Análise de agrupamento

Patrícia de Siqueira Ramos

UNIFAL-MG, *campus* Varginha

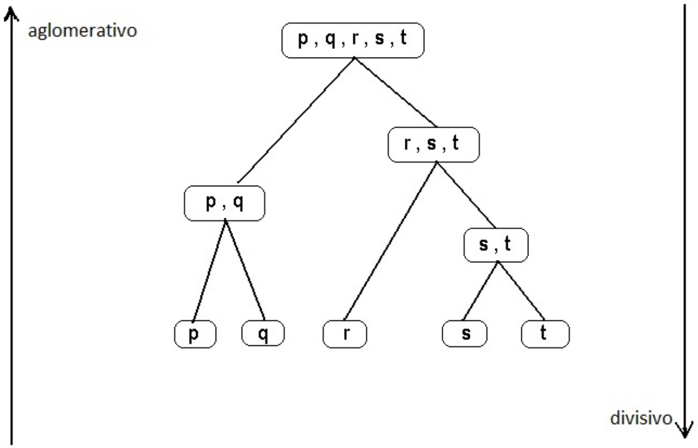
13 de Novembro de 2018

Métodos de agrupamento

a) Hierárquicos: o processo tem uma hierarquia em que subgrupos de grupos em um nível são agregados para formar novos grupos. Dendrogramas são utilizados. Tipos de métodos hierárquicos:

- aglomerativos: cada observação é um grupo no início e, a cada passo, grupos se fundem. Uma vez que um par de observações está em um grupo, o par não será mais separado
- divisivos: no início há um único grupo com todas as observações e, a cada passo, há subdivisões. Uma vez que um par de observações tenha sido separado, ele não estará mais no mesmo grupo

Ilustração dos métodos hierárquicos aglomerativos e divisivos para $n = 5$



Métodos de agrupamento

b) Não hierárquicos: grupos são formados pelo ajuste a algum critério em qualquer momento, movendo observações para dentro ou fora dos grupos. É mais difícil de usar pois o valor do número de grupos (k) deve ser predefinido.

a) Métodos de agrupamento hierárquicos

Aglomerativos (mais comuns)

- No início, $k = n$ grupos e, a cada passo, os elementos são agrupados até todos estarem em um único grupo ($k = 1$)

a) Métodos de agrupamento hierárquicos

Aglomerativos (mais comuns)

- No início, $k = n$ grupos e, a cada passo, os elementos são agrupados até todos estarem em um único grupo ($k = 1$)
- No estágio inicial há a menor variação interna (variância é zero pois há só um elemento)

a) Métodos de agrupamento hierárquicos

Aglomerativos (mais comuns)

- No início, $k = n$ grupos e, a cada passo, os elementos são agrupados até todos estarem em um único grupo ($k = 1$)
- No estágio inicial há a menor variação interna (variância é zero pois há só um elemento)
- A cada passo, os grupos são comparados por alguma medida de similaridade predefinida (grupos mais similares são combinados)

a) Métodos de agrupamento hierárquicos

Aglomerativos (mais comuns)

- No início, $k = n$ grupos e, a cada passo, os elementos são agrupados até todos estarem em um único grupo ($k = 1$)
- No estágio inicial há a menor variação interna (variância é zero pois há só um elemento)
- A cada passo, os grupos são comparados por alguma medida de similaridade predefinida (grupos mais similares são combinados)
- A escolha do número final k de grupos é subjetiva

a) Métodos de agrupamento hierárquicos

Aglomerativos (mais comuns)

- No início, $k = n$ grupos e, a cada passo, os elementos são agrupados até todos estarem em um único grupo ($k = 1$)
- No estágio inicial há a menor variação interna (variância é zero pois há só um elemento)
- A cada passo, os grupos são comparados por alguma medida de similaridade predefinida (grupos mais similares são combinados)
- A escolha do número final k de grupos é subjetiva
- Os 3 primeiros métodos de agrupamento hierárquicos aglomerativos que veremos são:
  - Ligação simples (vizinho mais próximo)
  - Ligação completa (vizinho mais distante)
  - Ligação média (distância média)

a) Métodos de agrupamento hierárquicos

Exemplo de dendrograma ($n = 9$, $p = 2$):

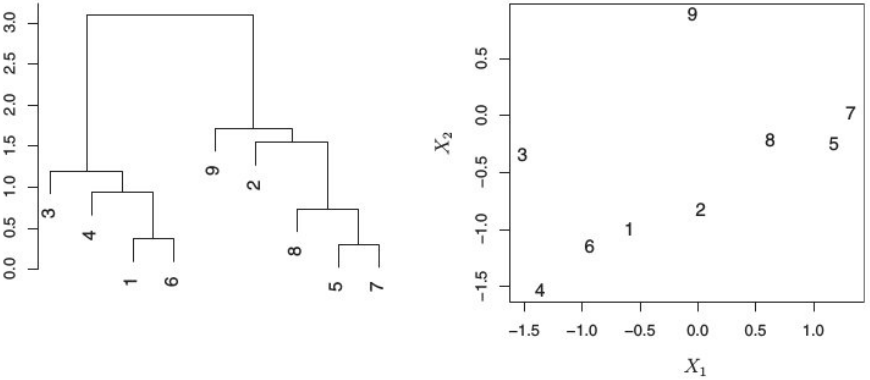
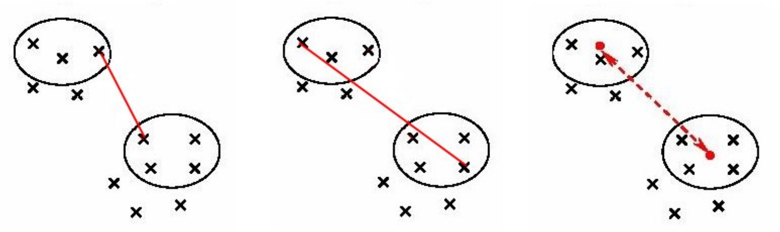


Ilustração de 3 métodos hierárquicos aglomerativos



Vizinho mais próximo Vizinho mais distante Centróide

i) Vizinho mais próximo (ligação simples)

- A distância entre os grupos é definida pelas observações mais próximas
- A cada passo, os dois grupos A e B mais similares em relação à distância

$$d_{AB} = min(d_{ij}), \qquad i \in A, j \in B$$

são unidos em um mesmo grupo

Ex.: Sejam as observações sobre renda e idade de 6 indivíduos

1	9,60	28
2	8,40	31
3	2,40	42
4	18,20	38
5	3,90	25
6	6,40	41

Aplicar o método do vizinho mais próximo usando a distância euclidiana para comparação dos grupos (apesar da distância de Mahalanobis ser mais indicada neste caso, a euclidiana é mais fácil de ser calculada manualmente).

Exemplo - vizinho mais próximo

- Passo 1:

1 - 2 - 3 - 4 - 5 - 6

- matriz de distâncias:

$$
\begin{array}{cc}
\begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} & \\
\begin{bmatrix}
- & & & & & \\
3,23 & - & & & & \\
15,74 & 12,53 & - & & & \\
13,19 & 12,04 & 16,29 & - & & \\
6,44 & 7,50 & 17,06 & 19,33 & - & \\
13,39 & 10,19 & 4,12 & 12,18 & 16,19 & -
\end{bmatrix} & \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array}
\end{array}
$$

- menor distância: $d_{12} = 3,23$, então unimos as observações 1 e 2

Exemplo - vizinho mais próximo

- Passo 2:

12 - 3 - 4 - 5 - 6

- matriz de distâncias:

$$\begin{array}{cc} \begin{array}{ccccc} 12 & 3 & 4 & 5 & 6 \end{array} & \\ \begin{bmatrix} - & & & & \\ 12,53 & - & & & \\ 12,04 & 16,29 & - & & \\ 6,44 & 17,06 & 19,33 & - & \\ 10,19 & 4,12 & 12,18 & 16,19 & - \end{bmatrix} & \begin{array}{c} 12 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \end{array}$$

- $min(d_{ij}) = d_{36} = 4,12$, então unimos as observações 3 e 6
- $d_{12;3} = min(d_{13}; d_{23}) = min(15,74; 12,53) = 12,53$
- e o mesmo para $d_{12;4}$, $d_{12;5}$, $d_{12;6}$

Exemplo - vizinho mais próximo

- Passo 3:

12 - 36 - 4 - 5

- matriz de distâncias:

$$\begin{array}{cc} \begin{array}{cccc} 12 & 36 & 4 & 5 \end{array} & \\ \begin{bmatrix} - & & & \\ 10,19 & - & & \\ 12,04 & 12,18 & - & \\ 6,44 & 16,19 & 19,33 & - \end{bmatrix} & \begin{array}{c} 12 \\ 36 \\ 4 \\ 5 \end{array} \end{array}$$

- $min(d_{ij}) = d_{12;5} = 6,44$, então unimos as observações 1, 2 e 5
- $d_{12;36} = min(d_{13}; d_{16}; d_{23}; d_{26}) =$ $min(15,74; 13,39; 12,53; 10,19) = 10,19$
- e o mesmo para $d_{12;4}$ etc.

Exemplo - vizinho mais próximo

- Passo 4:

125 - 36 - 4

- matriz de distâncias:

$$\begin{array}{cc} \begin{array}{ccc} 125 & 36 & 4 \end{array} & \\ \begin{bmatrix} - & & \\ 10,19 & - & \\ 12,04 & 12,18 & - \end{bmatrix} & \begin{array}{c} 125 \\ 36 \\ 4 \end{array} \end{array}$$

- $min(d_{ij}) = d_{125;36} = 10,19$, então unimos as observações 1, 2, 5 com 3 e 6
- $d_{125;36} = min(d_{13}; d_{16}; d_{23}; d_{26}; d_{53}; d_{56}) = 10,19$

Exemplo - vizinho mais próximo

- Passo 5:

12536 - 4

- matriz de distâncias:

$$\begin{array}{cc} \begin{array}{cc} 12536 & 4 \end{array} & \\ \begin{bmatrix} - & \\ 12,04 & - \end{bmatrix} & \begin{array}{c} 12536 \\ 4 \end{array} \end{array}$$

- $d_{12536;4} = min(d_{14}; d_{24}; d_{54}; d_{34}; d_{64}) = 12,04$

i) Vizinho mais próximo

Histórico do agrupamento (exemplo de idade e renda):

Passo	k	Fusão	Distância
1	5	(1,2)	3,23
2	4	(3,6)	4,12
3	3	(1,2), (5)	6,44
4	2	(1,2,5), (3,6)	10,19
5	1	(1,2,5,3,6), (4)	12,04

Fazer dendrograma.

ii) Vizinho mais distante (ligação completa)

- A similaridade entre os grupos é definida pelas observações mais distantes:

$$d_{AB} = \max(d_{ij}), \qquad i \in A, j \in B.$$

- A cada passo, essa distância é calculada para todos os pares de grupos e serão unidos os que tiverem menor valor de distância

Exemplo - vizinho mais distante

- Passo 1:

1 - 2 - 3 - 4 - 5 - 6

- matriz de distâncias:

$$\begin{array}{cc} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} & \\ \begin{bmatrix} - & & & & & \\ 3,23 & - & & & & \\ 15,74 & 12,53 & - & & & \\ 13,19 & 12,04 & 16,29 & - & & \\ 6,44 & 7,50 & 17,06 & 19,33 & - & \\ 13,39 & 10,19 & 4,12 & 12,18 & 16,19 & - \end{bmatrix} & \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \end{array}$$

- menor distância: $d_{12} = 3,23$, então unimos as observações 1 e 2

Exemplo - vizinho mais distante

- Passo 2:

12 - 3 - 4 - 5 - 6

- matriz de distâncias:

$$\begin{array}{cc} \begin{array}{ccccc} 12 & 3 & 4 & 5 & 6 \end{array} & \\ \begin{bmatrix} - & & & & \\ 15,74 & - & & & \\ 13,19 & 16,29 & - & & \\ 7,50 & 17,06 & 19,33 & - & \\ 13,39 & 4,12 & 12,18 & 16,19 & - \end{bmatrix} & \begin{array}{c} 12 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \end{array}$$

- $min(d_{ij}) = d_{36} = 4,12$, então unimos as observações 3 e 6
- $d_{12;3} = max(d_{13}; d_{23}) = max(15,74; 12,53) = 15,74$
- e o mesmo para $d_{12;4}$, $d_{12;5}$, $d_{12;6}$

Exemplo - vizinho mais distante

- Passo 3:

12 - 36 - 4 - 5

- matriz de distâncias:

$$
\begin{array}{cc}
\begin{array}{cccc} 12 & 36 & 4 & 5 \end{array} & \\
\begin{bmatrix}
- & & & \\
15,74 & - & & \\
13,19 & 16,29 & - & \\
7,50 & 17,06 & 19,33 & -
\end{bmatrix} & \begin{array}{c} 12 \\ 36 \\ 4 \\ 5 \end{array}
\end{array}
$$

- menor distância: $max(d_{ij}) = d_{12;5} = 7,50$, então unimos as observações 1, 2 e 5
- $d_{12;36} = max(d_{13}; d_{16}; d_{23}; d_{26}) = max(15,74; 13,39; 12,53; 10,19) = 15,74$
- e o mesmo para as outras

Exemplo - vizinho mais distante

- Passo 4:

125 - 36 - 4

- matriz de distâncias:

$$\begin{array}{cc} \begin{array}{ccc} 125 & 36 & 4 \end{array} & \\ \begin{bmatrix} - & & \\ 17,06 & - & \\ 19,33 & 16,29 & - \end{bmatrix} & \begin{array}{c} 125 \\ 36 \\ 4 \end{array} \end{array}$$

- menor distância: $max(d_{ij}) = d_{36;4} = max(d_{34}; d_{64}) = max(16,29; 12,18) = 16,29$, então unimos as observações 3 e 6 com 4
- exemplo: $d_{125;36} = max(d_{13}; d_{16}; d_{23}; d_{26}; d_{53}; d_{56}) = 17,06$

Exemplo - vizinho mais distante

- Passo 5:

125 - 364

- matriz de distâncias:

$$\begin{array}{cc}
\begin{array}{cc} 125 & 364 \end{array} & \\
\begin{bmatrix} - & \\ 19,33 & - \end{bmatrix} & \begin{array}{c} 125 \\ 364 \end{array}
\end{array}$$

- $d_{125;364} = max(d_{13}; d_{16}; d_{14}; d_{23}; d_{26}; d_{24}; d_{53}; d_{56}; d_{54}) = 19,33$

ii) Vizinho mais distante

Histórico do agrupamento (exemplo de idade e renda):

Passo	k	Fusão	Distância
1	5	(1,2)	3,23
2	4	(3,6)	4,12
3	3	(1,2), (5)	7,50
4	2	(3,6), (4)	16,29
5	1	(1,2,5), (3,6,4)	19,33

Fazer dendrograma.

iii) Distância média

- A distância entre dois grupos é a média das distâncias entre todos os pares de elementos dos dois grupos:

$$d_{AB} = \frac{1}{n_A n_B} \sum_{i \in A} \sum_{j \in B} d_{ij},$$

em que n_A e n_B são os números de observações nos grupos A e B.

Exemplo - distância média

- Passo 1:

1 - 2 - 3 - 4 - 5 - 6

- matriz de distâncias:

$$
\begin{array}{cc}
\begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} & \\
\begin{bmatrix}
- & & & & & \\
3,23 & - & & & & \\
15,74 & 12,53 & - & & & \\
13,19 & 12,04 & 16,29 & - & & \\
6,44 & 7,50 & 17,06 & 19,33 & - & \\
13,39 & 10,19 & 4,12 & 12,18 & 16,19 & -
\end{bmatrix} & \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array}
\end{array}
$$

- menor distância: $d_{12} = 3,23$, então unimos as observações 1 e 2

Exemplo - distância média

- Passo 2:

12 - 3 - 4 - 5 - 6

- matriz de distâncias:

$$\begin{array}{cc} \begin{array}{ccccc} 12 & 3 & 4 & 5 & 6 \end{array} & \\ \begin{bmatrix} - & & & & \\ 14,13 & - & & & \\ 12,62 & 16,29 & - & & \\ 6,97 & 17,06 & 19,33 & - & \\ 11,79 & 4,12 & 12,18 & 16,19 & - \end{bmatrix} & \begin{array}{c} 12 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \end{array}$$

- $min(d_{ij}) = d_{36} = 4,12$, então unimos as observações 3 e 6
- $d_{12;3} = (d_{13} + d_{23})/2 = (15,74 + 12,53)/2 = 14,13$
- e o mesmo para $d_{12;4}$, $d_{12;5}$, $d_{12;6}$

Exemplo - distância média

- Passo 3:

12 - 36 - 4 - 5

- matriz de distâncias:

$$\begin{array}{cc} \begin{array}{cccc} 12 & 36 & 4 & 5 \end{array} & \\ \begin{bmatrix} - & & & \\ 12,96 & - & & \\ 12,62 & 14,24 & - & \\ 6,97 & 16,62 & 19,33 & - \end{bmatrix} & \begin{array}{c} 12 \\ 36 \\ 4 \\ 5 \end{array} \end{array}$$

- menor distância: $d_{12;5} = (d_{15} + d_{25})/2 = 6,97$, então unimos as observações 1, 2 e 5
- $d_{12;36} = (d_{13} + d_{16} + d_{23} + d_{26})/4 = 12,96$
- e o mesmo para $d_{12;4}$ etc.

Exemplo - distância média

- Passo 4:

125 - 36 - 4

- matriz de distâncias:

$$\begin{array}{ccc|c} 125 & 36 & 4 & \\ \multicolumn{3}{c}{\left[\begin{array}{ccc} - & & \\ 14,18 & - & \\ 14,85 & 14,24 & - \end{array}\right]} & \begin{array}{c} 125 \\ 36 \\ 4 \end{array} \end{array}$$

- menor distância: $d_{125;36} = 14,18$, então unimos as observações 1, 2, 5 com 3 e 6
- $d_{125;36} = (d_{13} + d_{16} + d_{23} + d_{26} + d_{53} + d_{56}) = 14,18$

Exemplo - distância média

- Passo 5:

12536 - 4

- matriz de distâncias:

$$\begin{array}{cc} \begin{array}{cc} 12536 & 4 \end{array} & \\ \begin{bmatrix} - & \\ 14,61 & - \end{bmatrix} & \begin{array}{c} 12536 \\ 4 \end{array} \end{array}$$

- $d_{12536;4} = (d_{14} + d_{24} + d_{54} + d_{34} + d_{64}) = 14,61$

iii) distância média

Histórico do agrupamento (exemplo de idade e renda):

Passo	k	Fusão	Distância
1	5	(1,2)	3,23
2	4	(3,6)	4,12
3	3	(1,2), (5)	6,97
4	2	(1,2,5), (3,6)	14,18
5	1	(1,2,5,3,6), (4)	14,61

Fazer dendrograma.

iv) Centróide

- Neste método, a distância é medida entre os vetores de médias dos grupos, também chamados de centróides dos grupos
- Ex.: se temos $G_A = \{\mathbf{X}_{1\cdot}, \mathbf{X}_{3\cdot}, \mathbf{X}_{7\cdot}\}$ e $G_B = \{\mathbf{X}_{2\cdot}, \mathbf{X}_{6\cdot}\}$, os vetores de médias correspondentes são

$$\bar{\mathbf{X}}_A = \frac{1}{3}(\mathbf{X}_{1\cdot} + \mathbf{X}_{3\cdot} + \mathbf{X}_{7\cdot})$$
$$\bar{\mathbf{X}}_B = \frac{1}{2}(\mathbf{X}_{2\cdot} + \mathbf{X}_{6\cdot})$$

e a distância entre os grupos A e B é definida por

$$d^2_{AB} = (\bar{\mathbf{X}}_A - \bar{\mathbf{X}}_B)^T(\bar{\mathbf{X}}_A - \bar{\mathbf{X}}_B),$$

que é a distância euclidiana ao quadrado entre os vetores $\bar{\mathbf{X}}_A$ e $\bar{\mathbf{X}}_B$

- É um método que exige um tempo computacional maior do que os anteriores

Exemplo - centróide

- Passo 1:

1 - 2 - 3 - 4 - 5 - 6

- matriz de distâncias:

$$
\begin{array}{cc}
\begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} & \\
\begin{bmatrix}
- & & & & & \\
3,23 & - & & & & \\
15,74 & 12,53 & - & & & \\
13,19 & 12,04 & 16,29 & - & & \\
6,44 & 7,50 & 17,06 & 19,33 & - & \\
13,39 & 10,19 & 4,12 & 12,18 & 16,19 & -
\end{bmatrix} & \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array}
\end{array}
$$

- menor distância: $d_{12} = 3,23$, então unimos as observações 1 e 2

Exemplo - centróide

- Passo 2:

12 - 3 - 4 - 5 - 6

- matriz de distâncias:

$$
\begin{array}{cc}
\begin{array}{ccccc} 12 & 3 & 4 & 5 & 6 \end{array} & \\
\begin{bmatrix}
- & & & & \\
14,14 & - & & & \\
10,47 & 16,29 & - & & \\
6,80 & 17,06 & 19,33 & - & \\
11,79 & 4,12 & 12,18 & 16,19 & -
\end{bmatrix} & \begin{array}{c} 12 \\ 3 \\ 4 \\ 5 \\ 6 \end{array}
\end{array}
$$

- $min(d_{ij}) = d_{36} = 4,12$, então unimos as observações 3 e 6
- Para obter as distâncias entre o grupo $1-2$ com os outros, precisaremos obter:

Exemplo - centróide

Para obter $d_{12;3}$:

$G_{12} = \{\mathbf{X}_{1\cdot}, \mathbf{X}_{2\cdot}\}$ e $G_3 = \{\mathbf{X}_{3\cdot}\}$, os vetores de médias correspondentes são

$$\bar{\mathbf{X}}_{12} = \frac{1}{2}(\mathbf{X}_{1\cdot} + \mathbf{X}_{2\cdot}) = \frac{1}{2}\begin{bmatrix} 9,60 + 8,40 \\ 28 + 31 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 18 \\ 59 \end{bmatrix} = \begin{bmatrix} 9 \\ 29,5 \end{bmatrix}$$

$$\bar{\mathbf{X}}_{3} = \frac{1}{1}(\mathbf{X}_{3\cdot}) = \frac{1}{1}\begin{bmatrix} 2,40 \\ 42 \end{bmatrix} = \begin{bmatrix} 2,40 \\ 42 \end{bmatrix}$$

e a distância entre os grupos 12 e 3 é definida por

$$\begin{aligned} d^2_{12;3} =& (\bar{\mathbf{X}}_{12} - \bar{\mathbf{X}}_{3})^T(\bar{\mathbf{X}}_{12} - \bar{\mathbf{X}}_{3}) \\ =& \begin{bmatrix} 6,60 & -12,50 \end{bmatrix}\begin{bmatrix} 6,60 \\ -12,50 \end{bmatrix} = 199,81 \end{aligned}$$

$$d_{12;3} = \sqrt{199,81} = 14,14.$$

E o mesmo procedimento é adotado para as outras distâncias.

iv) Centróide

Histórico do agrupamento (exemplo de idade e renda):

Passo	k	Fusão	Distância
1	5	(1,2)	3,23
2	4	(3,6)	4,12
3	3	(1,2), (5)	6,80
4	2	(1,2,5), (3,6)	13,84
5	1	(1,2,5,3,6), (4)	13,00

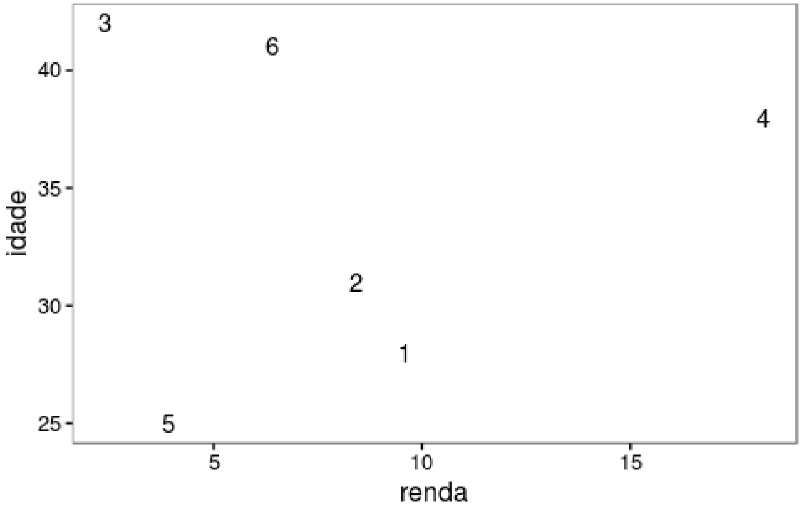
iv) Centróide

Histórico do agrupamento (exemplo de idade e renda):

Passo	k	Fusão	Distância
1	5	(1,2)	3,23
2	4	(3,6)	4,12
3	3	(1,2), (5)	6,80
4	2	(1,2,5), (3,6)	13,84
5	1	(1,2,5,3,6), (4)	13,00

- Agrupamento igual ao dos métodos do vizinho mais próximo e da ligação média (não é regra, nem sempre isso acontece)
- Diferente dos outros métodos, a distância no passo 5 foi menor do que o passo 4
- Isso pode ocorrer quando houver empates entre valores da matriz de distâncias (quanto maiores n e p, menor a probabilidade de acontecer)

Diagrama de dispersão - comparar os métodos



v) Ward

- Conhecido como o método de "mínima variância"
- A cada passo, calcula-se a soma de quadrados dentro de cada grupo (quadrado da distância euclidiana de cada observação do grupo em relação ao vetor de médias do grupo)
- Combinam-se os dois grupos que resultarem no menor valor de soma de quadrados
- O método de Ward e do centróide usam os vetores de médias amostrais como representantes da informação dos grupos, mas o de Ward leva em conta os tamanhos dos grupos que estão sendo comparados (tamanhos muito diferentes são penalizados)

v) Ward

Histórico do agrupamento (exemplo de idade e renda):

Passo	k	Fusão	Distância
1	5	(1,2)	3,23
2	4	(3,6)	4,12
3	3	(1,2), (5)	7,85
4	2	(3,6), (4)	16,48
5	1	(1,2,5), (3,6,4)	21,69

v) Ward

Histórico do agrupamento (exemplo de idade e renda):

Passo	k	Fusão	Distância
1	5	(1,2)	3,23
2	4	(3,6)	4,12
3	3	(1,2), (5)	7,85
4	2	(3,6), (4)	16,48
5	1	(1,2,5), (3,6,4)	21,69

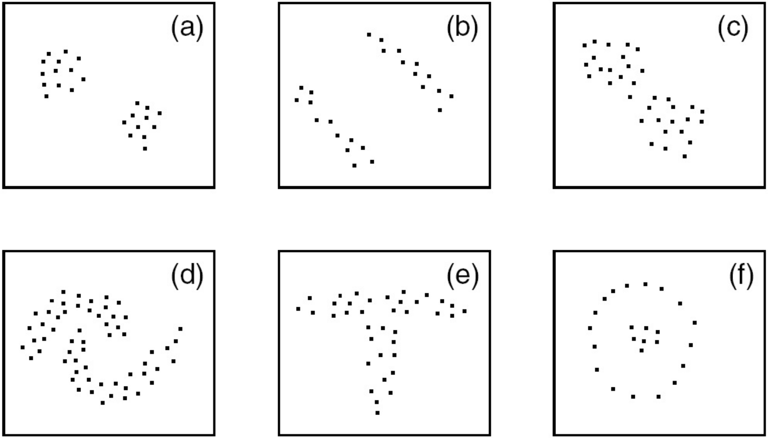
- Agrupamento igual ao do método do vizinho mais distante (não é regra, nem sempre isso acontece)
- Os valores das distâncias são um pouco diferentes dos outros métodos por utilizar vetores de médias

Resumo - métodos de agrupamento

- Há muitos métodos, mas nenhum é considerado o melhor
- Frequentemente, os diferentes métodos produzem resultados diferentes
- A maioria dos métodos produz grupos elipsóides, exceto vizinho mais próximo, que tende a não separar grupos próximos (efeito *chaining*)
- O vizinho mais distante tende a produzir grupos de mesmo diâmetro e a isolar valores discrepantes (*outliers*) nos primeiros passos
- Na distância média, grupos com variâncias internas próximas são obtidos e com partições melhores do que os métodos do vizinho mais próximo e mais distante
- O método de Ward tende a produzir grupos com números de elementos parecidos

Resumo - métodos de agrupamento

- Supor que existam duas variáveis e que seus diagramas de dispersão sejam os seguintes



Resumo - métodos de agrupamento

- A maioria dos métodos detectará dois grupos para a e b
- Alguns métodos podem ter problemas para identificá-los no caso c (por causa dos pontos intermediários)
- A maioria terá problemas com os casos d, e, f

Resumo - métodos de agrupamento

- A maioria dos métodos detectará dois grupos para a e b
- Alguns métodos podem ter problemas para identificá-los no caso c (por causa dos pontos intermediários)
- A maioria terá problemas com os casos d, e, f
- Boa prática: comparar resultados (se forem parecidos, há maior confiança que há aquela estrutura nos dados, senão, investigar o motivo)

Número de grupos na partição final

- Há muitos métodos para definir o número final k de grupos ou em qual passo o agrupamento deve ser interrompido, mas não há uma resposta exata para essa pergunta
- Alguns critérios podem ajudar:
  - análise da distância: quando se passa do passo i para o $i+1$, a similaridade entre os grupos decresce e a distância aumenta. Com os passos de fusão há pontos de salto maiores em relação aos demais, sugerindo o ponto de parada (se o n não for muito grande, avalia-se o dendrograma)

Número de grupos na partição final

- Há muitos métodos para definir o número final k de grupos ou em qual passo o agrupamento deve ser interrompido, mas não há uma resposta exata para essa pergunta
- Alguns critérios podem ajudar:
  - análise da distância: quando se passa do passo i para o $i+1$, a similaridade entre os grupos decresce e a distância aumenta. Com os passos de fusão há pontos de salto maiores em relação aos demais, sugerindo o ponto de parada (se o n não for muito grande, avalia-se o dendrograma)
- Outros critérios:
  - comportamento do nível de similaridade (medida que utiliza a distância entre os grupos)
  - análise da soma de quadrados entre os grupos: R^2
  - estatística pseudo-F (CALINSKI; HARABASZ, 1974) - método de Ward (correlação semiparcial)
  - estatística pseudo T^2 (DUDA; HART, 1973)
  - estatística CCC

Métodos hierárquicos divisivos

Esses métodos não serão vistos por não serem muito utilizados.