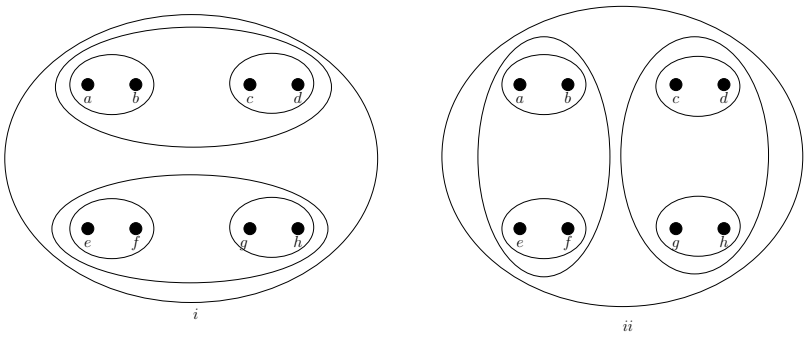
Análises classificatórias - exercícios (22/23)

Os exercícios assinalados com (*) foram parcialmente retirados ou modificados a partir de exercícios que aparecem na bibliografia citada nos slides. Alguns exercícios foram retirados de testes e exames da disciplina.

1. (*) Foram efetuadas duas análises classificatórias hierárquicas do conjunto de 8 pontos $X = \{a, b, c, d, e, f, g, h\}$ com a distância euclideana, tendo-se obtido as seguintes partições encadeadas de X.



   Indique, justificando, dois métodos hierárquicos que pudessem ter originado estas partições encadeadas e represente os respectivos dendrogramas (aproximados).

2. (*) Considere o conjunto de pontos na reta real,

$$X = \{0.2, 3, 4.2, 5, 5.9\}.$$

   (a) Efectue uma partição deste conjunto em dois grupos usando o método do vizinho mais afastado (*complete*) com a distância euclideana e represente o respectivo dendrograma. Comente.

   (b) Indique a respectiva matriz de distâncias cofenéticas.

   (c) Calcule os respectivos coeficientes de correlação cofenética de Pearson e Spearman.

3. (*) Efectue uma classificação hierárquica dos pontos

$$X = \{(1, 2), (2, 2), (4.5, 3), (6, 3)\},$$

   usando o método das distâncias médias entre grupos (*average*) e a distância de Manhattan.

4. (*) Aplique o método hierárquico do centroide ao conjunto de pontos do plano,

$$X = \{(0, 0), (8, 0), (4, 7.5)\}$$

   e represente o respectivo dendrograma. Comente.

5. A seguinte tabela contém os registos de presença (1) /ausência (0) relativos a 10 espécies de peixes em 4 bacias fluviais localizadas em África.

	SP1	SP2	SP3	SP4	SP5	SP6	SP7	SP8	SP9	SP10
OUEME	1	0	0	1	0	1	1	1	0	1
GAMBIE	1	0	1	0	1	1	0	0	0	1
GEBA	0	1	1	1	0	1	0	0	0	0
CRUBAL	0	1	0	0	1	1	0	0	0	0

Investigue se as bacias fluviais podem ser agregadas em grupos homogéneos quanto à presença das 10 espécies de peixes, utilizando o método hierárquico do vizinho mais afastado (*complete*) e uma medida de dissemelhança adequada.

6. A seguinte tabela contém as componentes de 5 vetores binários,a, b, c, d, e:

a:	1	0	0	1	1
b:	0	1	1	0	1
c:	1	0	0	0	1
d:	1	1	0	1	0
e:	0	1	1	0	0

Efetue análises classificatórias do conjunto dos vetores binários usando a distância de Manhattan e os métodos do vizinhos mais próximo (single) e mais afastado (complete). Comente.

7. Considere a tabela de contingência do slide 77.

   (a) Classifique os 5 países de acordo com o idioma principal falado em cada país, usando o método do vizinho mais afastado (complete) e uma distância apropriada.

   (b) Investigue se os 5 idiomas podem ser agregados em grupos homogéneos relativamente à sua distribuição pelos países, usando o método da inércia mínima (Ward) e uma distância adequada.

8. Classificou-se um conjunto formado por $N = 178$ vinhos com o algorítmo das k-médias móveis (k-means), considerando o número de grupos k a variar de 1 a 10. Foram obtidos os valores para das inércias intra-grupo (SSQ_w) em função do número de grupos descritos na tabela seguinte, tendo-se optado por formar 3 grupos.

k	1	2	3	4	5	6	7	8	9	10
SSQ_w	2301	1649	1271	1174	1116	1064	992	930	921	895

Posteriormente foram efetuadas classificações em 3 grupos com os métodos do vizinho mais próximo (*single*), do vizinho mais distante (*complete*), da média das distâncias entre grupos (*average*) e com o método da inércia mínima (*Ward*). As classificações em três grupos foram comparadas entre si usando o índice de Rand. Os resultados obtidos são apresentados na

tabela abaixo.

	single	complete	average	Ward
complete	0.3467			
average	0.9346	0.3495		
Ward	0.3445	0.8302	0.3448	
k−means	0.3460	0.8202	0.3467	0.9407097

(a) Justifque que a soma total dos quadrados das distâncias de cada ponto ao centro de gravidade da nuvem de pontos (SSQ_t) é igual ao valor da inércia intra-grupos (SSQ_w) para $k = 1$.

(b) Fundamente a opção de classificar os vinhos em 3 grupos usando dois critérios distintos.

(c) Sabendo que os métodos de classificação da inércia mínima (Ward) e das k-médias móveis (k-means) atribuem a mesma classe a 4767 pares de vinhos, quantos pares são classificados de forma discordante pelos dois métodos?

(d) Efetue uma análise classificatória hierárquica que permita agregar as classificações da tabela anterior em grupos homogéneos, usando uma medida de dissemelhança adequada e o método do vizinho mais afastado. Represente o respectivo dendrograma. e comente

9. Chama-se *diâmetro de um conjunto* C à maior dissemelhança entre pares de elementos de C, i.e., $\text{diam}(C) = \max_{x,y \in C} d(x, y)$. Chama-se *diâmetro* da partição $X = C_1 \cup \ldots \cup C_k$, ao maior dos diâmetros dos seus grupos, ou seja, a valor de $\max\{\text{diam}(C_1), \ldots, \text{diam}(C_k)\}$.

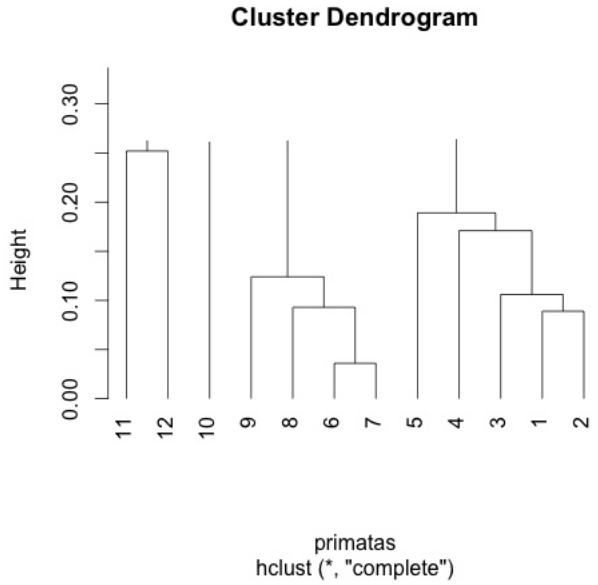
   No "DIMACS Workshop on Reticulated Evolution" organizado pela Universidade de Rutgers em Setembro de 2004, os investigadores P. Legendre e V. Makarenkov ilustraram um método para definir dissemelhanças entre espécies. Um exemplo apresentado pelos referidos autores diz respeito a dissemelhanças entre 12 espécies de primatas:

1. Homo sapiens	7. Macaca mulatta
2. Pan	8. Macaca fascicular.
3. Gorila	9. Macaca sylvanus
4. Pongo	10. Saimiri sciureus
5. Hylobatas	11. Tarsius syrichta
6. Macaca fuscata	12. Lemur catta

   Com base nestes dados obteve-se uma classificação hierárquica do conjunto dos primatas em 4 grupos, $C_1 = \{1, 2, 3, 4, 5\}$, $C_2 = \{6, 7, 8, 9\}$, $C_3 = \{10\}$ e $C_4 = \{11, 12\}$, usando o método do vizinho mais distante (*complete*) e a tabela de dissemelhanças abaixo, tendo-se obtido o dendrograma parcial abaixo

```
      1     2     3     4     5     6     7     8     9    10 11
2 0.089
```

```
3  0.104 0.106
4  0.161 0.171 0.166
5  0.182 0.189 0.189 0.188
6  0.232 0.243 0.237 0.244 0.247
7  0.233 0.251 0.235 0.247 0.239 0.036
8  0.249 0.268 0.262 0.262 0.257 0.084 0.093
9  0.256 0.249 0.244 0.241 0.242 0.124 0.120 0.123
10 0.273 0.284 0.271 0.284 0.269 0.289 0.293 0.287 0.287
11 0.322 0.321 0.314 0.303 0.309 0.314 0.316 0.311 0.319 0.320
12 0.308 0.309 0.293 0.293 0.296 0.282 0.289 0.298 0.287 0.285 0.252
```



(a) Indique os diâmetros (ver o exercício 7) dos grupos C_1, C_2, C_3 e C_4. Qual é o diâmetro da partição?

(b) Complete o dendrograma e comente a opção de formar 4 grupos.

(c) Considere a partição em 5 grupos de primatas definida pelo dendrograma.

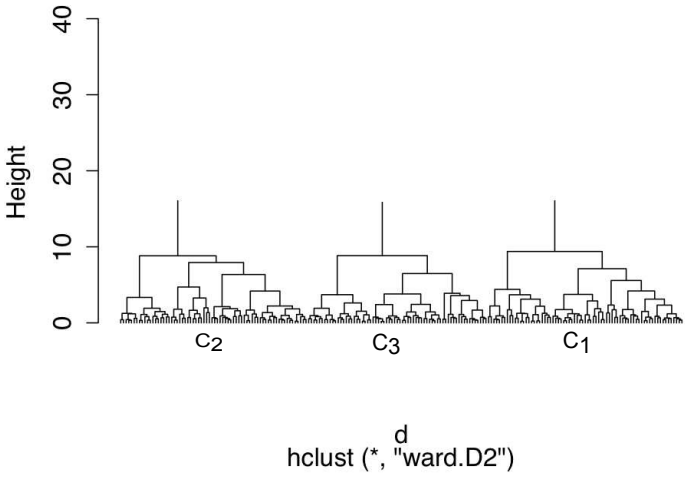
   i. Escreva estes grupos e indique o número de pares de primatas que seriam classificados de forma distinta pela nova partição em 5 grupos e pela partição em 4 grupos descrita acima.

   ii. Deduza da alínea anterior o valor do índice de RAND que se obteria comparando as duas partições.

10. Um estudo incidiu sobre sementes de três diferentes variedades de trigo: Kama, Rosa e Canadiano. Foram escolhidos ao acaso 70 sementes de cada uma das variedades, tendo sido observadas sete variáveis em cada semente:

Name	Description	Units
`Area`	Area (A)	mm^2
`Perimeter`	Perimeter (P)	mm
`Compacteness`	$\frac{4\pi A}{P^2}$	–
`Kernel_length`	Length	mm
`Kernel_width`	Width	mm
`asym_coeff`	Coefficient of assimetry	–
`length_kernel_groove`	Length of the groove	mm

Efetuou-se em seguida uma classificação hierárquica usando o método da inércia mínima (*ward*) com distância euclideana sobre a variáveis normalizadas do conjunto dos grãos de trigo. Efetuou-se depois um corte no dendrograma tendo sido obtido uma partição dos dados em 3 grupos, C_1, C_2 e C_3, contendo 73, 70 e 67 elementos, respectivamente, conforme indicado no dendrograma parcial abaixo.



Consolidou-se a partiç ao obtida anteriormente efectuando uma nova classficação sobre o mesmo conjunto de dados usando o método das k-médias móveis (*k-means*) em que as sementes iniciais foram os centroides dos grupos C_1, C_2 e C_3, tendo-se constatado que os dois métodos de classificação discordaram entre si relativamente a 1358 pares de grãos de trigo.

(a) Qual das classificações produziu grupos mais homogéneos? Justifique.

(b) Indique o valor do índice de RAND que se obteria comparando as duas partições.

(c) Sabendo que as distâncias entre os clusters C_1, C_2 e C_3, são dadas por

$$d(C_1, C_2) = 29.44, \qquad d(C_1, C_3) = 21.55, \qquad d(C_2, C_3) = 41.80,$$

complete o dendrograma indicando os custos de fusão dos grupos que agregar.

11. Um estudo envolveu gastrópodes marinhos (abalones) da espécie *Haliotis rubra*, tendo sido recolhidos ao acaso 4177 indivíduos: Para cada indivíviduo foram medidas 8 variáveis numéricas, e determinado o sexo (variável `Sex`), dentro de três categorias: macho (`M`), fêmea (`F`) e juvenil (`I`).

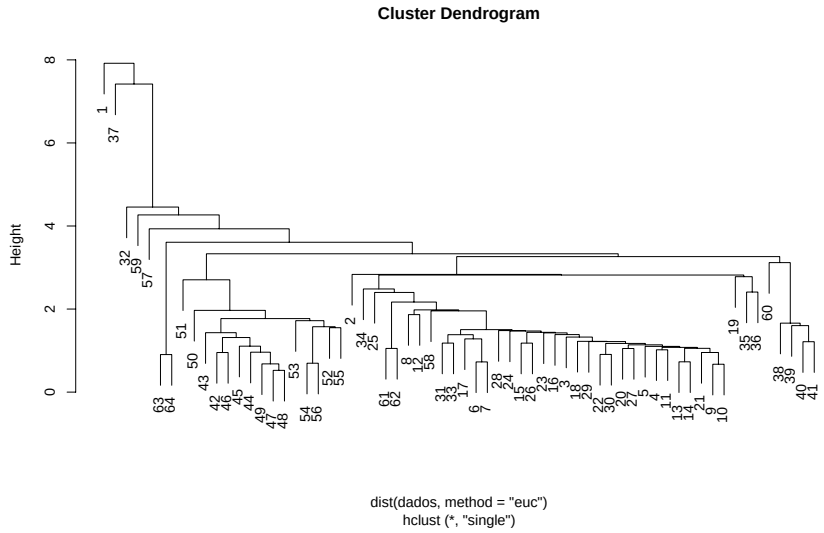
    Uma das oito variáveis numéricas é uma variável de contagens, a variável `Rings` que, através duma contagem de anéis, indica a idade. As restantes são variáveis contínuas: comprimento (`Length`); diâmetro (`Diameter`); e altura (`Height`); - todas em mm - e peso total do organismo (`Whole`); peso do animal sem a concha (`Shucked`); peso das vísceras (`Viscera`); e peso da concha seca (`Shell`) - estas últimas em g. Eis a matriz de correlações das variáveis numéricas e uma imagem dos gastrópodes.

	Length	Diameter	Height	Whole	Shucked	Viscera	Shell	Rings
Length	1.000	0.987	0.828	0.925	0.898	0.903	0.898	0.557
Diameter	0.987	1.000	0.834	0.925	0.893	0.900	0.905	0.575
Height	0.828	0.834	1.000	0.819	0.775	0.798	0.817	0.557
Whole	0.925	0.925	0.819	1.000	0.969	0.966	0.955	0.540
Shucked	0.898	0.893	0.775	0.969	1.000	0.932	0.883	0.421
Viscera	0.903	0.900	0.798	0.966	0.932	1.000	0.908	0.504
Shell	0.898	0.905	0.817	0.955	0.883	0.908	1.000	0.628
Rings	0.557	0.575	0.557	0.540	0.421	0.504	0.628	1.000

    (a) Usando uma medida de dissemelhança conveniente e o método do vizinho mais próximo (*single*) efectue uma classificação do conjunto das 6 variáveis contínuas, *comprimento*, *diâmetro*, *altura*, *peso total do organismo*, *peso do animal sem a concha* e *peso das vísceras* em grupos homogéneos e comente o resultado obtido. Represente também a respectiva matriz de distâncias cofenéticas.

    (b) Aplicando o método de agregação de Ward ao conjunto das 6 variáveis normalizadas com a distância euclideana obteve-se uma partição do conjunto do gastópodes em 2 grupos. A partição obtida foi depois comparada com a partição em 2 grupos definida pelas classes da variável `Sex`, *juvenil* e *não juvenil*, usando o índice de RAND, tendo-se obtido o valor de 0.6712376. Qual foi o número de pares de gastrópodes em que as duas classificações não coincidiram?

12. Foi efectuada uma análise classificatória sobre um conjunto de 64 observações normalizadas, usando o método hierárquico do vizinho mais próximo com a distância euclideana, tendo-se obtido o dendrograma abaixo. Sabe-se que a matriz de distâncias cofenéticas associada a este dendrograma contém os valores 7.42 e 7.92.



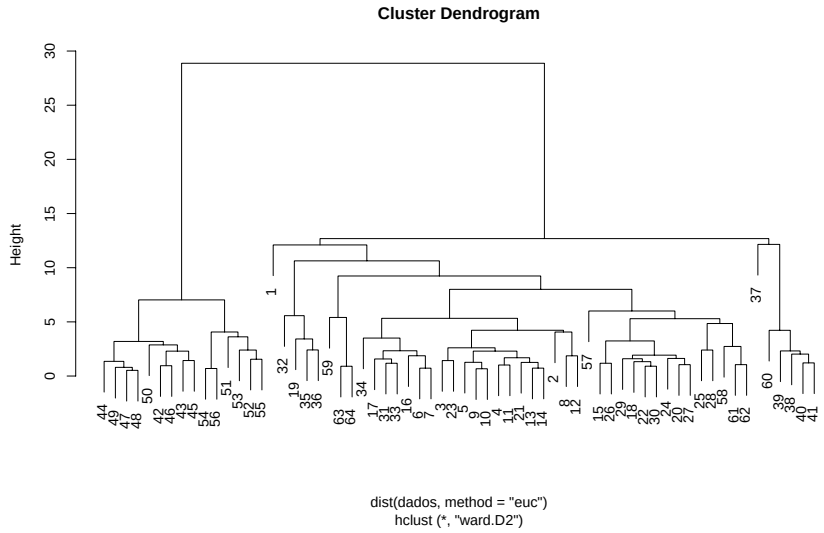
De acordo com a informação que foi disponibilizada qual(ais) da(s) seguinte(s) afirmação (ões) pode assegurar que está correcta:

(a) A distância entre as observações 1 e 37 é igual a 7.92

(b) A distância entre as observações 1 e 37 não é inferior a 7.92

(c) A distância entre as observações 37 e 32 é menor ou igual que a distância entre as observações 1 e 37.

Posteriormente foi efetuada uma análise classificatória do mesmo conjunto das 64 observações normalizadas usando o método hierárquico da inércia mínima (Ward) e a distância euclideana e foi aplicado um corte no respectivo dendrograma para se obter uma partição em 5 grupos. O dendrograma encontra-se representado a seguir e as distâncias entre os 5 grupos na tabela seguinte (com valores arredondados às duas casas decimais), onde a designação dos grupos segue a ordem dos grupos no dendrograma, da esquerda para a direita).



	C_1	C_2	C_3	C_4
C_2	15.58			
C_3	28.88	12.10		
C_4	16.81	12.13	12.28	
C_5	14.57	13.04	12.48	12.15

(a) Determine a distância cofenética entre as observações 1 e 37 para o método da inércia mínima. O que representa esta distância?

(b) Foi aplicado um processo de consolidação à partição em 5 grupos usando o método da inércia mínima, cujas sementes iniciais foram os centros de gravidade de cada um dos grupos. Constatou-se que se obtiveram as mesmas classes que as classes obtidas anteriormente pelo método da inércia mínima. O que conclui?

(c) Cada uma das 64 observações contém uma etiqueta que começa por código formado por 2 símbolos, como mostrado a seguir:

```
> groups
 [1] 4C 4C 4C 3A 4C 2A 4C 2A 2A 2A 2A 2A 2A 2A 2A 2A
 [17] 2A 2A 2A  2A 2A 2A 2A 2A 2A 2A 2A 2A 2A 2A 2A 2A
[33] 2A 2A 2A 2A 2A 2A 2A 4B 4B 2A 2A 2A 2A 2A 2A 2A
[49] 4A 4A 4A 4A 4A 4A 4A 4A 4A 4A 4A 4A 4A 4A 4A 4A
> table(groups)
2A 3A 4A 4B 4C
40  1 16  2  5
```

Sabe-se que o índice de RAND entre a classificação em 5 grupos dada pelas etiquetas **2A**, **3A**, **4A**, **4B** e **4C**, e a classificação em 5 grupos obtida pelo método da inércia mínima tem o valor de 0.5729167, e que existem 642 pares de medições que são classificados em grupos distintos por estas duas classificações. Determine o número de pares de medições que são classificados no mesmo grupo pelas duas classificações.

(d) Constatou-se mais tarde que uma das medições com a etiqueta **4B** estava mal catalogada, tendo-lhe sido atribuída a etiqueta **4D**. Determine o valor do índice de RAND entre a nova classificação em 6 grupos dada pelas etiquetas e a classificação anterior em 5 grupos dada pelo método da inércia mínima (Ward).

13. Considere uma classificação hierárquica. Seja d_{ij} a dissemelhança entre dois indivíduos genéricos i e j. Seja h_a a distância cofenética entre os indíviduos i e j na classificação com o método do vizinho mais afastado, e h_p a correspondente distância cofenética na classificação com o método do vizinho mais próximo. Mostre que se verifica a dupla desigualdade, $h_p \leq d_{ij} \leq h_a$.

14. Prove que no método do vizinho mais afastado os custos de fusão crescem monotonamente (sem usar a fórmula de *update* de Lance-Williams)

15. (a) Mostre que se na fórmula de Lance-Williams, se tem $\alpha_i, \alpha_j, \gamma \geq 0$ com $\alpha_i + \alpha_j + \beta \geq 1$ então os custos de fusão crescem monotonamente, isto é o respectivo dendrograma não possui inversões.

    (b) Utilizando a alínea anterior justifique que os custos de fusão para os métodos hierárquicos das distâncias médias entre grupos (*average*) e da inércia mínima (*Ward*) crescem monotonamente.

16. Prove a fórmula de update dada pela tabela de Lance-Williams para o método das distâncias médias entre grupos (average linkage).

17. Numa classificação usando o método das k-médias móveis (k-means) com 3 sementes um dos grupos finais é vazio. Em que condições poderá esta solução ser óptima?

18. O índice de RAND ajustado (ARI) entre duas partições $\mathcal{P}$ e $\mathcal{Q}$ de um mesmo conjunto finito X define-se como

$$ARI(\mathcal{P}, \mathcal{Q}) = \frac{RI(\mathcal{P}, \mathcal{Q}) - \mathrm{E}[RI]}{\max(RI) - \mathrm{E}[RI]},$$

onde $\max(RI) = 1$ e $\mathrm{E}[RI]$ é o valor esperado do índice de Rand (RI) considerando partições aleatoriamente escolhidas $\mathcal{P}'$ e $\mathcal{Q}'$ of X com o mesmo nú mero de grupos e a mesma partição de tamanhos que as partições $\mathcal{P}$ e $\mathcal{Q}$, respectivamente.

Utilize a fórmula acima para calcular $ARI(ab|cd, ab|c|d)$.