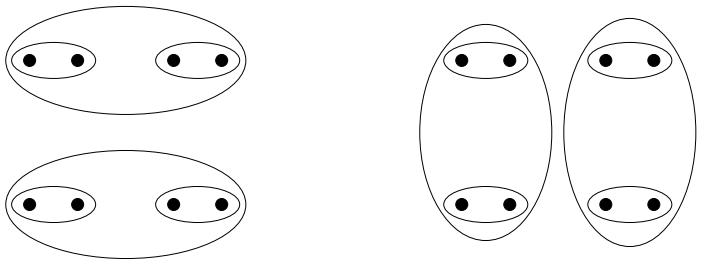
Exercícios de Análise Classificatória (18/19)

Os exercícios assinalados com (*) foram parcialmente retirados ou modificados a partir de exercícios que aparecem na bibliografia citada nos slides.

1. (*) As classificações hierárquicas seguintes foram obtidas aplicando dois métodos de agregação distintos à mesma configuração de 8 pontos, considerando distâncias euclideanas.



   Identifique, justificando, dois métodos de análise classificatória hierárquica que pudessem ter originado estes resultados.

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

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

   (a) Efectua uma partição deste conjunto em dois grupos usando o método do vizinho mais afastado (*complete*) com distância euclideana e represente o respectivo dendrograma. O que conclui?

   (b) Escreva a respectiva matriz de distâncias cofenéticas (também conhecidas por distâncias dendrogramáticas) e represente o respectivo diagrama de Sheppard.

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

   (d) Efectue uma nova classificação dos pontos com o método do vizinho mais afastado usando agora a matriz de distâncias cofenéticas como matriz de distâncias. O que observa?

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. As respectivas distâncias cofenéticas verificam a propriedade ultra-métrica? Justifique.

5. Prove que o método do vizinho mais afastado é monótono, ou seja, que os custos de fusão crescem monotonamente

6. Prove que uma classificação hierárquica com o método do vizinho mais próximo é invariante por transformações monótonas da matriz de dissemelhança.

7. Utilizando a fórmula de Lance-Williams mostre que os custos de fusão para o método hierárquico das distâncias médias entre grupos (*average*) crescem monotonamente.

8. Sobre um conjunto de K pontos distintos de $\mathbb{R}^n$ foi efectuada uma análise classificatória hierárquica pelo método do vizinho mais afastado com dissemelhança $d(x, y) = 1/\|x - y\|$, se $x \neq y$ e $d(x, x) = 0$, onde $\|x - y\|$ designa a distância euclideana usual. Se se efectuar um corte no dendrograma de altura h, o que pode afirmar sobre as distâncias euclideanas entre pares de elementos de um mesmo grupo? Justifique.

9. 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, $h_p \leq d_{ij} \leq h_a$.

10. Considere 3 grupos C_i, C_j e C_k numa análise classificatória dada pela fórmula de Lance-Williams e a dissemelhança $d_{ij,k}$ entre o grupo $C_{ij} = C_i \cup C_j$ e o grupo C_k. Mostre que:

    (a) $d_{ij,k}$ corresponde ao mínimo das dissemelhanças entre pontos de C_{ij} e de C_k no caso da fórmula de L-W para o método do vizinho mais próximo (*single*).

    (b) $d_{ij,k}$ corresponde à média das dissemelhanças entre pontos de C_{ij} e de C_k, no caso da fórmula de L-W para o método das distâncias médias entre grupos (*average*).

11. (*) Considere a matriz de dissemelhanças,

$$D = \begin{bmatrix} 0 & 1.8 & 2.4 & 2.3 \\ 1.8 & 0 & 2.5 & 2.7 \\ 2.4 & 2.5 & 0 & 1.2 \\ 2.3 & 2.7 & 1.2 & 0 \end{bmatrix}$$

    (a) Efectue uma análise classificatória hierárquica usando a fórmula de Lance-Williams para o método da inércia mínima (*ward*), com a matriz de dissemelhanças D e represente o respectivo dendrograma.

    (b) Repita a análise classificatória da alínea anterior usando a fórmula do "update" dado nos slides das aulas.

12. Mostre que a distância de Mahalanobis coincide com a distância euclideana das variáveis normalizadas, se as variáveis forem não correlacionadas.

13. Considere um conjunto X com N pontos em $\mathbb{R}^d$.

    (a) Considere $d = 1$ e prove que o centroide de X é o ponto que minimiza a soma dos quadrados das distâncias ao centroide,

    $$SSQ_X(y) = \sum_{x \in X} (x - y)^2.$$

    (b) Assumindo a validade do resultado anterior para d arbitrário e que a realocação apenas ocorre se SSQ_w baixar, conclua que a função SSQ_w decresce monotonamente durante o decurso do algorítmo das K-médias móveis aplicado a X, e portanto que o algorítmo converge (eventualmente para uma solução subóptima).

14. Numa classificação usando o método das K-médias móveis (K-means) com 3 sementes um dos grupos finais é vazio. Poderá esta solução ser óptima?

    (Sugestão: utilize o resultado do exercício anterior).

15. Classificou-se um grupo constituído por $N = 178$ vinhos com o algorítmo das K-médias, considerando o número de clusters k a variar de 1 a 10. Foram obtidos os seguintes valores para as inércias intra-grupo (SSQ_w) em função do número de clusters:

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 também 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 depois comparadas entre si usando o índice de Rand. Os resultados obtidos são apresentados na seguinte tabela.

	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 de quadrados total (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 critrios distintos.

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

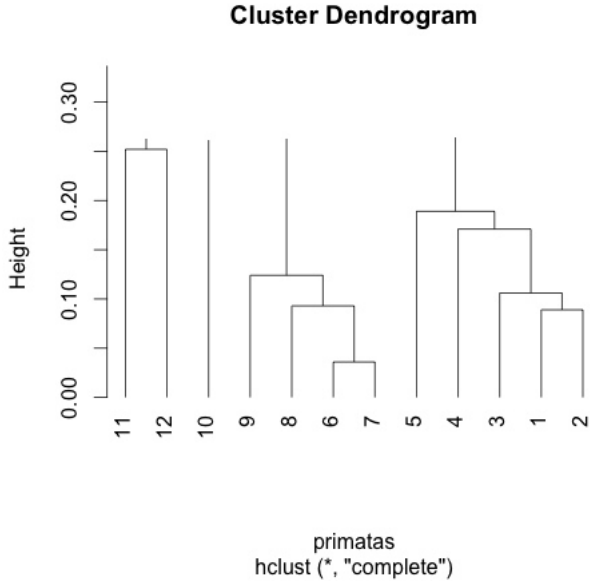
    (d) Efetue uma análise classificatória hierárquica que permita agregar as classificações da tabela anterior em grupos homogéneos, usando uma medida de dissemelhança adequada e o método do vizinho mais afastado. Represente o respectivo dendrograma.

16. 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 efectuou-se uma classificação hierárquica 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 dos grupos C_1, C_2, C_3 e C_4.
(o dimetro de um conjunto é a distância entre os seus elementos mais afastados)

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

(c) Represente como pontos num diagrama de Sheppard as relações entre as dissemelhanças entre os primatas *Homo sapiens*, *Macaca fuscata*, *Saimiri sciureus* e *Tarsius syrichta* e as respectivas distâncias cofenéticas dadas pelo método hierárquico.

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

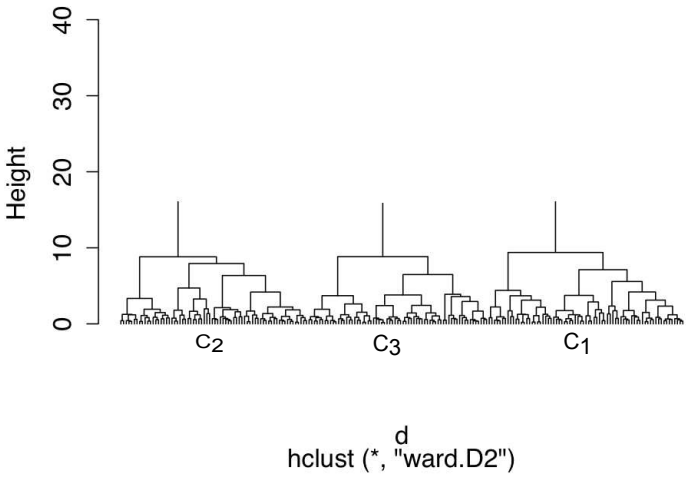
i. Escreva estes grupos e indique o número de pares de primatas que seriam classificados de forma discordante por esta partição e pela partição em 4 grupos descrita acima.

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

17. 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:

Nome	Descrição	Unidades
`Area`	Área (A) da superfície	mm^2
`Perimeter`	Perímetro (P)	mm
`Compacteness`	$\frac{4\pi A}{P^2}$	–
`Kernel_length`	Comprimento	mm
`Kernel_width`	Largura	mm
`asym_coeff`	Coeficiente de assimetria	–
`length_kernel_groove`	Comprimento do sulco	mm

Efetuou-se em seguida uma classificação hierárquica usando o método da inércia mínima (*ward*) com a 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.



Realizou-se posteriormente uma nova classficação sobre o mesmo conjunto de dados pelo método das K-médias móveis (*K-means*) utilizando como sementes iniciais 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) Para qual das classificações os grupos deverão ser mais homogéneos? Fundamente a sua resposta.

(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.

18. 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 aglomerativo do vizinho mais afastado (*complete*) e uma medida de dissemelhança adequada.

19. A seguinte tabela contém as componentes de 6 vetores binários.

a	1	0	0	0	0	1
b	1	0	1	0	1	1
c	0	0	1	1	0	1
d	1	1	0	0	0	1
e	1	0	1	0	1	0
f	0	0	1	1	0	0

Efectuou-se uma análise classificatória sobre o conjunto destes vetores binários em dois grupos usando a distância de Manhattan (para vetores binários é também conhecida por distância de *Hamming*) e o método dos K-medoides, considerando medoides iniciais d e f.

(a) Determine o valor da função objectivo na etapa inicial.

(b) Qual o custo de trocar o medoide d por e?

20. Um estudo envolve a observação de gastrópodes marinhos da espécie *Haliotis rubra*. Foram recolhidos ao acaso 4177 indivíduos, tendo sido medidas em cada indivíviduo 8 variáveis numéricas, bem como o sexo (variável `Sex`), com 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 so 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*, *dimetro*, *altura*, *peso total do organismo*, *peso do animal sem a concha* e *peso das vsceras* em grupos homogéneos e comente o resultado obtido. Represente ainda a respectiva matriz de distâncias cofenéticas.

(b) Usando as 6 variáveis da alínea anterior normalizadas e o método da inércia mínima (*ward*), efectuou-se uma partição do conjunto dos gastrpodes em dois grupos homogéneos. Esta partição foi posteriormente comparada com a classificação dada pela variável `Sex`, em *juvenil* e *não juvenil*, tendo-se obtido um índice de RAND de valor 0.6712376.

Qual foi o número de pares de gastrópodes em que as duas classificações não coincidiram?

(c) Considere os grupos $C_1, \ldots, C_k$ que são obtidos efectuando um corte de altura h num dendrograma definido pelo método do vizinho mais próximo.

Prove que a distância entre elementos pertencentes a grupos distintos C_i e C_j é superior ou igual a h.

(d) Utilizando a fórmula de Lance-Williams mostre que os custos de fusão para o método da inércia mínima crescem monotonamente.

No seguinte exercício utilize o programa informático R

1. Efectue uma análise classificatória hierárquica do conjunto dos lírios usando os métodos do vizinho mais próximo (*single*), do vizinho mais distante (*complete*), das distâncias médias entre grupos (*average*), do centroide e da inércia mínima (*ward*), com as distâncias euclideana, Manhattan, do máximo, de Canberra e de Mahalanobis. Quantos grupos os dendrogramas sugerem?

2. Calcule os respectivos coeficientes de correlação cofenética de Pearson e Spearman e comente.

3. Efetue o corte na árvore de modo a obter 3 grupos em cada um dos casos e compare entre si os resultados obtidos usando o indíce de RAND e com a classificação verdadeira em cada um dos três tipos de lírios (*setosa* [1:50], *versicolor* [51:100] e *virginica* [101:150]).

4. Efectue uma classificação em 3 classes do conjunto dos lírios com o algorítmo das K-médias (*kmeans*) considerando como sementes os centroides dos grupos formados com o método da inércia mínima. Compare os resultados obtidos e comente.

5. Efectue uma classificação em 3 classes do conjunto dos lírios com o algorítmo e dos partição em torno dos K-medoides (*PAM*).

6. Investigue se as classificações em 3 grupos obtidas nas alíneas anteriores e a classificação anteriormente conhecida para os lírios podem ser agregadas em grupos homogéneos usando uma medida de dissemelhança adequada.

7. Execute uma classificação hierárquica das variáveis do conjunto de dados dos lírios usando uma medida de dissemelhança d conveniente e interprete os grupos formados. Como interpreta geometricamente um grupo cujo custo de fusão seja inferior a um dado valor τ.

Código auxiliar da função RAND do Prof. Cadima

```
rand <- function(class1,class2){
  n <- length(class1)
  c <- as.dist(outer(class1,class1,"=="))
  d <- as.dist(outer(class2,class2,"=="))
  rand <- sum(c == d)/(n*(n-1)/2)
  rand
}
```