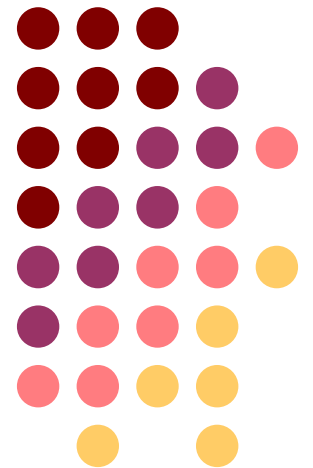


Métodos de Agrupamento Para

Dados Simbólicos: Hierarquico Divisivo

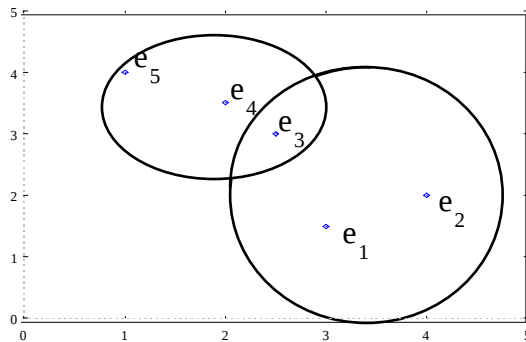


Renata Maria Cardoso Rodrigues de Souza

Estruturas classificatórias



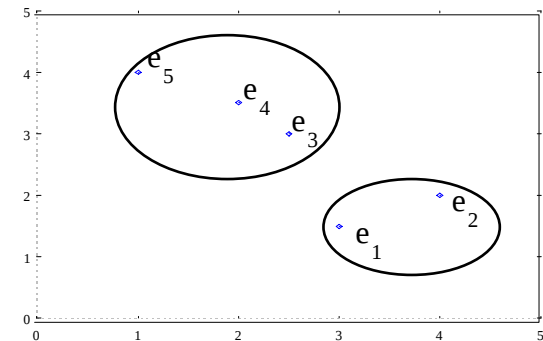
Cobertura



1) $\forall \ell = 1, \dots, K$ tem-se $P_\ell \neq \emptyset$

$$2) \bigcup_{\ell=1}^K P_\ell = \Omega$$

Partição



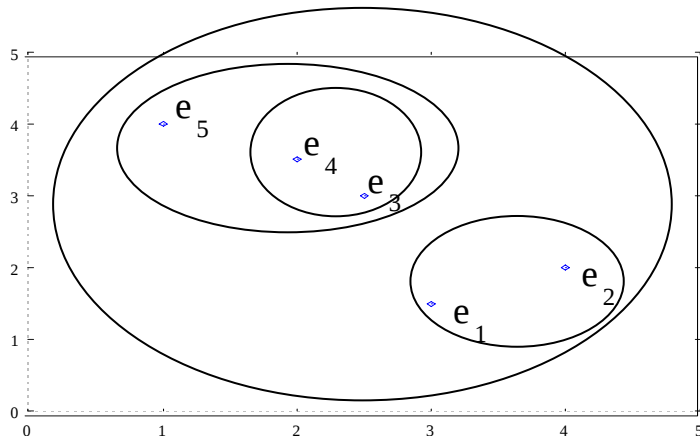
3) $\forall \ell, m = 1, \dots, K$ e $\ell \neq m$

então $P_\ell \cap P_m = \emptyset$

Estruturas Classificatórias



Hierarquia



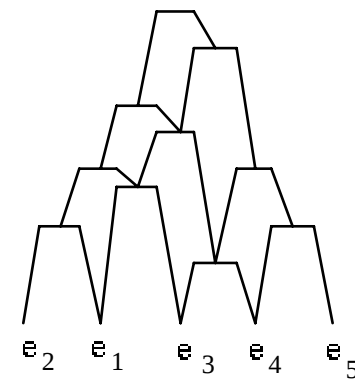
1) $\Omega \in H$

2) $\forall e \in \Omega$ então $\{e\} \in H$

3) $\forall h, h' \in H$ tem - se :

$$h \cap h' \neq \emptyset \Rightarrow h \subset h' \text{ ou } h' \subset h$$

Piramide



3) $\forall h, h' \in H$ tem - se $h \cap h' = \emptyset$ ou $h \cap h' \in H$

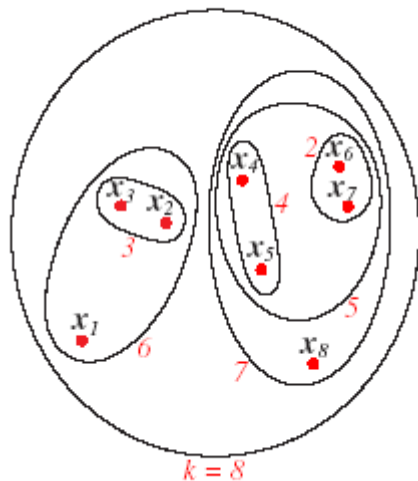
4) Existe uma ordem θ tal que

$$\forall h \in H, h \text{ é um intervalo de } \theta$$

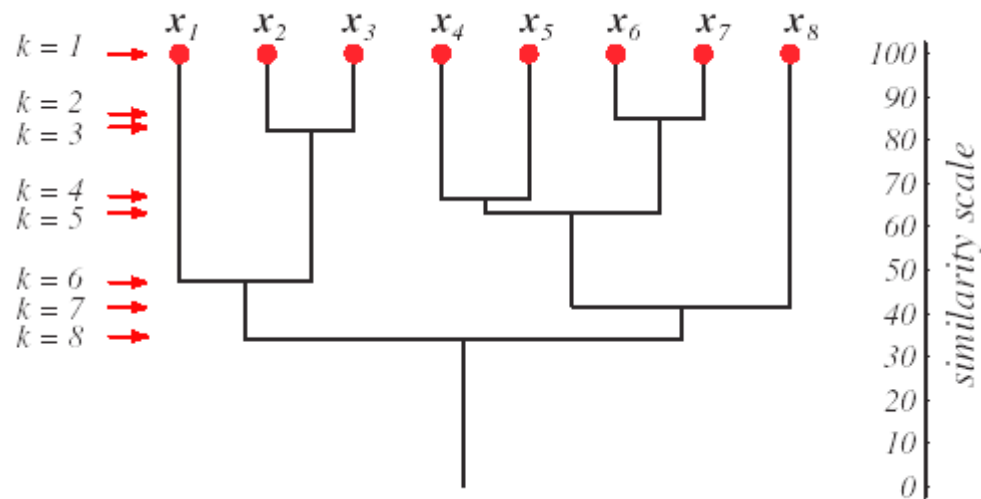
Classificação Hierarquica



Diagrama de Venn
sobre os dados bi-
dimensionais



Dendrograma: baseado em um
índice de heterogeneidade

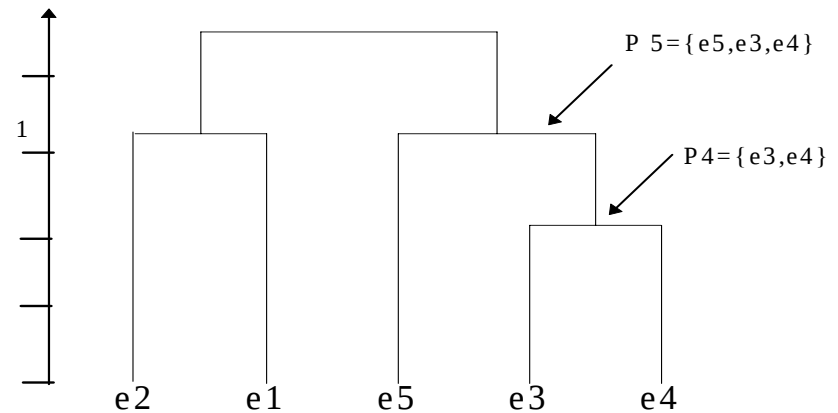


Métodos Hierárquicos



Parte-se de uma tabela de dados e calcula-se uma distância entre os indivíduos de Ω

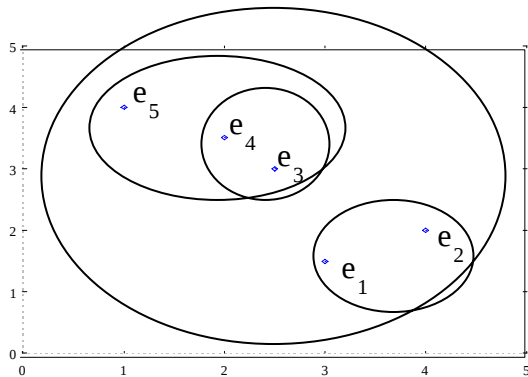
Os métodos ascendentes hierárquicos tem por objetivo a construção de uma sequencia de partições encaixadas chamada *hierarquia*. A representação gráfica dessas hierarquias é realizada por uma *arvore hierarquica* ou *dendrograma*.



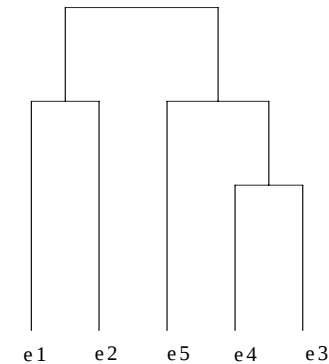
Hierarquia com índice



Hierarquia H



Hierarquia com índice (H, f)



$$1) E \in H$$

$$2) \forall e \in E \text{ então } \{e\} \in H$$

$$3) \forall h, h' \in H \text{ tem-se:}$$

$$h \cap h' \neq \emptyset \Rightarrow h \subset h' \text{ ou } h' \subset h$$

$$f : H \rightarrow \mathbb{R}^+$$

$$(1) \quad f(h) = 0 \text{ se e somente se } \text{card}(h) = 1$$

$$(2) \quad \forall h, h' \in H \quad h \subset h' \text{ e } h \neq h' \Rightarrow f(h) < f(h')$$

Método Divisivo

Marie Chavent et al (2000)



- ❑ Um método divisivo consiste em obter grupos de objetos em diferentes etapas produzindo uma hierarquia.
- ❑ O processo inicia com todos os objetos e divide sucessivamente cada grupo formando grupos menores de objetos.
- ❑ Uma divisão é definida otimizando um critério de minimização de variância intra-grupo
- ❑ Em cada etapa o agrupamento é obtido de forma monotética.

Notações



Ω = conjunto de todos os objetos (dados simbólicos)

$\{Y_1, \dots, Y_p\}$ – conjunto de variáveis simbólicas (multi-valor ordinal , intervalo ou modal)

$C = \{C_1, C_2\}$ uma bipartição associada a uma variável Y_i

n – número total de objetos

n_i – número total de objetos da classe i , $i=1,2$



Tabela de dados simbólicos

k	peso	altura
1	[30,40]	0.10 (alta) 0.30 (média) 0.60 (baixa)
2	[35,65]	0.20 (alta) 0.50 (média) 0.30 (baixa)
3	[20,30]	0.10 (alta) 0.80 (média) 0.10 (baixa)

Algoritmo



- ▮ Uma divisão é obtida por otimizando um critério de variância intra-classe (Inércia).
- ▮ O critério é minimizado com respeito a todas as bipartições induzidas por um conjunto de questões binárias de uma variável.
- ▮ Seleciona-se a bipartição onde a inércia inter-classe é máxima.

Algoritmo



Passo 1 : $P_1 = \Omega$; $v = 1$; $C = P_1$

Passo 2: Faça enquanto $v \leq V - 1$ (V é o nº de iterações)

Passo 3: Para cada C de P_v escolher $\{C_1, C_2\}$ induzida por questões binárias onde a **inércia intra-grupo** é mínima

$$W = I(C_1) + I(C_2) = \frac{1}{n \times n_1} \sum_{k, m \in C_1} \sum_{m > k} d_{km}^2 + \frac{1}{n \times n_2} \sum_{k, m \in C_2} \sum_{m > k} d_{km}^2$$

Passo 4: Escolher C de P_v tal que a **inércia inter-grupo** é máxima

$$\Delta(C) = I(C) - I(C_1) - I(C_2)$$

Passo 5: $P_{v+1} = P_v \cup \{C_1, C_2\} - \{C\}$

Passo 6: $v = v + 1$, vá para o Passo 2

Saída do algoritmo



A saída do método é uma hierarquia H constituída de $V+1$ classes

Cada classe $C_i \in H$ é organizada por $\Delta(C_i)$ no dendograma.

$$C_i \subset C_{i'} \Rightarrow \Delta(C_i) \leq \Delta(C_{i'})$$

Extensão do Critério de Variância Intra-grupo



- O critério usado para avaliar a qualidade de uma partição é uma extensão do critério soma de quadrados intra-grupo para o caso de uma matriz de distâncias.
- A extensão da soma de quadrados de erro (inércia) de uma classe

$$I(C_i) = \sum_{k \in C_i} ||\mathbf{x}_i - \boldsymbol{\mu}_i||^2$$

para uma matriz de distância D é dada por

$$I(C_i) = \frac{1}{n \times n_i} \sum_{k, m \in C_i} \sum_{m > k} d_{km}^2$$

onde n_i é o número de elementos do grupo i

- Uma generalização do critério de uma partição com m classes é dado por

$$W = \sum_{i=1}^m I(C_i)$$



Biparticionamento do grupo

- O objetivo é encontrar a bipartição $C_1 = (C_1^1, C_1^2)$ de menor soma de quadrados de erro.
- Uma escolha ótima é realizada avaliando $2^{n_i-1} - 1$ bipartições.
- Uma solução é escolher a bipartição entre as bipartições induzidas pelo conjunto de todas as possíveis questões binárias.

Questões Binárias – $[Y_j \leq c]$

$$q_c: \Omega \rightarrow \{V, F\}$$

A bipartição $\{C_1, C_2\}$ induzida por $q_c = [Y_j \leq c]$ é

$$C_1 = \{k \in C / q_c = V\}$$

$$C_2 = \{k \in C / q_c = F\}$$



Se Y_j é uma variável intervalo, então $Y_j(k) = [\alpha, \beta]$ e

$$q_c(k) = V \text{ se } m_k \leq c$$

$$q_c(k) = F \text{ se } m_k > c$$

onde
$$m_k = \frac{\alpha + \beta}{2}$$

Existirá $n_i - 1$ questões binárias onde c é o valor entre dois consecutivos valores de m_k .

Exemplo : $n = 3$

$Y_1 = \text{peso: } m_1 = 25 \text{ } m_2 = 35 \text{ e } m_3 = 50$

$c = 30$

Para $c = 30$ $q_c(1) = V$, $q_c(2) = F$ e $q_c(3) = F$

$C_1 = \{1\}$ e $C_2 = \{2, 3\}$



Se Y_j é uma variável ordinal ou modal então $Y_j(k) = \pi_k$ e

$$q_c(k) = V \text{ se } \sum_{x \leq c} \pi_k(x) \geq 1/2$$

$$q_c(k) = F \text{ se } \sum_{x \leq c} \pi_k(x) < 1/2$$

Existirá $|O_j| - 1$ questões binárias

Exemplo : c =baixa para a variável modal da tabela

Para c = baixa $q_c(1) = V$, $q_c(2) = F$ e $q_c(3) = F$

$C_1 = \{1\}$ e $C_2 = \{2,3\}$



Medida de Distância



■ Se Y_j é uma variável intervalo, então $Y_j(k) = [\alpha, \beta]$ e

$$d(k, m) = \left[\sum_{j=1}^p (\max\{|\alpha_k^j - \alpha_m^j|, |\beta_k^j - \beta_m^j|\})^2 \right]^{1/2}$$

Versões normalizadas

$$d(k, m) = \left[\sum_{j=1}^p \left[\frac{\max\{|\alpha_k^j - \alpha_m^j|, |\beta_k^j - \beta_m^j|\}}{m_j} \right]^2 \right]^{1/2}$$

$$\text{a)} \quad m_j^2 = \frac{1}{2n^2} \sum_{k=1}^n \sum_{m=1}^n (\max(|\alpha_k^j - \alpha_m^j|, |\beta_k^j - \beta_m^j|))^2$$

$$\text{b)} \quad m_j = |O_j|$$



Exemplo

$$a_1 = [Y_1 \in [30.40]] \quad a_2 = [Y_1 \in [35.65]]$$

$$d(1,2) = \left[(\max\{|30 - 35|, |40 - 65|\})^2 \right]^{1/2} = 25$$

Versão normalizada – $O_1 = [20,70]$

$$d(1,2) = 0.5$$



Se Y_j é uma variável modal então $Y_j(k) = \pi_k$

p=número de
variáveis

	1	...	T	Σ
1	f_{11}		f_{1T}	p
:	:		:	:
n	f_{n1}		f_{nT}	p
Σ	$F_{.1}$		$F_{.T}$	np

	1	...	T	Σ
1	f_{11}/np		f_{1T}/np	$p_{1.}$
:	:		:	:
n	f_{n1}/np		f_{nT}/np	$p_{n.}$
Σ	$p_{.1}$		$p_{.T}$	1

$$T = |O_1| + \dots + |O_p|$$

$$d^2(k, m) = \sum_{t=1}^T \frac{p_{..}}{p_{.t}} \left[\frac{p_{kt}}{p_{k.}} - \frac{p_{mt}}{p_{m.}} \right]^2 \quad \Rightarrow$$

$$d^2(k, m) = n \sum_{t=1}^T \frac{[f_{kt} - f_{mt}]^2}{f_{kt} + f_{mt}}$$

$$p_{kt} = \frac{f_{kt}}{np}$$

$$p_{k.} = \sum_{t=1}^T p_{kt}$$

$$p_{.t} = \sum_{k=1}^n p_{kt}$$

$$p_{..} = \sum_{k=1}^n \sum_{t=1}^T p_{kt} = 1$$



Exemplo

	baixa	média	alta	Σ
1	0.10	0.30	0.60	1
2	0.20	0.50	0.30	1
	0.30	0.80	0.90	2

$$d^2(1,2) = 2 \times \left[\frac{(0.10 - 0.20)^2}{0.10 + 0.20} + \frac{(0.30 - 0.50)^2}{0.30 + 0.50} + \frac{(0.60 - 0.30)^2}{0.60 + 0.30} \right] = 0.37$$

$$d(1,2) = 0.60$$

Tabela de óleos e graxas



k	Specific Gravity	Freezing point	Iodine Value	Saponification Value
1	[0.930,0.935]	[-27,-18]	[170,204]	[118,196]
2	[0.930,0.937]	[-5,-4]	[192,208]	[188,197]
3	[0.916,0.918]	[-6,-1]	[99,113]	[189,198]
4	[0.920,0.926]	[-6,-4]	[104,116]	[187,193]
5	[0.916,0.917]	[-21,-15]	[80,82]	[189,193]
6	[0.914,0.919]	[0,6]	[79,90]	[187,196]
7	[0.860,0.870]	[30,38]	[40,48]	[190,199]
8	[0.858,0.864]	[22,32]	[53,77]	[190,202]

1 – linseed oil

2 – perilla oil

3 – cottonseed oil

4 – sesam oil

5 – camelia oil

6 – olive oil

7 – beef tallow

8 – hog fat

Matriz de Dissimilaridade



	1	2	3	4	5	6	7	8
1	-	0.81	1.17	1.08	1.30	1.34	2.62	2.34
2	-	-	1.19	1	0.70	1.25	3.31	2.00
3	-	-	-	0.02	0.09	0.03	0.55	0.98
4	-	-	-	-	0.11	0.06	0.90	1.14
5	-	-	-	-	-	0.11	0.77	0.62
6	-	-	-	-	-	-	0.86	0.79
7	-	-	-	-	-	-	-	0.05
8	-	-	-	-	-	-	-	-

1 – linseed oil

2 – perilla oil

3 – cottonseed oil

4 – sesam oil

5 – camelia oil

6 – olive oil

7 – beef tallow

8 – hog fat

Passo 1



Calcular $W = I(C_1) + I(C_2)$ para todas as questões binárias e escolher a bipartição que minimize W

a) $Y = \text{specific gravity}$

$$m_1 = 0.9325$$

$$m_2 = 0.9335$$

$$m_3 = 0.9170$$

$$m_4 = 0.9230$$

$$m_5 = 0.9165$$

$$m_6 = 0.9165$$

$$m_7 = 0.8650$$

$$m_8 = 0.8610$$

$$c = \begin{cases} 0.86300 \\ 0.89075 \\ 0.91675 \\ 0.92000 \\ 0.92775 \\ 0.93330 \end{cases}$$



Para $c = 0.86300$

$$q_c(1) = F$$

$$q_c(2) = F$$

$$q_c(3) = F$$

$$q_c(4) = F$$

$$q_c(5) = F$$

$$q_c(6) = F$$

$$q_c(7) = F$$

$$q_c(8) = V$$

$$C_1: Y_1 \leq 0.86300 \quad C_2: Y_2 > 0.86300$$

$$C_1 = \{8\} \text{ e } C_2 = \{1,2,3,4,5,6,7\}$$

$$n_1 = 1 \quad n_2 = 7$$

$$I(C_1) = 0$$

$$I(C_2) = \frac{1}{n \times n_2} \sum_{k,m \in C_2} \sum_{k>m} d^2(k,m) = \frac{1}{8 \times 7} [0.81 + 1.17 + 1.08 + \dots + 0.77 + 0.86] = 0.344$$

$$W = 0.344$$



Para $c = 0.89075$

$$q_c(1) = F$$

$$q_c(2) = F$$

$$q_c(3) = F$$

$$q_c(4) = F$$

$$q_c(5) = F$$

$$q_c(6) = F$$

$$q_c(7) = V$$

$$q_c(8) = V$$

$$C_1: Y_1 \leq 0.8975 \quad C_2: Y_2 > 0.8975$$

$$C_1 = \{7,8\} \text{ e } C_2 = \{1,2,3,4,5,6\}$$

$$n_1 = 2 \quad n_2 = 6$$

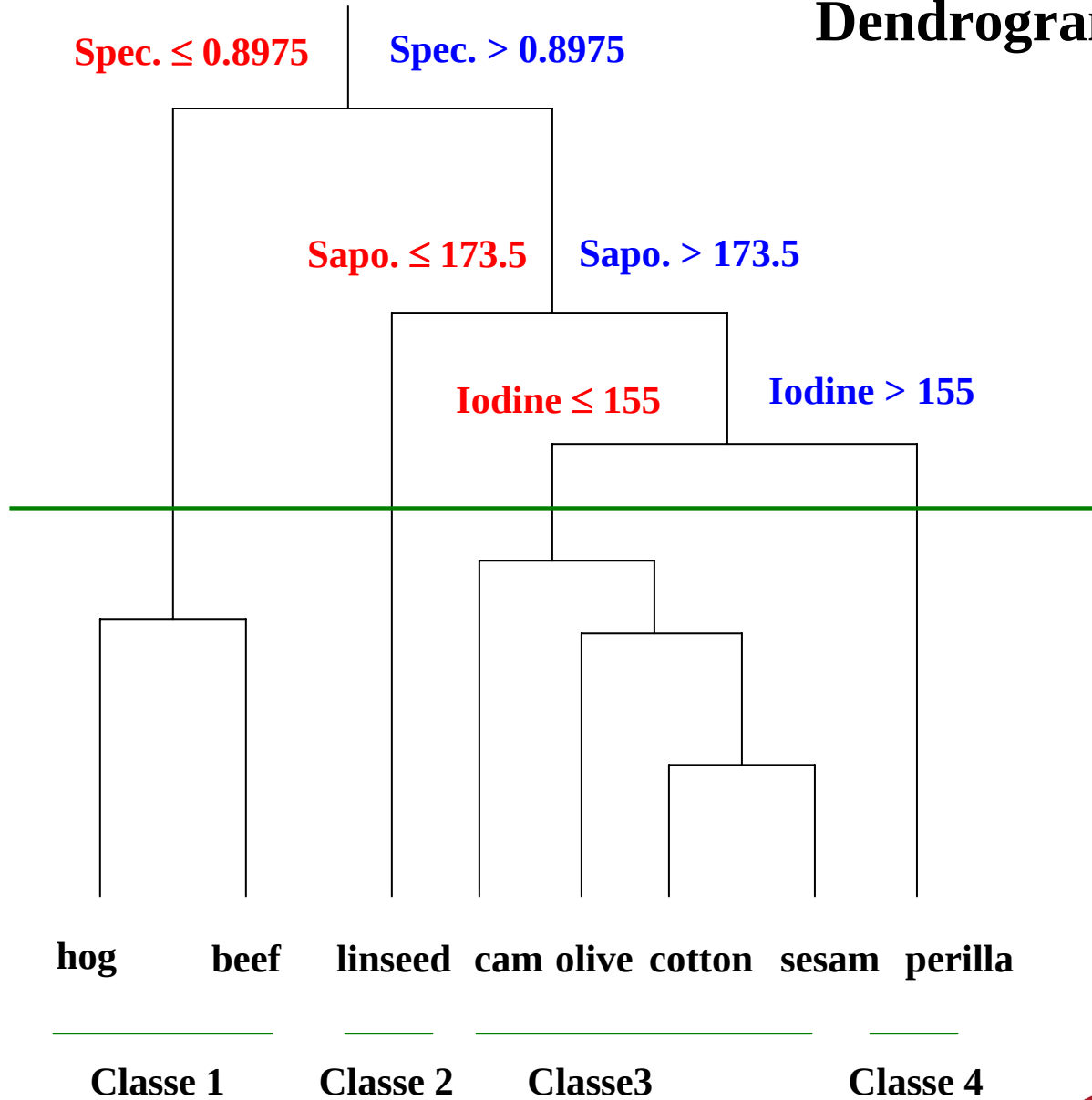
$$I(C_1) = \frac{0.05}{8 \times 2} = 0.003$$

$$I(C_2) = \frac{1}{6 \times 8} 10.86 = 0.21$$

$$W = 0.213$$



Dendrograma



Saída do algoritmo



$C_1 = \{\text{beef tallow, hog fat}\}$

$C_2 = \{\text{linseed}\}$

$C_3 = \{\text{camelia, olive, cottonseed, sesam}\}$

$C_4 = \{\text{perilla}\}$

Descrições

$C_1 = [\text{Spec.} \leq 0.8975]$

$C_2 = [\text{Spec.} > 0.8975] \wedge [\text{Sapo.} \leq 173.5]$

$C_3 = [\text{Spec.} > 0.8975] \wedge [\text{Sapo.} \leq 173.5] \wedge [\text{Iod.} \leq 155]$

$C_4 = [\text{Spec.} > 0.8975] \wedge [\text{Sapo.} \leq 173.5] \wedge [\text{Iod.} > 155]$