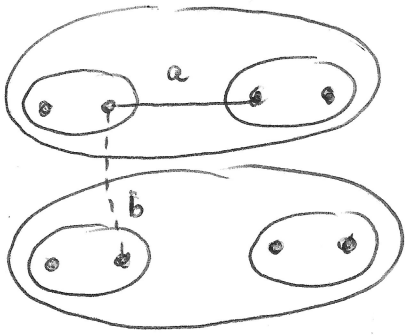
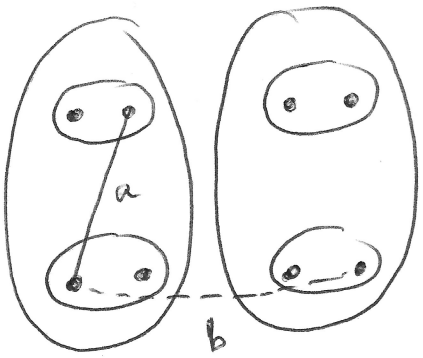
RESOLUÇÃO DE ALGUNS EXERCÍCIOS DE MMA

1.



$a < b$

método do vizinho mais próximo

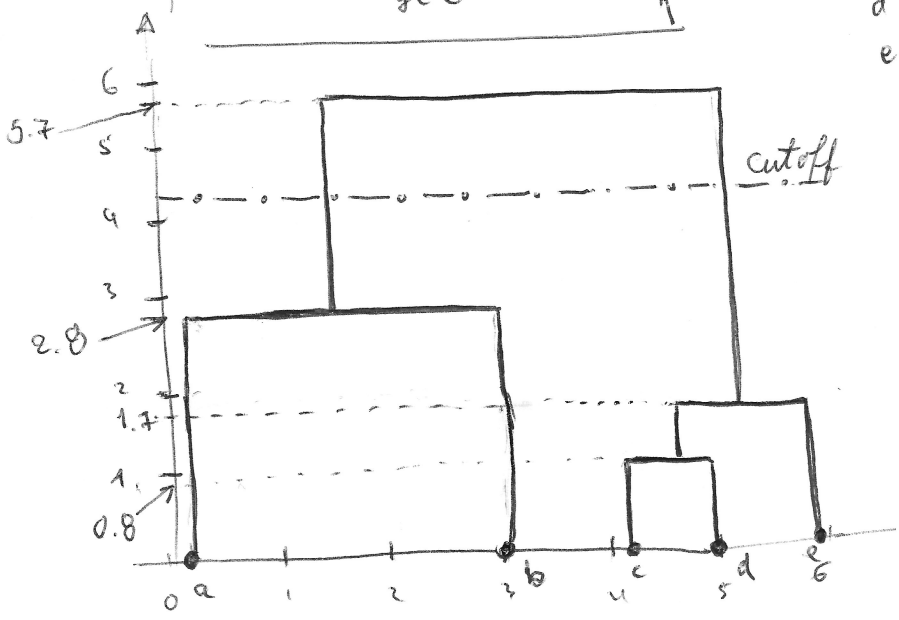


$a < b$

método do vizinho mais afastado

2. $X = \{a, b, c, d, e\} = \{0.2, 3, 4.2, 5, 5.9\}$

a) $d(C, C') = \max_{\substack{x \in C \\ y \in C'}} d(x, y)$



$d(a, \{c, d\}) = \max \{d(a,c), d(a,d)\}$

$= \max \{4, 4.8\} = 4.8$

etc...

	a	b	c	d
b	2.8	–	–	–
c	4	1.2		
d	4.8	2	0.8	
e	5.7	2.9	1.7	0.9

	a	b	{c,d}
b	2.8	–	–
{c,d}	4.8	2	–
e	5.7	2.9	1.7

	a	b
b	2.8	
{c,d,e}	5.7	2.9

	{a,b}
{c,d,e}	5.7

O dendrograma sugere 2 grupos: $\{a,b\}$, $\{c,d,e\}$

A presença do outlier **a** criou uma partição pouco "natural"

O método do vizinho mais afastado é sensível a observações atípicas!!

Exerc. repetir o procedimento com o método do vizinho mais próximo

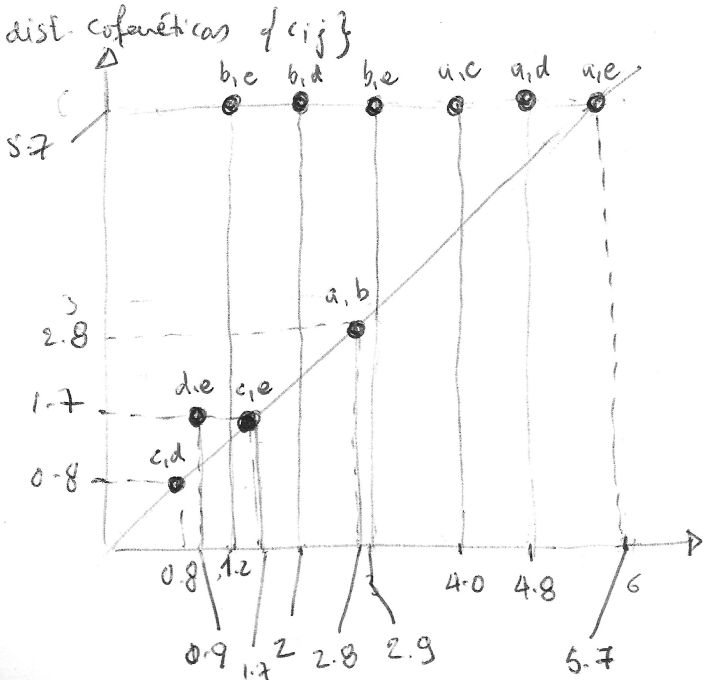
b)

	a	b	c	d
b	2.8	—	—	—
c	5.7	5.7	—	
d	5.7	5.7	0.8	
e	5.7	5.7	1.7	1.7

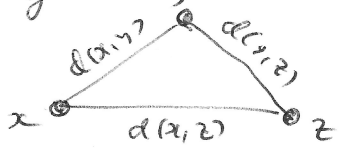
Nº PTS DIAGRAMA DE SHEPARD É DADO POR

$$\binom{N}{2} = \frac{N(N-1)}{2} = \frac{5\times 4}{2} = 10$$

onde $N = |X|$ = nº observações = 5



Obs as distâncias cofenéticas diferem uma verdadeira distância. Em particular verifica a desigualdade triangular



$$d(x,z) \leq d(x,y) + d(y,z)$$

De facto, neste caso definem uma ultra-métrica:

$$d(x,z) \leq \max\left(d(x,y), d(y,z)\right)$$

→ NO DIAGRAMA DE SHEPPARD TODAS AS DISTÂNCIAS COFENÉTICAS ≥ DIST. ORIGINAIS

⇓

DILATAÇÃO DO ESPAÇO DE ATRIBUTOS

c) $d_{ij} = (d_{12}, d_{13}, d_{14}, d_{15}, d_{23}, d_{24}, d_{25}, d_{34}, d_{35}, d_{45})$ (3)

$= (2.8, 4, 4.8, 5.7, 1.2, 2, 2.9, 0.8, 1.7, 0.9)$

$(c_{ij}) = (c_{12}, c_{13}, \dots c_{45}) =$

$= (2.8, 5.7, 5.7, 5.7, 5.7, 5.7, 5.7, 0.8, 1.7, 1.7)$

$$CPCC = cor(d_{ij}, c_{ij}) = \frac{\sum_{i<j} (d_{ij} - \bar{d})(c_{ij} - \bar{c})}{\sqrt{\sum_{i<j} (d_{ij} - \bar{d})^2 \cdot \sum_{i<j} (c_{ij} - \bar{c})^2}} = 0.6246$$

($\bar{d}, \bar{c}$ médias)

$CSCC = cor(rank(d_{ij}), rank(c_{ij})) = 0.7197016\dots$

CA

$0.8 < 0.9 < 1.2 < 1.7 < 2 < 2.8 < 2.9 < 4 < 4.8 < 5.7$

$1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8 \quad 9 \quad 10$

$(d_{ij}) = (2.8, 4, 4.8, 5.7, 1.2, 2, 2.9, 0.8, 1.7, 0.9)$

- $rank(d_{ij}) = (6, 8, 9, 10, 3, 5, 7, 1, 4, 2)$

$(c_{ij}) = (2.8, 5.7, 5.7, 5.7, 5.7, 5.7, 5.7, 0.8, 1.7, 1.7)$

$0.8 < 1.7 = 1.7 < 2.8 < 5.7 = 5.7 = 5.7 = 5.7 = 5.7 = 5.7$

$1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8 \quad 9 \quad 10$

2,5 médio

7,5 médio

- $rank(c_{ij}) = (4, 7.5, 7.5, 7.5, 7.5, 7.5, 7.5, 1, 2.5, 2.5)$

OBS se não existirem ranks iguais $\Rightarrow CSCC = 1 - \frac{6 \sum (d'_{ij} - c'_{ij})}{n^3 - n}$

onde $n = \binom{N}{2}$, $d'_{ij} = rank(d_{ij})$ e $c'_{ij} = rank(c_{ij})$

d) A partir da matriz das distâncias cofenéticas obtém-se o mesmo dendrograma que a partir das distâncias originais, logo a mesma classificação.

3. $X = \{ (1,2), (2,2), (4.5,3), (6,3) \} = \{ a, b, c, d \}$

- $d(\vec{x}, \vec{y}) = \sum_{i=1}^{n} |x_i - y_i|$ $\vec{x} = (x_1, \dots, x_n)$, $\vec{y} = (y_1, \dots, y_n)$

  ↑ Distância de Manhattan

por exemplo $d((1,2),(4.5,3)) = |1-4.5| + |2-3| = 4.5$

$d(a,c) = d_{13}$

- Distâncias médias entre grupos:

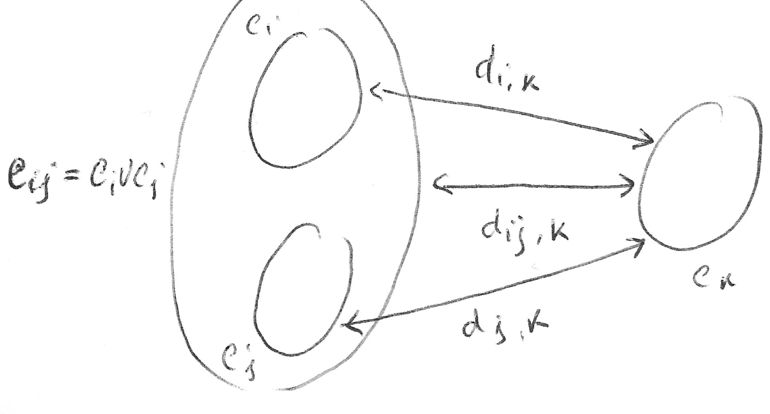
$d_{ik} = d(C_i, C_k)$ $n_i = |C_i|$: nº de elementos de C_i

$d_{jk} = d(C_j, C_k)$ $n_j = |C_j|$: nº el. de C_j

$d_{ij,k} = d(\underbrace{C_i \cup C_j}_{C_{ij}}, C_k)$ $n_{ij} = |C_i \cup C_j| = n_i + n_j$

$$d_{ij,k} = \alpha_i d_{ik} + \alpha_j d_{jk} = \frac{n_i}{n_i+n_j} d_{ik} + \frac{n_j}{n_i+n_j} d_{jk}$$

(ver a tabela de Lance-Williams para o método average)



$$d(c_i \cup c_j, c_k) = \frac{n_i}{n_i + n_j} d(c_i, c_k) + \frac{n_j}{n_i + n_j} d(c_j, c_k)$$

	a	b	c
b	1	–	–
c	4.5	3.5	–
d	6	5	1.5

- $d_{a \cup b, c}$ ($a \to i$, $b \to j$, $c \to k$) $= \frac{n_a}{n_a + n_b} d(a,c) + \frac{n_b}{n_a + n_b} d(b,c)$

$= \frac{1}{2} 4.5 + \frac{1}{2} 3.5 = 4.$

- $d_{a \cup b, d} = \frac{n_a}{n_a + n_b} d(a,d) + \frac{n_b}{n_a + n_b} d(b,d)$

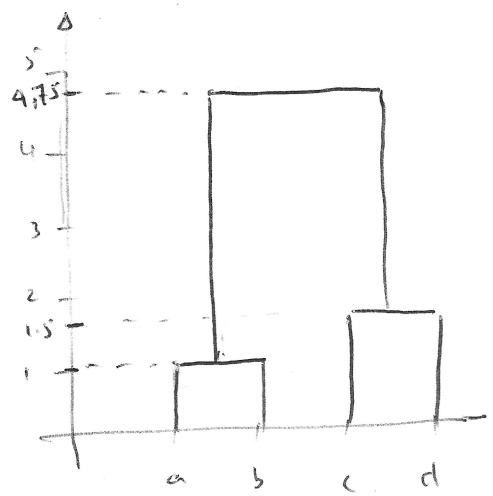
$= \frac{1}{2} 6 + \frac{1}{2} 5 = 5.5$

	{a,b}	c
c	4	
d	5.5	1.5

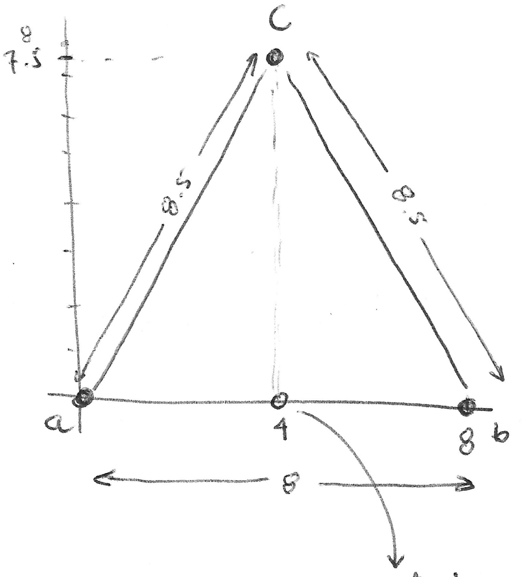
- $d_{c \cup d, \{a,b\}}$ ($c \to i$, $d \to j$, $\{a,b\} \to k$) $= \frac{n_c}{n_c + n_d} d_{c,\{a,b\}} + \frac{n_d}{n_c + n_d} d_{d,\{a,b\}}$

$= \frac{1}{2} 4 + \frac{1}{2} 5.5 = 4.75$

	{a,b}
{c,d}	4.75



4. $X = \{(0,0), (8,0), (4,7.5)\}$
   a b c



$d(a,b) = 8$

$d(a,c) = \sqrt{4^2 + 7.5^2} = 8.5 = d(b,c)$

$d(c, a \cup b) = d(x,y) = 7.5$

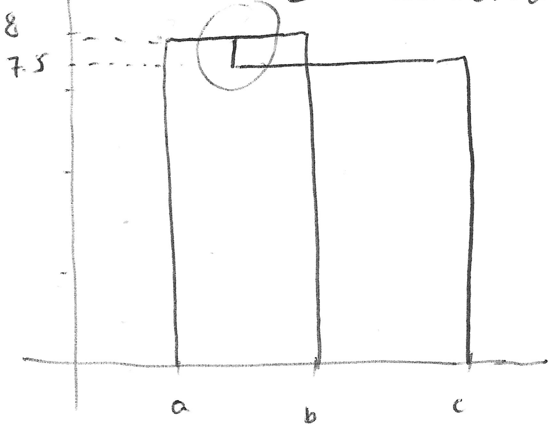
x = centroide de $\{c\}$ = $c = (4,7.5)$

y = centroide de $\{a,b\}$ = $(4,0)$

	a	b
b	8	–
c	8.5	8.5

⇓

	{a,b}
c	7.5



ocorreu uma inversão isto é os custos de fusão não crescem monótonamente

distâncias cofenéticas

	a	b
b	8	
c	7.5	7.5

OBS PROP ultramétrica
⇓
NÃO EXISTEM INVERSÕES

Tem-se:

$d(a,b) > d(a,c), d(b,c)$

logo $d(a,b) \geq \max(d(a,c), d(b,c))$

e logo não verifica a propriedade ultramétrica:

$d(x,z) \leq \max(d(x,y), d(y,z)) \quad \forall_{x,y,z}$

Vamos calcular $d(a \cup b, c)$ usando agora a fórmula de Lance-Williams. Temos por isso as distâncias ao quadrado como está indicado na tabela L-W:

$$d^2_{ij,k} = \frac{n_i}{n_i+n_j} d^2_{i,k} + \frac{n_j}{n_i+n_j} d^2_{j,k} - \frac{n_i n_j}{(n_i+n_j)^2} d^2_{i,j}$$

onde $d_{i,j} = d(c_i, c_j)$ $\quad n_k = |C_k| =$ nº elementos de C_k

$d_{i,k} = d(c_i, c_k)$ $\quad n_j = |C_j|$ ...

$d_{j,k} = d(c_j, c_k)$ $\quad n_i = |C_i|$ ...

d_{ij}	a	b
b	8	–
c	8.5	8.5

$\Longrightarrow$

d^2_{ij}	a	b
b	64	
c	72.25	72.25

$n_a = |data| = n_b = 1$

Assim,

$$d^2(a \cup b, c) = \frac{n_a}{n_a+n_b} d^2(a,c) + \frac{n_b}{n_a+n_b} d^2(b,c) - \frac{n_a n_b}{(n_a+n_b)^2} d^2(a,b) =$$

$$= \frac{1}{2} 72.25 + \frac{1}{2} 72.25 - \frac{1}{4} 64 = 56.25$$

$$\Longrightarrow d(a \cup b, c) = \sqrt{56.25} = 7.5 \;//$$

5. Temos que mostrar que após a aglomeração de 2 grupos $C_{ij} = C_i \cup C_j$

$$d_{ij,k} = d(C_{ij}, C_k) \geq d_{ij} = d(C_i, C_j)$$

para todo o grupo C_k distinto de C_i, C_j e C_{ij}

Por definição,

$$d_{ij,k} = \max_{\substack{x \in C_{ij} \\ z \in C_k}} d(x, y) =$$

$$= \max \left(\underbrace{\max_{\substack{x \in C_i \\ z \in C_k}} d(x, z)}_{d(C_i, C_k)}, \underbrace{\max_{\substack{y \in C_j \\ z \in C_k}} d(x, y)}_{d(C_j, C_k)} \right)$$

$$= \max \left(\underbrace{d(C_i, C_k)}_{d_{ik}}, \underbrace{d(C_j, C_k)}_{d_{jk}} \right).$$

Como C_i e C_j são aglomerados antes de

C_i e C_k e de C_j e C_k $\Rightarrow \begin{cases} d_{ij} \leq d_{ik} \\ d_{ij} \leq d_{jk} \end{cases}$

Logo $d_{ij,k} \geq d_{ij}$!

7. Tal como no exercício 5 Temos que mostrar que

$$\underbrace{d(c_i \cup c_j, c_k)}_{d_{ij,k}} \geq \underbrace{d(c_i \cup c_j)}_{d_{ij}}$$

para todo o c_k distinto de c_i, c_j e $c_i \cup c_j$

Ora pela fórmula de L-W para o método do "average"

$$d_{ij,k} = \frac{n_i}{n_i + n_j} d_{i,k} + \frac{n_j}{n_i + n_j} d_{j,k}$$

Como c_i e c_j foram agregados antes de

c_i e c_k e de c_j e c_k

$$d(c_i, c_j) \leq d(c_i, c_k), d(c_j, c_k) \Leftrightarrow \begin{cases} d_{i,k} \geq d_{ij} \\ d_{jk} \geq d_{ij} \end{cases}$$

(caso contrário, seria c_i agregado com o c_k
ou c_j " " o c_k)

$$\text{Logo } d_{ij,k} = \frac{n_i}{n_i + n_j} d_{i,k} + \frac{n_j}{n_i + n_j} d_{j,k}$$

$$\geq \frac{n_i}{n_i + n_j} d_{i,j} + \frac{n_j}{n_i + n_j} d_{i,j}$$

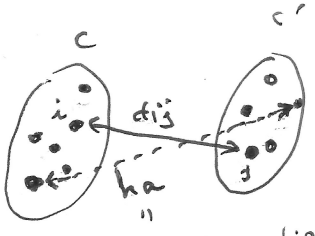
$$= \frac{n_i + n_j}{n_i + n_j} d_{i,j} = d_{i,j} \quad \square$$

9. Por definição a distância cofenética entre indivíduos i e j para o método do vizinho mais afastado é o custo de fusão quando i e j passam a fazer parte do mesmo cluster, ou seja, se $i \in C$ e $j \in C'$ (no passo anterior à fusão):

$$h_a = d(C, C') = \max_{\substack{x \in C \\ y \in C'}} d(x, y)$$

Como $i \in C$ e $j \in C'$

$$\max_{\substack{x \in C \\ y \in C'}} d(x, y) \geq d(i, j) = d_{ij}$$



$d(C, C')$: distância entre os vizinhos mais afastados

Logo $h_a \geq d_{ij}$.

Analogamente se prova que $h_p \leq d_{ij}$.

10. a) Pelo fórmula de L-W para o método do vizinho mais próximo (single)

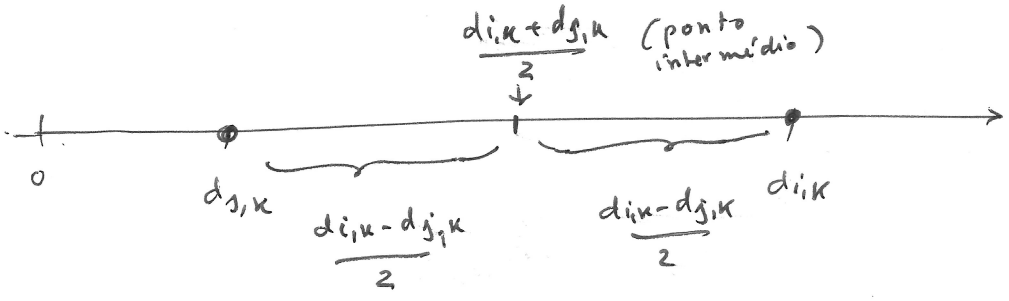
$$\alpha_i = \frac{1}{2} \; ; \quad \alpha_j = \frac{1}{2} \; ; \; \beta = 0 \; ; \; \gamma = -\frac{1}{2}$$

logo $$\boxed{d_{ij,k} = \frac{1}{2} d_{i,k} + \frac{1}{2} d_{j,k} - \frac{1}{2} |d_{i,k} - d_{j,k}|}$$

Suponhamos que $d_{i,k} \geq d_{j,k}$ (o caso $d_{j,k} \geq d_{i,k}$ é análogo) $\Rightarrow |d_{i,k} - d_{j,k}| = d_{i,k} - d_{j,k}$ e vem,

$$d_{ij,k} = \frac{1}{2} d_{i,k} + \frac{1}{2} d_{j,k} - \frac{1}{2} (d_{i,k} - d_{j,k})$$

$$= \frac{d_{i,k} + d_{j,k}}{2} - \frac{d_{i,k} - d_{j,k}}{2} = d_{j,k}$$



Analogamente se mostra que se $d_{i,k} \leq d_{j,k} \Rightarrow d_{ij,k} = d_{i,k}$

Logo $d_{ij,k} = \min(d_{i,k}, d_{j,k})$ que é precisamente a fórmula de recorrência (update) para o método do vizinho mais próximo c/ $c' = i$, $c'' = j$ e $c = k$ no slide ... ver)

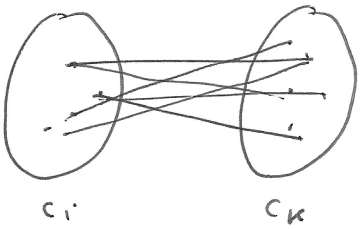
b) Pela fórmula de L-W para o método das distâncias médias (average) vem-se

$$\alpha_i = \frac{n_i}{n_i + n_j}, \quad \alpha_j = \frac{n_j}{n_i + n_j} \qquad \beta = \gamma = 0$$

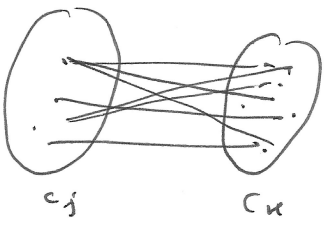
Logo,

$$\boxed{d_{ij,k} = \frac{n_i}{n_i + n_j} d_{i,k} + \frac{n_j}{n_i + n_j} d_{j,k}}$$

Ora, $d_{ik} = \dfrac{\sum_{y \in C_k} \sum_{x \in C_i} d(x,y)}{n_i n_k}$



$$d_{j,k} = \frac{\sum_{y \in C_k} \sum_{x \in C_j} d(x,y)}{n_j n_k}$$



- d_{ik} : média das distâncias entre pts de C_i e C_k
- d_{jk} : " " " " " " C_j e C_k

Então,

$$d_{ij,k} = \frac{n_i}{n_i + n_j} d_{i,k} + \frac{n_j}{n_i + n_j} d_{j,k} =$$

$$= \frac{n_i \sum_{y \in C_k} \sum_{x \in C_i} d(x,y)}{(n_i + n_j)(n_i n_k)} + \frac{n_j \sum_{y \in C_k} \sum_{x \in C_i} d(x,y)}{(n_i + n_j) n_j n_k}$$

$$= \frac{\sum_{y \in C_k} \left(\sum_{x \in C_i} d(x,y) + \sum_{x \in C_j} d(x,y) \right)}{n_k (n_i + n_j)} =$$

$$= \frac{\sum_{y \in C_k} \sum_{x \in C_i \cup C_j} d(x,y)}{n_k (n_i + n_j)} = \text{média das distâncias entre pts de } C_i \cup C_j \text{ e } C_k .$$

11. a) Pela fórmula de L-W para o método de variância mínima (Ward):

$$d^2_{ij,k} = \frac{n_i + n_k}{n_i + n_j + n_k} d^2_{i,k} + \frac{n_j + n_k}{n_i + n_j + n_k} d^2_{j,k} - \frac{n_k}{n_i + n_j + n_k} d^2_{i,j}$$

onde $d_{ij} = d(C_i, C_j)$ $\quad n_i = |C_i|$ etc...

$d_{ij,k} = d(C_i \cup C_j, C_k)$

$$D = [d_{ij}] = \begin{array}{c|cccc} & a & b & c & d \\ \hline a & 0 & 1.8 & 2.4 & 2.3 \\ b & 1.8 & 0 & 2.5 & 2.7 \\ c & 2.4 & 2.5 & 0 & 1.2 \\ d & 2.3 & 2.7 & 1.2 & 0 \end{array} \longrightarrow \begin{array}{c|ccc} d & a & b & c \\ \hline b & 1.8 & - & - \\ c & 2.4 & 2.5 & - \\ d & 2.3 & 2.7 & 1.2 \end{array}$$

$$\begin{array}{c|ccc} d^2 & a & b & c \\ \hline b & 3.24 & - & - \\ c & 5.76 & 6.25 & - \\ d & 5.29 & 7.29 & 1.44 \end{array}$$

d^2	a	b	c
b	3.24	–	–
c	5.76	6.25	–
d	5.29	7.29	(1.44) → menor

1ª fusão: cvd

- $d^2_{cvd,a} = d_{ij,k} = \frac{n_c + n_a}{n_c + n_d + n_a} d^2_{c,a} +$

$+ \frac{n_d + n_a}{n_c + n_d + n_a} d^2_{d,a} - \frac{n_a}{n_c + n_d + n_a} d^2_{c,d}$

$= \frac{2}{3} 5.76 + \frac{2}{3} 5.29 - \frac{1}{3} 1.44 = 6.88667$

- $d^2_{cvd,b} = \frac{n_c + n_b}{n_c + n_d + n_b} d^2_{c,b} + \frac{n_d + n_b}{n_c + n_d + n_b} d^2_{d,b} -$

$- \frac{n_b}{n_c + n_d + n_b} d^2_{c,d} =$

$= \frac{2}{3} 6.25 + \frac{2}{3} \times 7.29 - \frac{1}{3} 1.44 = 8.54\,4667$

após a 1ª fusão a matriz de proximidade vem:

d^2	a	b
b	(3.24)	–
{c,d}	6.88...	8.54...

→ mais pequeno

próxima fusão: a∪b

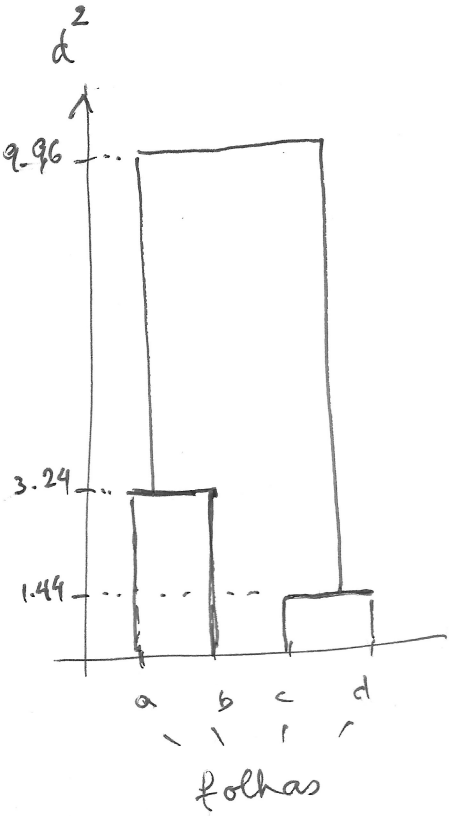
$$d^2_{a\cup b,\{c,d\}} = \frac{n_i + n_k}{n_i + n_j + n_k} d^2_{i,k} + \frac{n_j + n_k}{n_i + n_j + n_k} d^2_{j,k} -$$

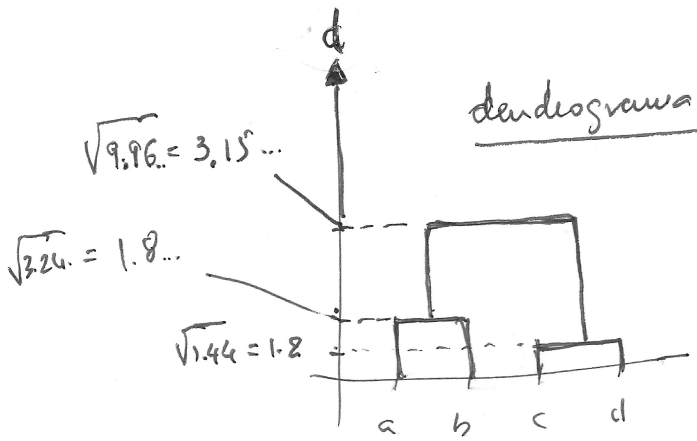
$$- \frac{n_k}{n_i + n_j + n_k} d^2_{i,j} = \frac{n_a + n_{\{c,d\}}}{n_a + n_b + n_{\{c,d\}}} d^2_{a,\{c,d\}}$$

$$+ \frac{n_b + n_{\{c,d\}}}{n_a + n_b + n_{\{c,d\}}} d^2_{b,\{c,d\}} - \frac{n_{\{c,d\}}}{n_a + n_b + n_{\{c,d\}}} d^2_{a,b}$$

$$= \frac{1+2}{1+1+2} \times 6.88\ldots + \frac{1+2}{1+1+2} \times 8.54 - \frac{2}{1+1+2} 3.24 = 9.96.$$

	$\{a,b\}$
$\{c,d\}$	9.96





dendrograma final

Pela fórmula do update do slide ...(var), o método da variância mínima (ward) procura agregar os clusters que minimizam o aumento da variância intra-grupos ie minimizar a estatística

$$\Delta_{ij} SSQ_W = e_{ij}^2 - e_i^2 - e_j^2$$

onde e_i^2 variância intra-grupos de C_i

e_j^2 " " " " C_j

e_{ij}^2 " " " " $C_{ij} = C_i \cup C_j$

Para calcular estas variâncias temos que recorrer à fórmula que apenas depende das distâncias entre elementos (e não dos centroides*):

$$e_K^2 = \sum_{x \in C_K} \|x - m_K\|^2 = \frac{\sum_{x,y \in C_K} \|x-y\|^2}{2 n_K}$$

onde m_K = centroide de C_K, $n_K = |C_K|$, $\|x-y\|^2 = d^2(x,y)$

(* não conhecemos os dados originais para podermos calcular os centroides)

d^2	a	b	c
b	3.24	–	–
c	5.76	6.25	–
d	5.29	7.29	1.44

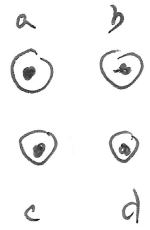
$$d^2(x,y) = \|x-y\|^2$$

$$e_k^2 = \frac{\sum_{x,y \in C_k} d^2(x,y)}{2\,m_k}$$

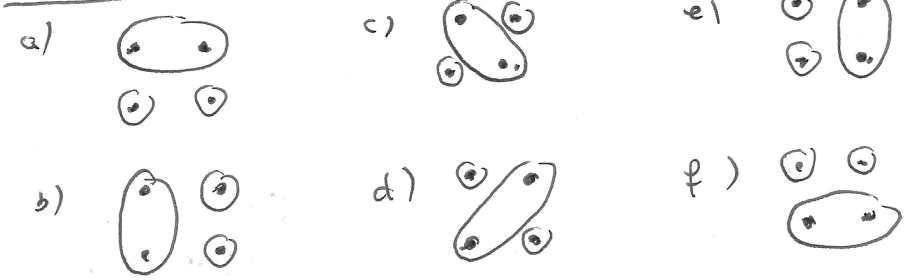
m_k = nº do grupo C_k, $h_k = |C_k|$

se $|C_k| = 1 \Rightarrow e_k^2 = 0$

partição inicial



2.ª partição ? 6 casos possíveis:



Custos de fusão:

a) $e^2_{[a,b]} = e_a^2 - e_b^2 = \frac{d^2(a,b) + d^2(b,a)}{2 \times 2}$

$e_a^2 = 0$, $e_b^2 = 0$

$= \frac{3.24}{2}$ //

b) $e^2_{[a,c]} - e_a^2 - e_c^2 = \frac{d^2(a,c) + d^2(c,a)}{2 \times 2}$

$= \frac{5.76 + 5.76}{4} = \frac{5.76}{2}$ //

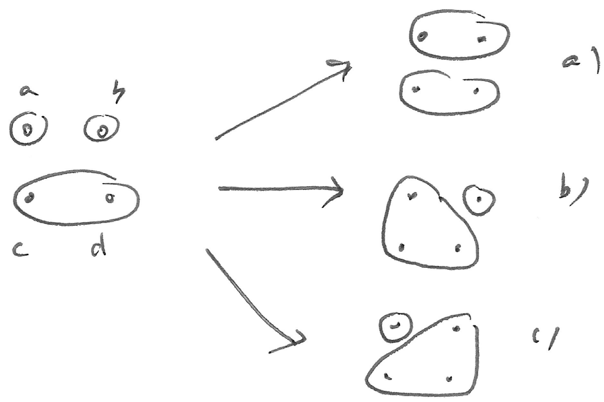
c) $e^2_{[a,d]} - e_a^2 - e_d^2 = \frac{5.29}{2}$ //

e) $e^2_{[b,c]} - e_b^2 - e_c^2 = \frac{6.25}{2}$ //

e) $e^2_{[b,d]} - e_b^2 - e_d^2 = \frac{7.29}{2}$ //

f) $e^2_{[c,d]} - e_c^2 - e_d^2 = \frac{1.44}{2}$ // ← menor!

⇒ 1ª fusão: c ∪ d c/ custo $\frac{1.44}{2} = 0.72$ (caso f))



Qual é a partição com menor incremento da inércia intra-grupo ?

OBS.:

aumento de inércia intra-grupo (custo de fusão) : $e^2_{ij} - e^2_i - e^2_j$

$C_i, C_j \longrightarrow C_{ij} = C_i \cup C_j$

a) $e^2_{\{a,b\}} - e^2_a - e^2_b = \frac{3.24}{2}$

$e^2_a = 0$, $e^2_b = 0$

b) $e^2_{\{a,c,d\}} - e^2_a - e^2_{\{c,d\}} =$

$\{a,c,d\} \to ij$, $a \to i$, $\{c,d\} \to j$

$$e^2_{\{a,c,d\}} = \frac{d^2(a,c) + d^2(a,d) + d^2(c,a) + d^2(c,d) + d^2(d,a) + d^2(d,c)}{2 \times 3}$$

$3 \to |\{a,c,d\}|$; $d^2(a,c)$; $d^2(a,d) = d^2(d,a)$; $d^2(c,d) = d^2(d,c)$

$$= \frac{d^2(a,c) + d^2(a,d) + d^2(c,d)}{3} = \frac{5.76 + 5.28 + 1.44}{2} = \frac{8.3266}{2}$$

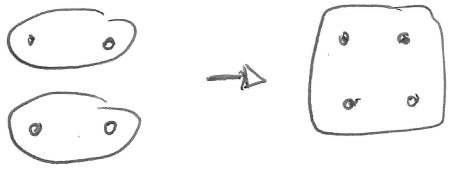
logo $e^2_{\{a,c,d\}} - e^2_a - e_{\{c,d\}} = \frac{8.3266... - 1.44}{2} = \frac{6.88...}{2}$

c) $e^2_{\{b,c,d\}} - e^2_b - e^2_{\{c,d\}} = \frac{9.9866...}{2} - \frac{1.44}{2} = \frac{8.54...}{2}$

2ª fusão: $a, b \longrightarrow a \cup b$

A partição c/ menor incremento da inércia intra-grupo é a partição a) c/ custo de fusão de $\frac{3.24}{2}$

3ª fusão:



custo da fusão, ou seja, aumento da inércia intra-grupo?

$$e^2_{\{a,b,c,d\}} - e^2_{\{a,b\}} - e^2_{\{c,d\}} = ?$$

$$? \qquad \frac{3.24}{2} \qquad \frac{1.44}{2}$$

Ora,

$$e^2_{\{a,b,c,d\}} = \frac{\sum_{x,y \in C_k} d^2(x,y)}{2 \times n_k} = \frac{\sum_{x,y \in C_k} d^2(x,y)}{4}$$

$$= \Big(d^2(a,b) + d^2(b,c) + d^2(a,d) + d^2(b,a) + d^2(b,c) + d^2(b,d)$$

$$+ d^2(c,a) + d^2(c,b) + d^2(c,d) + d^2(d,a) + d^2(d,b) + d^2(d,c) \Big) / 8$$

$$= \frac{d^2(a,b) + d^2(a,c) + d^2(a,d) + d^2(b,c) + d^2(b,d) + d^2(c,d)}{4}$$

$$= \frac{3.24 + 5.76 + 5.29 + 5.25 + 7.29 + 1.44}{4} = \frac{14.635}{2}$$

Assim,

$$e^2_{\{a,b,c,d\}} - e^2_{\{a,b\}} - e^2_{\{c,d\}} = \frac{9.955}{2}$$

Os custos de fusão dados pelo aumento de inércia mínima correspondem a metade dos custos de fusão ao quadrado dados pela fórmula de L-W.

12. $x = (x_1, \dots, x_N)$ duas observações de uma
$y = (y_1, \dots, y_N)$

população X com média μ e matriz de variâncias - covariâncias $\Sigma = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_N^2 \end{bmatrix}$

($\sigma_{ij} = 0$ se $i \neq j$ pois as variáveis são não correlacionadas)

$$\Rightarrow \underbrace{d_M(x,y)}_{\text{distância de Mahalanobis}} = \|x-y\|_M = \sqrt{(x-y)^T \Sigma^{-1} (x-y)} =$$

$$= \left(\begin{bmatrix} x_1 - y_1 & \cdots & x_N - y_N \end{bmatrix} \begin{bmatrix} \sigma_1^{-2} & & \\ & \ddots & \\ & & \sigma_N^{-2} \end{bmatrix} \begin{bmatrix} x_1 - y_1 \\ \vdots \\ x_N - y_N \end{bmatrix} \right)^{1/2} =$$

$$= \left(\begin{bmatrix} x_1 - y_1 & \cdots & x_N - y_N \end{bmatrix} \begin{bmatrix} \frac{(x_1 - y_1)}{\sigma_1^2} \\ \vdots \\ \frac{(x_N - y_N)}{\sigma_N^2} \end{bmatrix} \right)^{1/2} =$$

$$= \sqrt{\frac{(x_1 - y_1)^2}{\sigma_1^2} + \cdots + \frac{(x_N - y_N)^2}{\sigma_N^2}} \quad //$$

Por outro lado, as variáveis normalizadas são dadas por

$$x' = \left(\frac{x_1 - \mu_1}{\sigma_1}, \dots, \frac{x_N - \mu_N}{\sigma_N} \right)$$

$$y' = \left(\frac{y_1 - \mu_1}{\sigma_1}, \dots, \frac{y_N - \mu_N}{\sigma_N} \right)$$

e tem-se $x' - y' = \left(\frac{x_1 - y_1}{\sigma_1}, \dots, \frac{x_N - y_N}{\sigma_N} \right)$

Portanto,

$$d(x,y) = \|x' - y'\|_2 = \sqrt{\left(\frac{x_1 - y_1}{\sigma_1}\right)^2 + \dots + \left(\frac{x_N - y_N}{\sigma_N}\right)^2}$$

distância euclidiana

$$= \sqrt{\frac{(x_1 - y_1)^2}{\sigma_1^2} + \dots + \frac{(x_N - y_N)^2}{\sigma_N^2}} = d_M(x,y) \quad //$$

13. Seja $f(y) = \sum_{x \in X} (x - y)^2$ função real (quadrática)

Se f atinge um mínimo em $y_0 \Rightarrow f'(y_0) = 0$

Ora, $f'(y) = \left(\sum_{x \in X} (x - y)^2 \right)' = \sum_{x \in X} \left((x - y)^2 \right)'$

$$= \sum_{x \in X} 2(x - y)(-1) \quad //$$

Logo,

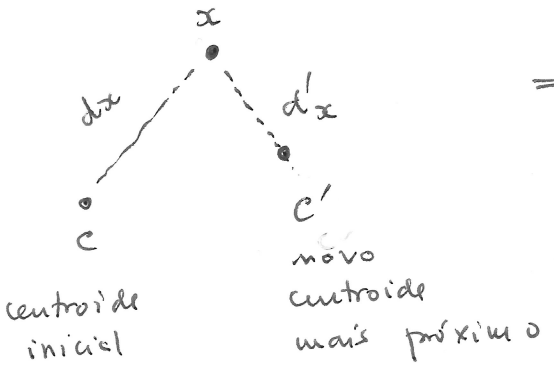
$$f'(y) = 0 \iff \sum_{x \in X} 2(x - y)(-1) = 0 \iff -\sum_{x \in X} (2x - 2y) = 0$$

$$\iff 2 \sum_{x \in X} x = 2 \sum_{x \in X} y = 2ny \quad \text{onde } n = |X|$$

$$\text{logo } y = \frac{\sum_{x \in X} x}{n} = \bar{x} \quad //$$

b) O algoritmo das k-médias consiste essencialmente em 2 passos que são repetidamente iterados:

i) Re-afectação dos pts aos centroides mais próximos:



$$\Rightarrow d_x \geq d'_x$$

$$\Rightarrow d_x^2 \geq (d'_x)^2$$

$$\Rightarrow \sum_{x \in X} (d'_x)^2 \leq \sum_{x \in X} d_x^2 \Rightarrow$$

Após a re-afectação a soma dos quadrados das distâncias de todos os pts aos centroides + próximos baixa.

$\therefore$ SSQw baixa!

ii) Recalcular os centroides dos novos clusters.

Após a re-afectação do ponto i) os grupos são alterados pelo que os seus centroides também são alterados pelo que SSQw baixa após este passo pois a SSQw(y) de cada grupo é atingida para y=centroide desse grupo.

Logo SSQ_W total baixa em todos os passos do algoritmo

Como há um nº finito de particiões o algoritmo irá parar ao fim de um nº finito de passos convergindo monotonamente (a decrescer) para algum valor (que pode corresponder a uma solução sub-optimal)

14. O método das K-médias móveis com K-grupos (K-means) procura definir uma partição $X = C_1 \cup \dots \cup C_K$ que minimize a estatística

$$SSQ_W = \sum_{j=1}^{K} e_j^2$$

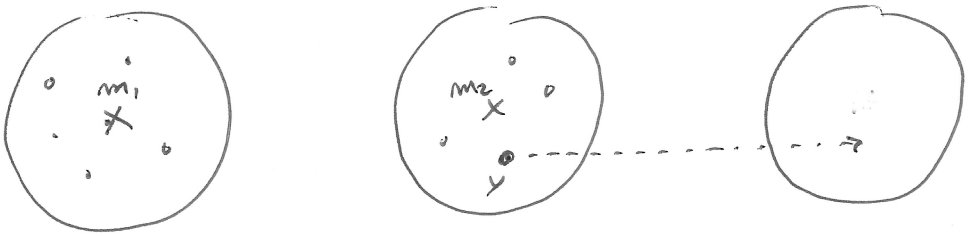
onde

$$e_j^2 = \sum_{x \in C_j} \|x - m_j\|^2$$

é a inércia do grupo C_j e m_j o seu centróide.

Suponhamos que $K = 3$ e que $C_3 = \emptyset$.

Vamos ainda admitir que existe um elemento y que não coincide com o ~~centroide~~ do seu grupo (nessa altura podemos supor $y \neq m_2$ (centroide do grupo 2)



C_1 C_2 C_3

Se transferirmos y para o grupo C_3 a soma dos quadrados $\sum_{\substack{x \in C_2 \\ x \neq y}} \|x - m_2\|^2$ diminui pois não contém a parcela positiva $\|y - m_2\|^2 > 0$ $(y \neq m_2)$.

Mas como agora C_2 tem menos um elemento temos que recalcular o seu centroide m_2' e a sua nova inércia intra-grupo,

$$SSQ_W = \sum_{\substack{x \in C_2 \\ x \neq y}} \|x - m_2'\|^2,$$

tq pela alínea 13 a) verifica $\sum_{\substack{x \in C_2 \\ x \neq y}} \|x - m_2'\|^2 \leq \sum_{\substack{x \in C_2 \\ x \neq y}} \|x - m_2$

Logo a inércia do grupo $C_2 \setminus \{y_4\}$ é inferior à inércia intragrupo de C_2 !

A inércia de C_3 mantém-se nula porque $C_3 = \{y\}$ só tem um elemento

e a inércia do grupo C_1 mantém-se também inalterada porque este grupo fica inalterado. Logo a inércia total intra-grupos

$$SSQ_W = \underbrace{\sum_{x \in C_1} \|x - m_1\|^2}_{\text{SSQw de } C_1 \text{ inalterado}} + \underbrace{\sum_{x \in C_2 \setminus \{y_4\}} \|x - m_2'\|^2}_{\text{SSQw de } C_2' = C_2 \setminus \{y_4\} \text{ baixa!}} + \underbrace{\|y - y\|^2}_{0}$$

baixa, pelo que a solução com um grupo vazio não pode ser ótima*

* com excepção dos casos em que os ~~✓~~ TODOS! elementos dos grupos C_1 e C_2 coincidam com os respectivos centroides. Nessa altura $SSQ_W = 0$ e não pode ser melhorada! VFF!

15. Pelo Teorema de Huygens, se $X = C_1 \cup C_2 \cup \dots \cup C_K$,

$$SSQ_t = SSQ_w + SSQ_b$$

onde • $SSQ_t = \sum_{x \in X} \| x - m_x \|^2$ m_x : centroide X

• $SSQ_w = \sum_{j=1}^{K} e_j^2$ onde

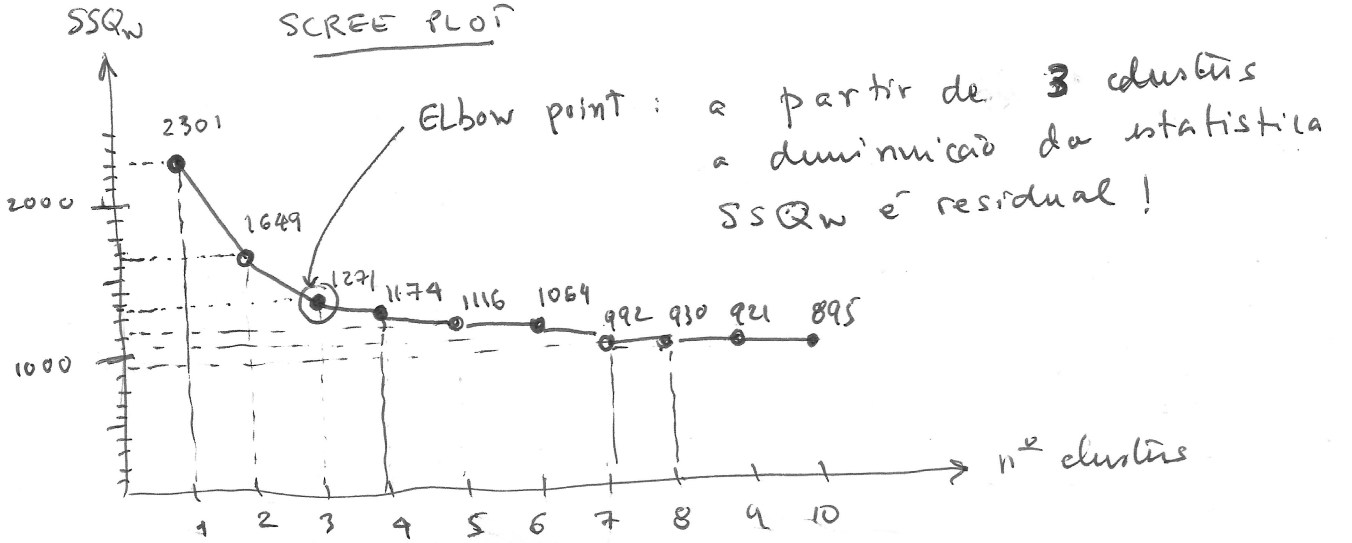
$$e_j^2 = \sum_{x \in C_j} \| x - m_j \|^2 \quad m_j : \text{centroide de } C_j$$

• $SSQ_b = \sum_{j=1}^{K} n_j \| m_j - m_x \|^2 \quad n_j = |C_j|$

a) Se $K=1 \Rightarrow X = C_1$ e $m_x = m_1$

$$\Rightarrow \begin{cases} SSQ_b = 0 \\ SSQ_t = SSQ_w = 2301 \end{cases} //$$

b) 2 critérios : scree plot SSQ_w e Calinski-Harabasz



Calinski-Harabasz (CH) $(k \geq 2)$ $N = 178$ vinhos

$$CH(k) = \frac{\frac{SSQ_b}{k-1}}{\frac{SSQ_w}{N-k}} = \frac{(SSQ_t - SSQ_w)/(k-1)}{SSQ_w/(N-k)} =$$

$$= \frac{2301 - SSQ_w}{SSQ_w} \cdot \frac{178-k}{k-1}$$

k	1	2	3	4	5	6	7	8	9	10
SSQ_w	2301	1649	1271	1174	1116	1064	992	930	921	895
SSQ_b	0	652	1030	1127	1185	1237	1309	1371	1380	1406
CH(k)	N.D	69,6	70.9	55.7	46.9	39.9	37,6	35,8	32.6	29,...

↓ máximo

O máximo para o C-H ocorre c/ 3 clusters

⇒ ambos os critérios sugerem 3 clusters!

c) $RAND = \frac{A+D}{\binom{N}{2}}$

A: nº de pares classificados no mesmo grupo por ambos os métodos = 4767

D: nº de pares classificados em grupos distintos por ambos os métodos = ?

nº total

$$\binom{N}{2} = \frac{N(N-1)}{2} = \frac{178 \times 177}{2} = 15753 \text{ de pares}$$

$$RAND = \frac{A+D}{\binom{N}{2}} \Rightarrow A+D = \binom{N}{2} \times RAND =$$

$$= 0.9407097 * 15753$$

$$= 14819$$

$$\therefore D = 14819 - A = 14819 - 4767 = 10052$$

pares classificados de forma discordante* por ambos os métodos.

(* isto é, em classes diferentes)

c) RAND toma valores entre 0 e 1 e é uma medida de semelhança

(RAND = 1 $\Leftrightarrow$ as 2 partições coincidem)

Consideramos a dissemelhança

$$d = 1 - RAND$$

d	s	e	A	W
c	0.653	—	—	—
A	0.065	0.650	—	—
W	0.6555	0.169	0.6552	—
K	0.654	0.179	0.6537	0.059

menor valor

1ª fusão: K ∪ W

	S	C	A
C	0.653	—	—
A	(0.065)	0.650	—
{W,K}	0.6555	0.179	0.6552

$d(S, W \cup K) =$

$= \max \{ d(S,W), d(S,K) \}$

$= \max \{ 0.6555, 0.654 \}$

$= 0.6555$ etc...

próxima fusão: A ∪ S

	{S,A}	C
C	0.653	—
{W,K}	0.6555	(0.179)

$d(C, \{S,A\}) =$

$= \max (d(C,A)\ d(C,S)\}$

$= \max (0.650, 0653)$

etc...

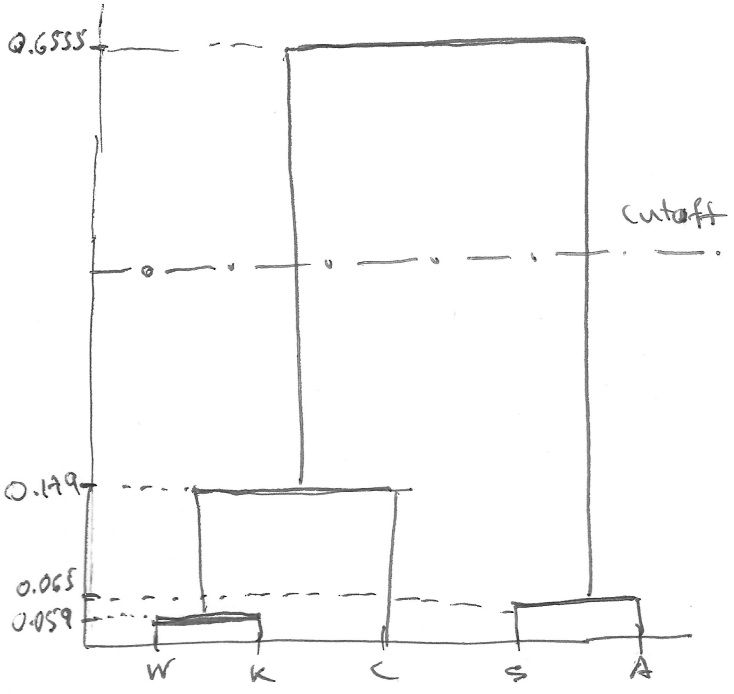
próxima fusão: $C \cup \{W,K\}$

	{S,A}
{W,K,C}	0.6555

$d(C \cup \{W,K\}, \{S,A\}) =$

$= \max \left(d(C, \{S,A\}), d(\{W,K\}, \{S,A\}) \right)$

$= \max (0.653, 0.6555) = 0.655$



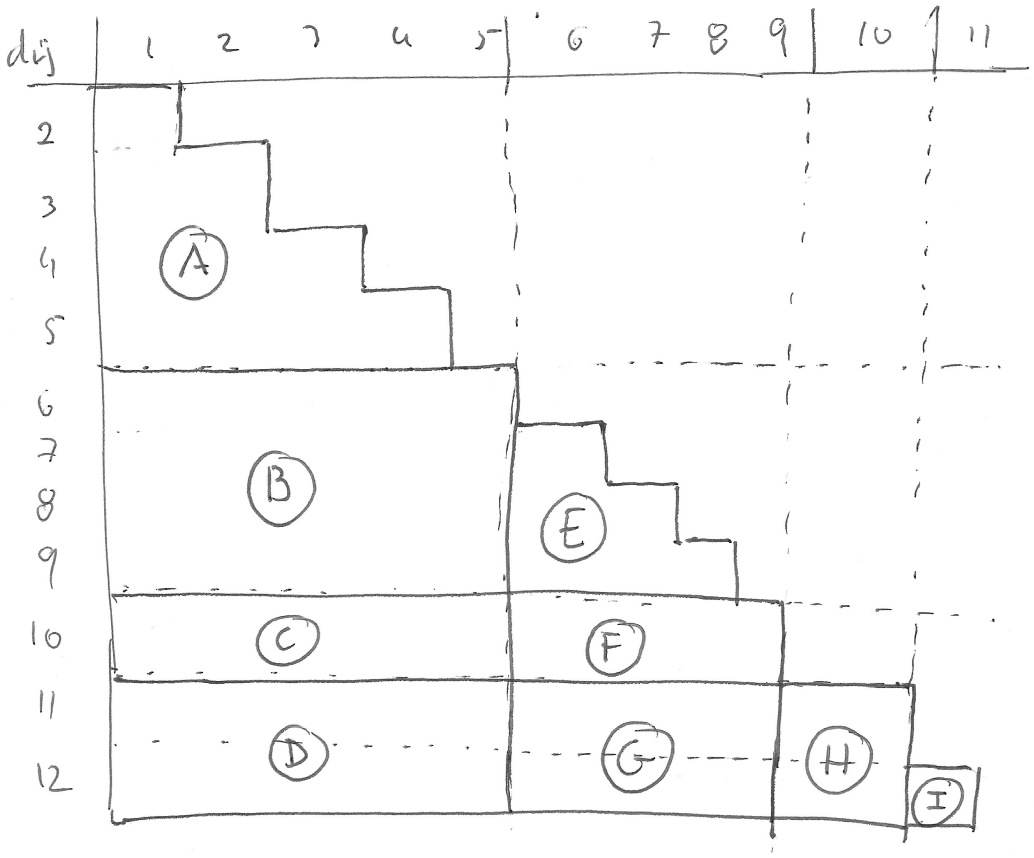
O dendrograma sugere 2 grupos

de classificação:

{ward, K-means, complete}

e {single, average} //

16.



a) $diam(C_1) = \max_{x,y \in C_1} d(x,y) = 0.189$ maior valor de (A)

$diam(C_2) = \max_{x,y \in C_2} d(x,y) = 0.124$ maior valor de (E)

$diam(C_3) = 0$, pois $C_3 = \{10\}$

$diam(C_4) = d_{11,12} = 0.252$

b) Por definição do método do vizinho mais afastado

$d(C_1, C_2) = \max_{\substack{x \in C_1 \\ y \in C_2}} d(x,y) = 0.268$ maior valor de (B)

$d(C_1, C_3) = \max_{x \in C_1, y \in C_3} d(x,y) = 0.284$ maior valor em (C)

$d(C_1, C_4) = \max_{x \in C_1, y \in C_4} d(x,y) = 0.322$ maior valor em (D)

$d(C_2, C_3) = \max_{x \in C_2, y \in C_3} d(x,y) = 0.293$ maior valor em (F)

$d(C_2, C_4) = \max_{x \in C_2, y \in C_4} d(x,y) = 0.319$ maior valor em (G)

$d(C_3, C_4) = \max_{x \in C_3, y \in C_4} d(x,y) = 0.320$ maior valor em (H)

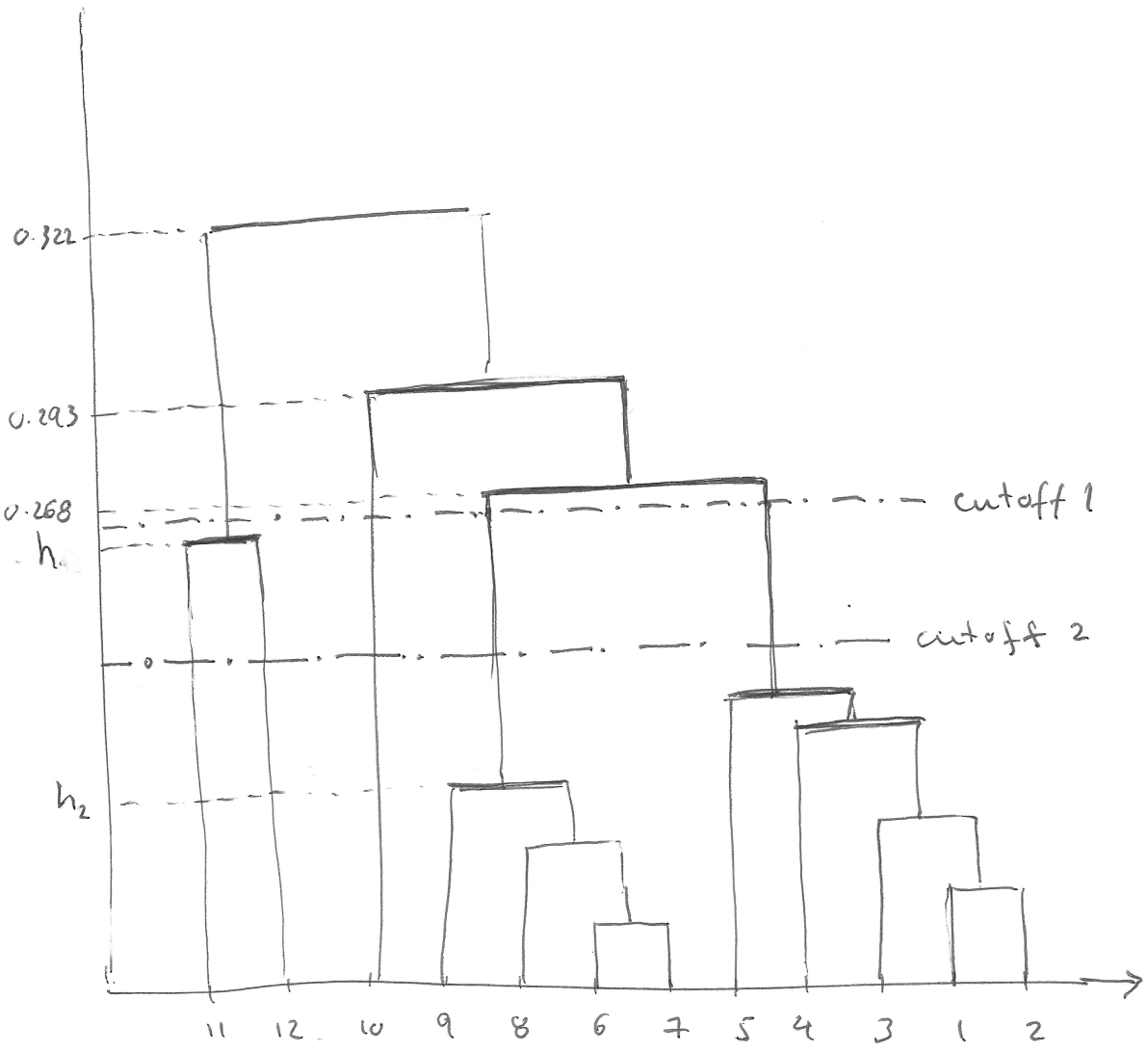
	C_1	C_2	C_3
C_2	0.268	—	—
C_3	0.284	0.293	—
C_4	0.322	0.319	0.320

⇒

	$C_1 \cup C_2$	C_3
C_3	0.293	
C_4	0.322	0.320

⇒

	$C_1 \cup C_2 \cup C_3$
C_4	0.322



A opção de formar 4 grupos (cutoff 1) não é correta pois corresponde a cortar o dendrograma entre as alturas h e 0.268 que estão muito próximas entre si. O dendrograma sugere um corte em 5 grupos (cutoff 2).

Ex mostre que $h_2 = \text{diam}(C_2) = 0.124$

OBS Pode-se mostrar que $d(C_i, C_j) =$ diâmetro de $C_i \cup C_j$, razão pela qual o método do vizinho mais afastado também se designa por método do diâmetro

c) distâncias originais

	1	7	10
7	0.233	–	–
10	0.273	0.293	–
11	0.322	0.316	0.320

distâncias cofenéticas

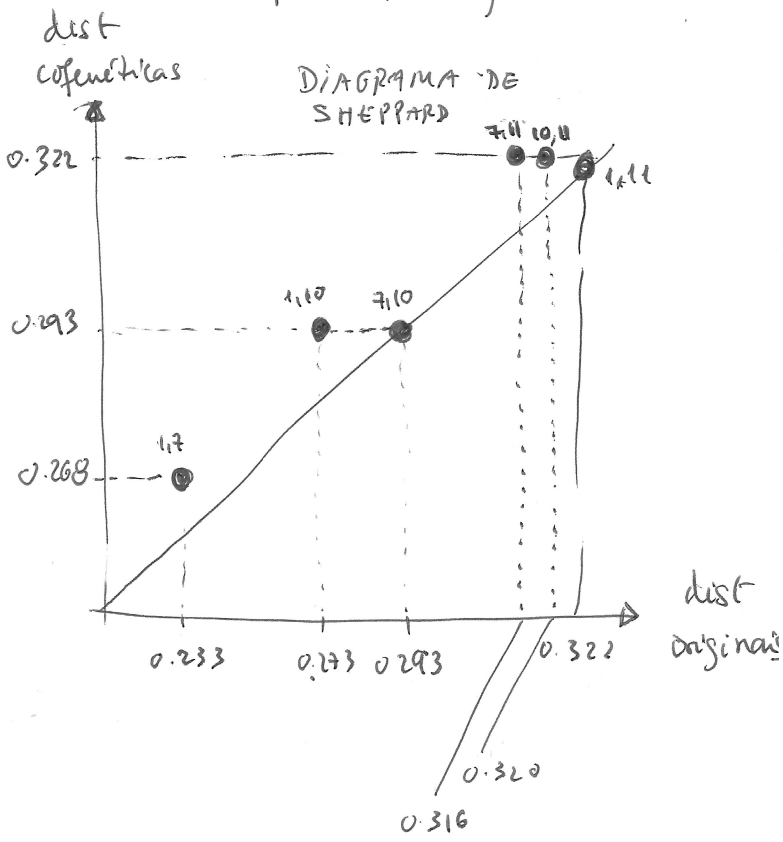
	1	7	10
7	0.268		
10	0.293	0.293	
11	0.322	0.322	0.322

1 – Homo Sapiens

7 – Macaca Fuscata

10 – Saimiri Sciureus

11 – Tarsius Syrichta



d) Partição em 4 grupos: 1 2 3 4 5 | 6 7 8 9 | 10 | 11 12

" " 5 grupos: 1 2 3 4 5 | 6 7 8 9 | 10 | 11 | 12

i) único par classificado de forma discordante 11, 12

ii) nº total de pares = $\binom{N}{2} = \binom{12}{2} = \frac{12 \times 11}{2} = 66$

A+D: pares classificados de forma concordante: 66 − 1 = 65

$$RAND = \frac{A+D}{\binom{N}{2}} = \frac{65}{66}$$

17. a) Ambos os métodos procuram minimizar a estatística SSQ_W, isto é, a heterogeneidade intra-grupo (variabilidade de cada grupo). Uma vez que o método das K-médias móveis foi aplicado usando como sementes iniciais os centroides obtidos pela análise classificatória com o método Ward e que o valor da estatística SSQ_W baixa em cada passo do algoritmo das K-médias móveis (kmeans) o valor desta estatística é inferior (ou igual) ao valor de SSQ_W obtido com o método da inércia mínima (ward), logo os grupos são mais homogéneos pela classificação c/ o kmeans.

b) $$RAND = \frac{A+D}{\binom{N}{2}} = \frac{\binom{N}{2} - (B+C)}{\binom{N}{2}} = \frac{21945 - 1358}{21945} = 0,9381...$$

onde,

$\binom{N}{2} = \frac{210 \times 209}{2}$, $A+D$: nº pares em que ambos os métodos concordaram

$B+C$: " " " " " discordaram

c)

d_{ij}	c_1	c_2
c_2	29.44	–
c_3	21.55	41.80

$\Longrightarrow$

d_{ij}^2	c_1	c_2
c_2	866.71	
c_3	(464.40)	1747.24

menor valor

próxima fusão: $c_{13} = c_1 \cup c_3$ cl custo $d_{13}^2 = 464.4$

última fusão: $c_{13} \cup c_2 = c_{13,2}$

custo de fusão $d(c_{13}, c_2) = ?$

Pela fórmula de L-W. para o método do Ward

com $i = 1$, $j = 3$, $k = 2$ vem

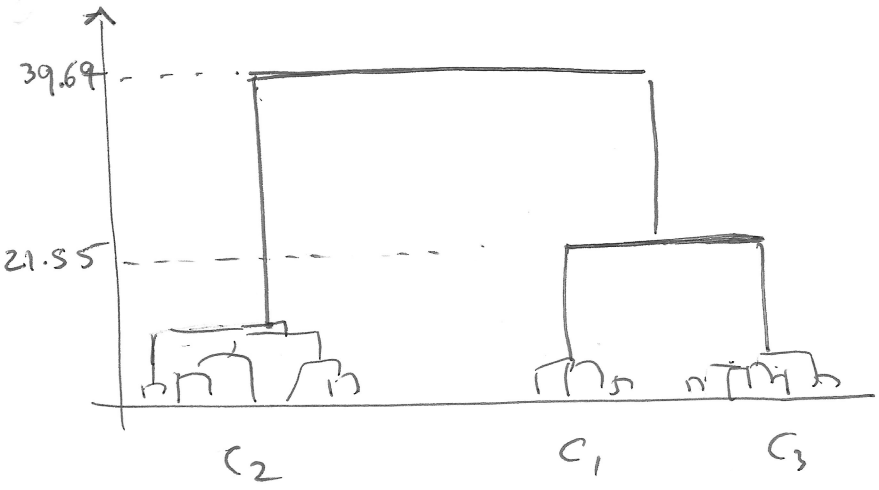
$$d^2(c_{13}, c_2) = d_{13,2}^2 = \frac{n_1 + n_2}{n_1 + n_2 + n_3} d_{1,2}^2 + \frac{n_3 + n_2}{n_1 + n_2 + n_3} d_{3,2}^2$$

$$- \frac{n_2}{n_1 + n_2 + n_3} d_{1,3}^2 = \frac{73+70}{210} 866.71$$

$$+ \frac{67+70}{210} 1747.24 - \frac{70}{210} 464.40 = 1575.25$$

logo $d^2(c_3, c_2) = 1575.25 \Rightarrow d(c_{13}, c_2) = \sqrt{1575.25} = 39.6$

Dendrograma final:



18. Para dados binários de presença (1) / ausência (0) podemos utilizar o coeficiente de semelhança de Jaccard,

$$J = \frac{A}{A+B+C}$$, para comparar as bacias 1 e 2

onde A: nº de espécies presentes em ambas as bacias

B: " " " presentes na bacia 1 e ausentes da bacia 2;

C: " " " ausentes da bacia 1 e presentes na bacia 2.

Por exemplo $\delta(\text{OUEME}, \text{GAMBIE}) = \frac{3}{3+3+2} = \frac{3}{8}$

J	OV	GA	GE
GA	3/8	—	—
GE	2/9	2/7	—
CR	1/8	2/6	2/5

→

D=1-J	OV	GA	GE
GA	5/8	—	—
GE	7/9	5/7	—
CR	7/8	4/6	3/5

OBS

$J = \frac{A}{A+B+C}$ medida de semelhança → $D = 1 - J = \frac{B+C}{A+B+C}$

medida de dissemelhança

Pelo método de classificação hierárquico do vizinho mais afastado:

	OV	GA	GE
GA	5/8		
GE	7/9	5/7	
CR	7/8	4/6	3/5
max:	7/8	5/7	menor valor

1ª fusão: GE ∪ CR

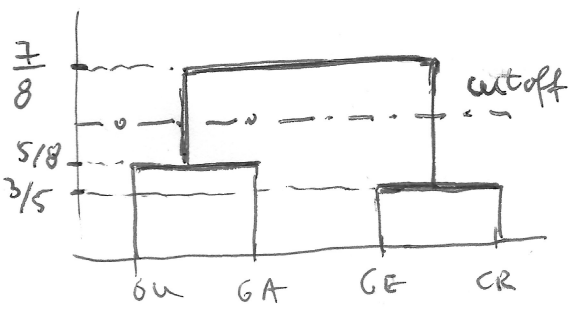
→

	OV	GA
GA	5/8	—
{GE, CR}	7/8	5/7

max {3/8, 5/7} = 7/8

	{OV, GA}
→ {GE, CR}	7/8

2ª Fusão: GA ∪ OV



A análise classificatória com o método do vizinho mais afastado e a dissemelhança 1 − J sugere os 2 grupos de bacias: {OUEME, GAMBIE} e {GEBA, CRUBAL}