Aprendizagem de Dados Simbólicos e/ou Numéricos

Francisco de A.T. de Carvalho

Agrupamento (Clustering)

- Métodos usados para a construção de grupos de objetos
  - com base nas semelhanças e diferenças entre os mesmos
  - de tal maneira que os grupos obtidos são os mais homogêneos e bem separados possíveis.
- Duas grandes classes de problemas em classificação
  - classificação não supervisionada
  - classificação supervisionada

Agrupamento (Clustering)

- Classificação não supervisionada
  - se propõe a encontrar classes homogêneas a partir de um conjunto de indivíduos
  - Objetivo: os indivíduos semelhantes devem pertencer a mesma classe
  - É um objetivo intuitivo mas não é uma definição precisa da noção de classe

Agrupamento (Clustering)

- Agrupar para que?

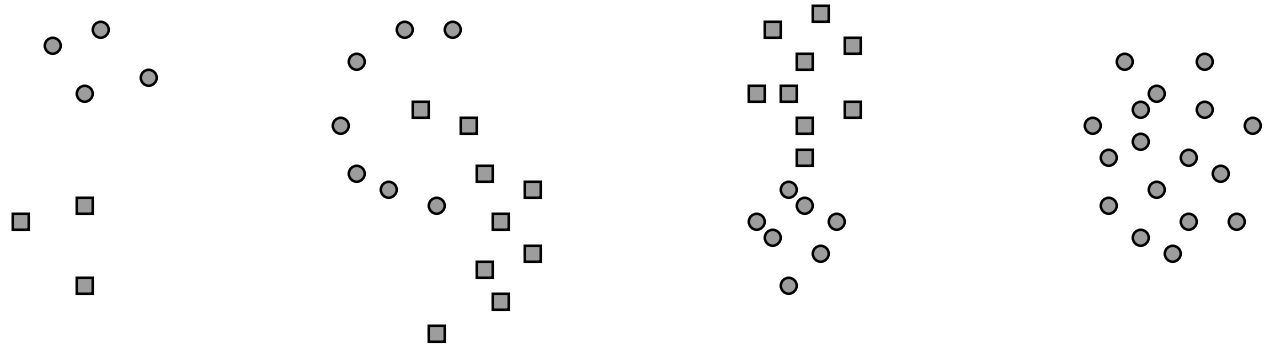
- Existe classes “naturais” e o desafio é encontra-las
- Deseja-se construir as classes segundo estruturas classificatórias (impostas)
- Encontrar classes úteis para o usuário
- Simplificação dos dados
- Geração de Hipóteses
- Predição com base nos grupos formados

Agrupamento (Clustering)

- O que é um grupo? Não existe uma única definição satisfatória

- Coesão interna
- Isolamento externo

Agrupamento (Clustering)



(a) (b) (c) (d)

- a) Grupos coesos e isolados
- b) Grupos isolados mas não coesos
- c) Grupos coesos com vários pontos intermediários
- d) Não existência de grupos “naturais”

Principais Etapas

- a) aquisição dos dados
  - 1) Seleção das observações (indivíduos, objetos, casos, itens)
  - 2) Seleção das variáveis (caracteres, descritores) e das correspondentes escalas
  - 3) Construção da Tabela de Dados
- b) Pré-processamento dos dados
  - 1) Mudança de escala
  - 2) Normalização
  - 3) Extração de caracteres

Principais Etapas

- c) Construção da Tabela de Dados
- d) Cálculo da Proximidade
  - 1) Escolha de um Índice de Proximidade
  - 2) Construção da Matriz de Proximidades
- e) Seleção de um Algoritmo de Formação de Grupos em função do tipo de agrupamento desejado
- f) Análise e Interpretação dos Resultados

Conceitos Básicos

- *Indivíduo*
- Ω: conjunto das indivíduos (população, amostra)
- $\omega \in \Omega$: indivíduo (especimen) ou grupo de indivíduos (espécie)

- *Variáveis*
- A cada característica (escolhida pelo usuário ou por um especialista), pode-se associar uma ou mais variáveis:

$$Y_i : \Omega \rightarrow D_i$$
$$\omega \rightarrow Y_i(\omega)$$

- D_i: Domínio da variável y_i

Conceitos Básicos

- As variáveis podem ser
  - quantitativas
    - contínuas (ex, Peso, Altura)
    - discretas (ex, numero de antenas, número de filhos)
  - qualitativas (ex, sexo, grau de instrução)
  - binárias (ex, presença de asas)

- com escala
  - nominal (ex, sexo (masculino, feminino))
  - ordinal (ex, Grau de instrução{primário, segundário, superior})
  - intervalar (ex, grau celsius)
  - proporcional (ex, grau kelvin, idade)

Tabela de Dados

N objetos ou individuos $\Omega=\{\omega_1,\ldots, \omega_i ,\ldots, \omega_N \}$

p descritores $Y=\{Y_1,\ldots, Y_j,\ldots, Y_p\}$

A cada objeto ω_i de Ω é associado um vetor de descrição $\left(x_i^1,...,x_i^j,...,x_i^p\right)$ representando as p medidas

A cada variável ou parametro Y_j é associado um vetor $\left(x_1^j,...,x_i^j,...,x_N^j\right)$

que representa o conjunto de valores observados de Ω sobre Y_j

	Y_1	…	Y_j	…	Y_p
w_1	x_1^1	…	x_1^j	…	x_1^p
…		…		…	
w_i	x_i^1	…	x_i^j		x_i^p
…		…		…	
w_N	x_N^1	…	x_N^j	…	x_N^p

Tabela de Dados

Tipos de Tabelas

quantitativas
qualitativas
binárias
heterogêneas

Exemplo:

Nome	Cobertura do Corpo	Cavidades do Coração	Temperatura do Corpo	Fertilização
mamífero	pelos	4	regulada	interna
pássaro	penas	4	regulada	interna
réptil	pele seca	4 imperfeitas	não regulada	interna
anfíbio	pele úmida	3	não regulada	externa
peixe	escamas	2	não regulada	externa

Índices de proximidade

Índices de Proximidade

- Similaridade
- Dissimilaridade

Índice de Similaridade

É uma função

$$s : \Omega \times \Omega \rightarrow R^{+}$$

$$(\omega_a, \omega_b) \rightarrow s(\omega_a, \omega_b)$$

Índices de proximidade

Índice de Similaridade

tal que

- $s(\omega_a, \omega_b) = s(\omega_b, \omega_a), \forall (\omega_a, \omega_b) \in \Omega \times \Omega$
- $s(\omega_a, \omega_a) = s(\omega_b, \omega_b) = s_{\max} > s(\omega_a, \omega_b),$

$\forall (\omega_a, \omega_b) \in \Omega \times \Omega, \text{com } \omega_a \neq \omega_b$

Quanto mais próximo dois indivíduos mais elevado é o valor da medida de similaridade entre eles

Índices de proximidade

Índice de Dissimilaridade

É uma função

$$d : \Omega \times \Omega \rightarrow R^{+}$$

$$(\omega_a, \omega_b) \rightarrow d(\omega_a, \omega_b)$$

Índices de proximidade

Índice de Dissimilaridade

tal que

- $d(\omega_a, \omega_b) = d(\omega_b, \omega_a), \forall (\omega_a, \omega_b) \in \Omega \times \Omega$

- $d(\omega_a, \omega_a) = 0, \forall \omega_a \in \Omega$

$$d(\omega_a, \omega_b) = s_{\max} - s(\omega_a, \omega_b)$$

Quanto mais próximos dois indivíduos menor é o valor da medida de dissimilaridade entre eles

Índices de proximidade

Exemplos de Índices de Proximidade

a) Tabelas de variáveis quantitativas

$$d(\omega_a,\omega_b)=\left[\sum_{j=1}^{p}\left|y_j(\omega_a)-y_j(\omega_b)\right|^{\lambda}\right]^{\frac{1}{\lambda}},\lambda\geq 1$$

b) Tabelas de variáveis binárias

	ω_a	1	0
ω_b			
1		x	y
0		z	w

$$s(\omega_a,\omega_b)=\frac{x}{x+y+z}\ \text{(Jaccard)}$$

Índices de proximidade

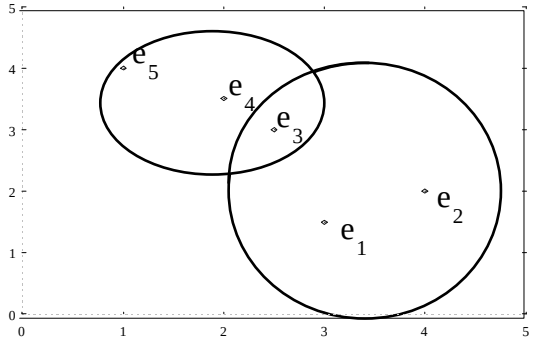
Outros aspectos relativos aos índices de proximidade

- Escala das Variáveis
- Correlação entre as Variáveis
- Descrições heterogêneas (Variáveis de diferentes tipos)
- Índices de proximidade entre padrões descritos por strings ou árvores
- Índices de proximidade dependentes do contexto
- Índices de proximidade conceptual

Classificação Hierárquica 1/

Estruturas classificatórias

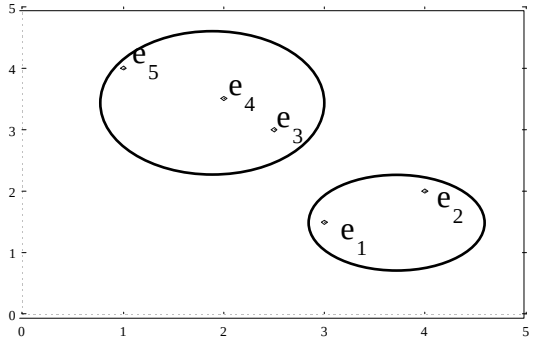
Cobertura



$)\forall \ell = 1, \cdots, K$ tem-se $P_\ell \neq \varnothing$

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

Partição



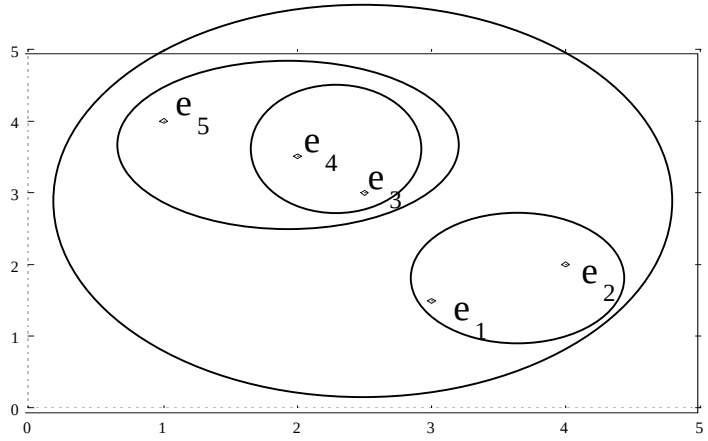
$3) \forall \ell, m = 1, \cdots, K$ e l ≠ m

então $P_l \cap P_m = \varnothing$

Classificação Hierárquica 2/

Estruturas Classificatórias

Hierarquia



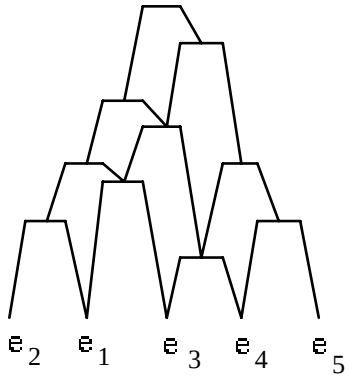
1) $\Omega \in H$

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

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

$h \cap h' \neq \varnothing \Rightarrow h \subset h'$ ou $h' \subset h$

Piramide



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

4) Existe uma ordem θ tal que

$\forall h \in H$, h é um intervalo de θ

Classificação Hierárquica 3/

- *Métodos de Agrupamento*

Em Taxinomia Numérica distingue-se três grupos de métodos
Técnicas de Otimização

Objetivo: obter uma partição. Número de grupos fornecido pelo usuário
Técnicas hierárquicas

Objetivo: obter uma hierarquia (ou uma pirâmide)
Pode-se obter uma partição "cortando-se" a hierarquia em um determinado nível.

Classificação Hierárquica 4/

- *Métodos de Agrupamento*

Técnicas de Cobertura

Objetivo: obter grupos que eventualmente podem partilhar indivíduos.

Outros Aspectos Relativos aos Métodos de Agrupamento

Métodos Aglomerativos versus Métodos Divisivos

Métodos Monotéticos versus Métodos Politeticos

Classificação Hierárquica 5/

- *Outros Aspectos Relativos aos Métodos de Agrupamento*

  Agrupamento Hard versus Agrupamento Fuzzy

  Métodos Incrementais versus Métodos não Incrementais

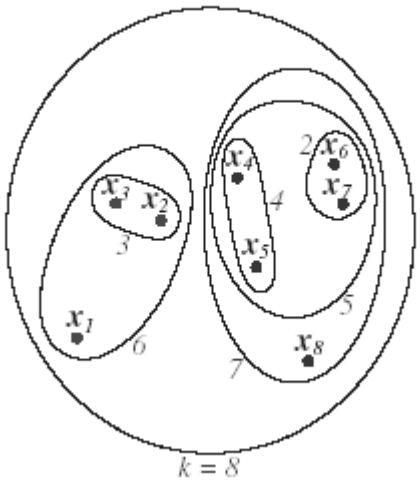
  Métodos Paramétricos versus Métodos não Paramétricos

  Métodos Geométricos versus Métodos não Geométricos

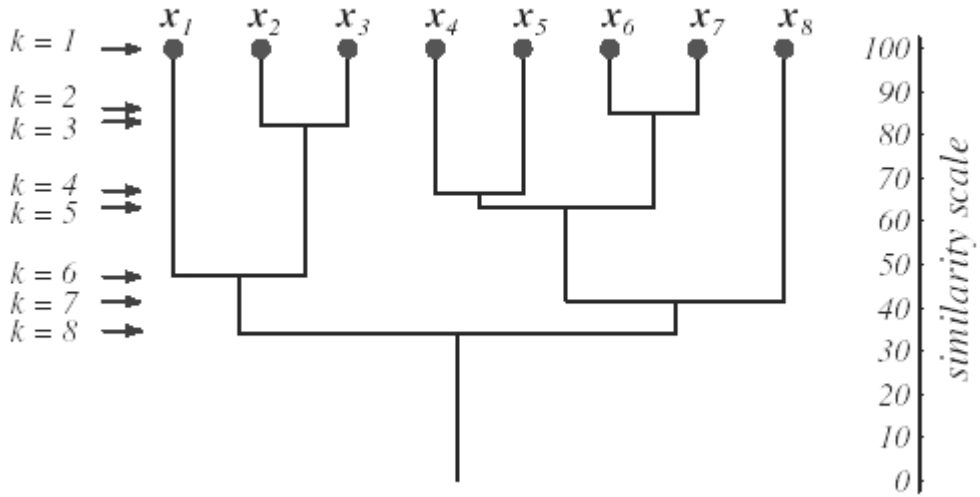
Classificação Hierárquica 6/

Classificação Hierarquica

Diagrama de Venn



Dendograma

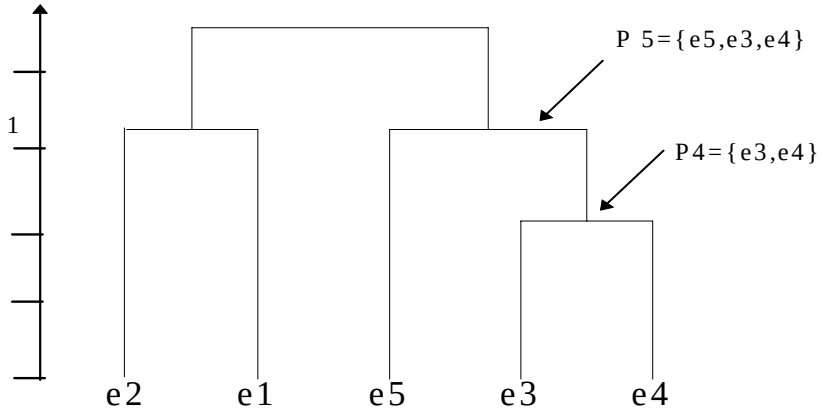


Classificação Hierárquica 7/

Métodos Hierárquicos Aglomerativos

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

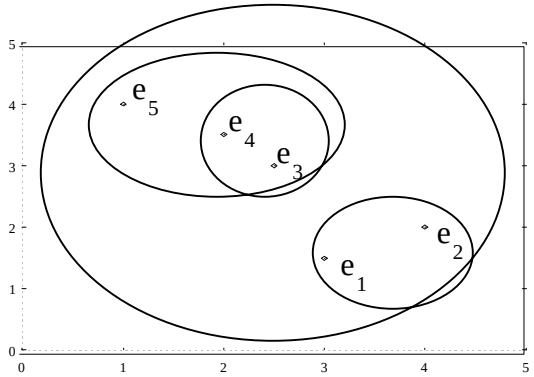
Os métodos ascendentes hierárquicos tem por objetivo a construção de uma sequencia de partições encaixadas chamada *hierarquia*. A representação gráfica dessas hierarquias é realisada por uma *arvore hierarquica* ou *dendrograma*.



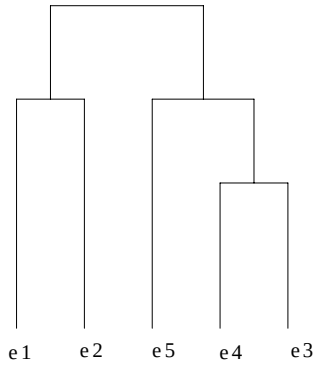
Classificação Hierárquica 8/

Métodos Hierárquicos Aglomerativos

Hierarquia H



Hierarquia com indice (H,f)



1) $E \in H$

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

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

$h \cap h' \neq \varnothing \Rightarrow h \subset h'$ ou $h' \subset h$

$f : H \rightarrow \Re^{+}$

(1) $f(h) = 0$ se e somente se $card(h)=1$

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

Classificação Hierárquica 9/

- Obtenção de um índice de dissimilaridade à partir de uma Hierarquia com Índice
  - Pode-se associar a uma hierarquia com índice (H,f) um índice de dissimilaridade σ: $\Omega \times \Omega \rightarrow \Re^+$ da seguinte maneira
    - $\sigma(\omega_k, \omega_l) = Min_{h \in H} \{f(h) \mid \omega_k \in h \text{ e } \omega_l \in h\}$
      (é a altura do nó mais baixo que contém ω_k e ω_l)

Classificação Hierárquica 10/

- Índice de Agregação entre grupos de indivíduos
  - Para construir uma Hierarquia é necessário um índice de proximidade entre grupos
  - Um índice de agregação entre de grupos de indivíduos é uma função

$$\begin{aligned} \delta : P(\Omega) \times P(\Omega) &\rightarrow \Re^{+} \\ (h_1, h_2) &\rightarrow \delta(h_1, h_2) \end{aligned}$$

tal que

$$1) \forall h_1, h_2 \in P(\Omega), \delta(h_1, h_2) \geq 0$$

$$2) \forall h_1, h_2 \in P(\Omega), \delta(h_1, h_2) = \delta(h_2, h_1)$$

Classificação Hierárquica 11/

- Índice de Agregação entre grupos de indivíduos
  - Relação entre f e δ

    f é um indice sobre a hierarquia H, δ é um indice de agregação entre classes

    $$\forall h_1, h_2 \in H,\ f(h_1 \cup h_2) = \delta(h_1, h_2)$$

  - Observação

    Para garantir que uma hierarquia com índice não apresente inversões, pode-se definir f da seguinte maneira:

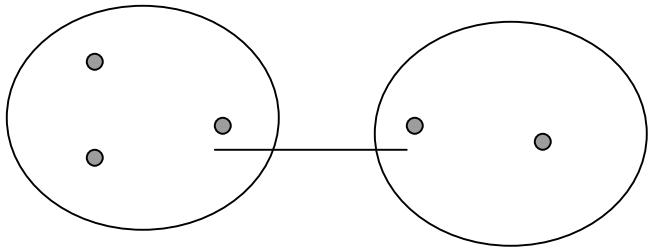
    $$\forall h_1, h_2 \in H,\ f(h_1 \cup h_2) = Max\{ f(h_1 \cup h_2), f(h_1), f(h_2)\}$$

Classificação Hierárquica 12/

- Principais Índice de Agregação utilizados
  - Os índices de agregação geralmente são construídos a partir do índice de dissimilaridade entre os indivíduos de Ω escolhido pelo usuário
  - Índice da ligação simples ou do vizinho mais próximo

Classificação Hierárquica 13/

- Índice da ligação simples ou do vizinho mais próximo



$$\delta(h_1,h_2) = \underset{\substack{\omega_i \in h_1 \\ \omega_j \in h_2}}{Min}\, d(\omega_i,\omega_j)$$

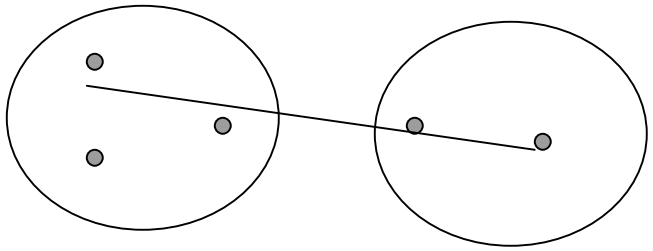
- Fórmula de recorrência

$$\delta(h, h_1 \cup h_2) = Min\{\delta(h,h_1),\delta(h,h_2)\} =$$

$$= \frac{1}{2}(\delta(h,h_1)+\delta(h,h_2)) - \frac{1}{2}|\delta(h,h_1)-\delta(h,h2)|$$

Classificação Hierárquica 14/

- Índice da ligação completa ou do vizinho mais longe



$$\delta(h_1,h_2) = \underset{\substack{\omega_i \in h_1 \\ \omega_j \in h_2}}{Max}\, d(\omega_i,\omega_j)$$

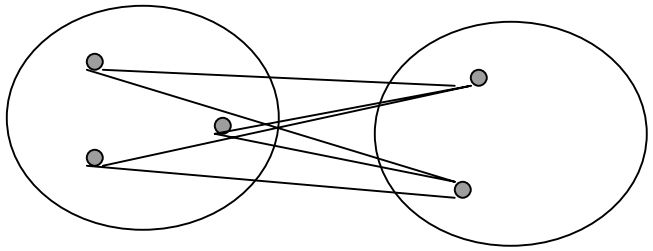
- Fórmula de recorrência

$$\delta(h, h_1 \cup h_2) = Max\{\delta(h,h_1),\delta(h,h_2)\} =$$

$$= \frac{1}{2}(\delta(h,h_1)+\delta(h,h_2)) + \frac{1}{2}|\delta(h,h_1)-\delta(h,h2)|$$

Classificação Hierárquica 15/

- Índice da média das distâncias



$$\delta(h_1,h_2)=\frac{1}{|h_1||h_2|}\sum_{\substack{\omega_i\in h_1\\ \omega_j\in h_2}} d(\omega_i,\omega_j)$$

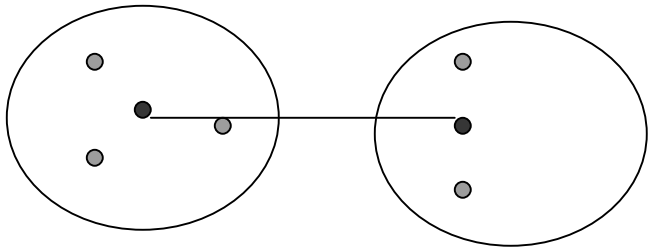
- Fórmula de recorrência

$$\delta(h,h_1\cup h_2)=\frac{1}{|h||h_1\cup h_2|}\sum_{\substack{\omega_i\in h_1\cup h_2\\ \omega_j\in h}} d(\omega_i\cup\omega_j)=$$

$$=\frac{|h_1|}{|h_1|+|h_2|}\delta(h,h_1)+\frac{|h_2|}{|h_1|+|h_2|}\delta(h,h_2)$$

Classificação Hierárquica 16/

- Índice de agregação dos centros de gravidade



$$\delta(h1_1, h_2) = d^2(G(h_1), G(h_2))$$

- Fórmula de recorrência

$$\delta(h, h_1 \cup h_2) = \frac{|h_1|}{|h_1| + |h_2|}\delta(h, h_1) + \frac{|h_2|}{|h_1| + |h_2|}\delta(h, h_2) -$$

$$- \frac{|h_1||h_2|}{(|h_1| + |h_2|)^2}\delta(h_1, h_2)$$

Classificação Hierárquica 17/

- Índice de minimização do aumento da variância (Ward)

$$\delta(h_1,h_2) = \frac{|h_1||h_2|}{|h_1|+|h_2|} d^2(G(h_1),G(h_2))$$

- Fórmula de recorrência

$$\delta(h,h_1 \cup h_2) = \frac{|h|+|h_1|}{|h|+|h_1|+|h_2|}\delta(h,h_1) + \frac{|h|+|h_2|}{|h|+|h_1|+|h_2|}\delta(h,h_2) -$$

$$- \frac{|h|}{|h|+|h_1|+|h_2|}\delta(h_1,h_2)$$

Classificação Hierárquica 18/

- Fórmula geral de recorrência de Lance e Williams

$$\delta(h, h_1 \cup h_2) = \alpha_1 \, \delta(h, h_1) + \alpha_2 \, \delta(h, h_2) +$$
$$+ \beta \, \delta(h_1, h_2) + \gamma \left| \delta(h, h_1) - \delta(h, h_2) \right|$$

Método	α_1	α_2	β	γ
Single – Link	*1 / 2*	*1 / 2*	*0*	*– 1 / 2*
Complet – Link	*1 / 2*	*1 / 2*	*0*	*1 / 2*
UPGMA	$\lvert h_1 \rvert / (\lvert h_1 \rvert + \lvert h_2 \rvert)$	$\lvert h_2 \rvert / (\lvert h_1 \rvert + \lvert h_2 \rvert)$	*0*	*0*
UPGMC	$\lvert h_1 \rvert / (\lvert h_1 \rvert + \lvert h_2 \rvert)$	$\lvert h_2 \rvert / (\lvert h_1 \rvert + \lvert h_2 \rvert)$	$-\lvert h_1 \rvert \lvert h_2 \rvert / (\lvert h_1 \rvert + \lvert h_2 \rvert)^2$	*0*
Ward	$(\lvert h \rvert + \lvert h_1 \rvert) / (\lvert h \rvert + \lvert h_1 \rvert + \lvert h_2 \rvert)$	$(\lvert h \rvert + \lvert h_2 \rvert) / (\lvert h \rvert + \lvert h_1 \rvert + \lvert h_2 \rvert)$	$-\lvert h \rvert / (\lvert h \rvert + \lvert h_1 \rvert + \lvert h_2 \rvert)$	*0*

Classificação Hierárquica 19/

- Inversão e Monotonicidade
  - Para que não ocorra inversão no dendograma é necessário que a fórmula de recorrência do índice de agregação seja monotonica, isto é,

$$\delta(h, h_1 \cup h_2) \geq \delta(h_1, h_2)$$

  - Considere a fórmula geral de recorrência:

$$\delta(h, h_1 \cup h_2) = \alpha_1 \, \delta(h, h_1) + \alpha_2 \, \delta(h, h_2) +$$
$$+ \beta \, \delta(h_1, h_2) + \gamma \left| \delta(h, h_1) - \delta(h, h_2) \right|$$

Classificação Hierárquica 20/

- Inversão e Monotonicidade
  - Se

$$\alpha_1 \geq 0, \alpha_2 \geq 0, \alpha_1 + \alpha_2 + \beta \geq 1 \; e \; \gamma \geq -Min(\alpha_1, \alpha_2)$$

então a fórmula de recorrência é monotonica

- Observação: dos métodos apresentados anteriormente apenas o UPGMC não é monotonico

Classificação Hierárquica 21/

- *Algoritmo Geral de Agrupamento Hierárquico Aglomerativo*

Passo 1: Iniciar o agrupamento formado por grupos unitários

Passo 2: Encontre, no agrupamento corrente, o par de grupos de dissimilaridade mínima

Passo 3: Construa um novo grupo pela fusão desse par de grupos de dissimilaridade mínima

Classificação Hierárquica 22/

- *Algoritmo Geral de Agrupamento Hierárquico Aglomerativo*

  Passo 4: Atualize a matriz de dissimilaridades:

  - suprima as linhas e as colunas correspondentes aos grupos fusionados e
  - adicione uma linha e uma coluna correspondente as dissimilaridades entre o novo grupo e os grupos antigos

- Passo 5: Se todos os objetos estão grupados, pare; senão vá para o passo 2

Classificação Hierárquica 23/

- *Exemplo*

E01:(Sono=Pouco,T=Carro,Conic=Sim,Alcool=Não,Sair=Não,Fome=Sim)
E02:(Sono=Pouco,T=Carona,Conic=Não,Alcool=Não,Sair=Sim,Fome=Sim)
E03:(Sono=Sim,T=Carro,Conic=Não,Alcool=Sim,Sair=Sim,Fome=Não)
E04:(Sono=Sim,T=Outros,Conic=Sim,Alcool=Sim,Sair=Sim,Fome=Não)

Classificação Hierárquica 24/

- *Exemplo*

Matriz de Dissimilaridades

$$D = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0 & 3 & 5 & 5 \\ & 0 & 4 & 5 \\ & & 0 & 2 \\ & & & 0 \end{pmatrix} \end{array}$$

Classificação Hierárquica 25/

- *Exemplo*

Passo 1: $C1$={E01}, $C2$={E02}, $C3$={E03}, $C4$={E04}

Passo 2: dmin = 2 $\Rightarrow$ $C5$= $C3 \cup C4$ = {E03,E04}

Passo 3:

$$D = \begin{array}{c} \\ 1 \\ 2 \\ 5 \end{array} \begin{pmatrix} 0 & 3 & 5 \\ & 0 & 4 \\ & & 0 \end{pmatrix}$$

Classificação Hierárquica 26/

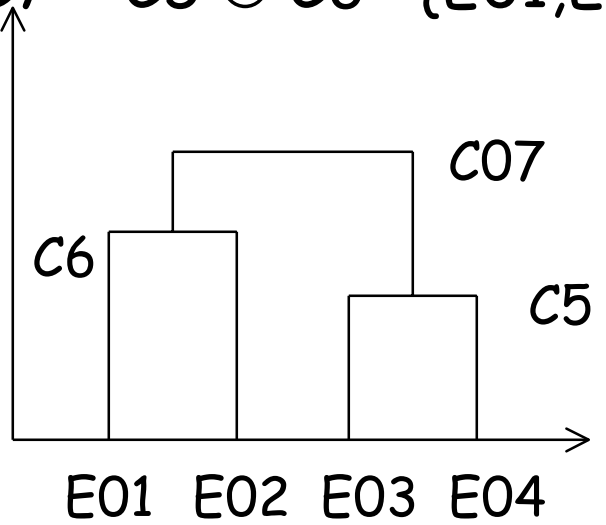
- *Exemplo*

Passo 4: dmin = 3 $\Rightarrow$ C6= C1 $\cup$ C2 ={E01,E02}

Passo5

$$D = \begin{matrix} 5 \\ 6 \end{matrix} \begin{pmatrix} 0 & 4 \\ & 0 \end{pmatrix}$$

Passo 6: dmin = 4 $\Rightarrow$ C7 = C5 $\cup$ C6 ={E01,E02,E03,E04}



Classificação Hierárquica 27/

- *Métodos Hierárquicos Divisivos*
  - Métodos Politéticos x Métodos Monotéticos
- Algoritmo da divisão pela média das distâncias
  - Passo 1

$$R = \{1, \ldots, n\}$$

  - Passo 2

$$A = R \; e \; B = \emptyset$$

$$\forall i \in A \quad calcule \; d(i, A \setminus \{i\}) = \frac{1}{|A| - 1} \sum_{\substack{j \in A \\ j \neq i}} d(i, j)$$

Classificação Hierárquica 28/

- Algoritmo da divisão pela média das distâncias
  - Passo 2

$$Encontre\ i' \in A\ tal\ que\ d(i', A\setminus\{i'\}) = \max_{i \in A} d(i, A\setminus\{i\})$$

$$Faça\ A = A\setminus\{i'\}\ e\ B = B \cup \{i'\}$$

  - Passo 3

$$\forall i \in A\ calcule\ \ d(i, A\setminus\{i\}) - d(i, B) = \frac{1}{|A| - 1} \sum_{\substack{j \in A \\ j \neq i}} d(i, j) - \frac{1}{|B|} \sum_{h \in B} d(i, h)$$

$$Encontre\ \ i' \in A\ tal\ \ que\ \ d(i', A\setminus\{i'\}) - d(i', B) = \max_{i \in A} \{ d(i, A\setminus\{i\}) - d(i, B)\}$$

Classificação Hierárquica 29/

- Algoritmo da divisão pela média das distâncias
  - Passo 4

$$Se\ d(i', A\setminus\{i'\}) - d(i', B) > 0$$

$$A = A\setminus\{i'\}\ e\ B = B\cup\{i'\}$$

$$Re torne\ \ ao\ \ PASSO\ \ 3$$

  - Passo 5

$$Calcule\ dim(Q) = \max_{\substack{j\in Q \\ h\in Q}} d(j,h)\ onde\ Q\ é\ um\ grupo$$

$$Escolha\ Q^*\ tal\ que\ dim(Q^*) = Max\ dim(Q)$$

$$Faça\ L = dim(Q^*), R = Q^*\ e\ va\ \ para\ o\ PASSO\ 2$$

  - Passo 6: Pare quando todos os grupos forem unitários

Classificação Hierárquica 30/

- Algoritmo da divisão pela média das distâncias
  - Exemplo: agrupe os pontos

	Altura	*Peso*	*Idade*
A	*180*	*79*	*30*
B	*175*	*75*	*25*
C	*170*	*70*	*28*
D	*167*	*63*	*21*
E	*180*	*71*	*18*
F	*165*	*60*	*28*

Classificação Hierárquica 31/

- Métodos Divisivos (Monotético)

Todas as variáveis são binárias

- Passo 1

$$R = \{1, \ldots, n\} \, (indivíduos)$$
$$Y = \{y_1, \ldots, y_p\} \, (variáveis)$$

Classificação Hierárquica 32/

- Métodos Divisivos (Monotético)
  Passo 2

$$\forall j \; calcule \; C(y_j) = \sum_{k \neq j} c(y_j, y_k) \; onde$$

$$c(y_j, y_k) = \left| a_{jk} d_{jk} - b_{jk} c_{jk} \right| \quad e$$

j \ k	1	0
1	a_{jk}	b_{jk}
0	c_{jk}	d_{jk}

Classificação Hierárquica 33/

- Métodos Divisivos (Monotético)
  Passo 2

$$Selecione\ t\ tal\ que\ C(y_t) = \underset{y_j \in Y}{Max}\, C(y_j)$$

$$Faça\ A = \{ i \mid y_t(i) = 1 \}, B = \{ i \mid y_t(i) = 0 \}\ e\ Y = Y \setminus \{ y_t \}$$

- Passo 3: Repita o passo 2 para cada grupo
- Passo 4: O processo de separação continua até
  - só restar grupos unitários ou
  - as variáveis restantes são incapazes de separar os grupos ou
  - não restam mais variáveis

Classificação Hierárquica 34/

- Métodos Divisivos (Monotético)
  - Exemplo

	1	2	3	4	5	6
A	1	1	0	1	1	0
B	1	1	0	0	0	1
C	1	1	1	1	1	0
D	1	1	1	1	1	0
E	0	0	0	1	0	1
F	0	0	0	0	0	0
G	0	0	1	1	0	1
H	0	0	1	1	1	0