Análise de Agrupamento (*Cluster analysis*)

Anderson Rodrigo da Silva

Exemplos de aplicações de análise de agrupamento

- Pesquisas de mercado
  - Agrupamento de “cidades-teste”
- Bancos de germoplasma
  - Caracterização
  - Estudos de divergência ou diversidade genética
- Biologia
  - Agrupamento de espécies ou unidades de conservação
- Educação
  - Agrupamento de escolas, professores, alunos...

Exemplo 1: cães pré-históricos da Tailândia

Escavações na Tailândia produziram ossos caninos cobrindo um período em torno de 3500 a.C. até o presente. Entretanto, a origem desses cães é incerta. Para tentar esclarecer, medidas de espécimes disponíveis foram tomadas:

Grupo	LM	AMAPM	CPM	LPM	CPTM	CPQM
cão moderno	9.7	21	19.4	7.7	32	36.5
chacal dourado	8.1	16.7	18.3	7	30.3	32.9
lobo chinês	13.5	27.3	26.8	10.6	41.9	48.1
lobo indiano	11.5	24.3	24.5	9.3	40	44.6
cuon	10.7	23.5	21.4	8.5	28.8	37.6
dingo	9.6	22.6	21.1	8.3	34.4	43.1
cao pre-historico	10.3	22.1	19.1	8.1	32.2	35

LM: largura da mandíbula, AMAPM: Altura da mandíbula abaixo do primeiro molar, CPM: comprimento do primeiro molar, LMP: largura do primeiro molar, CPTM: comprimento do primeiro ao terceiro molar, CPQM: comprimento do primeiro ao quarto molar

Medidas de distâncias multivariadas

- Tipos de dados: valores de p variáveis tomados em n objetos ("matriz X").
- As medidas devem ser escolhidas de acordo com os tipos de variáveis.
  - Quantitativas: euclidiana, euclidiana média, Mahalanobis, Manhattan, etc.
  - Padrão binário: coeficiente de Jaccard, coeficiente de Roger, etc.
  - Padrão multicategórico: coeficiente de coincidência simples, dissimilaridade de Cole-Rodgers
  - Para os 3 tipos, simultaneamente: coeficiente de Gower (1971)
- É recomendável que se faça uma padronização das variáveis de modo que estas sejam igualmente importantes na determinação das distâncias.

Exemplo 1: cães pré-históricos da Tailândia

Matriz de distâncias euclidianas

	(1) c_md	(2) chc_	(3) lb_c	(4) lb_n	(5) uon	(6) ing	(7) c_p
(1) c_md	0						
(2) chc_	6	0					
(3) lb_c	19	24	0				
(4) lb_n	13	19	6	0			
(5) Uon	5	9	18	14	0		
(6) Ing	7	13	13	7	8	0	
(7) c_p	2	7	19	14	5	9	0

Tipos de métodos de agrupamento

- Hierárquicos aglomerativos
  - Vizinho mais próximo (ligação simples)
  - Vizinho mais distante (ligação completa)
  - UPGMA (ligação média)

- Não hierárquicos
  - Algoritmo K-médias
  - Tocher, Tocher modificado

Métodos hierárquicos

Resumo dos passos:

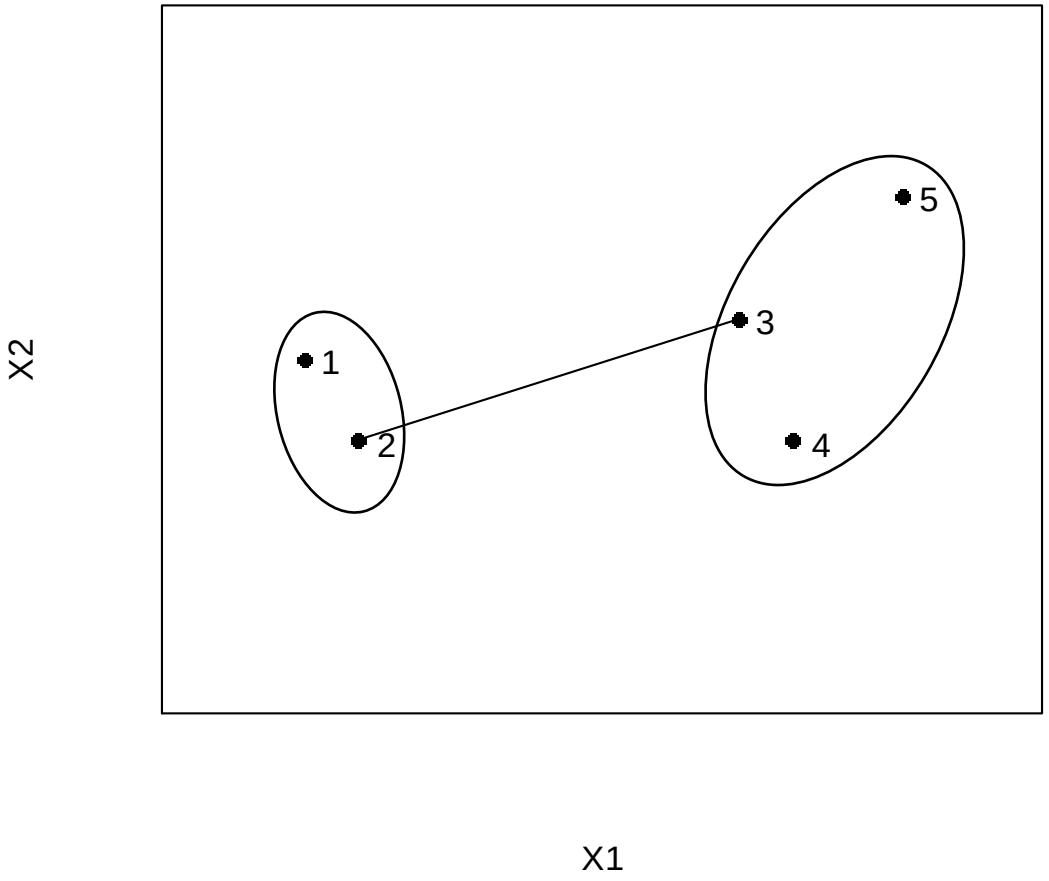
1) Cada indivíduo constitui um cluster de tamanho 1 → n clusters.

2) Em cada estágio do algoritmo pares de "entidades" são combinados e constituem um novo conglomerado.

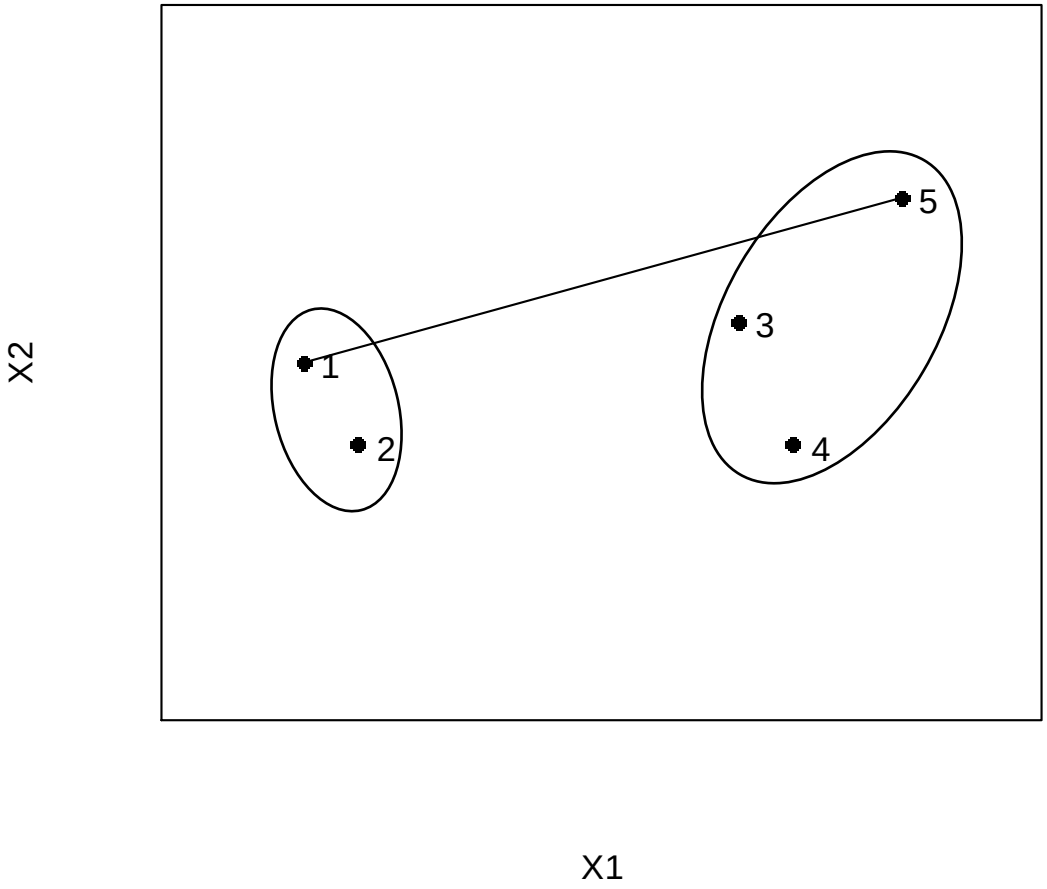
3) Propriedade de hierarquia: cada novo conglomerado é um agrupamento de conglomerados antes formados.

4) Construção do dendrograma ou árvore da "história" do agrupamento.

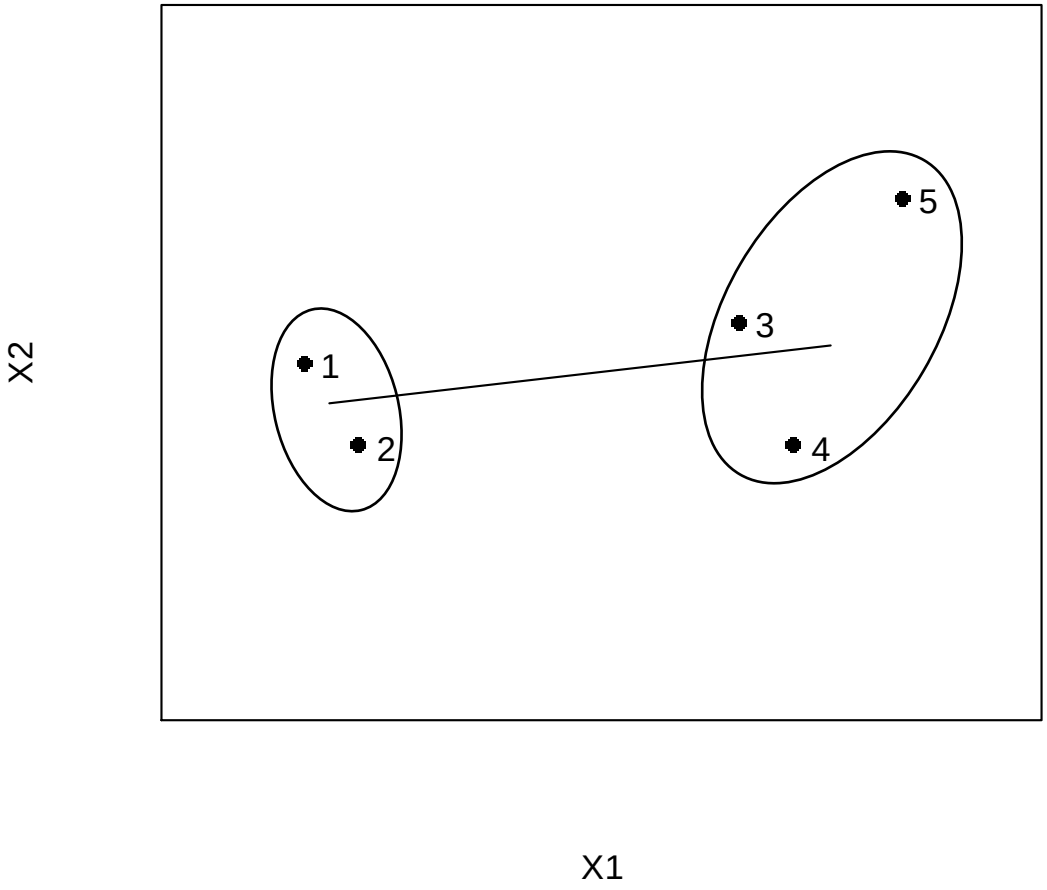
Método do vizinho mais próximo



Método do vizinho mais distante



Método da ligação média (UPGMA)



Exemplo 2 (p.141, Manly 2005)

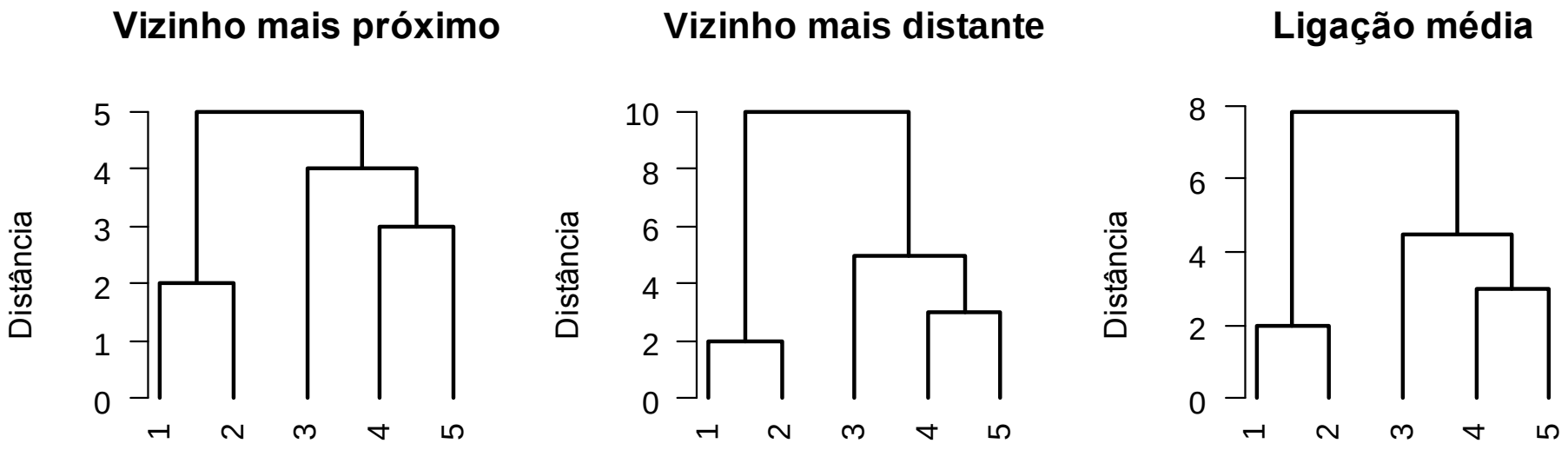
Tabela 9.1 - Matriz de distâncias entre cinco objetos.

Objeto	Objeto				
	1	**2**	**3**	**4**	**5**
1	0				(Sim.)
2	2	0			
3	6	5	0		
4	10	9	4	0	
5	9	8	5	3	0

Fonte: Manly, 2008

Método	Função objetivo
Vizinho mais próximo	$d_{ij,k} = \text{mín}(d_{ik}, d_{jk})$
Vizinho mais distante	$d_{ij,k} = \text{máx}(d_{ik}, d_{jk})$
Ligação média	$d_{ij,k} = \text{média}(d_{ik}, d_{jk})$

Dendrogramas



Critérios para encontrar o número de grupos

1) Comportamento dos níveis de fusão

2) Nível de similaridade

3) Alguns critérios objetivos: R^2, Pseudo F, Pseudo T^2, Mojena (1977), etc.

Correlação cofenética

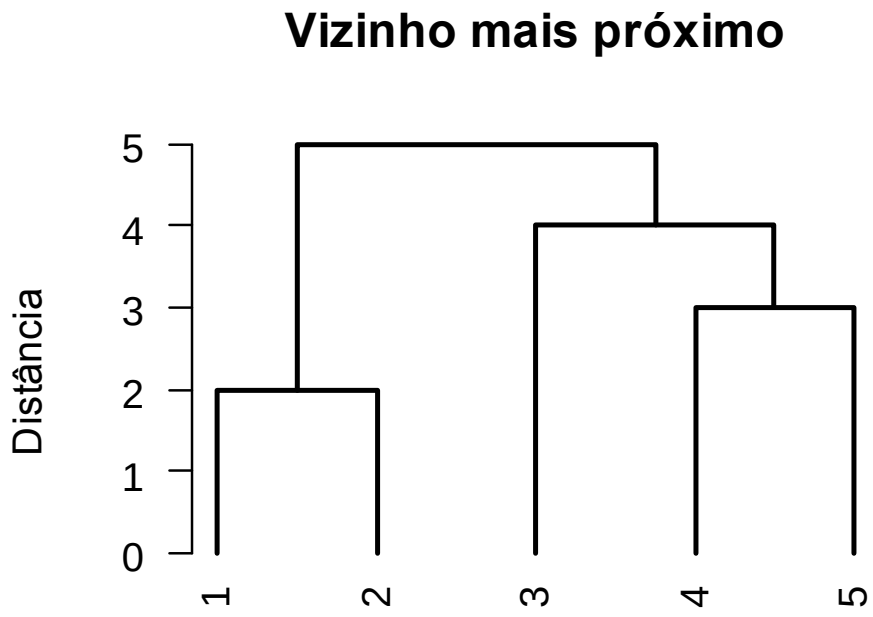
Distâncias originais

	1	2	3	4
2	2			
3	6	5		
4	10	9	4	
5	9	8	5	3

Cor = 0.82

Distâncias cofenéticas

	1	2	3	4
2	2			
3	5	5		
4	5	5	4	
5	5	5	4	3



Exercícios

1) Construa um dendrograma pelo método do vizinho mais distante a partir da matriz de distancias euclideanas dos dados de medidas das mandíbulas de cães; Determine grupos de cães; Avalie a qualidade do agrupamento

2) No R, construa a matriz de distancias multivariadas dos dados proteinas.txt (do site, http://arsilva.weebly.com/uploads/2/1/0/0/21008856/proteinas.txt). Encontre grupos de países semelhantes em relação a fonte proteica base da alimentação.

Exemplo ACP vs AG

Matriz de dados (simulados) padronizados de 10 objetos e 4 variáveis.

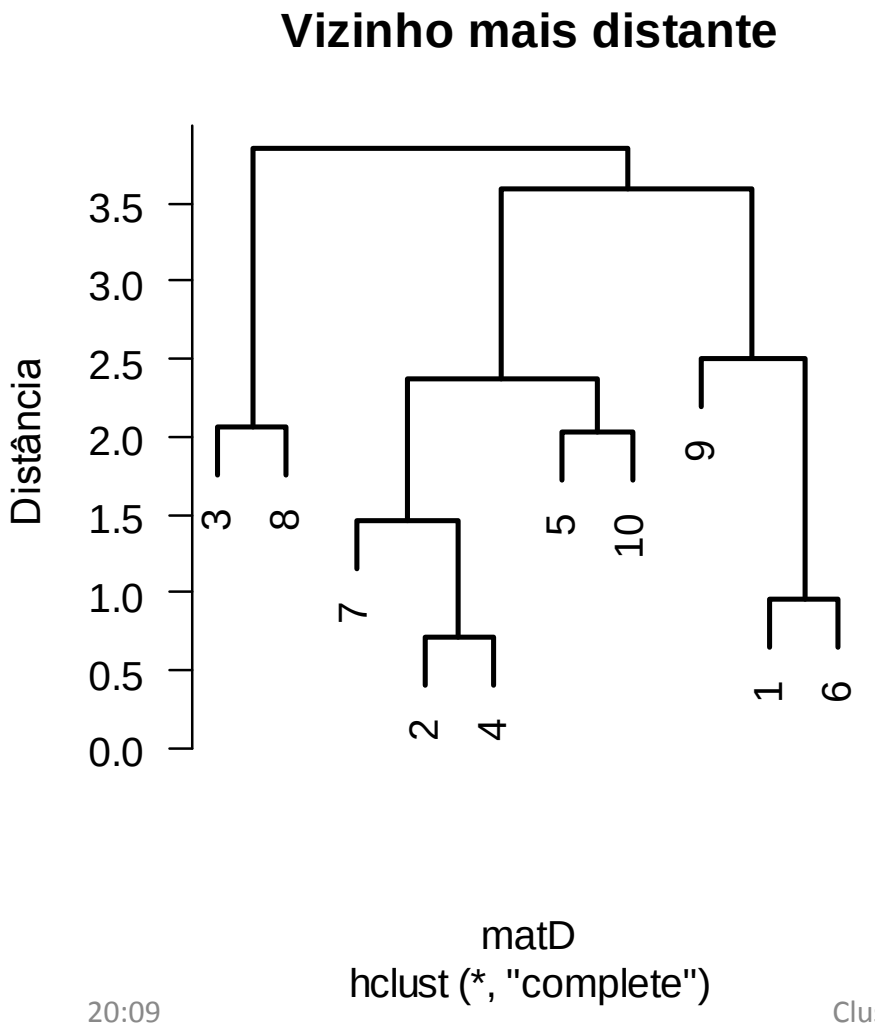
	x1	x2	x3	x4
[1,]	-0.14	0.17	-0.44	1.58
[2,]	-0.04	-0.03	0.29	0.16
[3,]	1.01	1.88	0.72	-0.28
[4,]	-0.16	0.24	0.46	0.79
[5,]	-2.16	0.70	0.19	-0.22
[6,]	0.50	-0.02	0.23	1.39
[7,]	-0.76	-0.14	0.59	-0.49
[8,]	0.78	0.32	2.00	0.14
[9,]	0.75	0.12	-1.84	0.00
[10,]	-1.10	-0.59	-0.86	-0.73

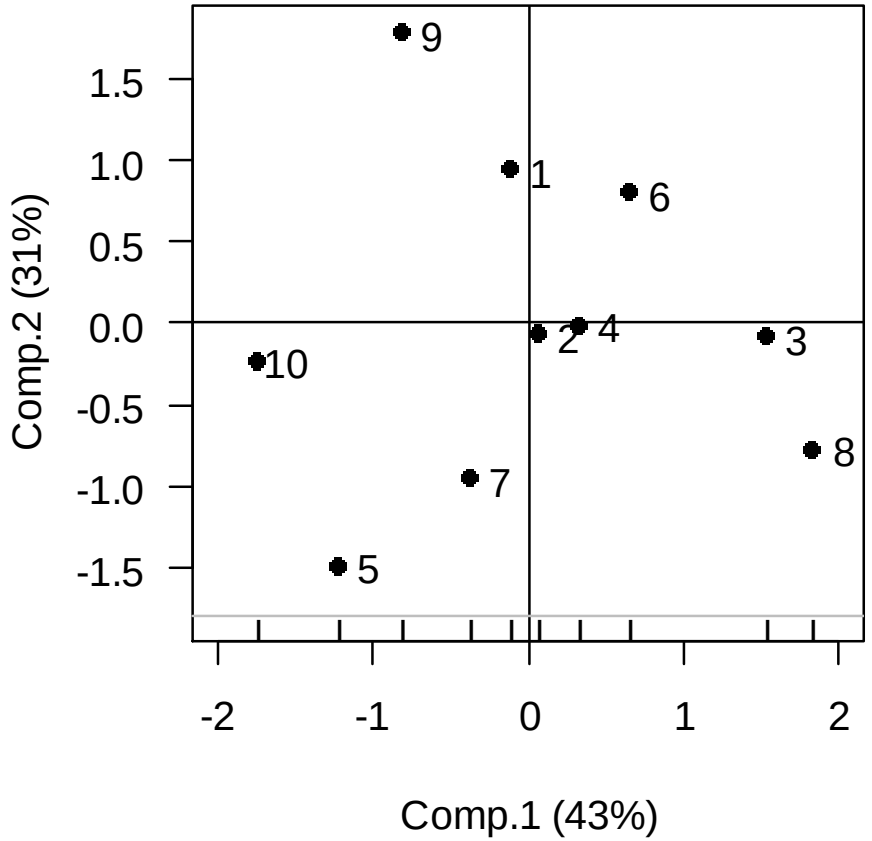
Exemplo ACP vs AG

Matriz de distâncias euclidianas entre 10 objetos.

	1	2	3	4	5	6	7	8	9
2	1.62								
3	3.01	2.26							
4	1.21	0.72	2.29						
5	2.83	2.27	3.42	2.30					
6	0.96	1.35	2.62	0.96	3.19				
7	2.42	1.01	2.69	1.47	1.71	2.29			
8	2.99	1.93	2.07	1.92	3.49	2.21	2.22		
9	2.29	2.28	3.13	2.60	3.60	2.51	2.92	3.85	
10	2.65	1.88	3.64	2.37	2.04	2.93	1.58	3.65	2.33

Exemplo ACP vs AG





Algoritmo k-médias

- Não hierárquico
- Processo iterativo
- Resumo dos passos:
  1) Escolhe-se k centróides para iniciar o processo de partição.
  2) Cada um dos n objetos é comparado com cada centróide, em geral usando a distância euclidiana. O elemento é alocado ao grupo cuja distância é a menor.
  3) Recalcula-se os valores dos centróides para os novos grupos e repete-se o passo 2.
  4) Os passos 2 e 3 são repetidos até que nenhuma realocação seja necessária.

Para análises no R

- Pacote: *stats*
- Funções: *hclust, cophenetic*
- Argumentos

```
hclust(d, method = "single", ...)
cophenetic(x)
```