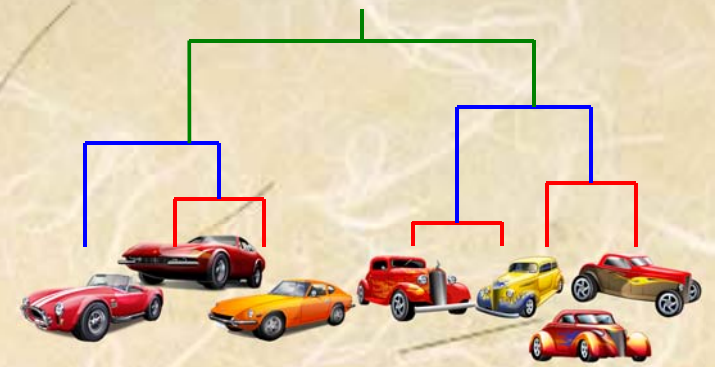
- 予めクラスタ数を決め (or決まっていたり), **クラスタリング**を行う



例: 3つのクラスタに分類

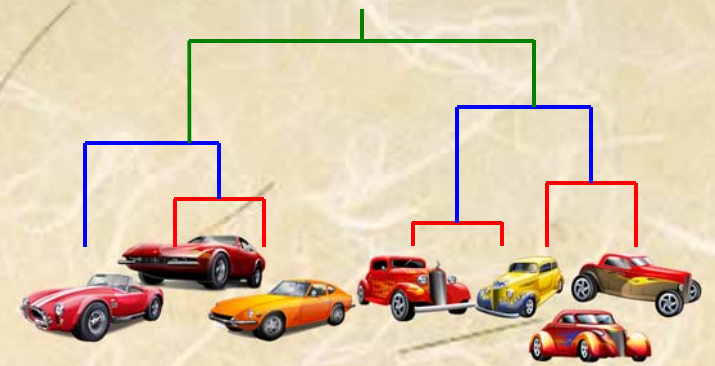
1. クラスタ分析概要

- 例：階層的的方法（対象の属性が2つの場合）



1. クラスタ分析概要

- どうやって類似度を測るか？



									
			3	1	2	3	4	6	6
	x_1	x_2	1	2	3	5	5	5	3
	3	1							
	1	2							
	2	3							
	3	5							
	4	5							
	6	5							
	6	3							

2. 類似度の測定

類似度は尺度により距離や相関で測る
(距離: 近いほうが類似)
(相関: 高いほうが類似)

● 距離【**間隔尺度**】

- ユークリッド距離
- ユークリッド平方距離
- 重み付きユークリッド距離
- マンハッタン距離
- ミンコフスキー距離
- マハラノビス汎距離

● 相関【**間隔尺度**】

- Pearsonの積率相関係数
- ベクトル内積

● 相関【**順序尺度**】

- Spearmanの順位相関係数
- Kendallの順位相関係数

● 距離【**名義尺度** [0, 1]】

- 類似比
- 一致係数
- Russel-Rao係数
- Rogers-Tanimoto係数
- Hamann係数
- ファイ係数

● 変量間類似度【**名義尺度**】

- 平均平方根一致係数
- グッドマン・クラスカル λ

2. 類似度の測定

データと尺度

				比率尺度	比率尺度		
			間隔尺度	間隔尺度	間隔尺度		
			順序尺度	順序尺度	順序尺度	順序尺度	
名義尺度	名義尺度	名義尺度	名義尺度	名義尺度	名義尺度	名義尺度	
学籍番号	氏名	性別	生年月日	身長	体重	問題発見技法成績	...
1	文教太郎	男	1987.5.6	175cm	69kg	B	...
2	湘南花子	女	1988.1.4	163cm	48kg	AA	...
3	⋮	⋮	⋮	⋮	⋮	⋮	



2. 類似度の測定



● 個体間類似度

- ユークリッド距離

(cf. l_2 -ノルム)

- マンハッタン距離

(cf. l_1 -ノルム)

- ミンコフスキー距離

(cf. l_p -ノルム)

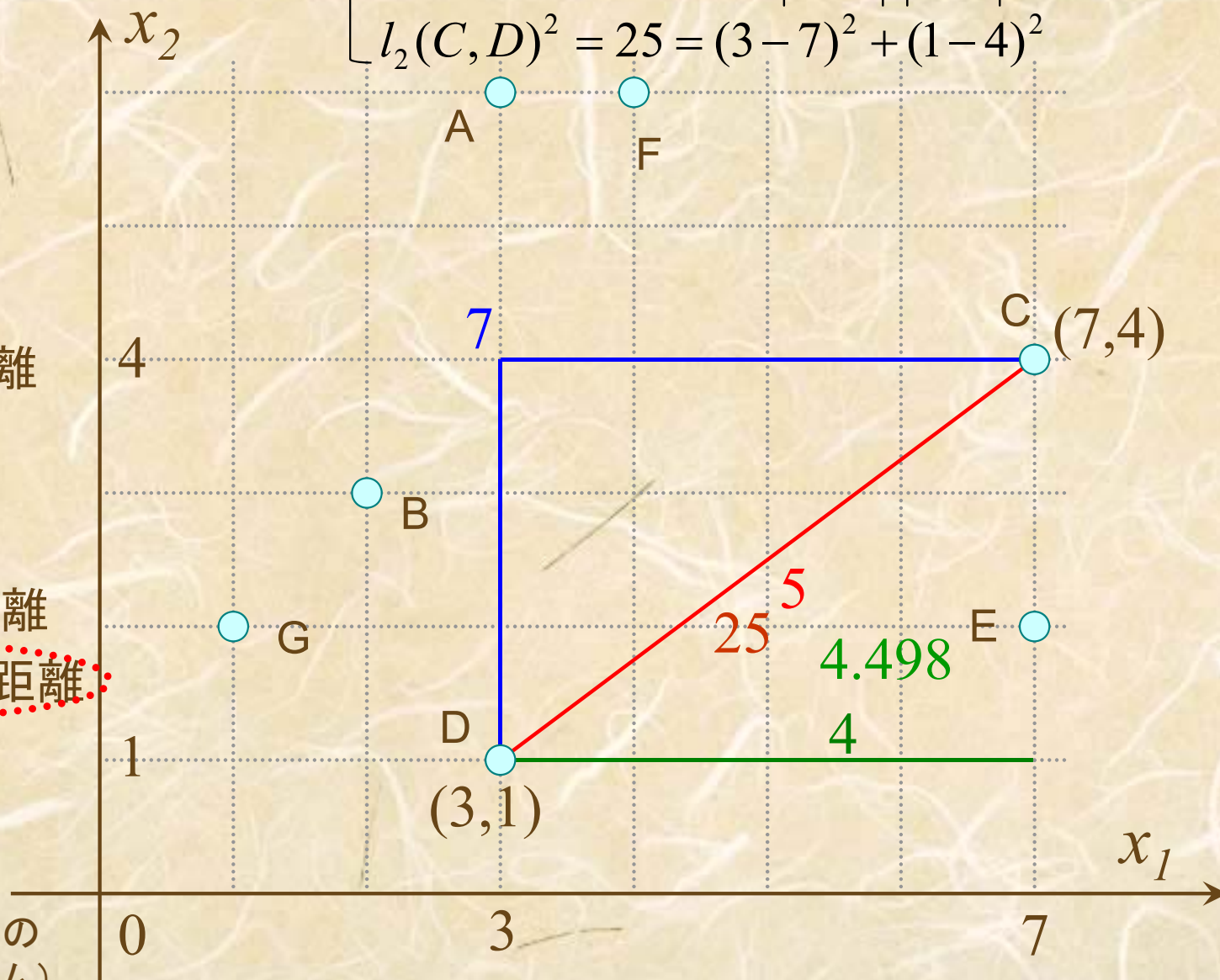
(cf. l_∞ -ノルム)

- マハラノビス汎距離

- ユークリッド平方距離

クラスター分析でよく使われる

(注: 各ノルムとは2変量の差ベクトルに対するノルム)



$$l_2(C, D) = 5 = \sqrt{(3-7)^2 + (1-4)^2}$$

$$l_1(C, D) = 7 = |3-7| + |1-4|$$

$$l_3(C, D) = 4.498 = \sqrt[3]{|3-7|^3 + |1-4|^3}$$

$$l_\infty(C, D) = 4 = \max\{|3-7|, |1-4|\}$$

$$l_2(C, D)^2 = 25 = (3-7)^2 + (1-4)^2$$

2. 類似度の測定

● 個体間類似度

- ユークリッド距離
(cf. l_2 -ノルム)

- マンハッタン距離
(cf. l_1 -ノルム)

- ミンコフスキー距離
(cf. l_p -ノルム)
(cf. l_∞ -ノルム)

● マハラノビス汎距離

2変量版 $\mathbf{x} = (x_1, x_2)$

$$D \equiv \sqrt{\frac{u_1^2 + u_2^2 - 2\rho u_1 u_2}{1 - \rho^2}}$$

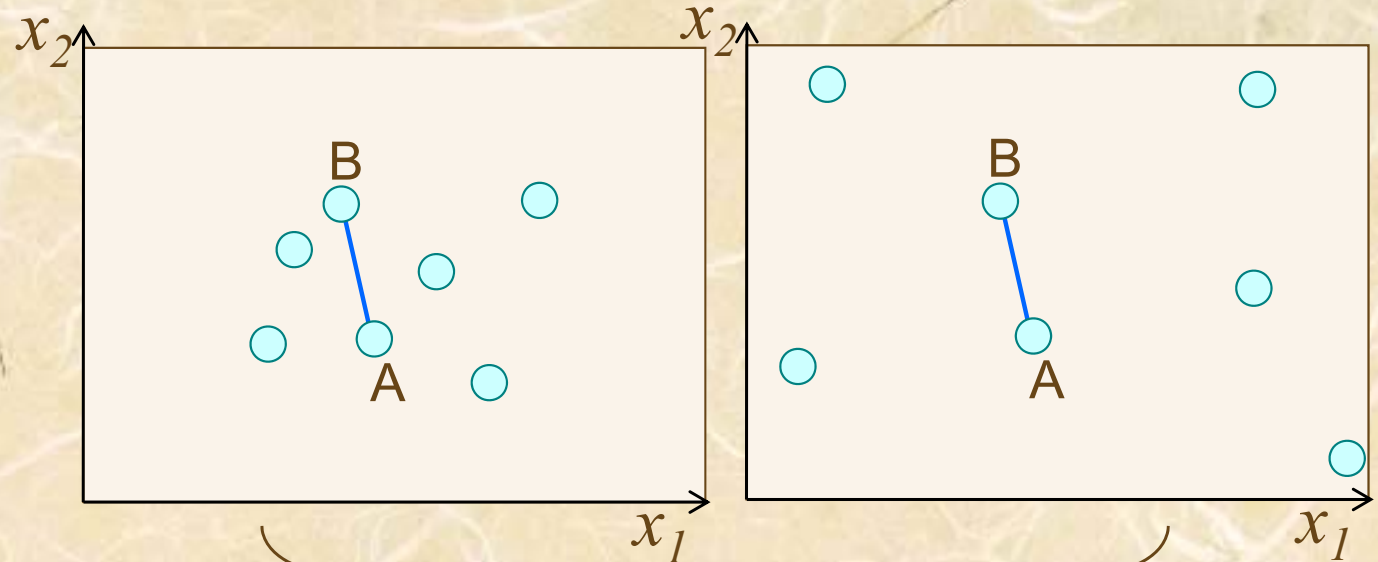
多変量版 $\mathbf{x} = (x_1, \dots, x_m)$

$$D \equiv (\mathbf{x}_p - \mathbf{x}_q)^T \Sigma^{-1} (\mathbf{x}_p - \mathbf{x}_q)$$

u_1, u_2 は x_1, x_2 の標準化変量で,
 $u_1 = \frac{x_1 - \mu_1}{\sigma_1}, u_2 = \frac{x_2 - \mu_2}{\sigma_2}$

μ_1, μ_2 はそれぞれ x_1, x_2 の平均
 σ_1, σ_2 はそれぞれ x_1, x_2 の標準偏差
 ρ は x_1, x_2 の相関係数

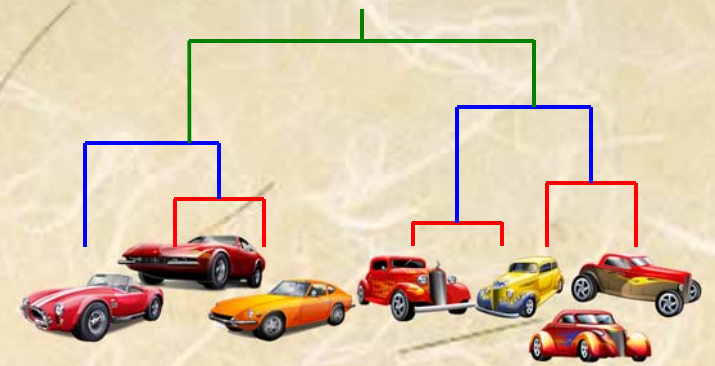
Σ は $\mathbf{x}_p, \mathbf{x}_q$ の分散共分散行列

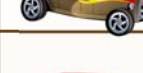


左側の対象内での、A-B間距離と
 右側の対象内でのA-B間距離が
異なる! (ユークリッド距離などでは同じ)

2. 類似度の測定

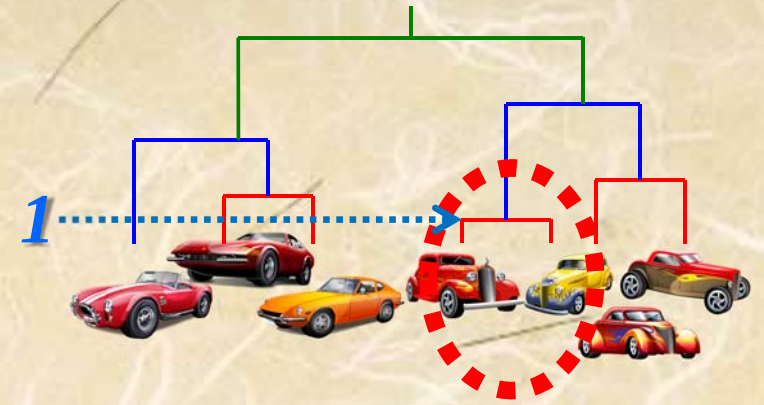
- どうやって類似度を測るか？
 - ・ 例：ユークリッド平方距離



									
			3	1	2	3	4	6	6
	x_1	x_2	1	2	3	5	5	5	3
	3	1		5	5	16	17	25	13
	1	2			2	13	18	34	26
	2	3				5	8	20	16
	3	5					1	9	13
	4	5						4	8
	6	5							4
	6	3							

2. 類似度の測定

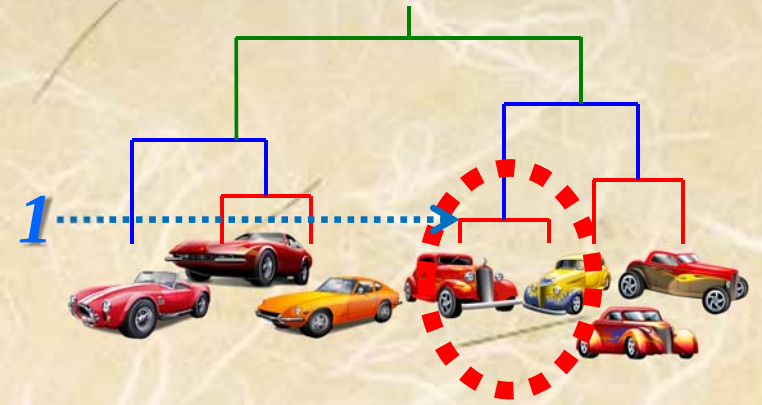
- どうやって類似度を更新するか？



									
			3	1	2	3	4	6	6
	x_1	x_2	1	2	3	5	5	5	3
	3	1		5	5	16	17	25	13
	1	2			2	13	18	34	26
	2	3				5	8	20	16
	3	5					1	9	13
	4	5						4	8
	6	5							4
	6	3							

2. 類似度の測定

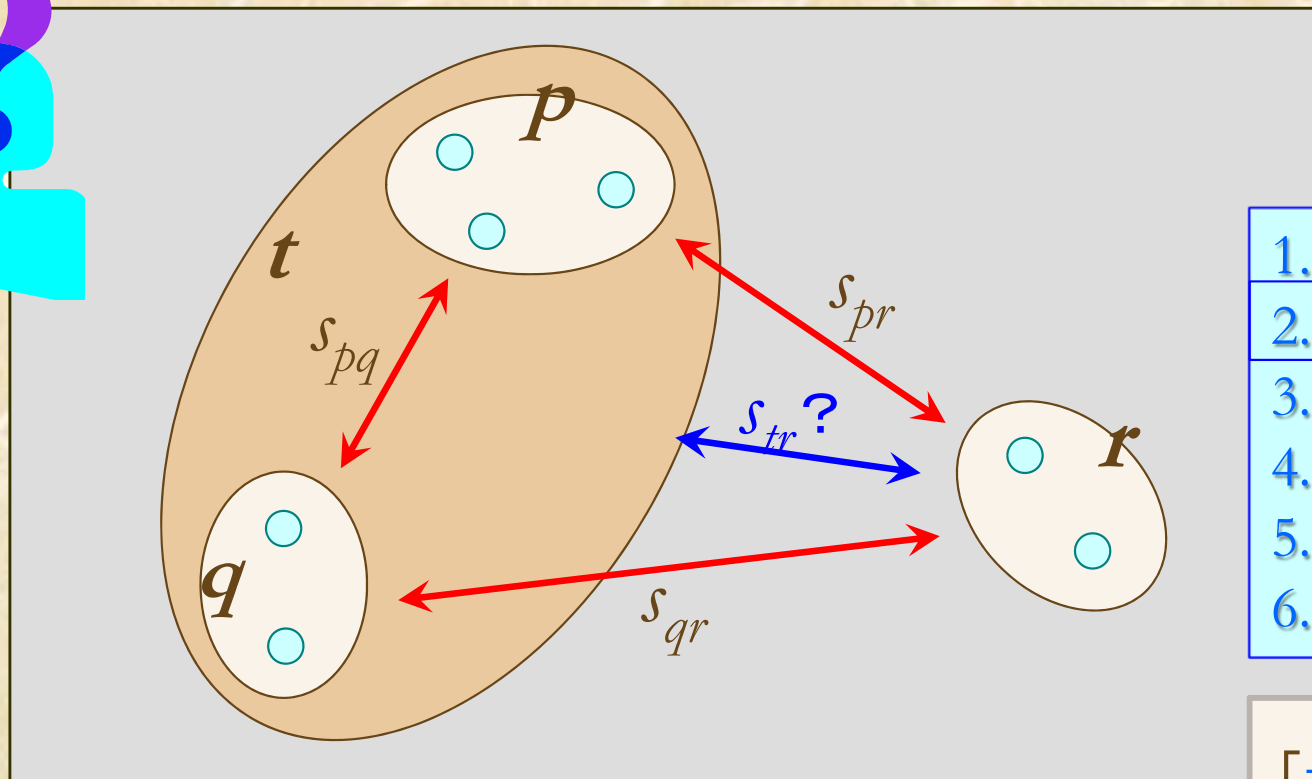
- どうやって類似度を更新するか？



								
			3	1	2	3,4	6	6
	x_1	x_2	1	2	3	5,5	5	3
	3	1		5	5	16,17	25	13
	1	2			2	13,18	34	26
	2	3				5,8	20	16
	3,4	5,5				1	9,4	13,8
	6	5						4
	6	3						

3. クラスタ化の方法

- 新たなクラスタ生成時の類似度の更新方法
 - クラスタ p , クラスタ q が一つのクラスタ t になる場合, 他のクラスタ r との類似度をどう更新する?



(s_{pr} : クラスタ p , r の類似度)

1. 最短距離法
2. 最長距離法
3. 群平均法
4. 重心法
5. 中央値法
6. ワード法

「最短」か「最長」か
何らかの「平均」

3. クラスタ化の方法

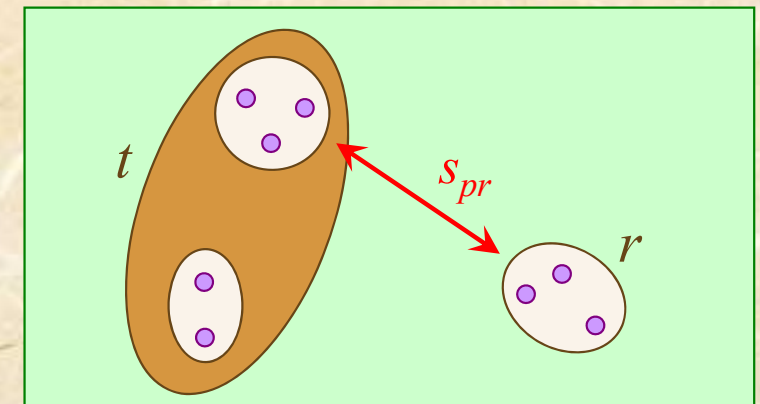
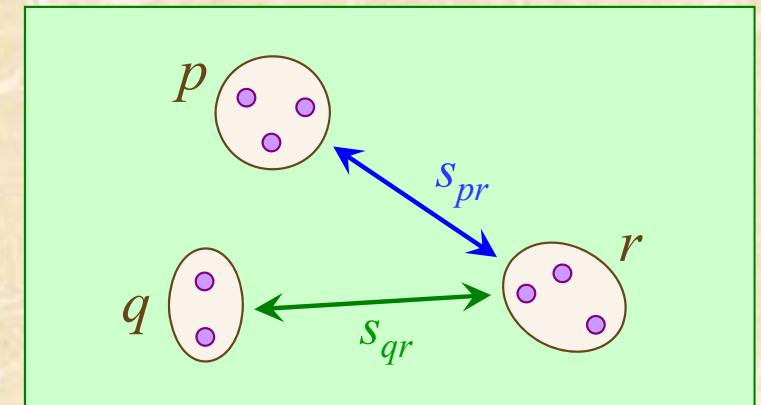
1. 最短距離法 (nearest neighbor method)

〔単連結法 (single linkage method)〕

$$s_{tr} = \min \{s_{pr}, s_{qr}\}$$

あるクラスタにおいて、クラスタ内の各対象が、そのクラスタ外の任意の対象よりも、そのクラスタ内の**少なくとも1つの対象とより近接**している。

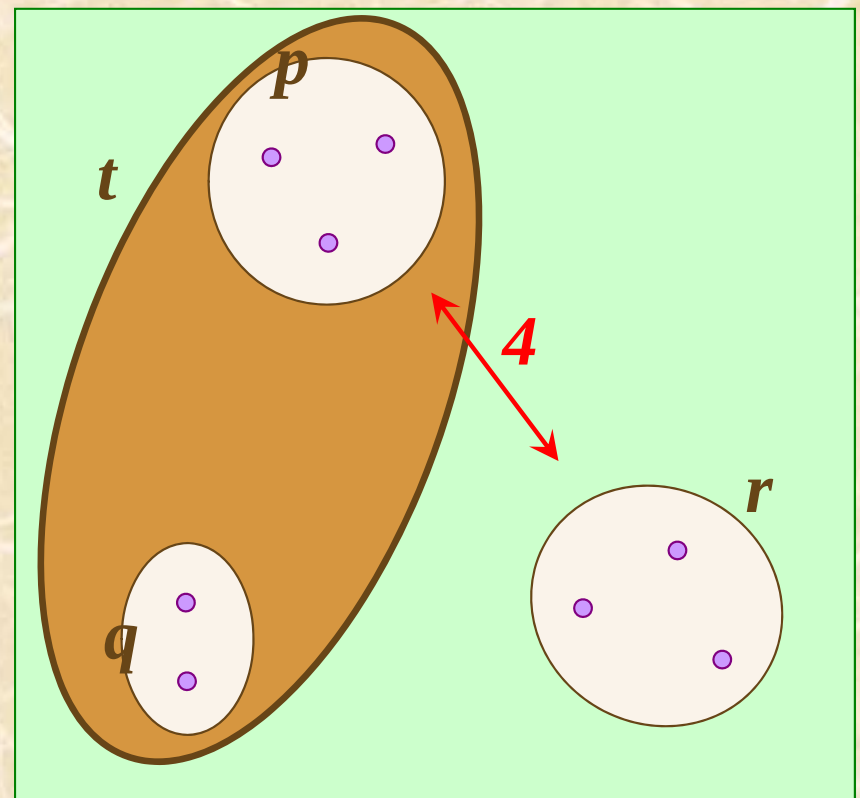
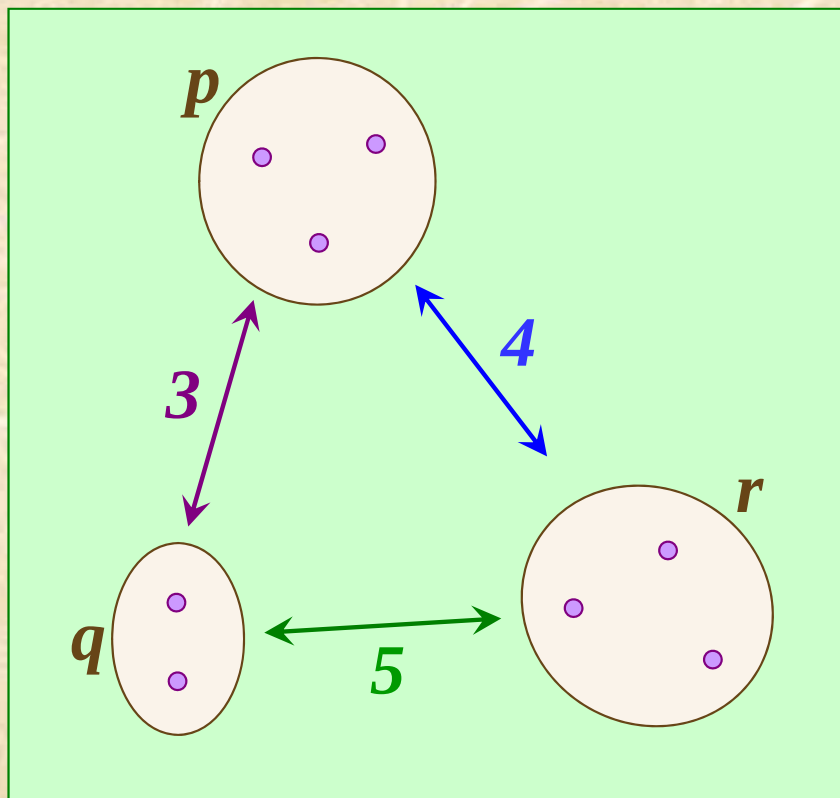
※類似度は、対象間の類似度の大小関係だけで決まる。
よって、類似度(距離)は**順序尺度**ならばよい。



3. クラスタ化の方法

1. 最短距離法

$$s_{tr} = \min \{s_{pr}, s_{qr}\}$$



3. クラスタ化の方法

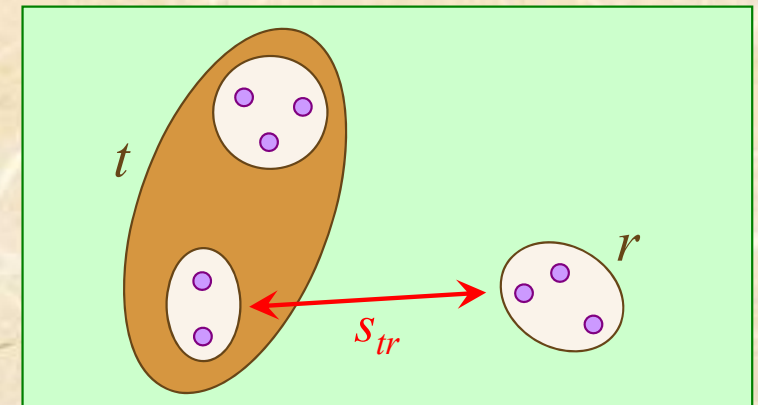
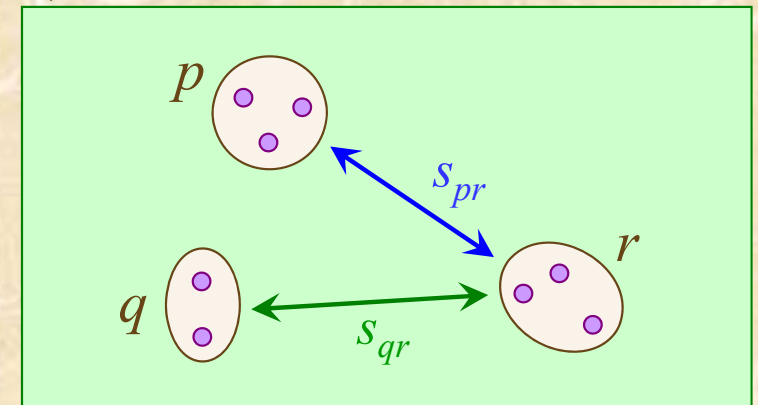
2. 最長距離法 (furthest neighbor method)

〔完全連結法 (complete linkage method)〕

$$s_{tr} = \max \{s_{pr}, s_{qr}\}$$

あるクラスタにおいて、クラスタ内の**全ての対象が**、そのクラスタ外の任意の対象との距離よりも**常に近接**している。

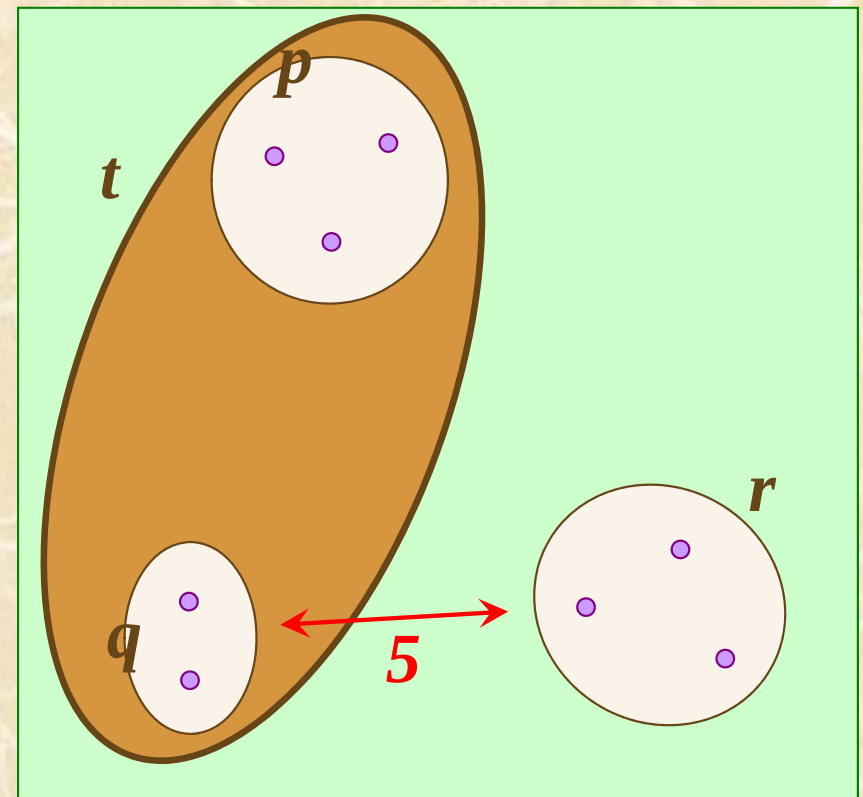
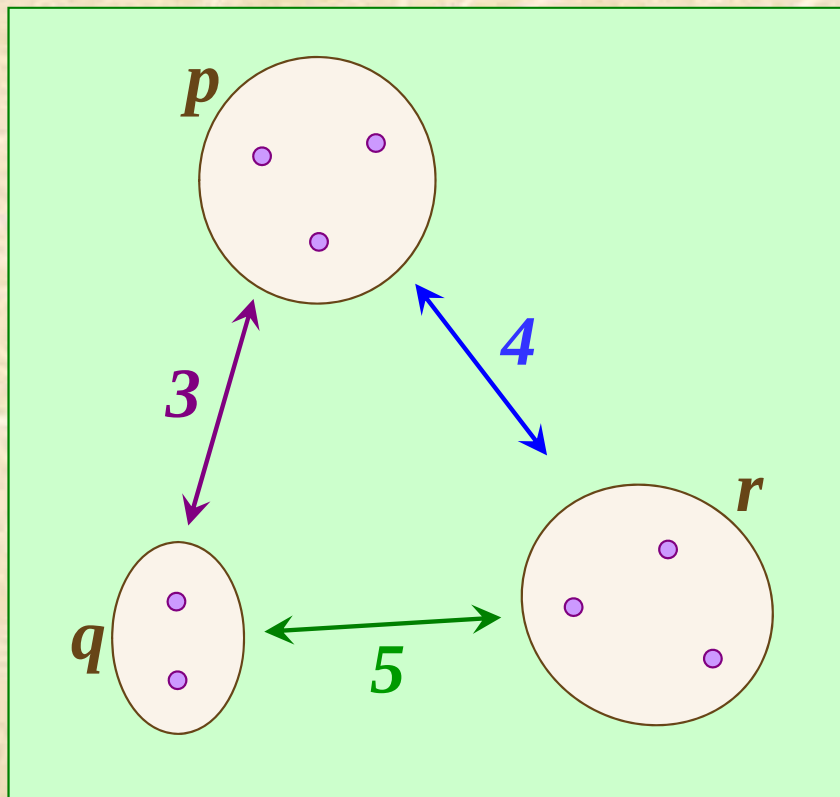
※類似度は、対象間の類似度の大小関係だけで決まる。
よって、類似度(距離)は**順序尺度**ならばよい。



3. クラスタ化の方法

2. 最長距離法

$$s_{tr} = \max \{s_{pr}, s_{qr}\}$$

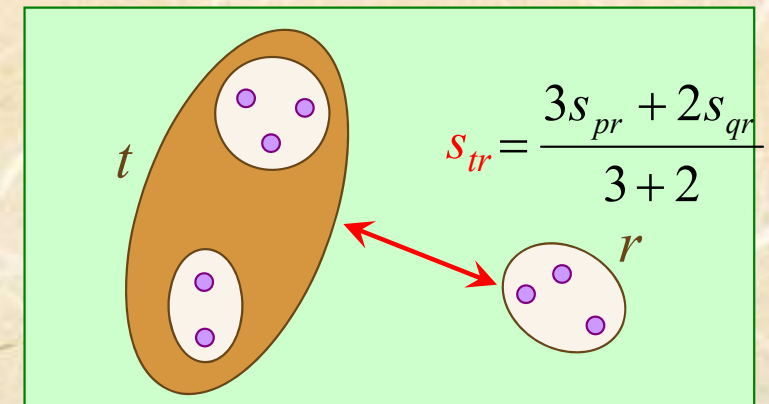
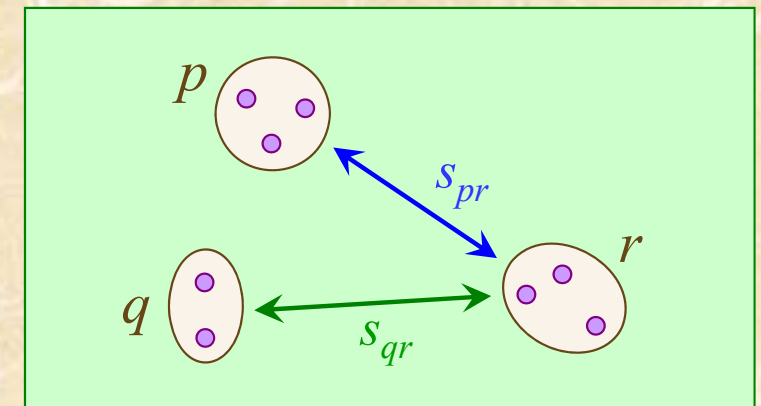


3. クラスタ化の方法

3. 群平均法 (group average method)

$$s_{tr} = \frac{n_p}{n_p + n_q} s_{pr} + \frac{n_q}{n_p + n_q} s_{qr}$$

$\left\{ \begin{array}{l} n_p : \text{クラスタ } p \text{ に含まれる対象数} \\ n_q : \text{クラスタ } q \text{ に含まれる対象数} \end{array} \right.$

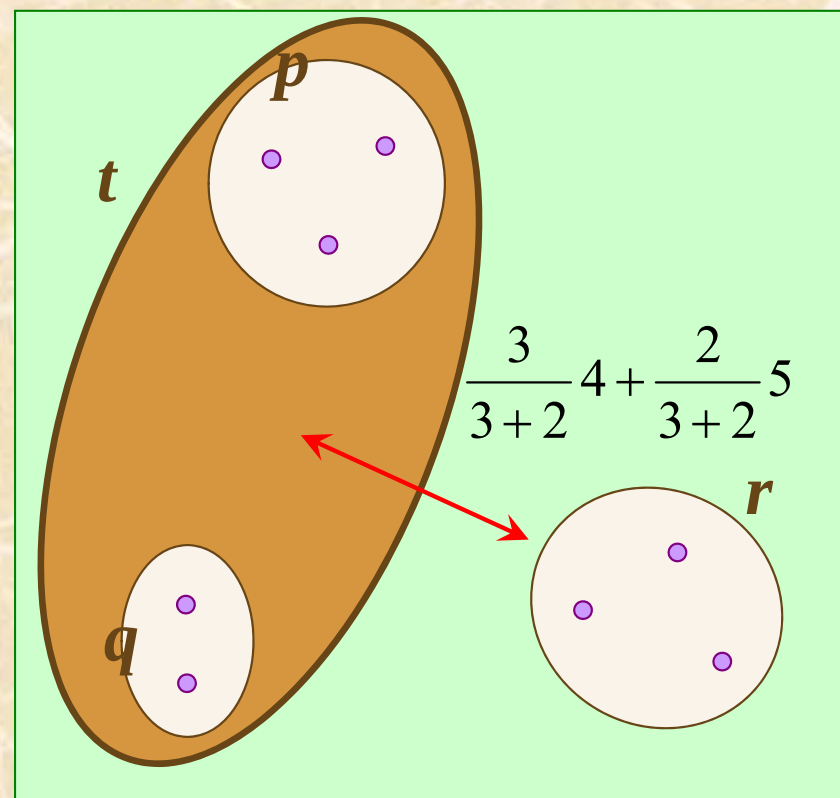
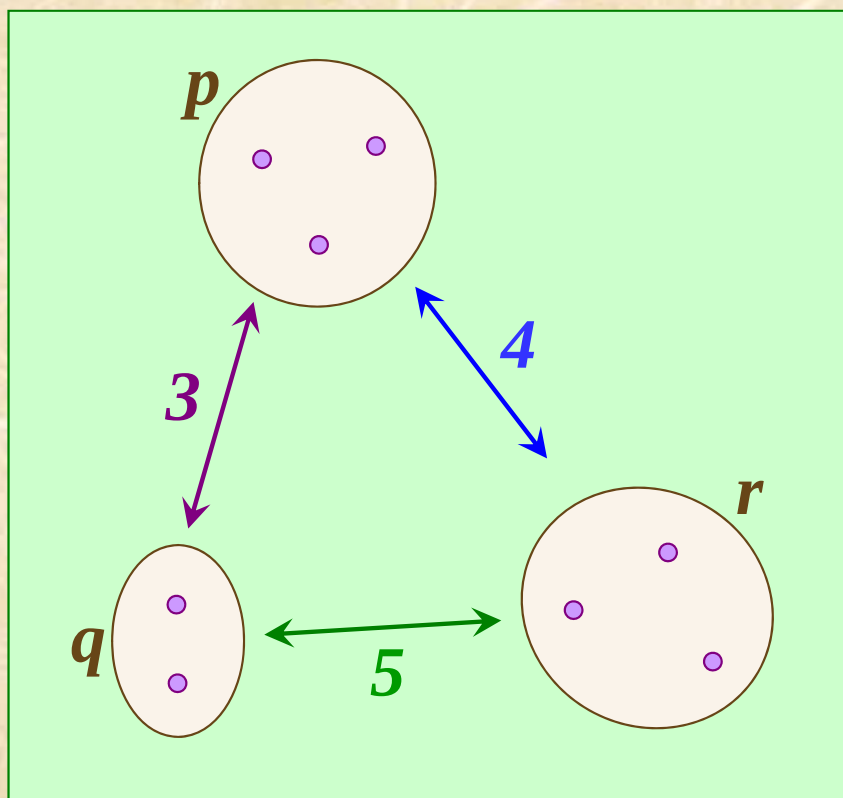


※類似度は、**間隔尺度**ならばOK

3. クラスタ化の方法

3. 群平均法

$$S_{tr} = \frac{n_p}{n_p + n_q} S_{pr} + \frac{n_q}{n_p + n_q} S_{qr}$$



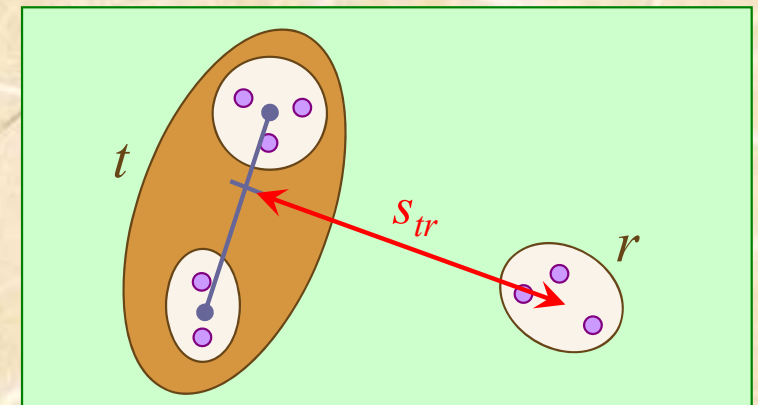
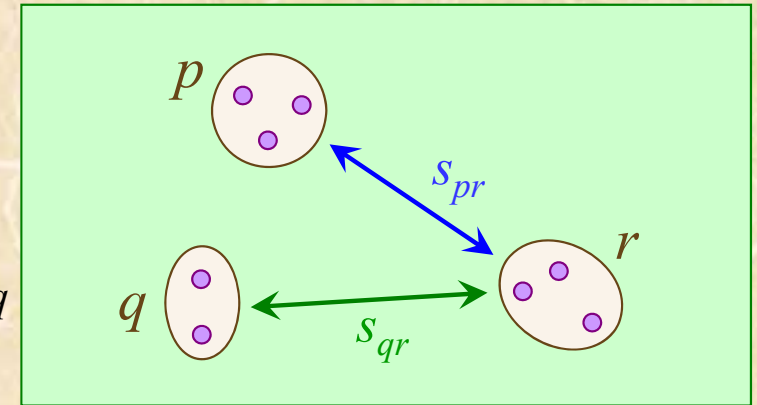
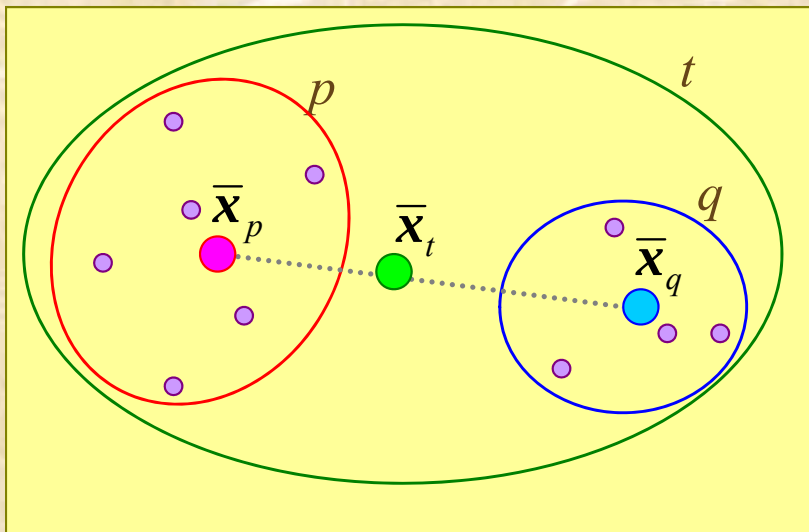
3. クラスタ化の方法

4. 重心法 (centroid method)

$$S_{tr} = \frac{n_p}{n_p + n_q} S_{pr} + \frac{n_q}{n_p + n_q} S_{qr} - \frac{n_p n_q}{(n_p + n_q)^2} S_{pq}$$

$\left\{ \begin{array}{l} n_p : \text{クラスタ } p \text{ に含まれる対象数} \\ n_q : \text{クラスタ } q \text{ に含まれる対象数} \end{array} \right.$

※導出過程より, 類似度 S_{tr} はユークリッド平方距離の時のみ妥当. → cf.ファイル「クラスタ分析ノート.pdf」



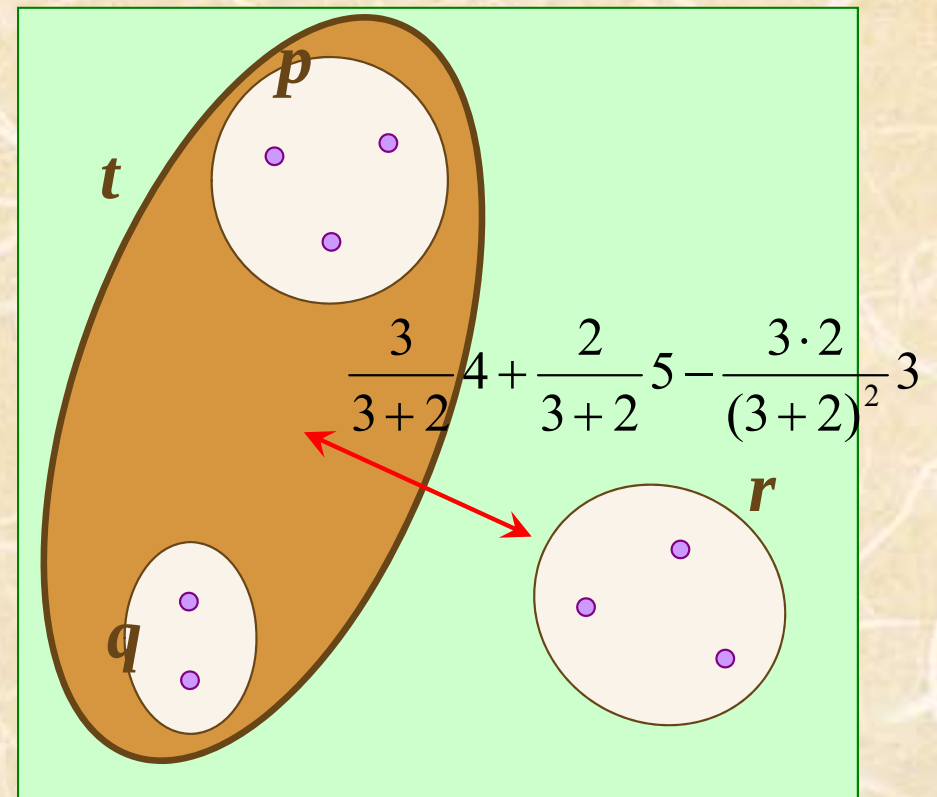
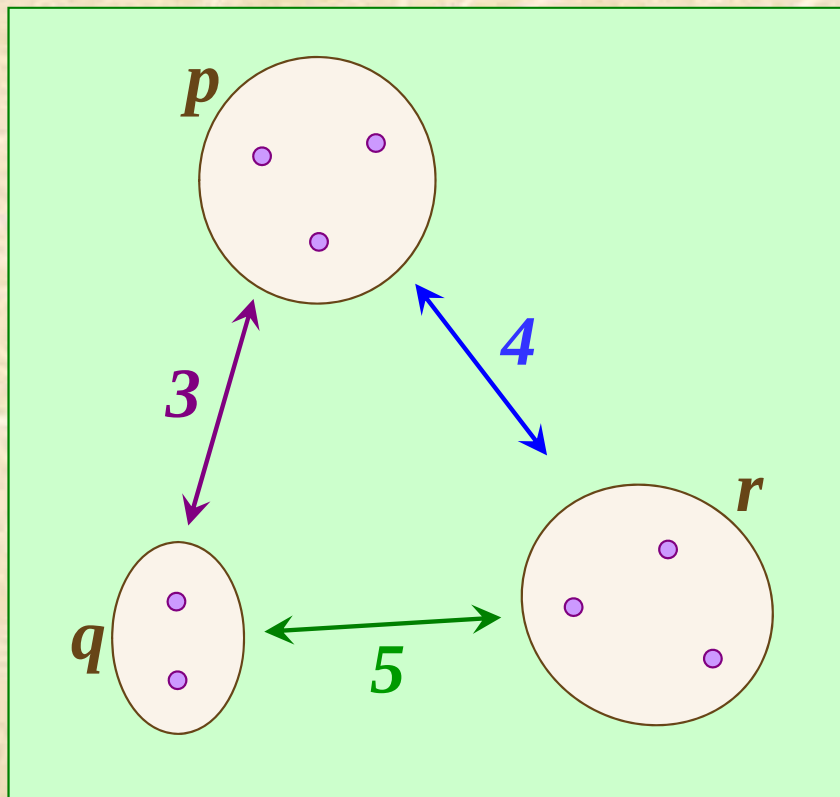
$$\bar{x}_t = \frac{n_p \bar{x}_p + n_q \bar{x}_q}{n_p + n_q}$$

※ $\mathbf{x}$ はベクトル

3. クラスタ化の方法

4. 重心法

$$s_{tr} = \frac{n_p}{n_p + n_q} s_{pr} + \frac{n_q}{n_p + n_q} s_{qr} - \frac{n_p n_q}{(n_p + n_q)^2} s_{pq}$$



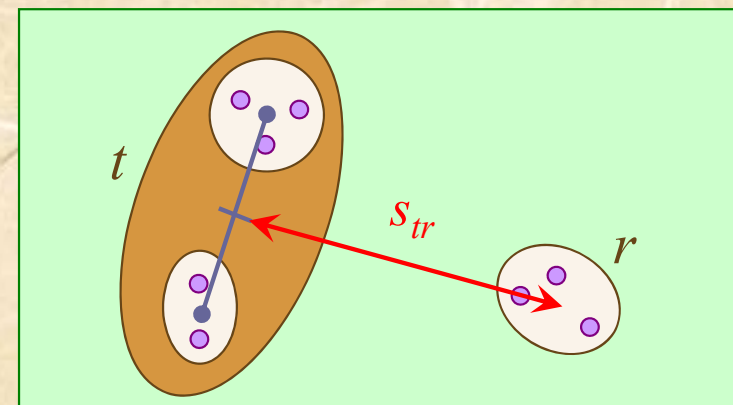
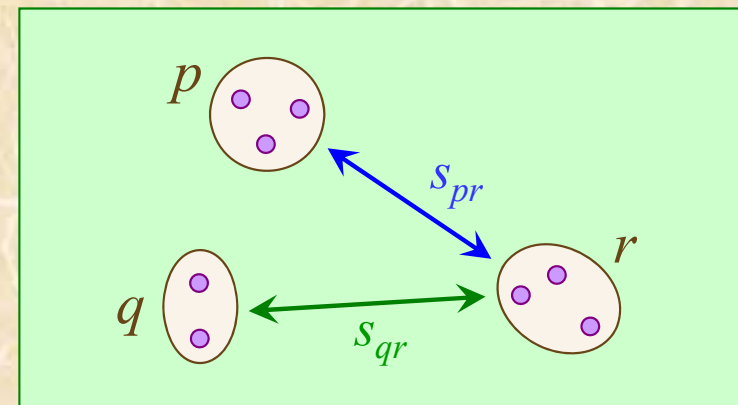
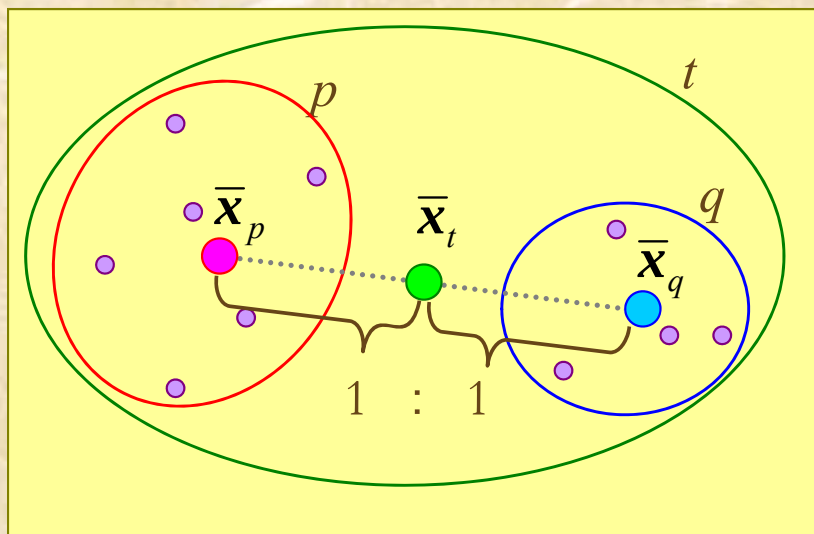
3. クラスタ化の方法

5. 中央値法 (median method)

$$s_{tr} = \frac{1}{2} s_{pr} + \frac{1}{2} s_{qr} - \frac{1}{4} s_{pq}$$

(重心法の簡易版, 重心の代わりに中央値を取る
重心法で $n_p := 1, n_q := 1$ に相当)

※導出過程より, 類似度 s_{tr} はユークリッド平方距離の時のみ妥当. → cf.ファイル「クラスタ分析ノート.pdf」



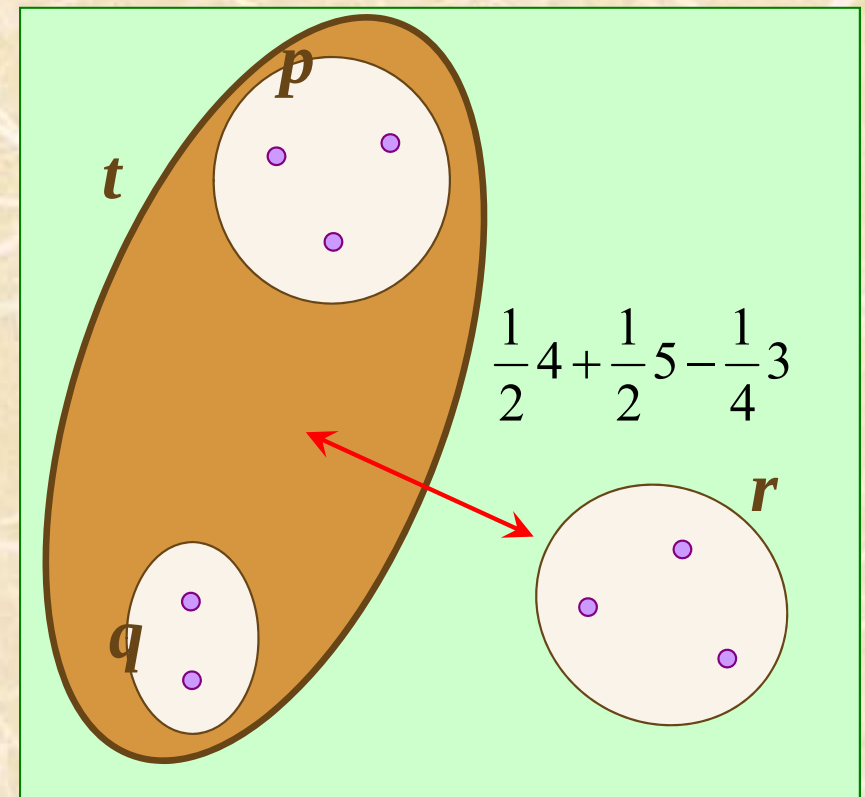
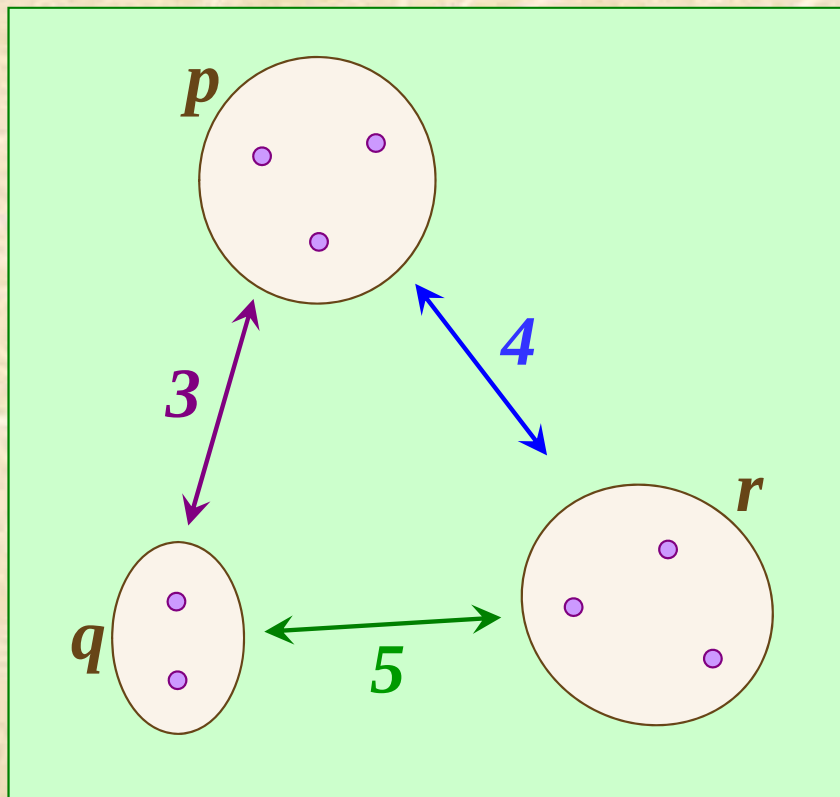
$$\bar{\mathbf{x}}_t = \frac{\bar{\mathbf{x}}_p + \bar{\mathbf{x}}_q}{2}$$

※ $\mathbf{x}$ はベクトル

3. クラスタ化の方法

5. 中央値法

$$s_{tr} = \frac{1}{2} s_{pr} + \frac{1}{2} s_{qr} - \frac{1}{4} s_{pq}$$

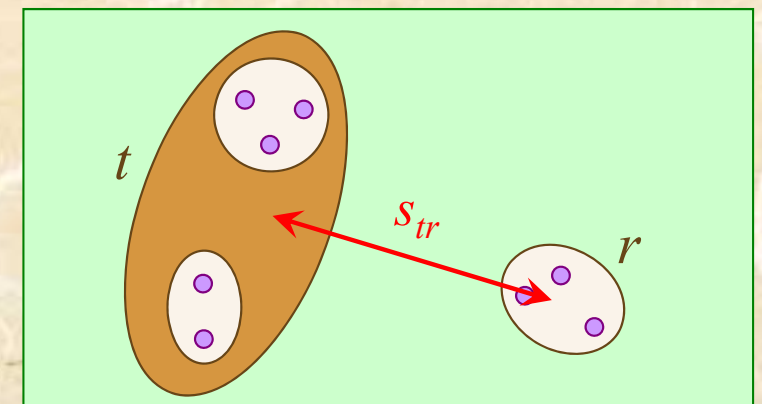
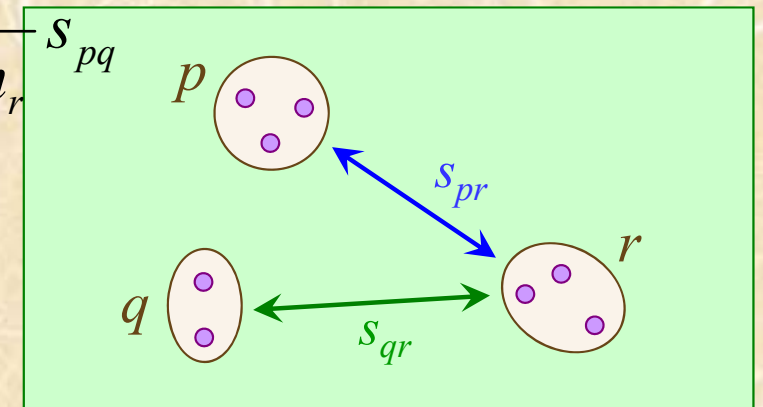


3. クラスタ化の方法

6. ウォード法 (Ward method)

$$S_{tr} = \frac{n_p + n_r}{n_p + n_q + n_r} S_{pr} + \frac{n_q + n_r}{n_p + n_q + n_r} S_{qr} - \frac{n_r}{n_p + n_q + n_r} S_{pq}$$

$\left\{ \begin{array}{l} n_p : \text{クラス} p \text{ に含まれる対象数} \\ n_q : \text{クラス} q \text{ に含まれる対象数} \\ n_r : \text{クラス} r \text{ に含まれる対象数} \end{array} \right.$

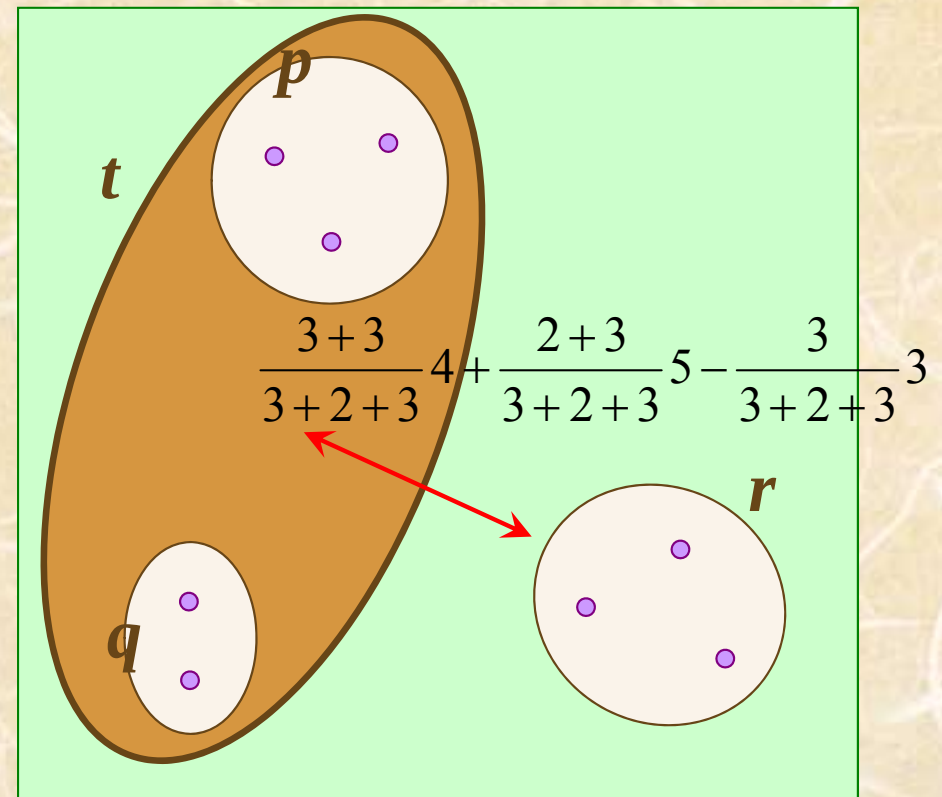
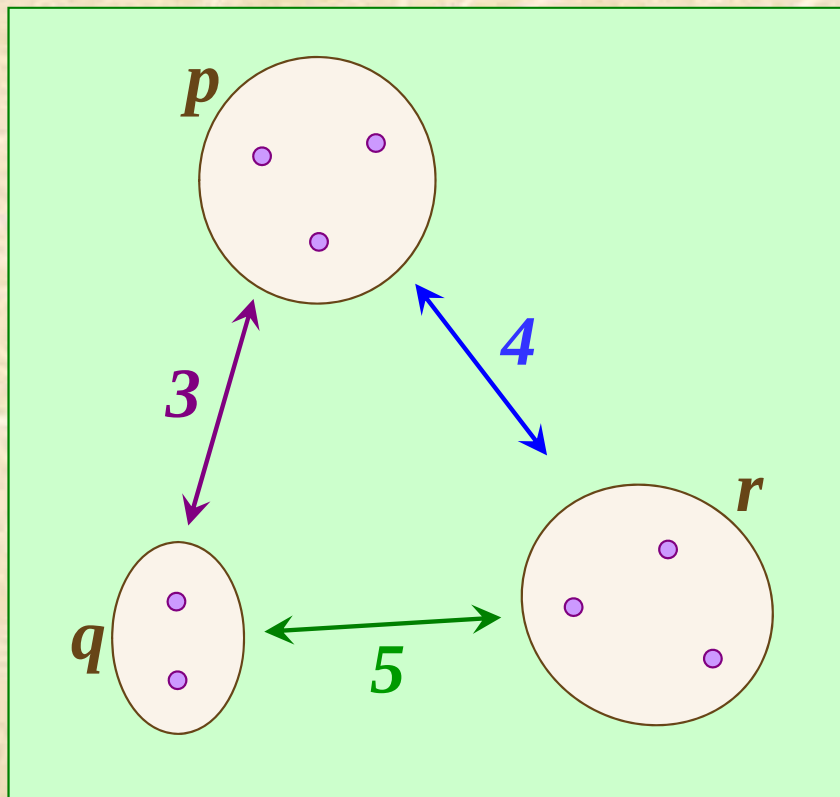


※導出過程より, 類似度 S_{tr} は
ユークリッド平方距離の時のみ妥当.
→ cf.ファイル「クラスタ分析ノート.pdf」

3. クラスタ化の方法

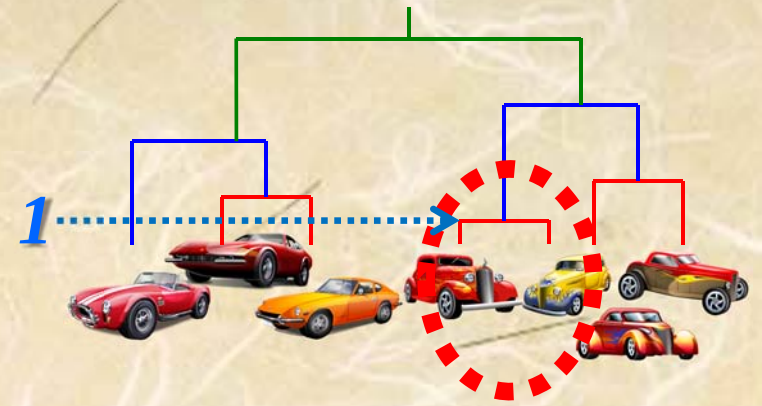
6. ウォード法

$$s_{tr} = \frac{n_p + n_r}{n_p + n_q + n_r} s_{pr} + \frac{n_q + n_r}{n_p + n_q + n_r} s_{qr} - \frac{n_r}{n_p + n_q + n_r} s_{pq}$$



3. クラスタ化の方法

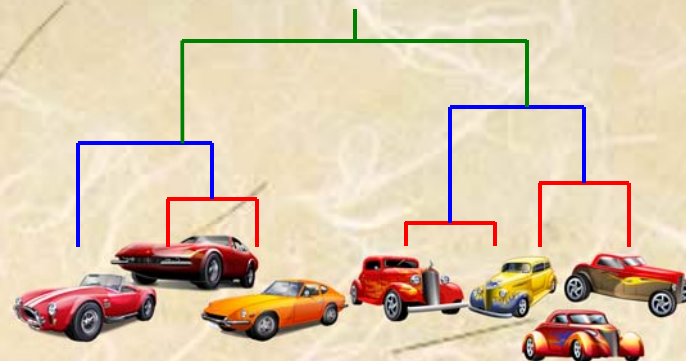
- どうやって類似度を更新するか？



								
			3	1	2	3:4	6	6
	x_1	x_2	1	2	3	5:5	5	3
	3	1		5	5	16:17	25	13
	1	2			2	13:18	34	26
	2	3				5:8	20	16
	3:4	5:5				1	9:4	13:8
	6	5						4
	6	3						

3. クラスタ化の方法

- どうやって類似度を更新するか？

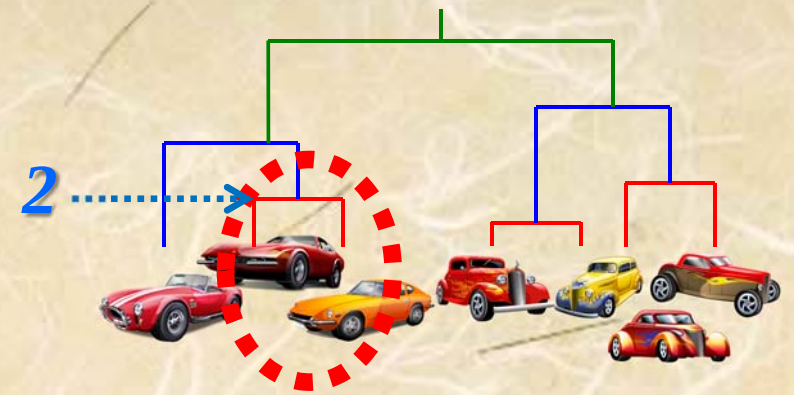


	x_1	x_2	3	1	2	3:4	$\frac{1+1}{1+1+1}16 + \frac{1+1}{1+1+1}17 - \frac{1}{1+1+1}1$		
			1	2	3	5:5	$\frac{1+1}{1+1+1}13 + \frac{1+1}{1+1+1}18 - \frac{1}{1+1+1}1$		
	3	1		5	5	21.7	25	13	
	1	2			2	20.3	34	26	
	2	3				8.3	20	16	
	3:4	5:5	$\frac{1+1}{1+1+1}5 + \frac{1+1}{1+1+1}8 - \frac{1}{1+1+1}1$			1	8.3	13.7	
	6	5	$\frac{1+1}{1+1+1}$	$\frac{1+1}{1+1+1}$	$\frac{1+1}{1+1+1}$			4	
	6	3	$\frac{1+1}{1+1+1}9 + \frac{1+1}{1+1+1}4 - \frac{1}{1+1+1}1$						

$$\frac{1+1}{1+1+1}13 + \frac{1+1}{1+1+1}8 - \frac{1}{1+1+1}1$$

3. クラスト化の方法

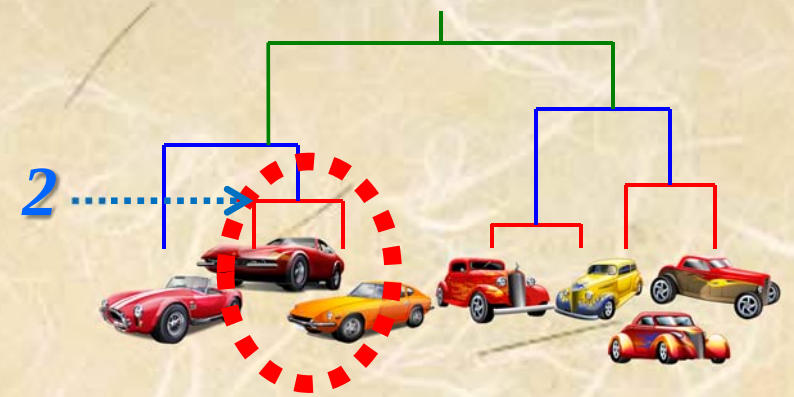
- どうやって類似度を**更新**するか？



						 		
	x_1	x_2						
				5	5	21.7	25	13
					2	20.3	34	26
						8.3	20	16
 							8.3	13.7
								4
								

3. クラスタ化の方法

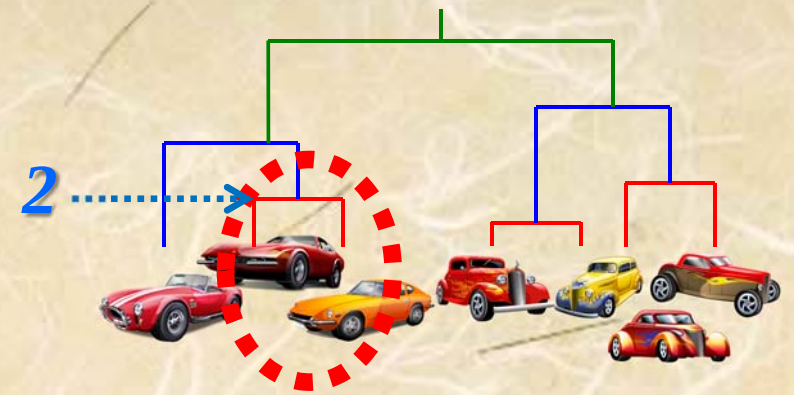
- どうやって類似度を**更新**するか？



							
	x_1	x_2					
				5:5	21.7	25	13
				2	20.3:8.3	34:20	26:16
						8.3	13.7
							4
							

3. クラスト化の方法

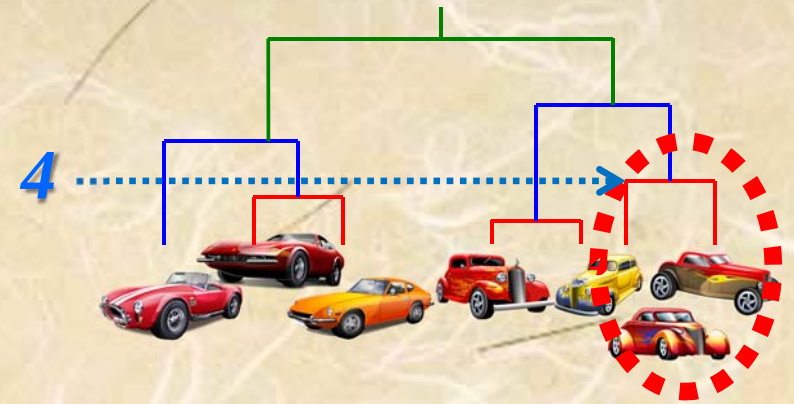
- どうやって類似度を**更新**するか？



					 		
				$\frac{1+1}{1+1+1}$	$5 + \frac{1+1}{1+1+1}$	$5 - \frac{1}{1+1+1}$	2
	x_1	x_2				$\frac{1+2}{1+1+2}$	$20.3 + \frac{1+2}{1+1+2}$
				6	21.7	25	13
 				2	20.5	35.3	27.3
 						8.3	13.7
			$\frac{1+1}{1+1+1}$	$34 + \frac{1+1}{1+1+1}$	$20 - \frac{1}{1+1+1}$		4
							

3. クラスタ化の方法

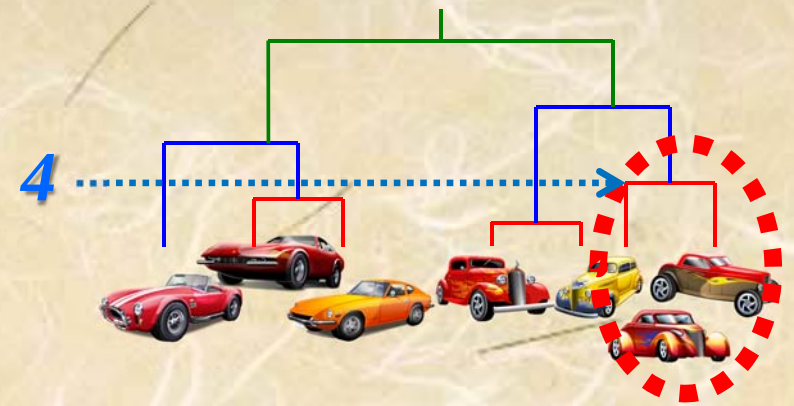
- どうやって類似度を**更新**するか？



							
	x_1	x_2					
				6	21.7	25	13
					20.5	35.3	27.3
						8.3	13.7
							4
							

3. クラスタ化の方法

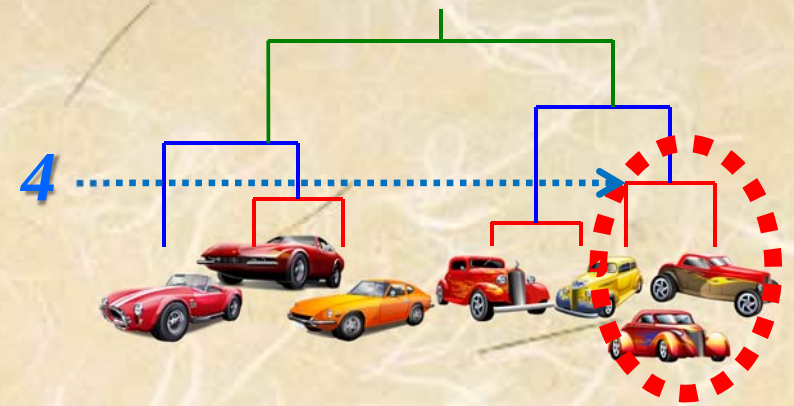
- どうやって類似度を**更新**するか？



						
	x_1	x_2				
				6	21.7	25:13
					20.5	35.3:27.3
						8.3:13.7
						4

3. クラスタ化の方法

- どうやって類似度を**更新**するか？



					 	 
			$\frac{1+1}{1+1+1}25 + \frac{1+1}{1+1+1}13 - \frac{1}{1+1+1}4$			
	x_1	x_2				
				6	21.7	24
 					20.5	45
 						14.5
 						4

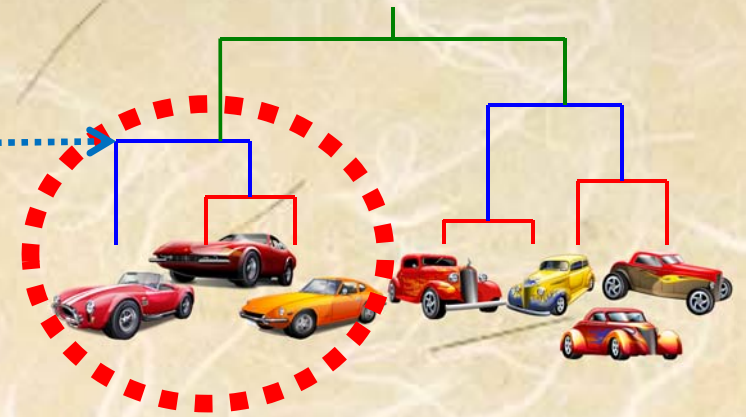
$$\frac{1+2}{1+1+2}35.3 + \frac{1+2}{1+1+2}27.3 - \frac{2}{1+1+2}4$$

$$\frac{1+2}{1+1+2}8.3 + \frac{1+2}{1+1+2}13.7 - \frac{2}{1+1+2}4$$

3. クラスタ化の方法

6

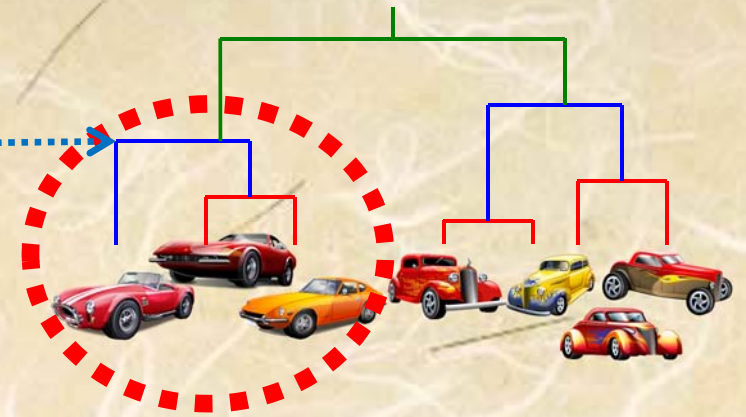
- どうやって類似度を**更新**するか？



						
	x_1	x_2				
				6	21.7	24
					20.5	45
						14.5
						

3. クラスタ化の方法 6

- どうやって類似度を**更新**するか？

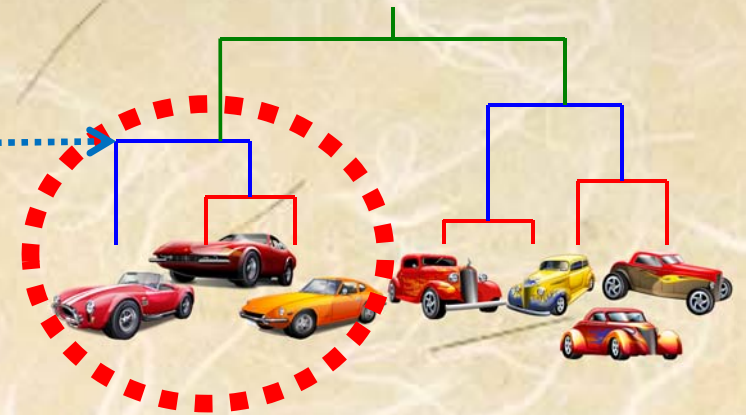


			 	 	 
	x_1	x_2			
  			6	21.7	24
				20.5	45
 					14.5
 					

3. クラスタ化の方法

6

- どうやって類似度を**更新**するか？



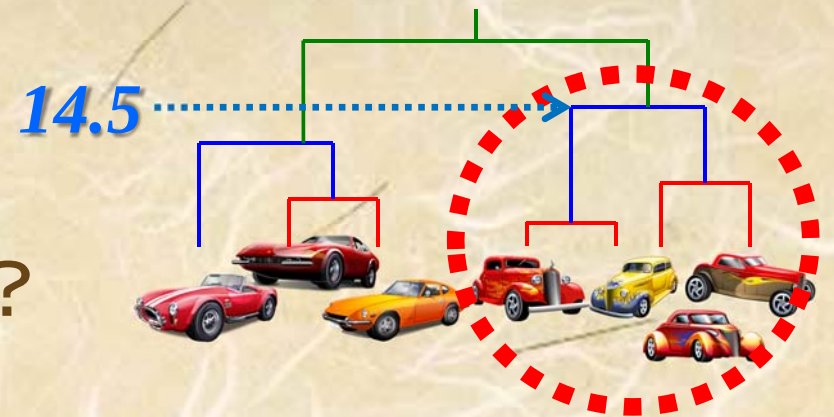
					
	x_1	x_2			
					
			6	27	48
					14.5
					

$$\frac{1+2}{1+2+2} 21.7 + \frac{2+2}{1+2+2} 20.5 - \frac{2}{1+2+2} 6$$

$$\frac{1+2}{1+2+2} 24 + \frac{2+2}{1+2+2} 45 - \frac{2}{1+2+2} 6$$

3. クラスタ化の方法

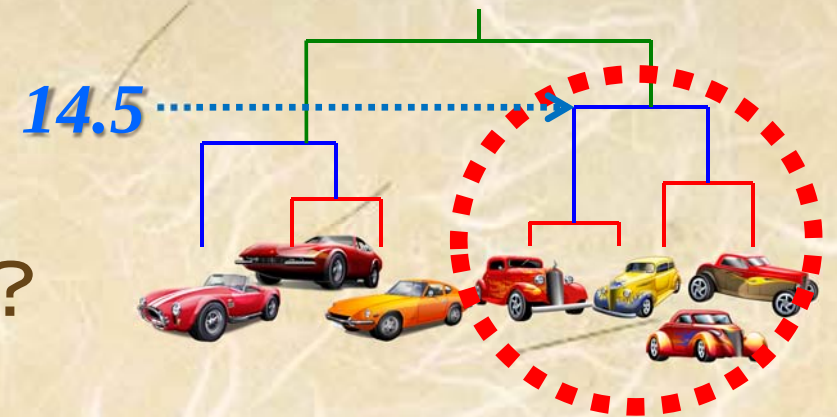
- どうやって類似度を**更新**するか？

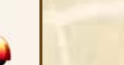


			 	 	 
	x_1	x_2			
  				27	48
 					14.5
 					

3. クラスタ化の方法

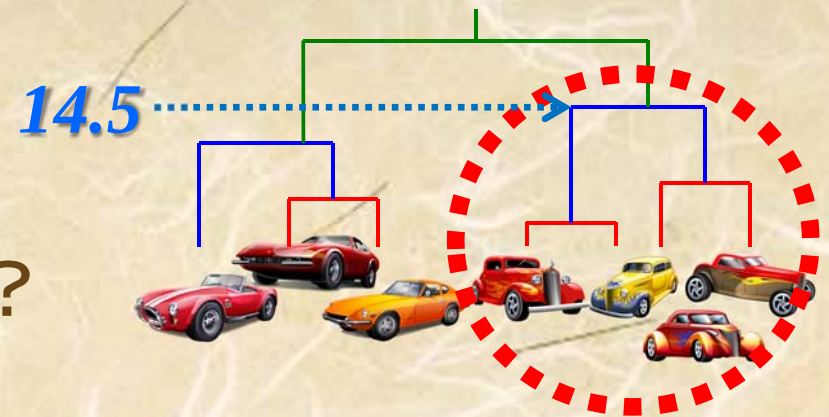
- どうやって類似度を**更新**するか？



			  	    
	x_1	x_2		
  				27 48
   				14.5

3. クラスタ化の方法

- どうやって類似度を**更新**するか？

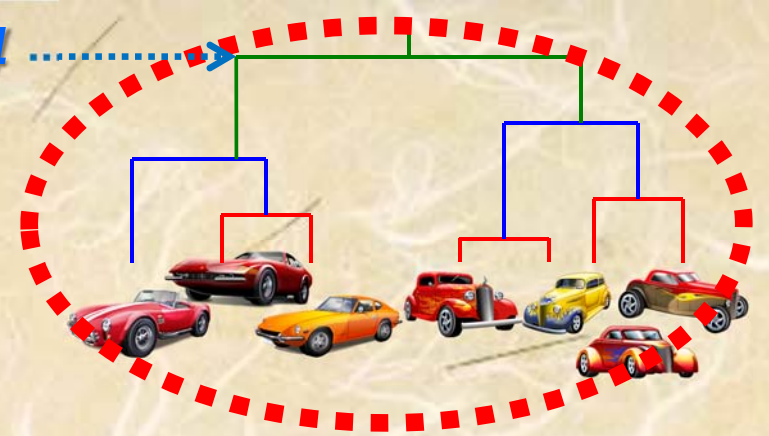


				
	x_1	x_2		
				47.4
 				
   				14.5

$$\frac{2+3}{2+2+3}27 + \frac{2+3}{2+2+3}48 - \frac{3}{2+2+3}14.5$$

3. クラスタ化の方法

- どうやって類似度を**更新**するか？



			 	   
	x_1	x_2		
  				
   				

4. クラスタ分析の実施

- Excelを用いて計算するクラスタ分析: 例2

- 対象: 5人の学生
- 対象の属性: 7つ

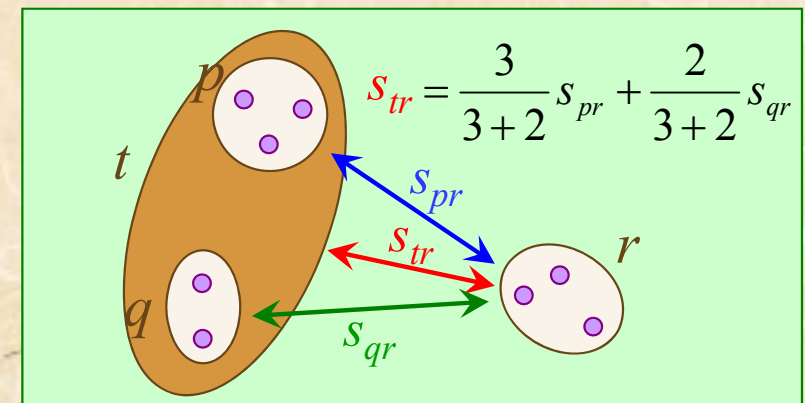
	属性1	属性2	属性3	属性4	属性5	属性6	属性7
太郎	13	12	7	1	13	13	12
次郎	6	5	8	4	9	5	15
三郎	13	14	5	15	2	19	17
四郎	13	5	8	7	9	3	13
五郎	1	18	6	1	3	1	20

- 距離: ユークリッド平方距離

$$l_2(Taro, Jiro)^2 = (13 - 6)^2 + (12 - 5)^2 + \dots + (12 - 15)^2$$

- クラスタ間の類似度更新方法: 群平均法

$$s_{tr} = \frac{n_p}{n_p + n_q} s_{pr} + \frac{n_q}{n_p + n_q} s_{qr}$$



4. クラスタ分析の実施

● Excelで計算によるクラスタ分析: 例2

	属性1	属性2	属性3	属性4	属性5	属性6	属性7
太郎	13	12	7	1	13	13	12
次郎	6	5	8	4	9	5	15
三郎	13	14	5	15	2	19	17
四郎	13	5	8	7	9	3	13
五郎	1	18	6	1	3	1	20



類似度の測定: ユークリッド平方距離による

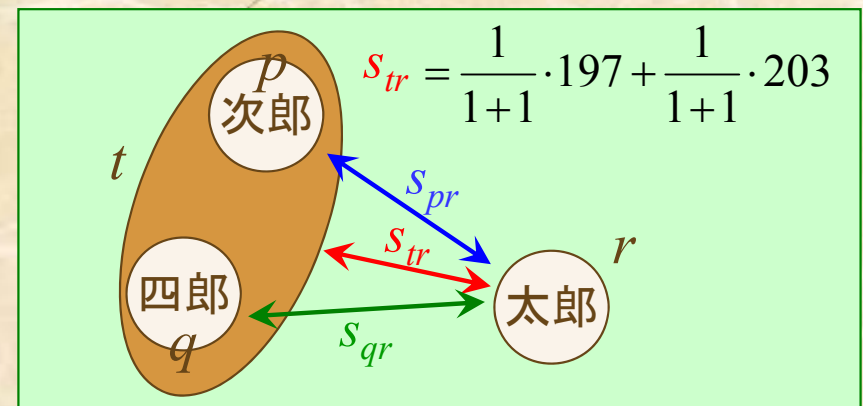
	太郎	次郎	三郎	四郎
次郎	197			
三郎	386	509		
四郎	203	66	475	
五郎	489	234	691	442

$$l_2(Taro, Jiro)^2 = 197 = (13 - 6)^2 + \dots + (12 - 15)^2$$



類似度の更新: 群平均法による

	太郎	次&四	三郎
次&四	200		
三郎	386	492	
五郎	489	363	691



$$s_{tr} = \frac{n_p}{n_p + n_q} s_{pr} + \frac{n_q}{n_p + n_q} s_{qr}$$

4. クラスタ分析の実施

● Excelで計算によるクラスタ分析: 例2

	太郎	次&四	三郎
次&四	200		
三郎	386	492	
五郎	439	363	691

↓ 類似度の更新: 群平均法による

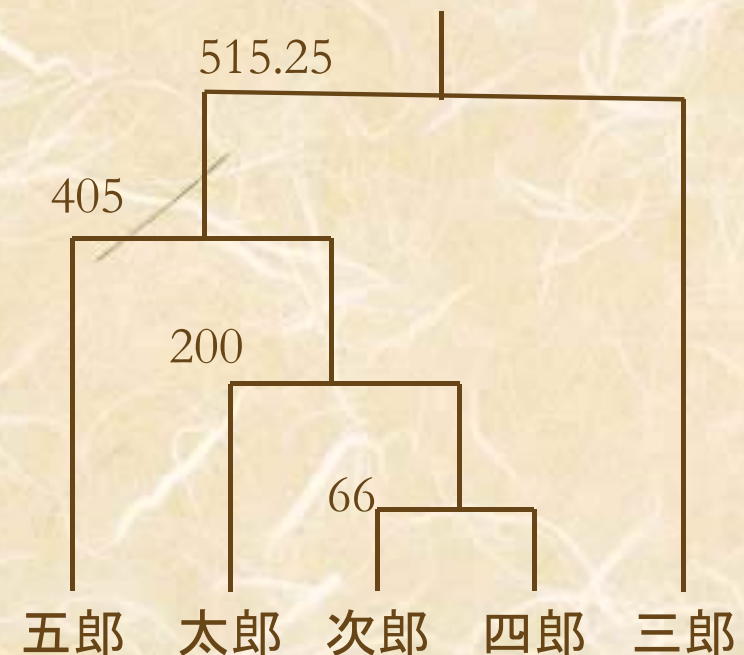
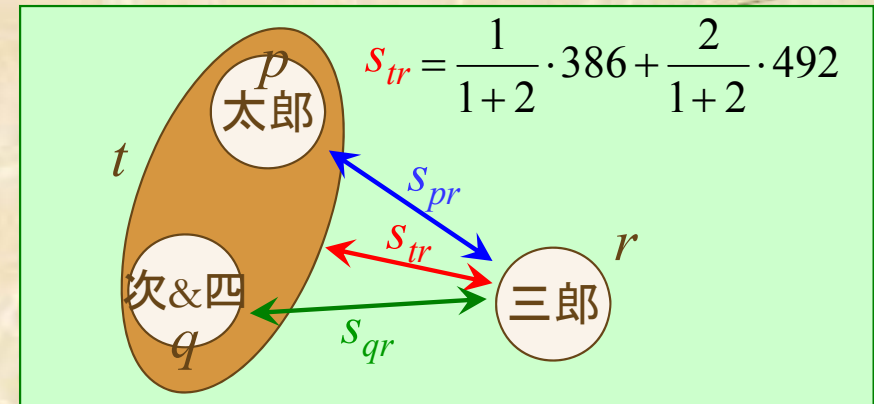
	太&(次&四)	三郎
三郎	456.67	
五郎	405	691

	太&(次&四)	三郎
三郎	456.67	
五郎	405	691

↓ 類似度の更新: 群平均法による

	五&(太&(次&四))
三郎	515.25

$$s_{tr} = \frac{n_p}{n_p + n_q} s_{pr} + \frac{n_q}{n_p + n_q} s_{qr}$$

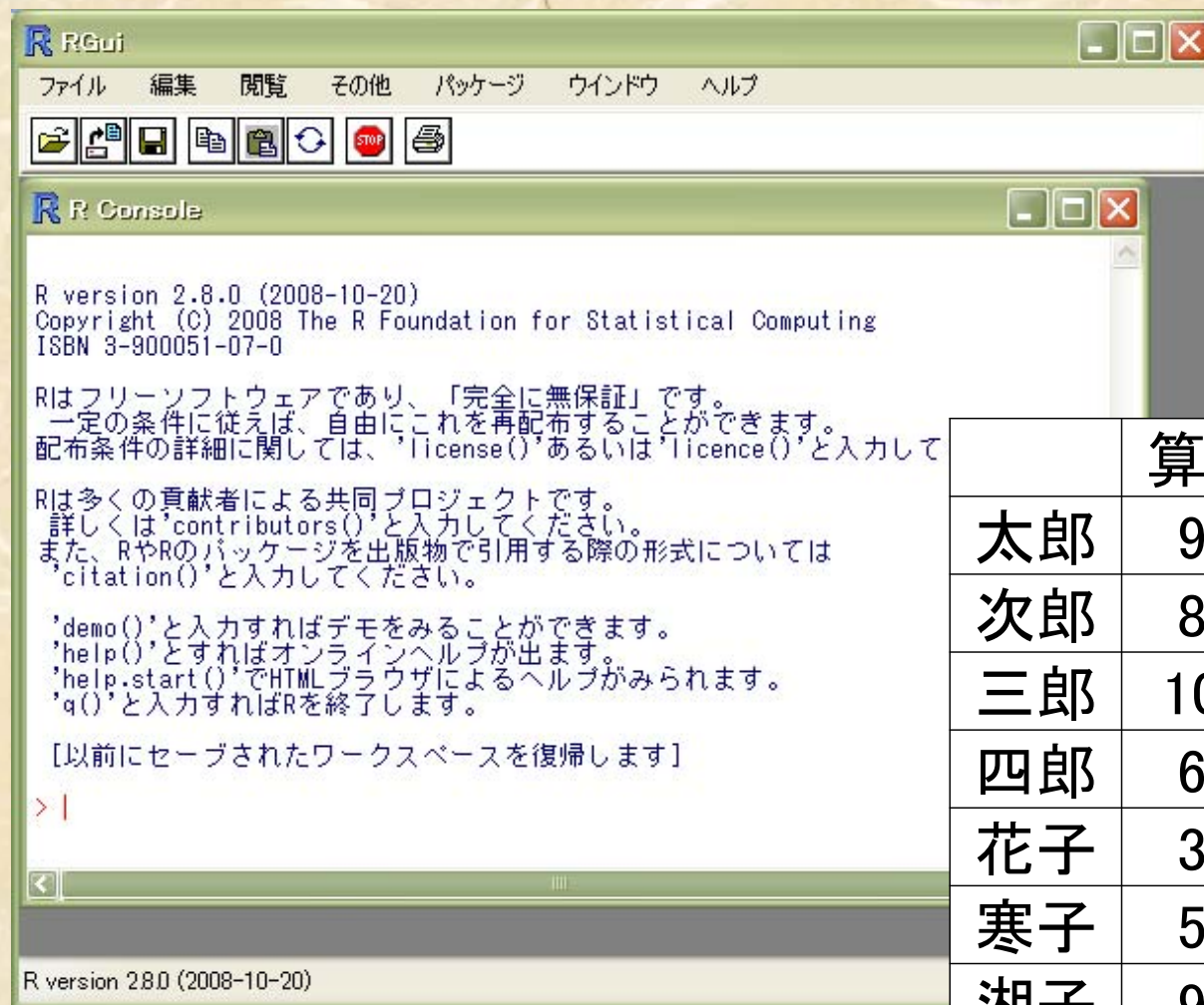


樹形図(デンドログラム)

4. クラスタ分析の実施

● Rによるクラスタ分析: 1. 起動画面とデータファイル

R起動時画面



データをcsvファイルで
用意
(Excelやeditorで作成)

ファイル「data-seiseki.csv」

	算数	理科	国語	英語	社会
太郎	90	100	70	90	30
次郎	80	60	70	70	20
三郎	100	40	30	70	80
四郎	60	30	40	80	80
花子	30	60	80	90	90
寒子	50	60	40	30	60
湘子	90	100	90	80	70

4. クラスタ分析の実施

● Rによるクラスタ分析: 2. クラスタ分析の実施例

```
> seiseki <- read.csv("C:/data-seiseki.csv", header=T, row.names=1)
```

← csvファイルを読み込み、
変数seisekiに格納

```
> seiseki
```

← 変数seisekiの中身確認

	算数	理科	国語	英語	社会
太郎	90	100	70	90	30
次郎	80	60	70	70	20
三郎	100	40	30	70	80
四郎	60	30	40	80	80
花子	30	60	80	90	90
寒子	50	60	40	30	60
湘子	90	100	90	80	70

```
> seiseki.d <- dist(seiseki, "manhattan")
```

← 対象間の類似度を
manhattan距離で測定し、
変数seiseki.dに格納

```
> seiseki.d
```

← 変数seiseki.dの中身確認

	太郎	次郎	三郎	四郎	花子	寒子
次郎	80					
三郎	180	140				
四郎	190	150	70			
花子	170	150	170	120		
寒子	200	140	140	110	150	
湘子	70	130	150	160	140	190

```
> (seiseki.hc <- hclust(seiseki.d, "ward"))
```

← ward法でクラスタ分析を
実施し、変数seiseki.hcに
格納

```
Call:
```

```
hclust(d = seiseki.d, method = "ward")
```

```
Cluster method : ward  
Distance       : manhattan  
Number of objects: 7
```

クラスタ化: ward法

← 類似度: manhattan距離 を確認!

対象の数: 7

```
> plot(seiseki.hc, hang=-1)
```

← 結果を樹形図で表示

```
> |
```

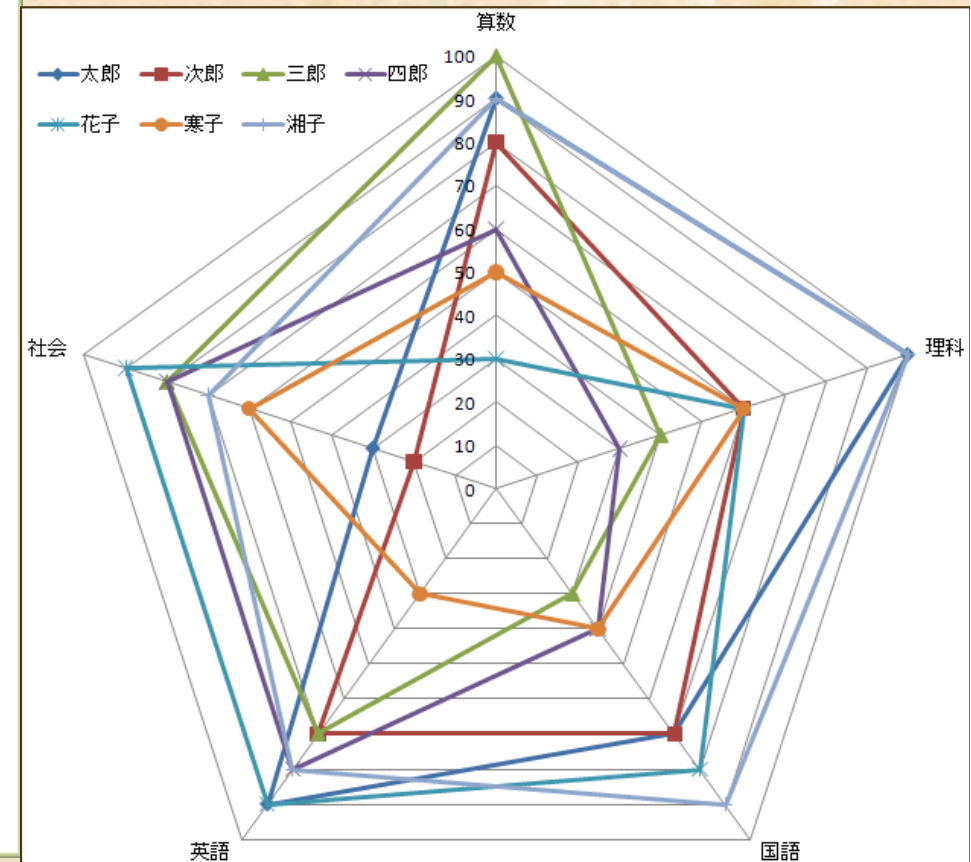
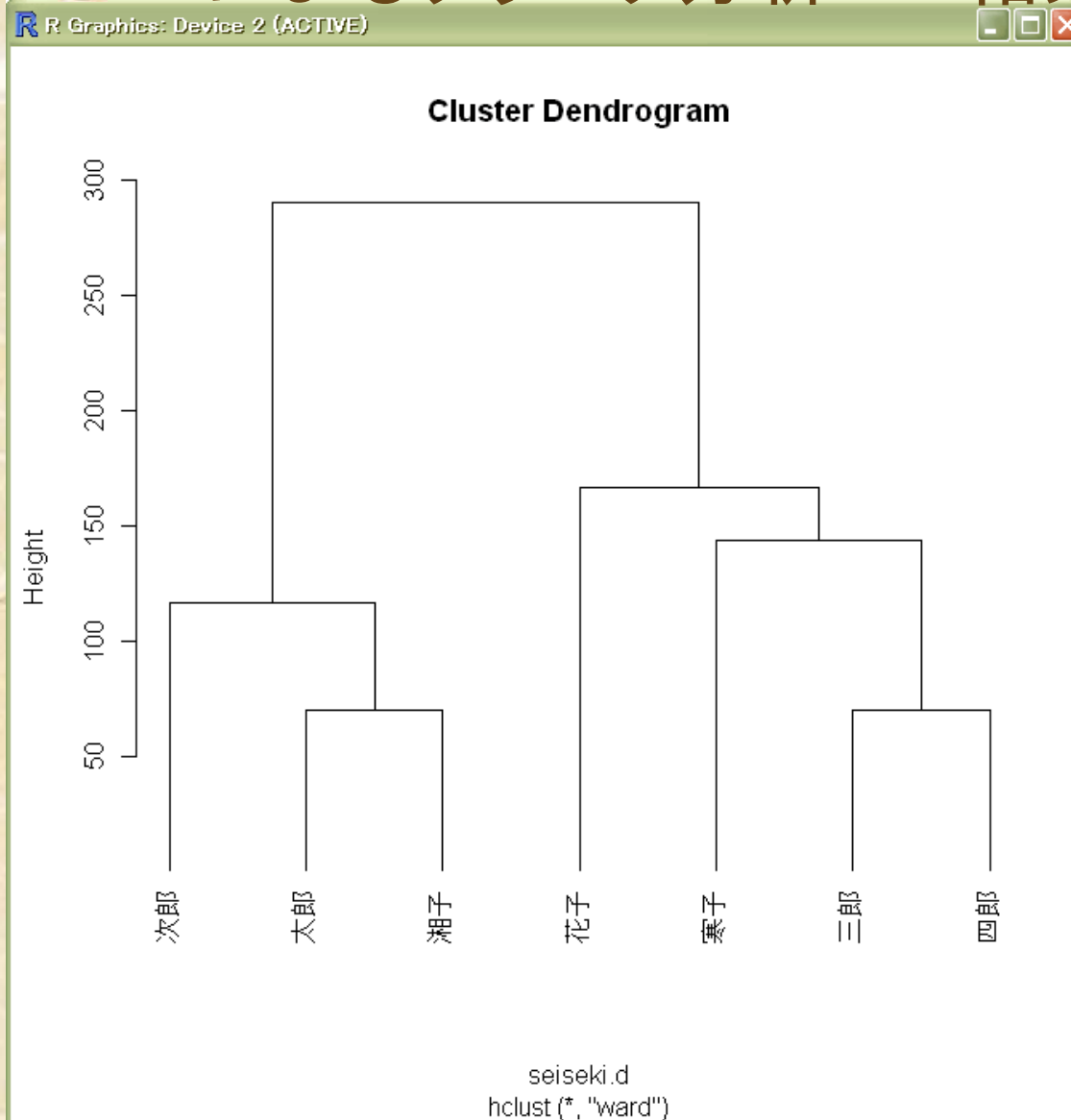
注) ward法を用いる場合、距離はユークリッド平方距離を使うのが妥当

4. クラスタ分析の実施

R によるクラスタ分析: 3. 結果

cf. 元データ

	算数	理科	国語	英語	社会
太郎	90	100	70	90	30
次郎	80	60	70	70	20
三郎	100	40	30	70	80
四郎	60	30	40	80	80
花子	30	60	80	90	90
寒子	50	60	40	30	60
湘子	90	100	90	80	70



4. クラスタ分析の実施

● Rによるクラスタ分析: 4.手法選択について

● 距離の測定: 関数dist() 【書式: dist(data, “method”)】

- methodの部分に距離の測定方法を指定

- euclidean ... ユークリッド距離 (l_2 ノルム) ex) dist(data) ← 指定無しだとこれ
- manhattan ... マンハッタン距離 (l_1 ノルム) ex) dist(data, “manhattan”)
- minkowski ... ミンコフスキー距離 (l_p ノルム) ex) dist(data, “minkowski”, p=4)
- maximum ... l_∞ ノルム ex) dist(data, “maximum”)

注) ユークリッド平方距離は、ユークリッド距離の計算後、2乗する

● クラスタ化の方法: 関数hclust() 【書式: hclust(data.d, “method”)】

- methodの部分にクラスタ化の方法を指定

- single ... 最短距離法 ex) hclust(data.d, “single”)
- complete ... 最長距離法 ex) hclust(data.d, “complete”)
- average ... 群平均法 ex) hclust(data.d, “average”)
- centroid ... 重心法 ex) hclust(data.d^2, “centroid”)
- median ... 中央値法 ex) hclust(data.d, “median”)
- ward ... ウォード法 ex) hclust(data.d^2, “ward”)

注) この2つの手法では「ユークリッド平方距離」を用いる
(data.dがユークリッド距離の計算結果でその2乗を使用)

4. クラスタ分析の実施

● R commanderによるクラスタ分析

名前	国語	英語	数学	理科	社会
太郎	57	47	62	73	72
次郎	77	38	58	71	49
三郎	53	43	42	50	52
四郎	39	54	56	61	56
茅ヶ子	57	56	73	59	58
寒子	55	80	69	84	53
藤子	77	44	65	69	79
塚子	59	55	67	56	68
鎌子	46	52	45	56	60

元データ[* .csv]

7% R コマンダー

ファイル 編集 データ 統計量 グラフ モデル 分布 ツール ヘルプ

データセット
新しいデータセット...
データセットのロード...
データセットの結合...
データのインポート
パッケージ内のデータ
アクティブデータセット
アクティブデータセット内の変数の管理

データセットの編集 データセットを表示 モデル: <アクティブモデルなし>

テキストファイルまたはクリップボード, URL から...
SPSSデータセットから...
Minitabデータセットから...
STATAデータセットから...
Excel または Access, dBase のデータセットから...

7% ファイルまたはクリップボード, URL からテキストデ...

データセット名を入力: Dataset

ファイル内に変数名あり: ☒

欠測値の記号: NA

データファイルの場所

ローカルファイルシステム ☒
クリップボード ☐
インターネットの URL ☐

フィールドの区切り記号

空白 ☐
カンマ ☒
タブ ☐
その他 ☐ 指定:

小数点の記号

ピリオド[.] ☒
カンマ[,] ☐

OK キャンセル ヘルプ

左上のような「*.csv」ファイルの
データ読込例

7% R コマンド

ファイル 編集 データ 統計量 グラフ モデル 分布 ツール ヘルプ

R データセット: Dataset データセットの編集 データセットを表示 モデル: <アクティブモデルなし>

スクリプトウィンドウ

```
Dataset <-  
  read.table("C:/Users/libitum/Documents/Dat/講義2012/春:問題発見技法/データセット/国語_英語_数学_理科_社会.csv",  
    header=TRUE, sep=",", na.strings="NA", dec=",",  
    library(relimp, pos=4)  
showData(Dataset, placement='-20+200', font=ge,  
  maxwidth=80, maxheight=30)
```

出カウィンドウ

> Dataset <-

実行

Dataset

	名前	国語	英語	数学	理科	社会
1	太郎	57	47	62	73	72
2	次郎	77	38	58	71	49
3	三郎	53	43	42	50	52
4	四郎	39	54	56	61	56
5	茅ヶ子	57	56	73	59	58
6	寒子	55	80	69	84	53
7	藤子	77	44	65	69	79
8	塚子	59	55	67	56	68
9	鎌子	46	52	45	56	60

「アクティブデータセット」
の「ケース名の設定」法

7% R コマンド

ファイル 編集 データ 統計量 グラフ モデル 分布 ツール ヘルプ

R データセット: 新しいデータセット... データセットのロード... データセットの結合... データのインポート... パッケージ内のデータ... アクティブデータセット... アクティブデータセット内の変数の管理...

スクリプトウィンドウ

```
Dataset <-  
  read.table("C:/Users/libitum/Documents/Dat/講義2012/春:問題発見技法/データセット/国語_英語_数学_理科_社会.csv",  
    header=TRUE, sep=",", na.strings="NA", dec=",",  
    library(relimp, pos=4)  
showData(Dataset, placement='-20+200', font=ge,  
  maxwidth=80, maxheight=30)
```

出カウィンドウ

> Dataset <-
+ read.table("C:/Users/libitum/Documents/Dat/講義2012/春:問題発見技法/データセット/国語_英語_数学_理科_社会.csv",
+ header=TRUE, sep=",", na.strings="NA", dec=",", library(relimp, pos=4)
> library(relimp, pos=4)

データセットを表示

モデル: <アクティブモデルなし>

ケース名の設定...

アクティブデータセットの選択...

アクティブデータセットを新しくする

アクティブデータセットのヘルプ (可能なら)

アクティブデータセット内の変数

アクティブデータセットの部分集合を抽出...

アクティブデータセット内の変数の集計...

アクティブデータセットから行を削除...

アクティブデータセット内の変数を積み重ねて結合...

欠測値のあるケースを削除...

アクティブデータセットの保存...

アクティブデータセットのエクスポート...

7% Dataset

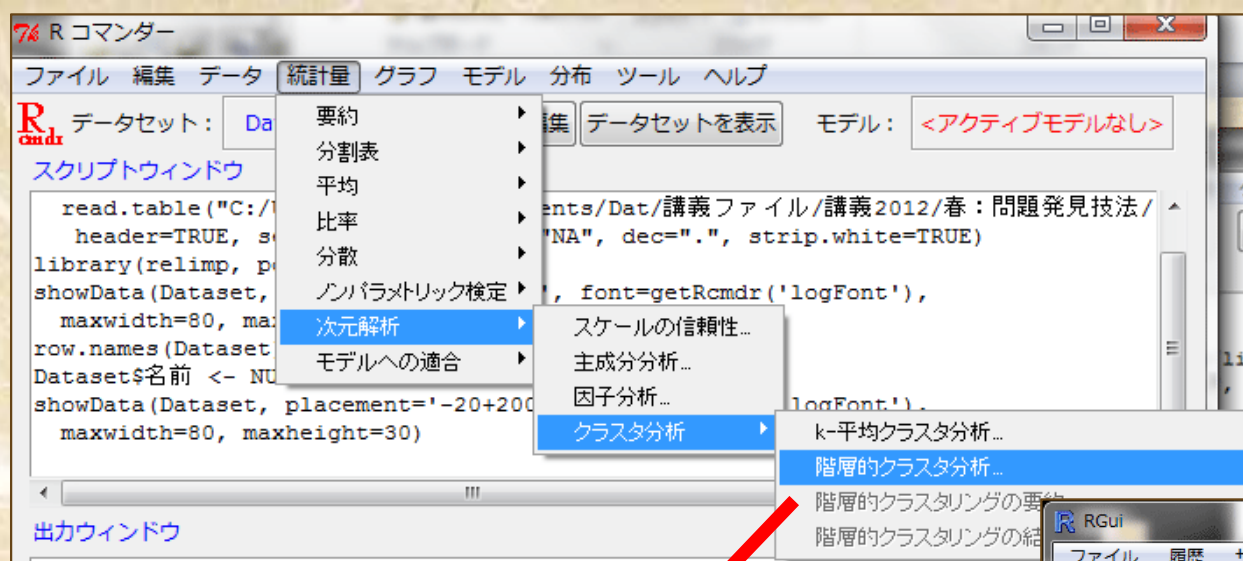
	国語	英語	数学	理科	社会
太郎	57	47	62	73	72
次郎	77	38	58	71	49
三郎	53	43	42	50	52
四郎	39	54	56	61	56
茅ヶ子	57	56	73	59	58
寒子	55	80	69	84	53
藤子	77	44	65	69	79
塚子	59	55	67	56	68
鎌子	46	52	45	56	60

7% ケースの名前を設定

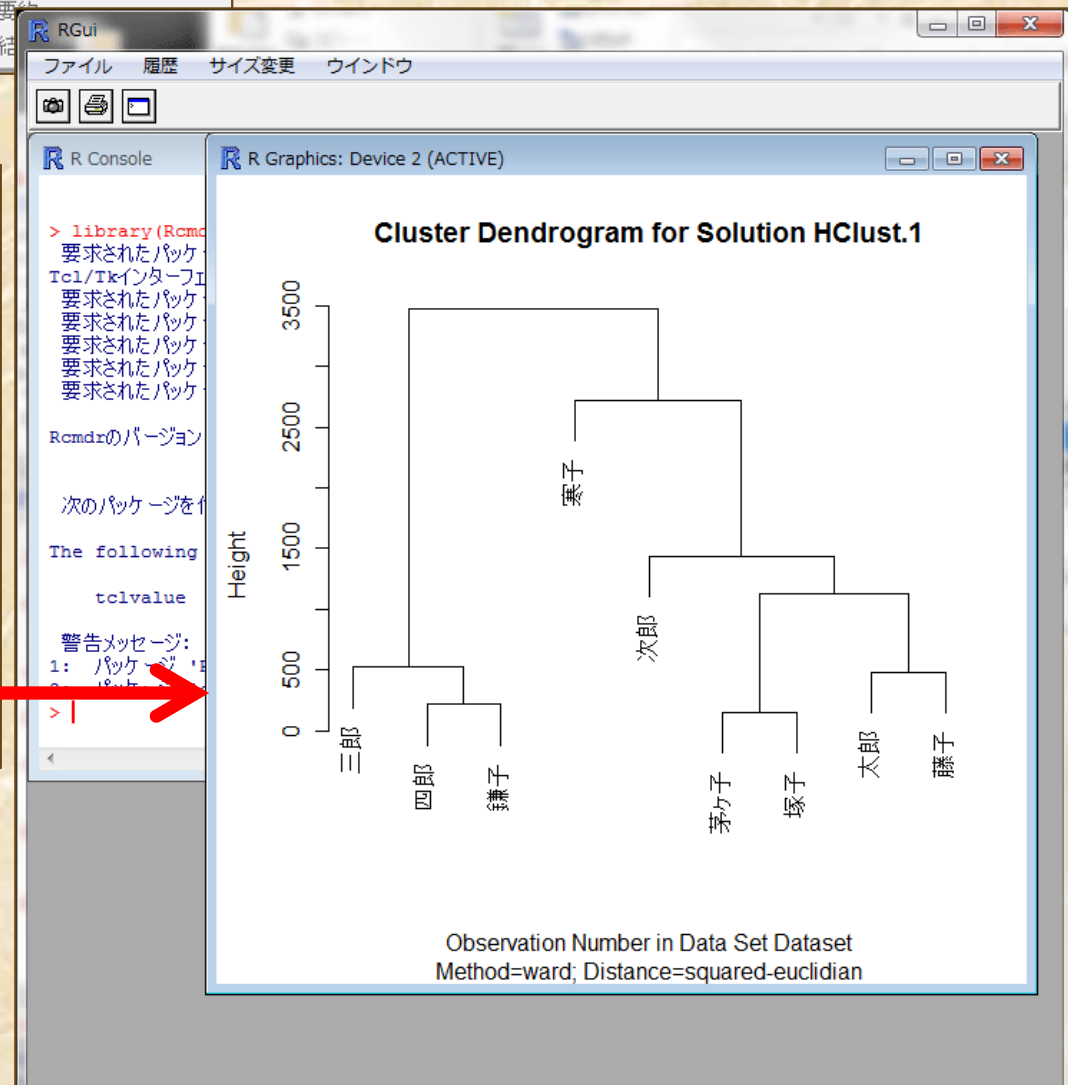
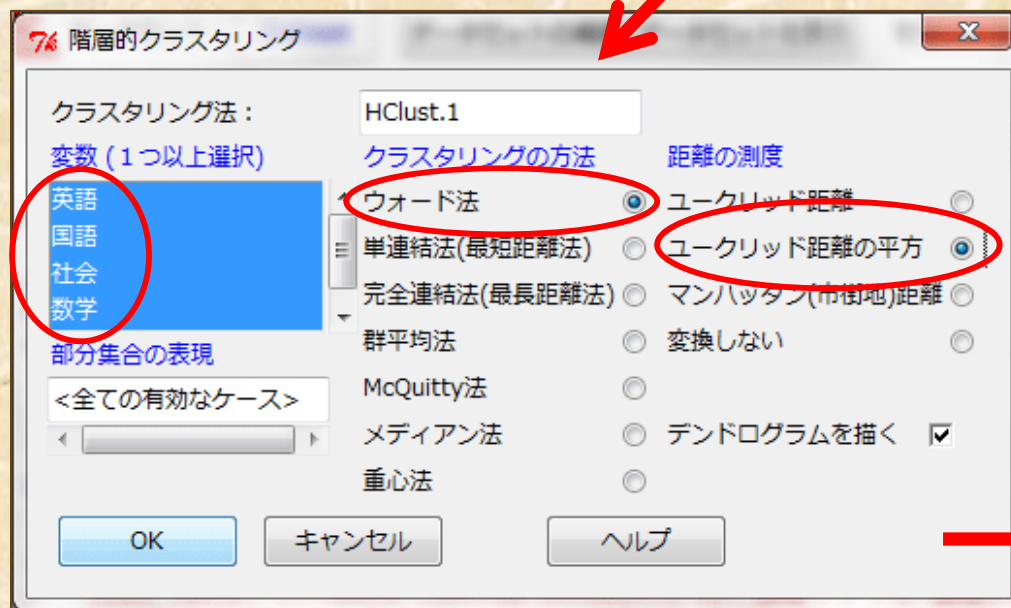
行名を含む変数を選択

社会
数学
名前
理科

OK キャンセル ヘルプ



階層的クラスター分析の実手順例



5. クラスター分析実施上の注意点

● クラスター分析の長所

- 探索的手法なので、データ構造を事前に知らなくてよい
- あらゆる種類のデータに適用可能: 数値・カテゴリー
- 適用が簡単

● クラスター分析の短所

- どんな属性値を選んだらいいのか？
- どの類似度(距離)測定法を選んだらいいのか？
- どのクラスタ化更新法を選んだらいいのか？
- データのスケールリング
- 結果の解釈が困難な可能性がある

迷ったらとりあえず
「ユークリッド平方距離」
で

迷ったらとりあえず
「ワード法」
で

6. 非階層的クラスタ分析

- K-means法

- 事前にクラスタ数をKとしてクラスタリングを行う



例：3つのクラスタ(K=3)に分類したい！としよう

6. 非階層的クラスタ分析

● K-means法

Step0: Kを決める

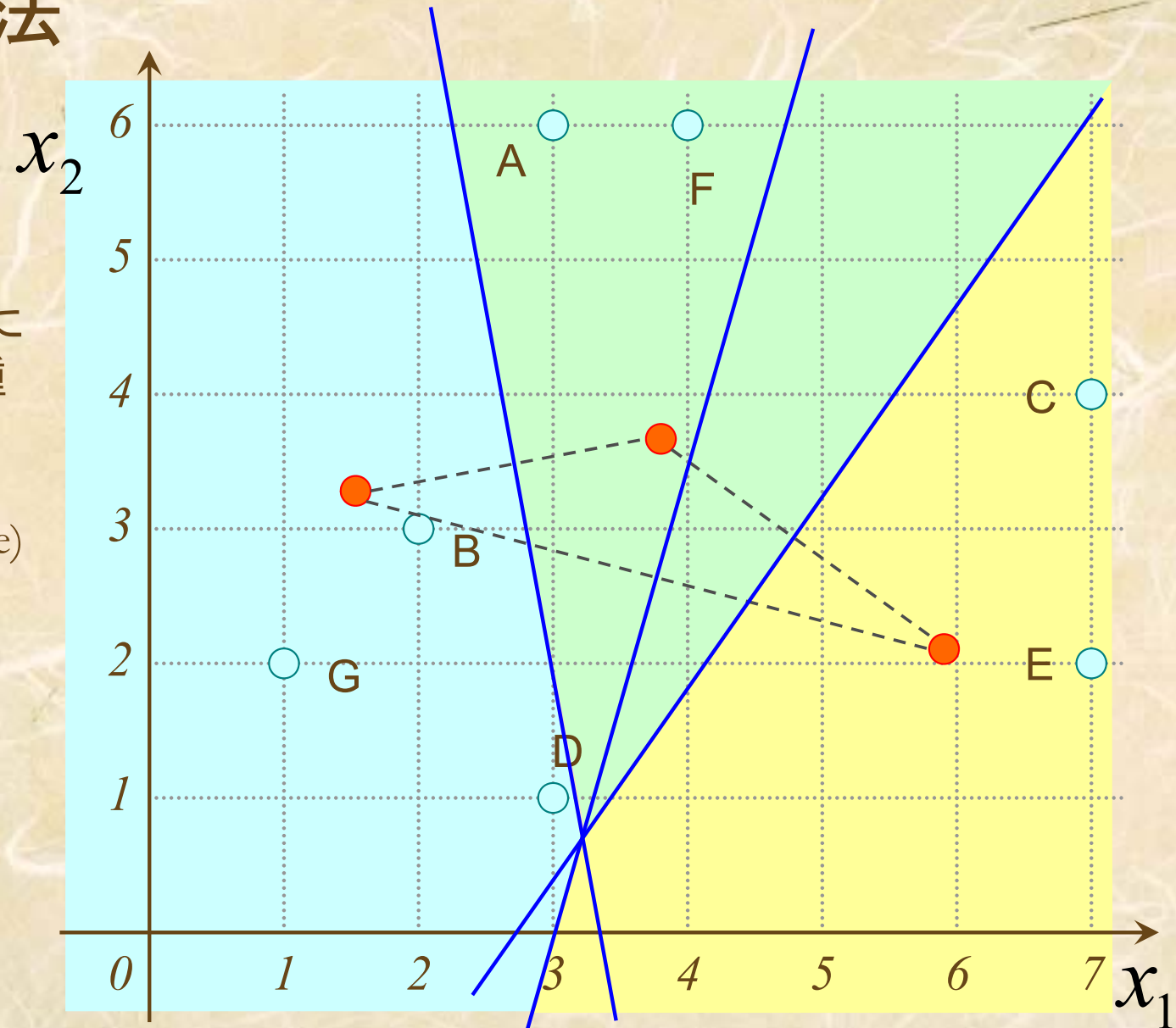
(ex. $K:=3$)

Step1: K個の種を置く

Step2: 何らかの距離により、もっとも近い種に含まれるよう境界線で分ける.

(ex. Euclidean distance)

(cf. Voronoi diagrams)



6. 非階層的クラスタ分析

● K-means法

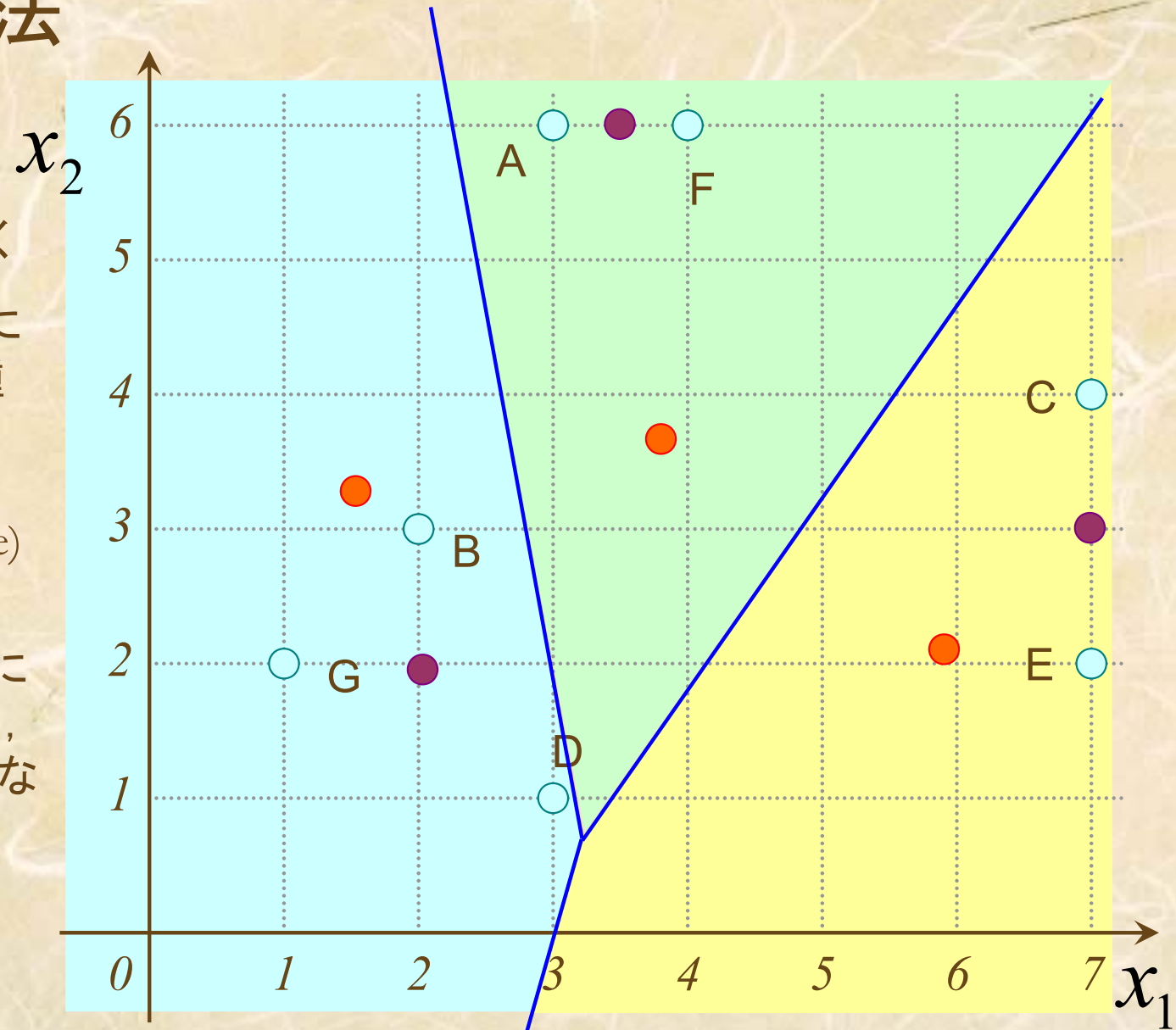
Step0: Kを決める
(ex. $K=3$)

Step1: K個の種を置く

Step2: 何らかの距離により、もっとも近い種に含まれるよう境界線で分ける.

(ex. Euclidean distance)
(cf. Voronoi diagrams)

Step3: 各クラスごとに何らかの距離により、重心を計算し、新たな種とする.



6. 非階層的クラスタ分析

● K-means法

Step0: Kを決める
(ex. $K=3$)

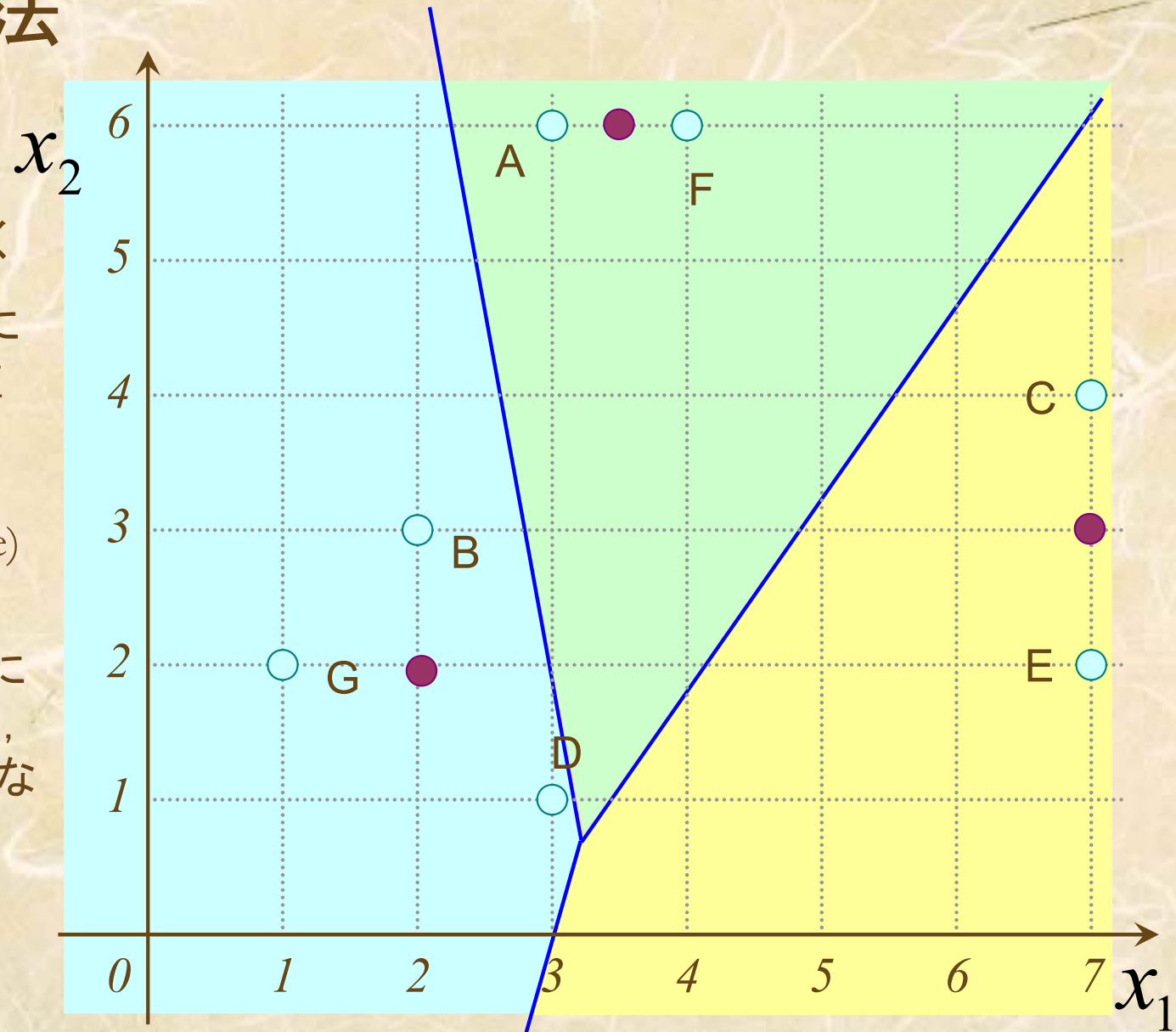
Step1: K個の種を置く

Step2: 何らかの距離により、もっとも近い種に含まれるよう境界線で分ける.

(ex. Euclidean distance)
(cf. Voronoi diagrams)

Step3: 各クラスごとに何らかの距離により、重心を計算し、新たな種とする.

Step2-4 をクラスが更新されなくなるまで繰り返す



6. 非階層的クラスタ分析

● K-means法

Step0: Kを決める

(ex. $K=3$)

Step1: K個の種を置く

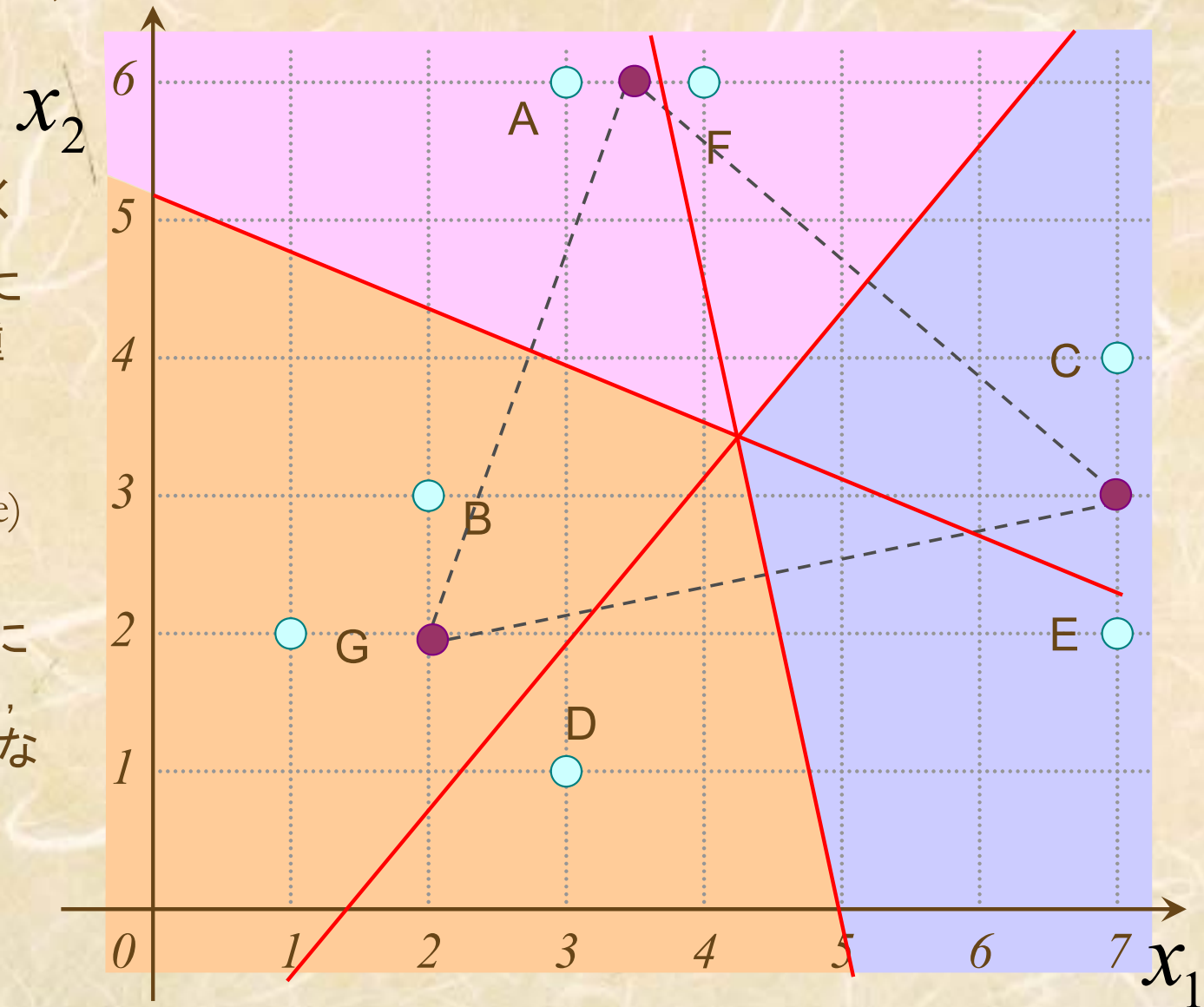
Step2: 何らかの距離により、もっとも近い種に含まれるよう境界線で分ける.

(ex. Euclidean distance)

(cf. Voronoi diagrams)

Step3: 各クラスごとに何らかの距離により、重心を計算し、新たな種とする.

Step2-4 をクラスタが更新されなくなるまで繰り返す



6. 非階層的クラスタ分析

K-means法

Step0: Kを決める
(ex. $K=3$)

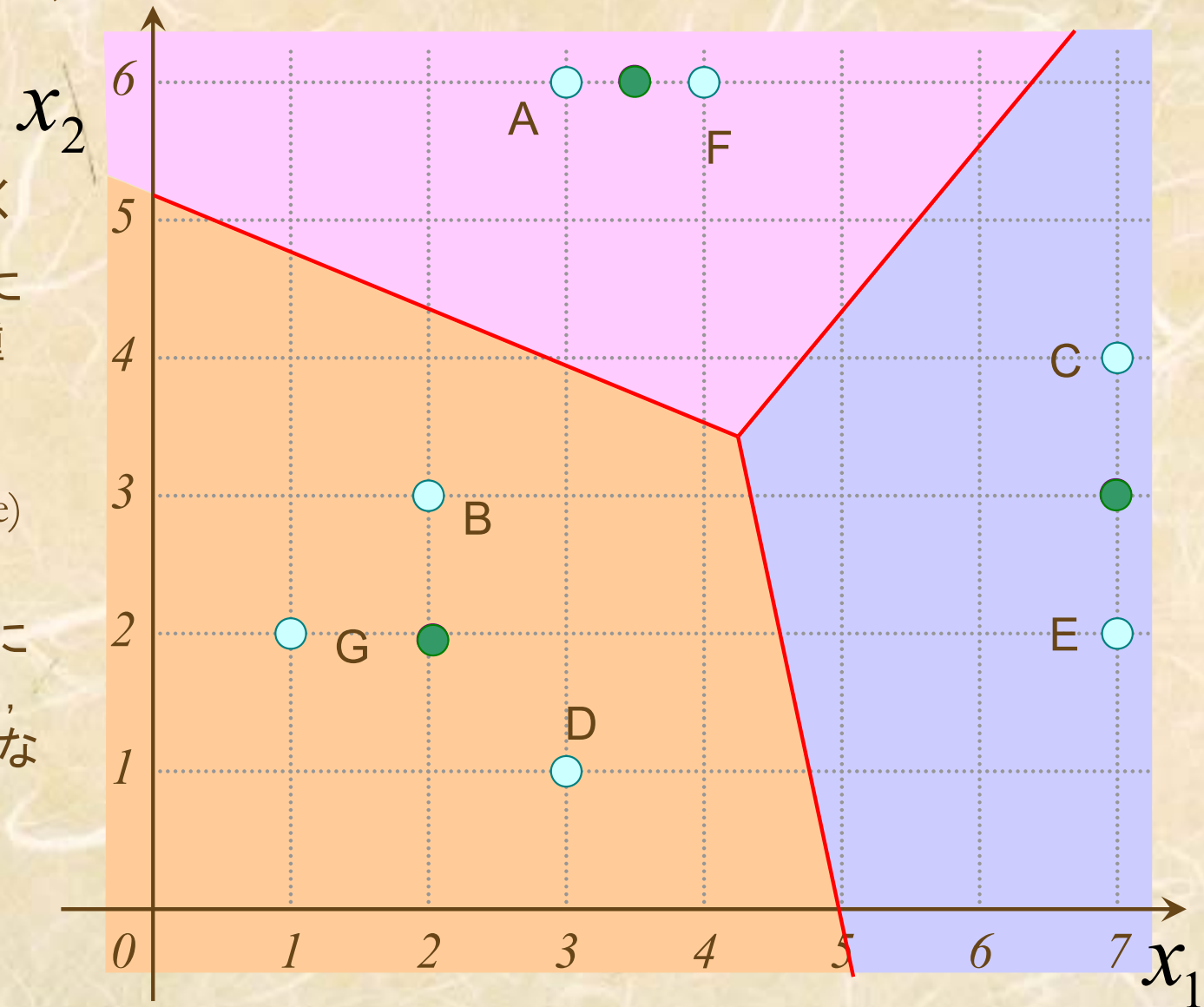
Step1: K個の種を置く

Step2: 何らかの距離により、もっとも近い種に含まれるよう境界線で分ける.

(ex. Euclidean distance)
(cf. Voronoi diagrams)

Step3: 各クラスごとに何らかの距離により、重心を計算し、新たな種とする.

Step2-4 をクラスが更新されなくなるまで繰り返す



7. クラスタ分析の実施

● Rによるクラスタ分析: 4. K-means法による結果

```
> (seiseki.km <- kmeans(seiseki, 3))  
K-means clustering with 3 clusters of sizes 1, 3, 3  
  
Cluster means:  
      算数      理科      国語 英語      社会  
1 30.00000 60.00000 80.00000  90 90.00000  
2 70.00000 43.33333 36.66667  60 73.33333  
3 86.66667 86.66667 76.66667  80 40.00000  
  
Clustering vector:  
太郎 次郎 三郎 四郎 花子 寒子 湘子  
  3    3    2    2    1    2    3  
  
Within cluster sum of squares by cluster:  
[1]    0 3600 3000  
  
Available components:  
[1] "cluster" "centers" "withinss" "size"  
> |
```

K-means法でクラスタ数を3として分析を実施し、
変数seiseki.kmに格納

cf. 元データ

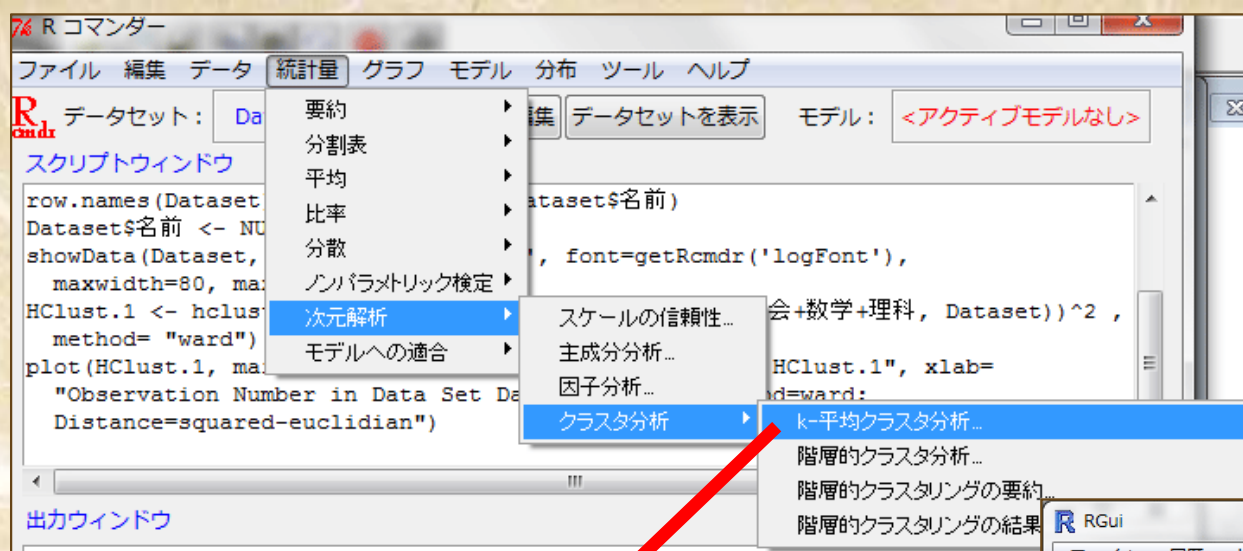
	算数	理科	国語	英語	社会
太郎	90	100	70	90	30
次郎	80	60	70	70	20
三郎	100	40	30	70	80
四郎	60	30	40	80	80
花子	30	60	80	90	90
寒子	50	60	40	30	60
湘子	90	100	90	80	70

結果:

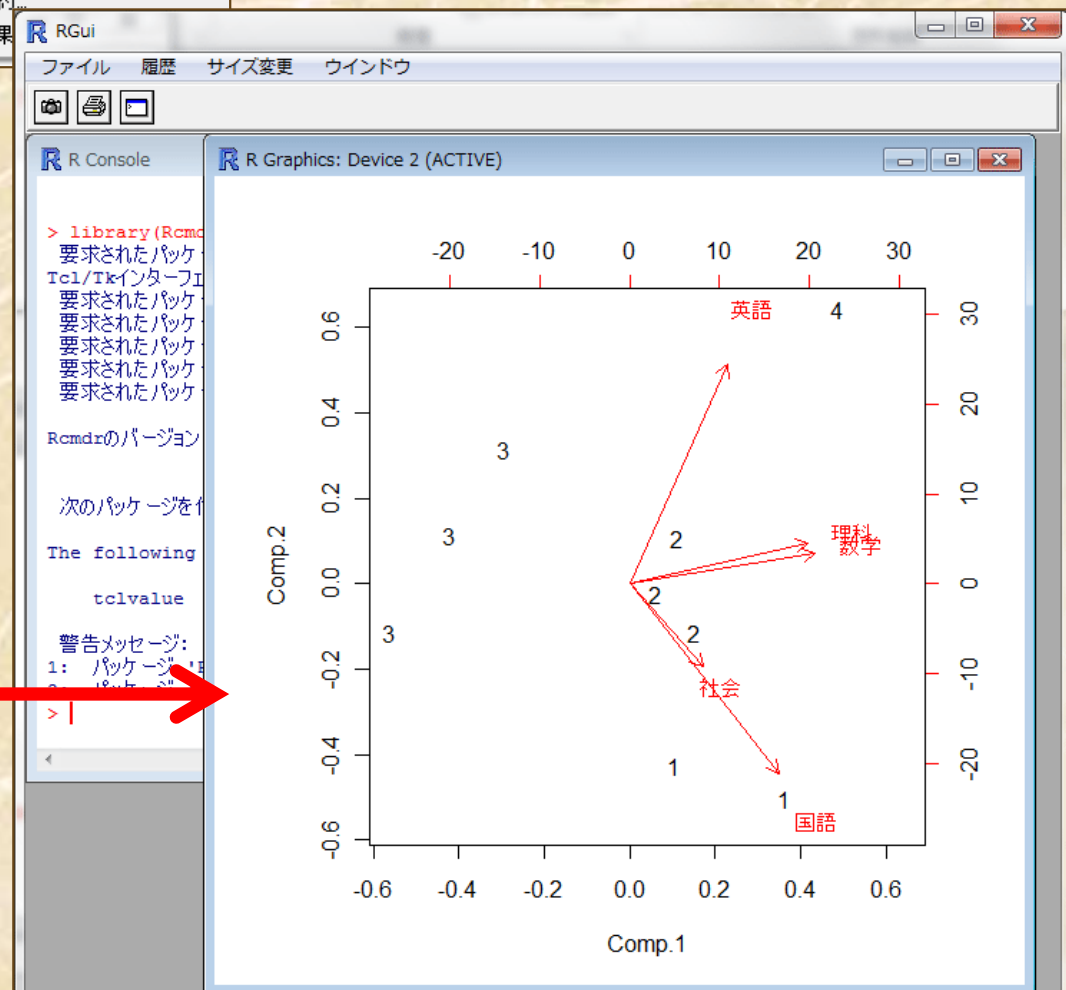
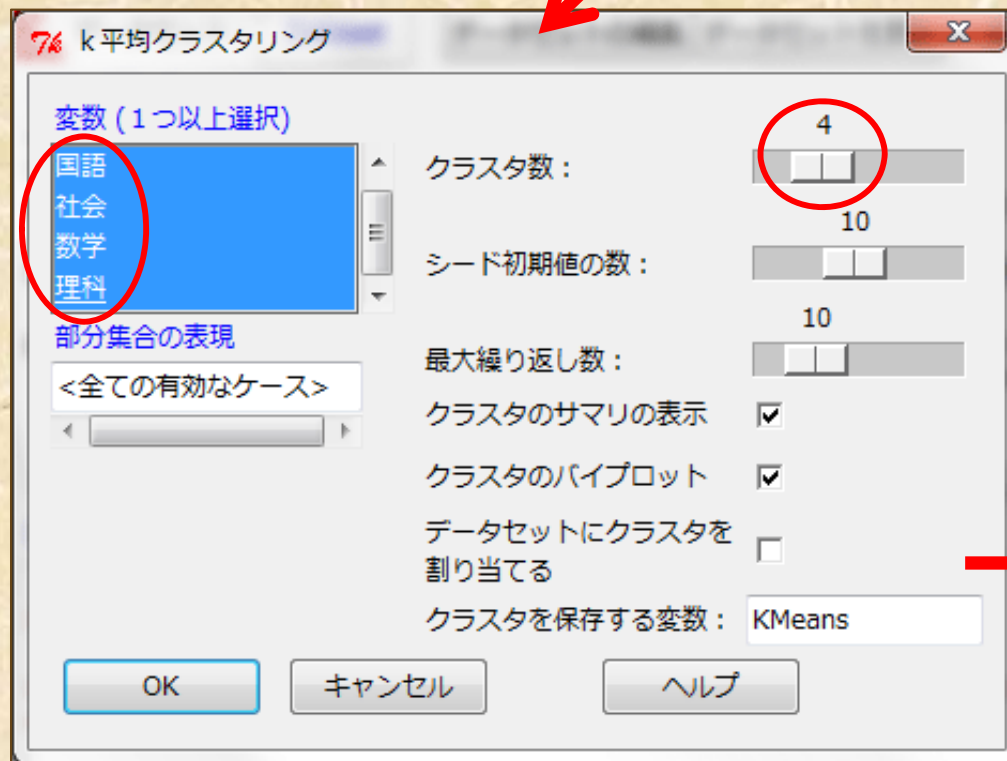
cluster1: 花子

cluster2: 三郎, 四郎, 寒子

cluster3: 太郎, 次郎, 湘子

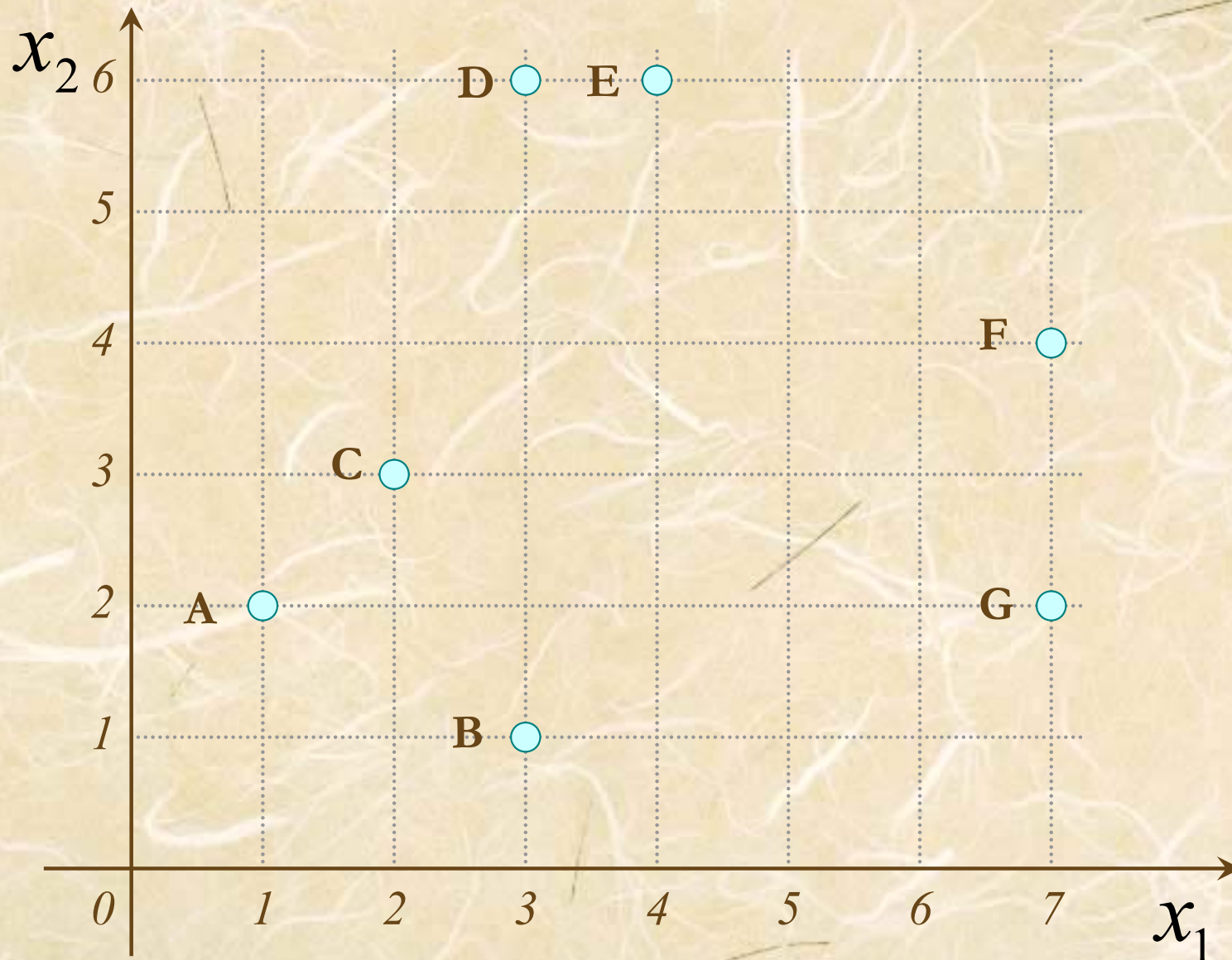


R commander による
非階層的クラスター分析
(K-means法)の実施手順
例



演習

- 類似度をマンハッタン距離で測定し, クラスタ間の類似度更新に**最短距離法**を用いてクラスタ分析をしよう!



参考文献

- 田中豊・脇本和昌『多変量統計解析法』現代数学社(1983)
- 河口至商『多変量解析入門Ⅱ』森北出版(1978,2005)
- 青木繁伸『Rによる統計解析』オーム社(2009)
- 荒木孝治『RとRコマンダーではじめる多変量解析』日科技連(2007)
- 金明哲『Rによるデータサイエンス』森北出版(2007)
- 新納浩幸『Rで学ぶクラスタ解析』オーム社(2007)