Análise de Agrupamento

Cluster Analysis

Objetivo

- Classificar objetos, itens ou indivíduos de acordo com as suas semelhanças.
- Os objetos semelhantes são alocados em um mesmo grupo e, portanto, aqueles que pertencem a diferentes grupos são considerados dissimilares.
- A semelhança é quantificada por meio de medidas de similaridade/dissimilaridade.

Características

- Os objetos são agrupados em um número desconhecido de grupos.
- O número de grupos não é conhecido *a priori.*
  - Os componentes dos grupos não são conhecidos.

Os métodos de agrupamentos

- **Hierárquicos** – os elementos-objetos de um grupo tornam-se elementos do grupo superior, constituindo um série hierarquizada.
  - Aglomerativos
    - Inicia-se com *n* grupos, cada um contendo um dos objetos e termina-se com um único grupo contendo todos os objetos.
    - Em cada passo, cada objeto ou grupo é reunido com outro objeto ou grupo.
  - Divisivos
    - Inicia-se com um único grupo contendo os *n* objetos e termina-se com *n* grupos de um único objeto cada.

Os métodos de agrupamentos

- **Não hierárquicos**
  - Define-se o número *k* de grupos inicialmente e, então, aloca-se os n objetos aos k grupos de maneira otimizada. Nestes casos, sempre usa-se uma alocação arbitrária para iniciar o processo e interativamente buscar a alocação ótima.
  - No agrupamento ótimo, os objetos de um determinado grupo possuem distâncias mínimas para o centróide (média do grupo), em relação aos centróides dos demais grupos.

Problemas nos agrupamentos

1. Escolha da medida de proximidade mais adequada para quantificar a semelhança entre os objetos.
2. Determinação do método que será utilizado para construir os grupos.

Com base nas medidas de proximidades, os objetos são alocados aos grupos de forma que as distâncias entre os grupos são grandes e os elementos dentro dos grupos estão o mais próximo possível uns dos outros.

Agrupamentos hierárquicos

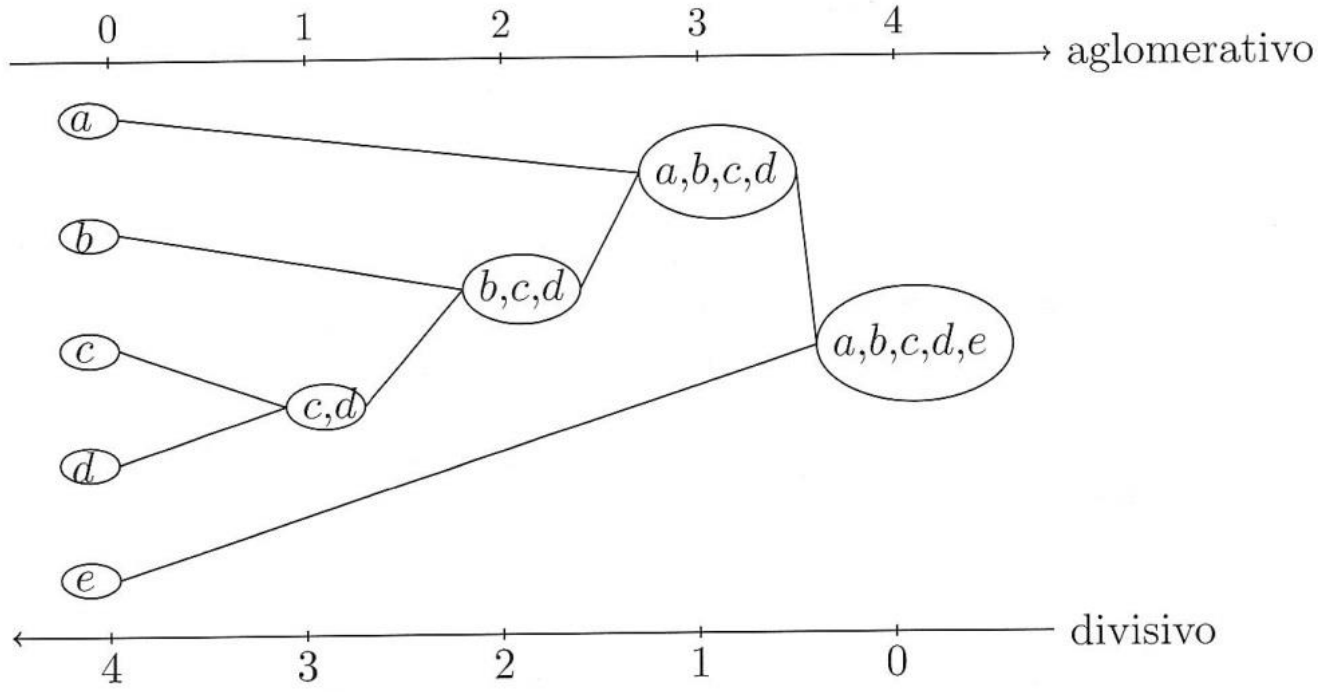

Figura 9.2. Distinção entre os métodos hierárquicos aglomerativos e divisivos, mostrando os sentidos de formação dos grupos.

Métodos de agrupamento

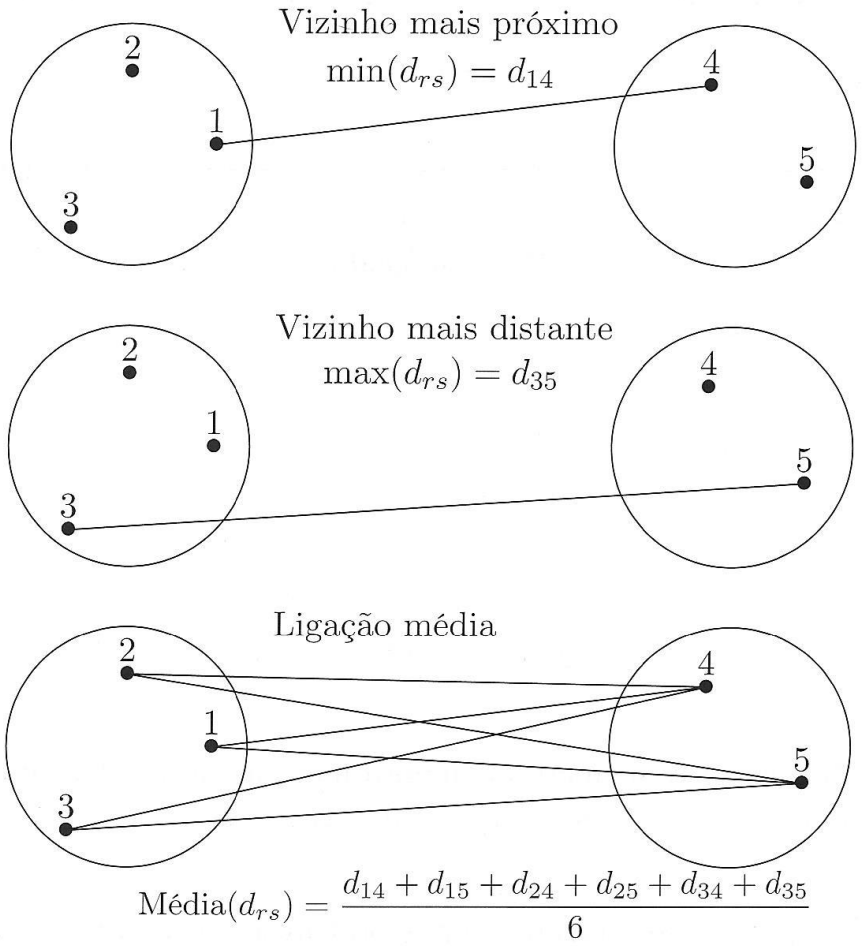

Figura 9.3. Distância entre dois grupos para os métodos de agrupamento do vizinho mais próximo, vizinho mais distante e ligação média.

Algoritmo

1. Iniciar a análise com n grupos, cada um deles com um único objeto, considerando a matriz D simétrica n x n de dissimilaridades (distâncias);
2. Identificar na matriz D o par de grupos mais similar R e S (menor distância);
3. Fundir os grupos mais similares RS e rearranjar as distâncias da matriz D da seguinte forma:
    1. Acrescentar uma nova linha e uma nova coluna com as distâncias entre o grupo recém-formado RS e os grupos remanescentes.
4. Repetir os passos 2 e 3 até que todos os objetos formem um único grupo.

Exemplo

	Variáveis							
Objetos	1	2	3	4	5	6	7	8
1	1	1	1	1	0	1	1	1
2	0	1	0	0	1	0	0	1
3	0	1	0	0	1	0	1	1
4	0	0	0	0	1	1	0	0
5	0	0	0	0	1	0	0	0

Como fazer a matriz de distâncias?

Tabela 9.3. Tabela de contingência para dados binários retratando o número de concordâncias e de discordâncias entre a i-ésima e k-ésima variáveis.

		Variável k		
		1	0	Total
Variável	1	a	b	$a+b$
i	0	c	d	$c+d$
Total		$a+c$	$b+d$	$n=a+b+c+d$

	Variáveis							
Objetos	1	2	3	4	5	6	7	8
1	1	1	1	1	0	1	1	1
2	0	1	0	0	1	0	0	1
3	0	1	0	0	1	0	1	1
4	0	0	0	0	1	1	0	0
5	0	0	0	0	1	0	0	0

Matriz de distâncias entre os objetos

A matriz de distâncias $\boldsymbol{D} = [d_{rs}]$ entre os objetos, sendo $d_{rs} = \sqrt{(b+c)/p}$, é dada por

Distância de Sokal
p = número de variáveis

Agrupamento hierárquico do vizinho mais próximo – *single linkage*

A matriz de distâncias $\boldsymbol{D} = [d_{rs}]$ entre os objetos, sendo $d_{rs} = \sqrt{(b+c)/p}$, é dada por

$$\boldsymbol{D} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0{,}0000 & 0{,}8660 & 0{,}7906 & 0{,}9354 & 1{,}0000 \\ 0{,}8660 & 0{,}0000 & 0{,}3536 & 0{,}6124 & 0{,}5000 \\ 0{,}7906 & 0{,}3536 & 0{,}0000 & 0{,}7071 & 0{,}6124 \\ 0{,}9354 & 0{,}6124 & 0{,}7071 & 0{,}0000 & 0{,}3536 \\ 1{,}0000 & 0{,}5000 & 0{,}6124 & 0{,}3536 & 0{,}0000 \end{bmatrix} \end{array}.$$

Agrupamento hierárquico do vizinho mais próximo – *single linkage*

A matriz de distâncias $\boldsymbol{D} = [d_{rs}]$ entre os objetos, sendo $d_{rs} = \sqrt{(b+c)/p}$, é dada por

$$\boldsymbol{D} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0{,}0000 & 0{,}8660 & 0{,}7906 & 0{,}9354 & 1{,}0000 \\ 0{,}8660 & 0{,}0000 & 0{,}3536 & 0{,}6124 & 0{,}5000 \\ 0{,}7906 & 0{,}3536 & 0{,}0000 & 0{,}7071 & 0{,}6124 \\ 0{,}9354 & 0{,}6124 & 0{,}7071 & 0{,}0000 & 0{,}3536 \\ 1{,}0000 & 0{,}5000 & 0{,}6124 & 0{,}3536 & 0{,}0000 \end{bmatrix} \end{array}.$$

$$\boldsymbol{D} = \begin{array}{c} \\ 23 \\ 1 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} 23 & 1 & 4 & 5 \end{array} \\ \begin{bmatrix} 0{,}0000 & 0{,}7906 & 0{,}6124 & 0{,}5000 \\ 0{,}7906 & 0{,}0000 & 0{,}9354 & 1{,}0000 \\ 0{,}6124 & 0{,}9354 & 0{,}0000 & 0{,}3536 \\ 0{,}5000 & 1{,}0000 & 0{,}3536 & 0{,}0000 \end{bmatrix} \end{array};$$

Agrupamento hierárquico do vizinho mais próximo – *single linkage*

$$D = \begin{array}{c} \\ 23 \\ 1 \\ 4 \\ 5 \end{array} \begin{bmatrix} 0{,}0000 & 0{,}7906 & 0{,}6124 & 0{,}5000 \\ 0{,}7906 & 0{,}0000 & 0{,}9354 & 1{,}0000 \\ 0{,}6124 & 0{,}9354 & 0{,}0000 & 0{,}3536 \\ 0{,}5000 & 1{,}0000 & 0{,}3536 & 0{,}0000 \end{bmatrix}; \quad \begin{array}{cccc} 23 & 1 & 4 & 5 \end{array}$$

$$D = \begin{array}{c} 45 \\ 23 \\ 1 \end{array} \begin{bmatrix} 0{,}0000 & 0{,}5000 & 0{,}9354 \\ 0{,}5000 & 0{,}0000 & 0{,}7906 \\ 0{,}9354 & 0{,}7906 & 0{,}0000 \end{bmatrix}; \quad \begin{array}{ccc} 45 & 23 & 1 \end{array}$$

Agrupamento hierárquico do vizinho mais próximo – *single linkage*

$$D = \begin{array}{c} \\ 45 \\ 23 \\ 1 \end{array} \begin{bmatrix} 0{,}0000 & 0{,}5000 & 0{,}9354 \\ 0{,}5000 & 0{,}0000 & 0{,}7906 \\ 0{,}9354 & 0{,}7906 & 0{,}0000 \end{bmatrix}; \qquad \begin{array}{ccc} 45 & 23 & 1 \end{array}$$

$$D = \begin{array}{c} \\ 2345 \\ 1 \end{array} \begin{bmatrix} 0{,}0000 & 0{,}7906 \\ 0{,}7906 & 0{,}0000 \end{bmatrix}; \qquad \begin{array}{cc} 2345 & 1 \end{array}$$

Agrupamento hierárquico do vizinho mais próximo – *single linkage*

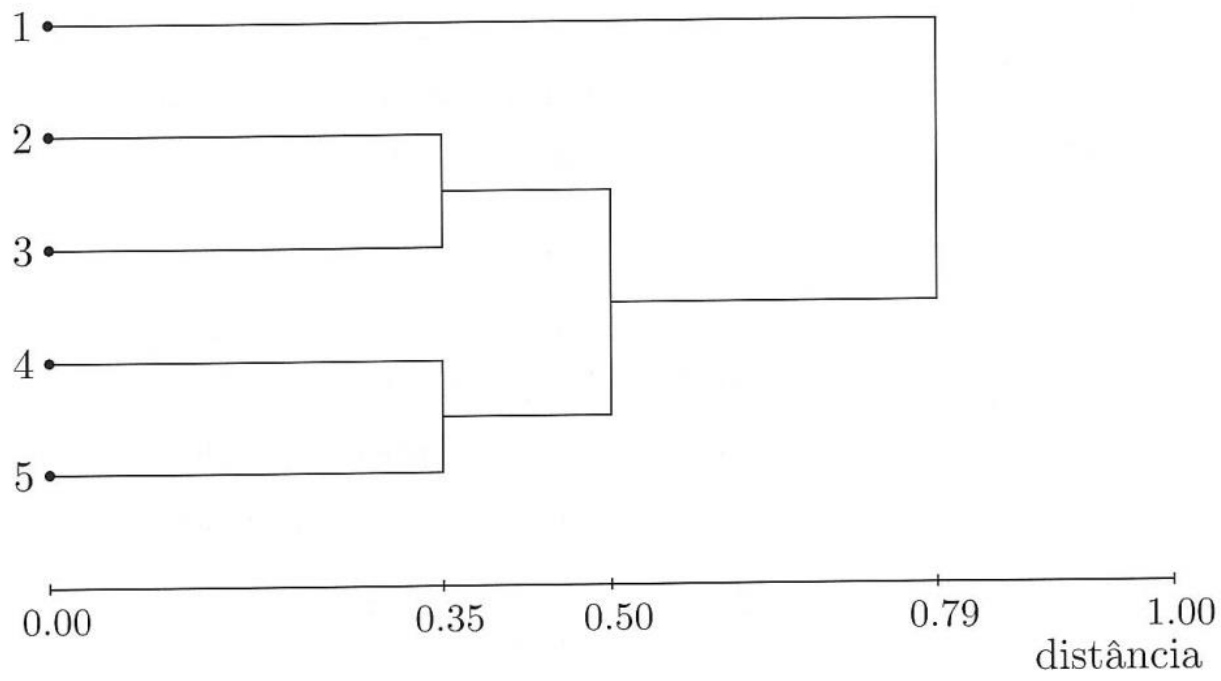

Figura 9.4. Dendrograma para o agrupamento de 5 objetos mensurados a partir de dados binários e da distância de Sokal como medida de dissimilaridade, utilizando o método do vizinho mais próximo.

Exercício

- Calcular os agrupamentos pelos métodos do vizinho mais distante e da ligação média

Agrupamento hierárquico do vizinho mais distante – *complete linkage*

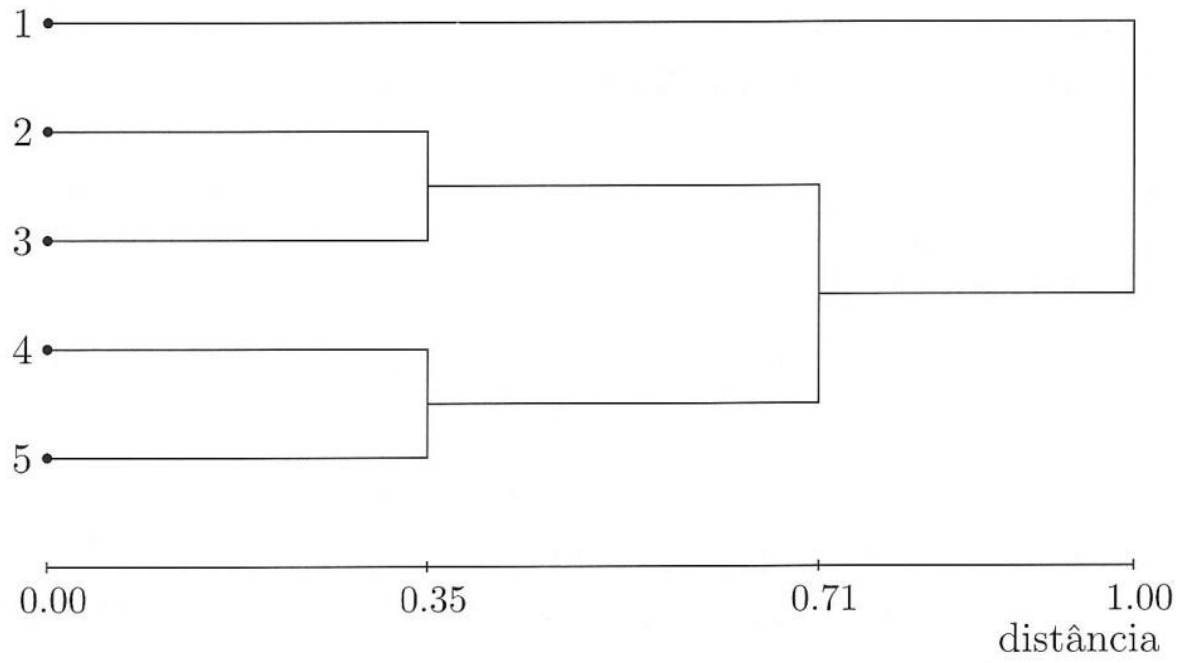

Figura 9.5. Dendrograma para o agrupamento de 5 objetos mensurados a partir de dados binários e da distância de Sokal como medida de dissimilaridade, utilizando o método do vizinho mais distante.

Agrupamento hierárquico da ligação média – *average linkage ou Unweighted Pair-Group Average: UPGMA*

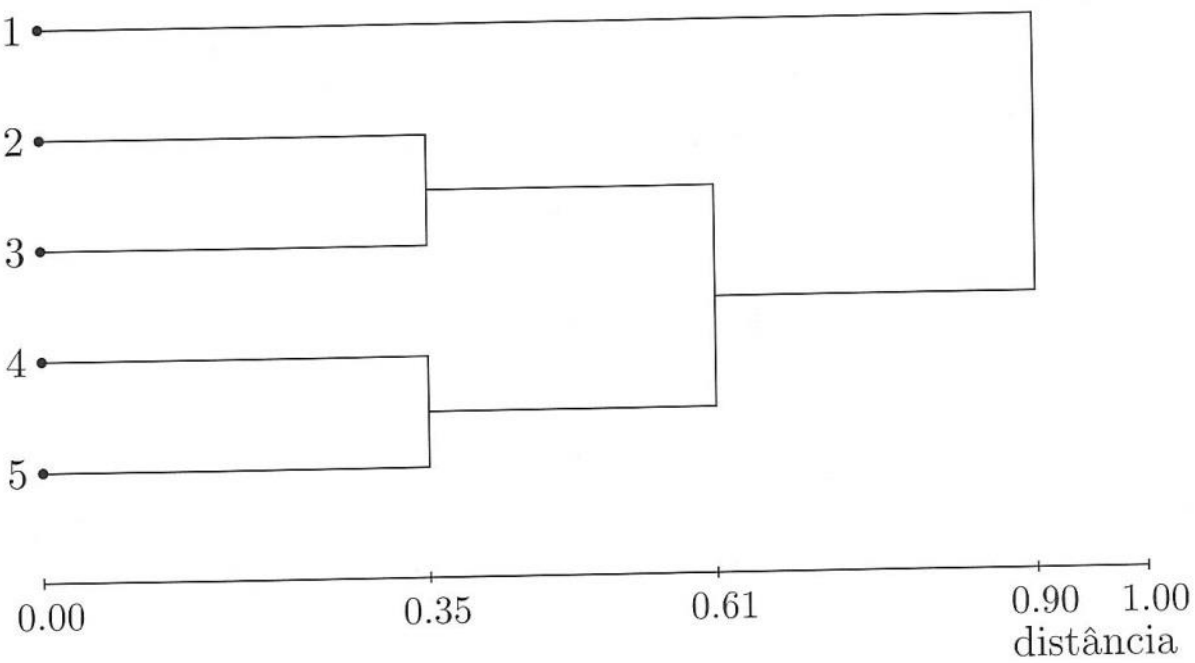

Figura 9.6. Dendrograma para o agrupamento de 5 objetos mensurados a partir de dados binários e da distância de Sokal como medida de dissimilaridade, utilizando o método da ligação média.

Agrupamento hierárquico de Ward

- Método da variância mínima
- Um grupo será reunido a um outro se essa reunião proporcionar o menor aumento da variância intragrupo.
- A variância intragrupo será calculada para todas as alternativas de aglomeração, escolhendo a que proporciona o menor aumento na variância.

Agrupamento hierárquico de Ward

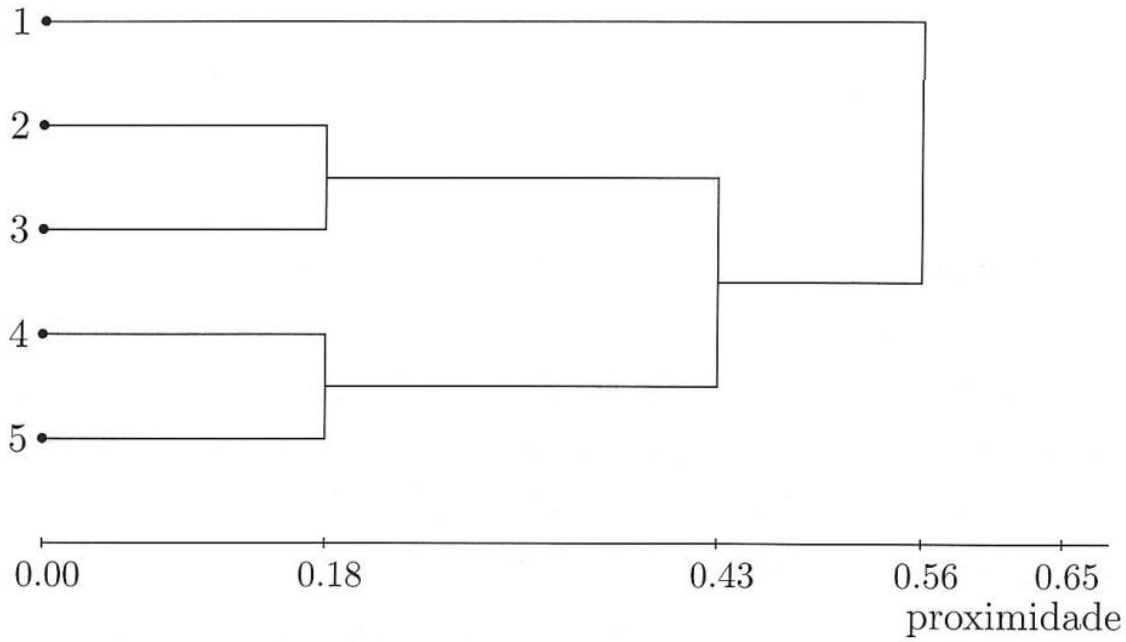

Figura 9.8. Dendrograma para o agrupamento de 5 objetos mensurados a partir de dados binários e da proximidade baseada na distância de Sokal como medida de dissimilaridade, utilizando o método de Ward (1963).

Determinação do número de grupos

- Nos métodos hierárquicos podemos escolher o número de grupos *k*, escolhendo um ponto de corte na escala de distância e cortando os ramos do dendrograma.
  - Mas qual o critério?

Como escolher o número de grupos?

- O número de grupos deve otimizar a qualidade do ajuste do agrupamento.
- A ideia é buscar a maior amplitude nas distâncias de junção dos grupos formados.

Como escolher o número de grupos

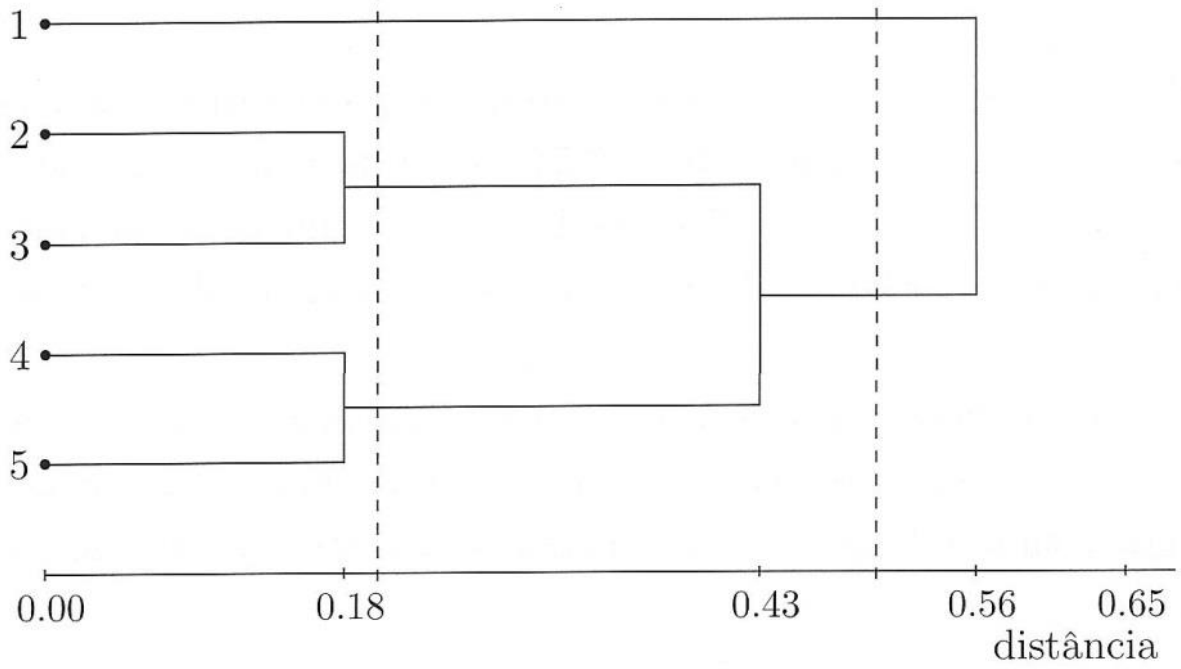

Figura 9.9. Dendrograma para o agrupamento de 5 objetos mensurados a partir de dados binários e da distância de Sokal como medida de dissimilaridade, utilizando o método de Ward e destacando dois diferentes cortes.

Como escolher o número de grupos

- Verifique que na figura anterior na migração de 5 grupos para 3 grupos, a amplitude foi 0,18; de 3 grupos para 2 grupos, foi 0,25; e de 2 grupos para 1 grupo, 0,13. Nesse caso devemos escolher k=3 grupos.

Validação do agrupamento

- Validar o agrupamento significa certificar-se que os grupos realmente diferem uns dos outros.
- Para métodos hierárquicos pode-se utilizar o **coeficiente de correlação cofenética** (*C*).
- Esse coeficiente mede o grau de preservação das distâncias emparelhadas pelo dendrograma resultante do agrupamento em relação às distâncias originais.
- A ideia é obter a correlação entre as distâncias observadas entre os objetos e as distâncias recuperadas da análise de agrupamento.
- A matriz de distâncias recuperadas é denominada matriz cofenética.

Matriz cofenética

- O primeiro grupo a ser aglomerado foi dado pelos objetos 2 e 3. Assim, a proximidade entre eles foi 0,35 (distância de junção dos grupos);
- Depois aglomerou-se os objetos 4 e 5 a 0,35 de distância;
- No estágio seguinte os grupos (23) e (45) foram aglomerados na distância 0,50; Portanto, a proximidade recuperada entre os objetos 2 e 4, 2 e 5, 3 e 4, 3 e 5 são todas iguais a 0,50.
- Com isso montamos a matriz cofenética de proximidades.

Agrupamento pelo vizinho mais próximo

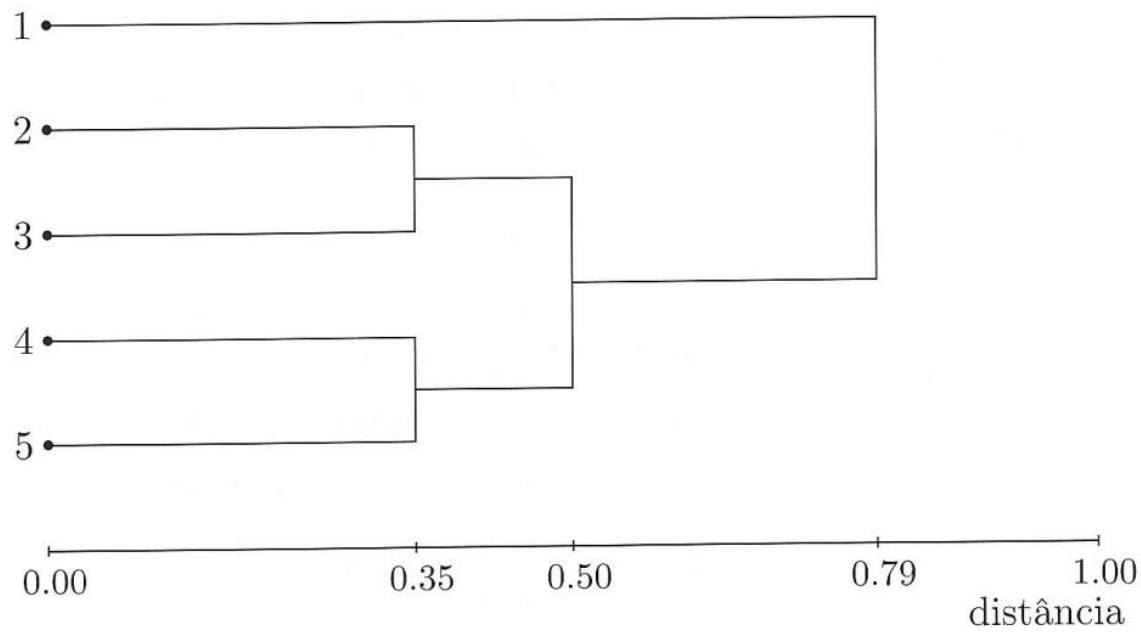

Figura 9.4. Dendrograma para o agrupamento de 5 objetos mensurados a partir de dados binários e da distância de Sokal como medida de dissimilaridade, utilizando o método do vizinho mais próximo.

Matriz cofenética

Distâncias recuperadas

	1	2	3	4	5
1	0	0,79	0,79	0,79	0,79
2		0	0,35	0,5	0,5
3			0	0,5	0,5
4			0	0	0,35
5					0

A matriz de distâncias $\boldsymbol{D} = [d_{rs}]$ entre os objetos, sendo $d_{rs} = \sqrt{(b+c)/p}$, é dada por

$$\boldsymbol{D} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0{,}0000 & 0{,}8660 & 0{,}7906 & 0{,}9354 & 1{,}0000 \\ 0{,}8660 & 0{,}0000 & 0{,}3536 & 0{,}6124 & 0{,}5000 \\ 0{,}7906 & 0{,}3536 & 0{,}0000 & 0{,}7071 & 0{,}6124 \\ 0{,}9354 & 0{,}6124 & 0{,}7071 & 0{,}0000 & 0{,}3536 \\ 1{,}0000 & 0{,}5000 & 0{,}6124 & 0{,}3536 & 0{,}0000 \end{bmatrix} \end{array}.$$

Cálculo do coeficiente de correlação cofenética

Objetos		Distância recuperada	Distância original
1	2	0,79	0,86
1	3	0,79	0,79
1	4	0,79	0,93
1	5	0,79	1
2	3	0,35	0,35
2	4	0,5	0,61
2	5	0,5	0,5
3	4	0,5	0,70
3	5	0,5	0,61
4	5	0,35	0,35
Correlação cofenética (>0,8)			**0,9404**

Estudo Dirigido III

1. Faça uma Análise de Agrupamento com a matriz SimilaridadePCA.
2. Pesquisar sobre o uso de *Bootstrap*.