Análise de agrupamento

Patrícia de Siqueira Ramos

UNIFAL-MG, *campus* Varginha

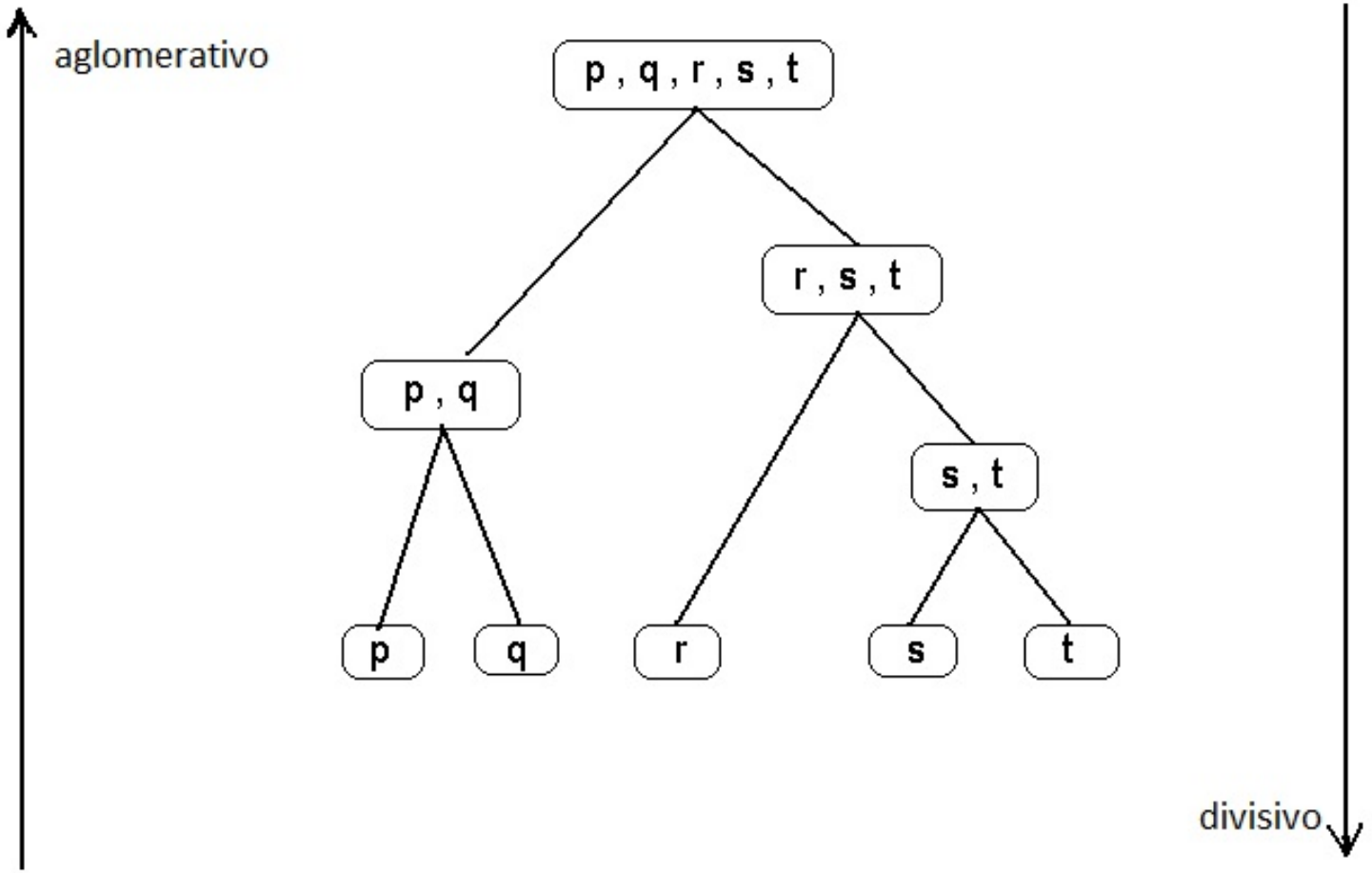
3 de Junho de 2016

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)

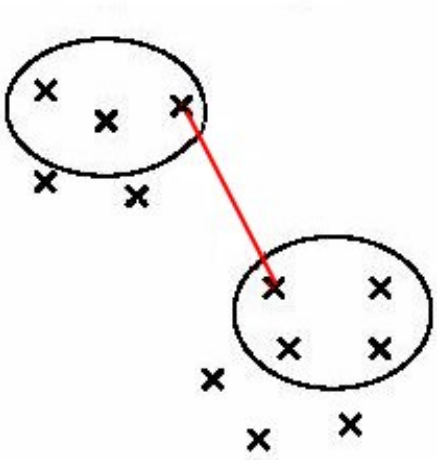
- No início, $k = n$ grupos e, a cada passo, os elementos são agrupados até todos estarem em um único grupo ($k = 1$)
- No estágio inicial há a menor variação interna (variância é zero pois há só um elemento)
- A cada passo, os grupos são comparados por alguma medida de similaridade predefinida (grupos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

a) Métodos de agrupamento hierárquicos

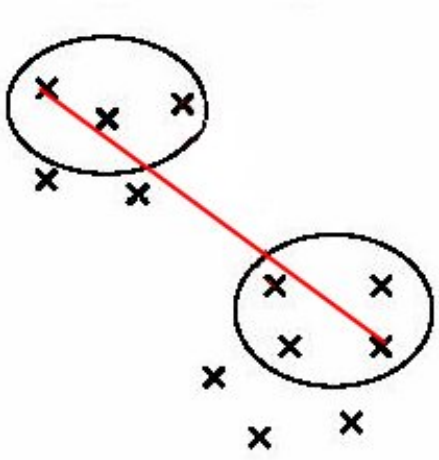
Aglomerativos (mais comuns)

- 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)

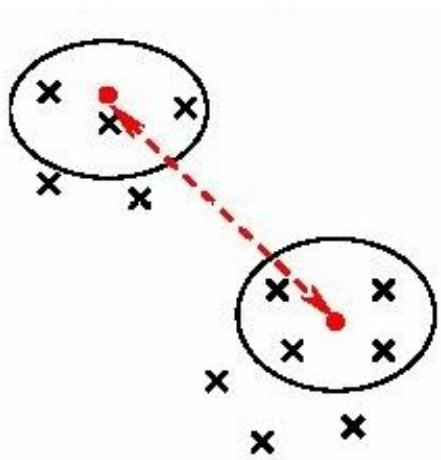
Ilustração de 3 métodos hierárquicos aglomerativos



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).

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

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.

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

$$d_{AB} = (\bar{\mathbf{x}}_A - \bar{\mathbf{x}}_B)^T(\bar{\mathbf{x}}_A - \bar{\mathbf{x}}_B),$$

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

iv) Centróide

Histórico do agrupamento (exemplo de idade e renda):

Passo	k	Fusão	Distância
1	5	(1,2)	10,44
2	4	(3,6)	17,00
3	3	(1,2), (5)	46,26
4	2	(1,2,5), (3,6)	190,66
5	1	(1,2,5,3,6), (4)	166,61

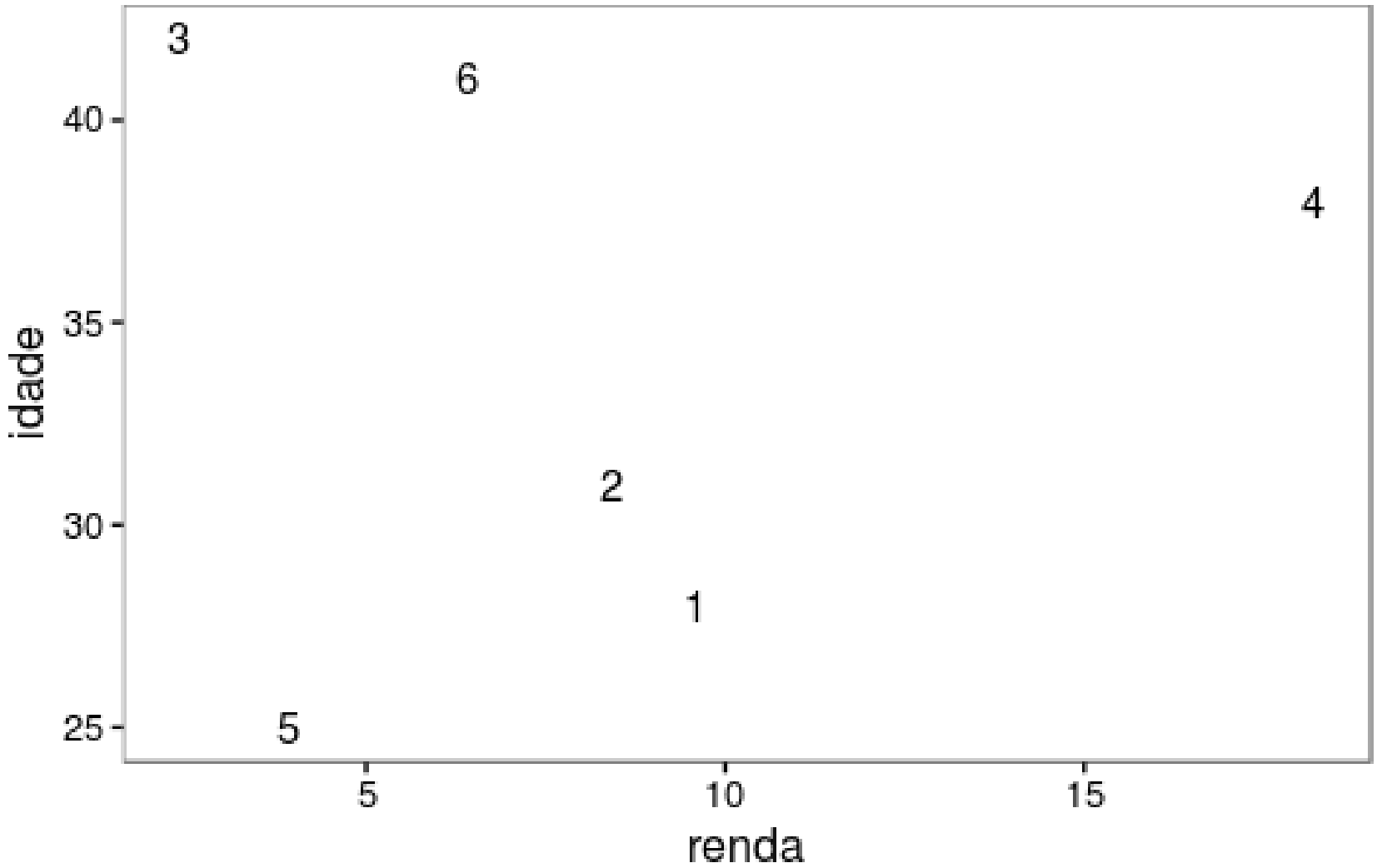
iv) Centróide

Histórico do agrupamento (exemplo de idade e renda):

Passo	k	Fusão	Distância
1	5	(1,2)	10,44
2	4	(3,6)	17,00
3	3	(1,2), (5)	46,26
4	2	(1,2,5), (3,6)	190,66
5	1	(1,2,5,3,6), (4)	166,61

- Agrupamento igual ao dos métodos do vizinho mais próximo e da ligação média
- 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)	10,44
2	4	(3,6)	17,00
3	3	(1,2), (5)	61,68
4	2	(3,6), (4)	270,25
5	1	(1,2,5), (3,6,4)	465,00

v) Ward

Histórico do agrupamento (exemplo de idade e renda):

Passo	k	Fusão	Distância
1	5	(1,2)	10,44
2	4	(3,6)	17,00
3	3	(1,2), (5)	61,68
4	2	(3,6), (4)	270,25
5	1	(1,2,5), (3,6,4)	465,00

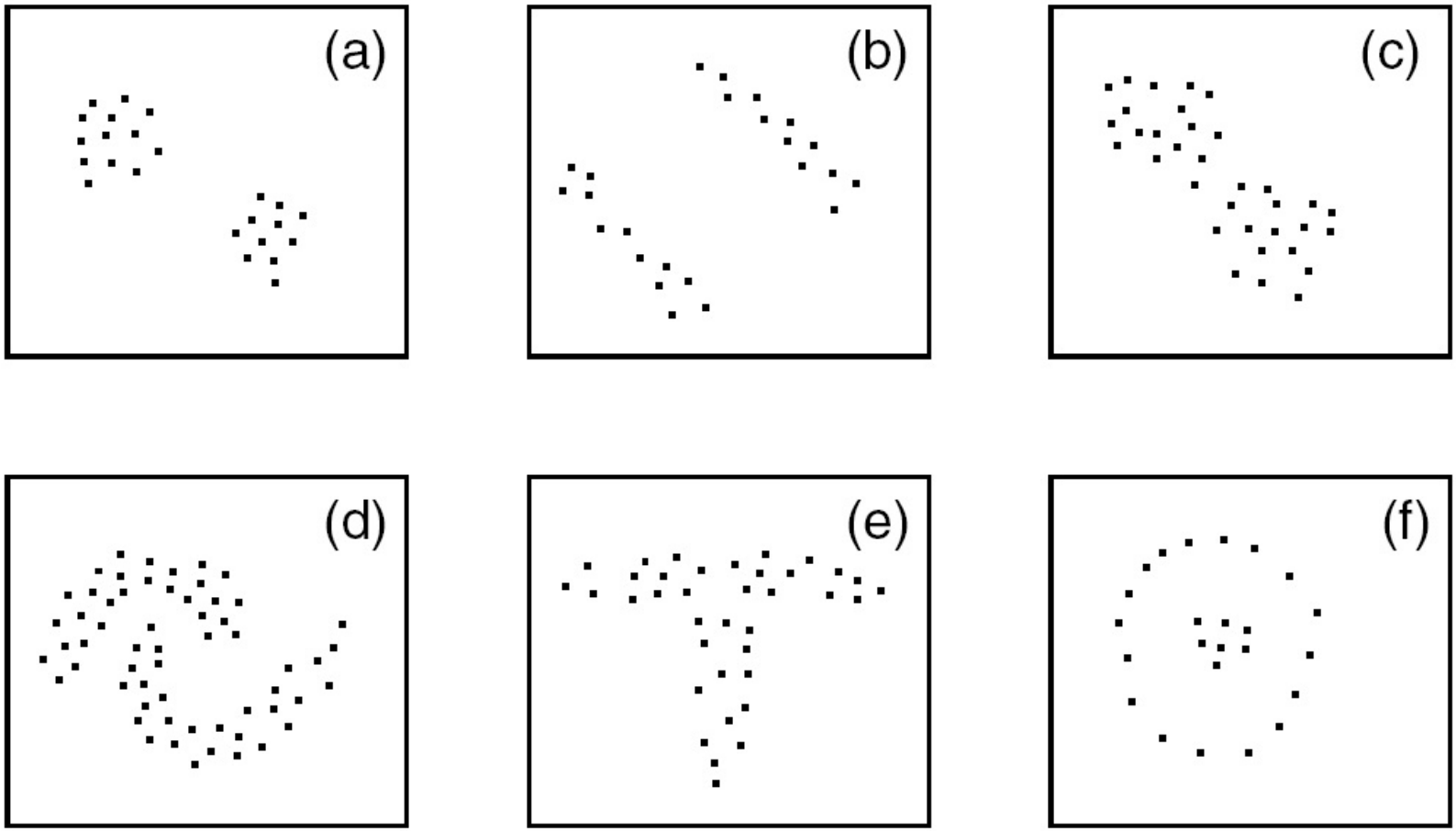
- Agrupamento igual ao do método do vizinho mais distante
- Os valores das distâncias são maiores do que dos outros métodos porque utiliza as distâncias euclidianas ao quadrado (como o centróide)

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) - Incluir a fórmula e breve explicação, por ser o mais famoso
  - método de Ward (correlação semiparcial)
  - estatística pseudo T^2 (DUDA; HART, 1973)
  - estatística CCC
- Pacote `NbClust` do `R` retorna o número de grupos ideal segundo 26 critérios

Métodos hierárquicos divisivos

Esses métodos não serão vistos por não serem muito utilizados.