O que é Agrupamento?

Conhecido também por *aprendizado não-supervisionado* e, às vezes, chamado de *classificação* por estatísticos e de *segmentação* por pessoas de *marketing*

- O que é agrupamento?
- Algumas aplicações
- Definição formal e complexidade computacional
- Objetivos e características
- Grupos naturais
- Medidas de (dis)similaridade
- A tarefa de agrupamento e desafios
- Agrupamento particional
  - K-Médias
  - C-Médias Nebuloso
- Agrupamento hierárquico
  - Single-link
- Medidas de avaliação

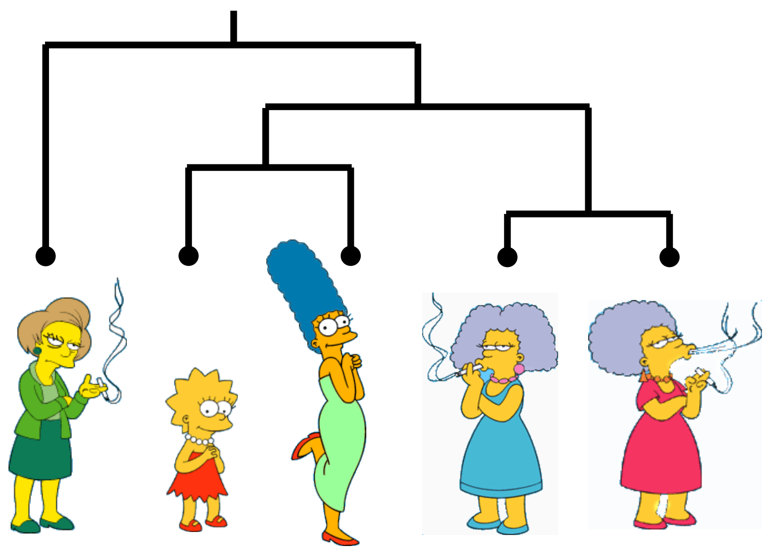
Dois Tipos de Agrupamento

- **Algoritmos Hierárquicos:**

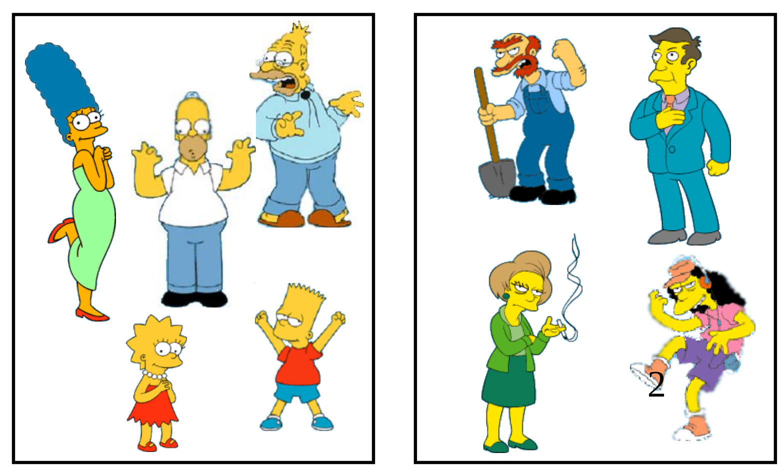
➢ Os objetos são organziados em uma estrutura hierárquica chamada *dendrograma*, onde é possível visualizar diferentes quantidades de grupos simultaneamente

➢ Obtêm uma solução com todas as variações de *k*

Hierárquico



Particional

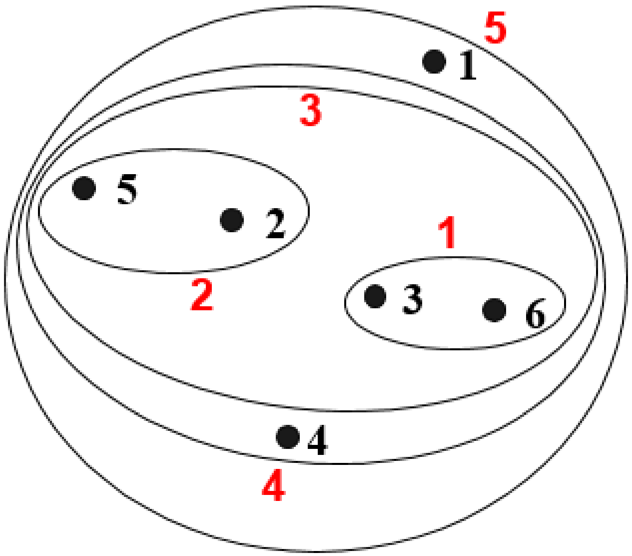


Agrupamento hierárquico

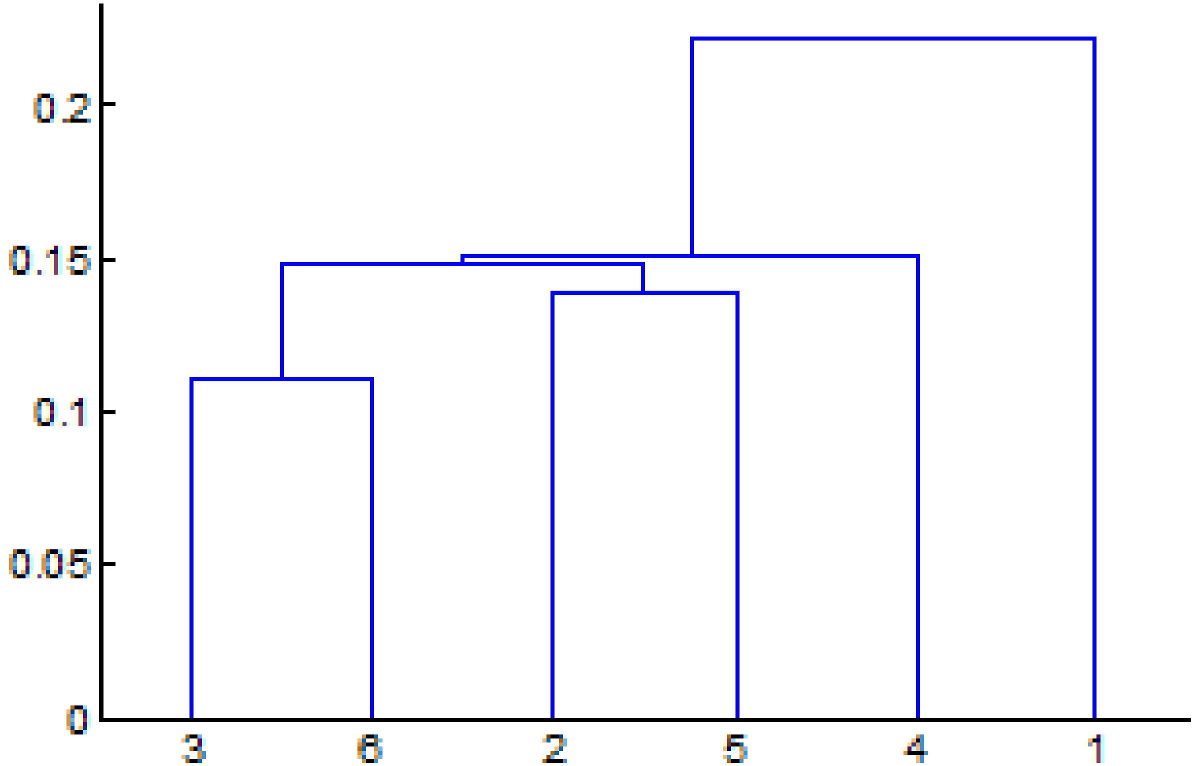
• Os métodos hierárquicos são caracterizados por *sucessivas divisões* ou *fusões* ***hierárquicas*** dos dados, geralmente apresentando como resultado um *dendrograma*, o qual representa os possíveis agrupamentos de dados. Os métodos hierárquicos agrupam os objetos dentro de uma árvore de grupos, podendo ser:

– *Aglomerativos*: inicialmente cada objeto pertence a um grupo e os objetos se unem sucessivamente em grupos até que um critério de parada seja atingido. Cada passo combina os 2 grupos mais similares

Agrupamento aglomerativo



Grupos aninhados

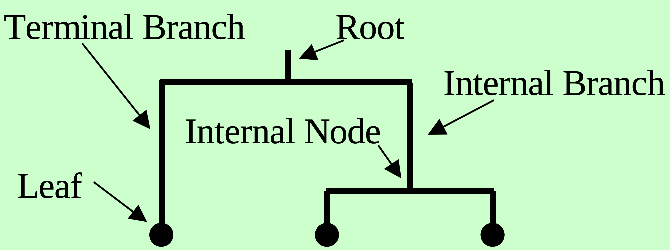


Dendrograma

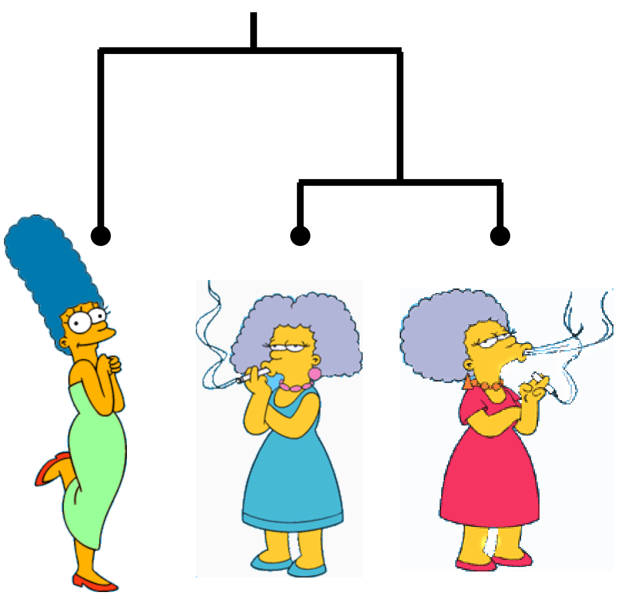
Agrupamento hierárquico

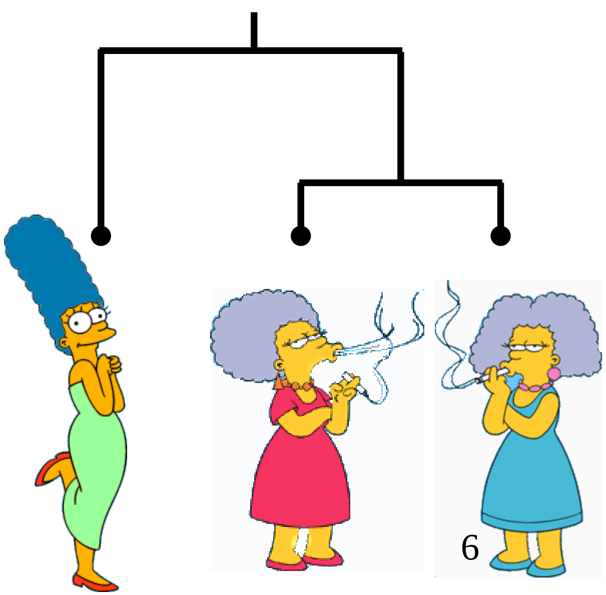
- *Divisivos*: no início do processo de agrupamento todos os objetos fazem parte do mesmo grupo, que é dividido sucessivamente em grupos menores até que um critério de parada seja atingido. Cada passo divide o grupo menos homogêneo em 2 novos grupos

Uma Ferramenta Útil para Resumir as Similaridades

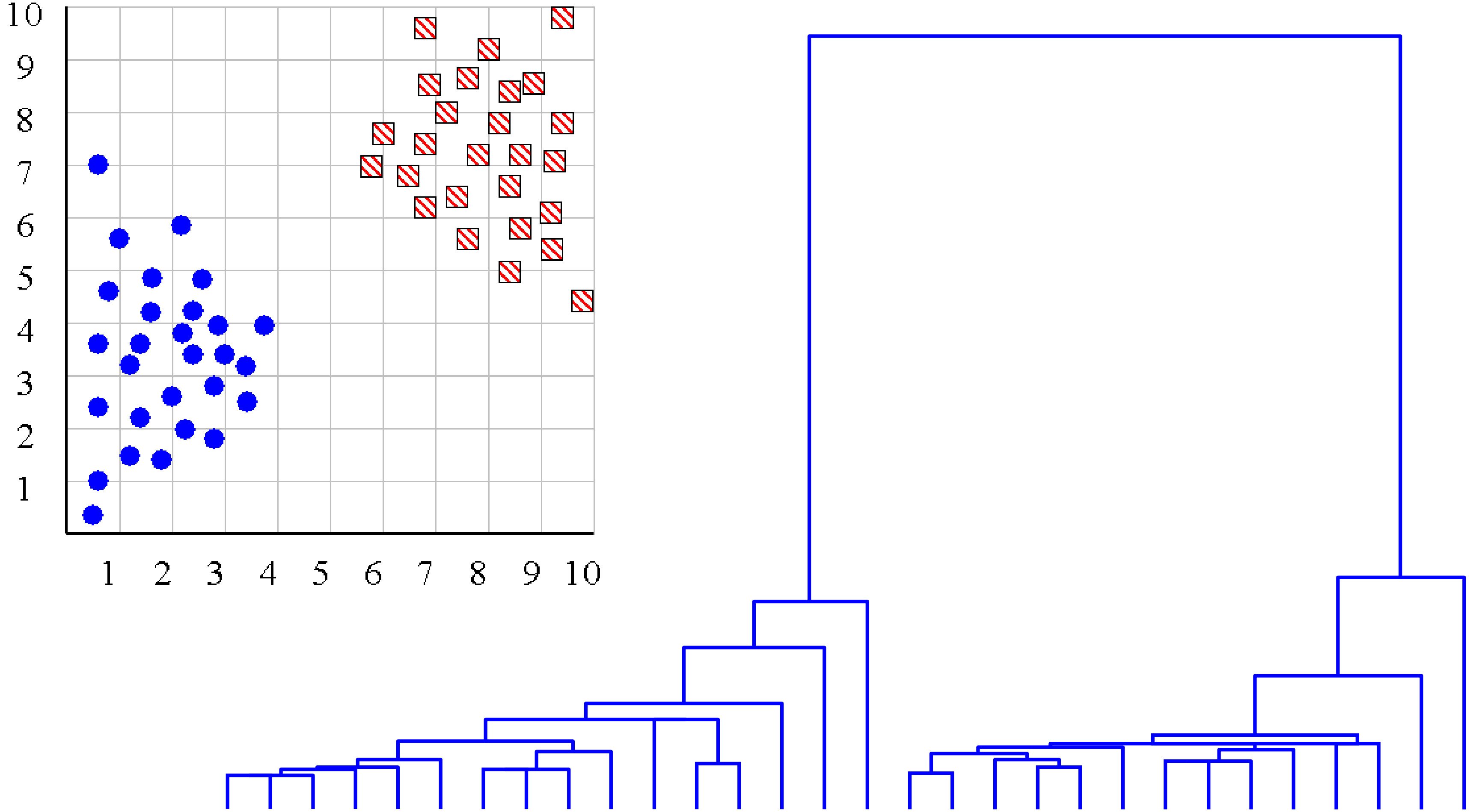


A **similaridade** entre dois objetos em um dendograma é **representada pela altura** do nó interno mais baixo que eles compartilham.

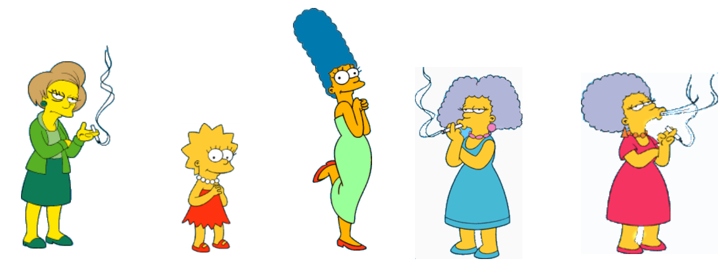


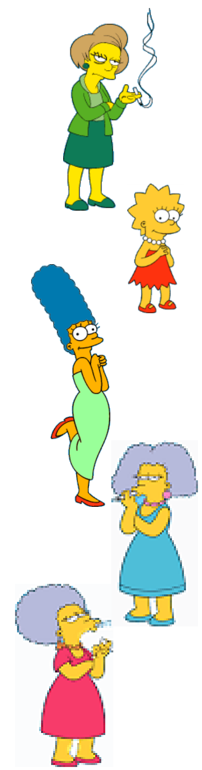


Nós podemos olhar no dendograma para determinar o número "correto" de agrupamentos. Nesse caso, a existência de duas árvores bem separadas é um forte indicativo de dois *clusters*. (Infelizmente, raramente as coisas são assim tão claras.)



Nós começamos com uma matriz de distâncias que contém as distâncias entre cada par de objetos no nosso banco de dados.





0	8	8	7	7
	0	2	4	4
		0	3	3
			0	1
				0

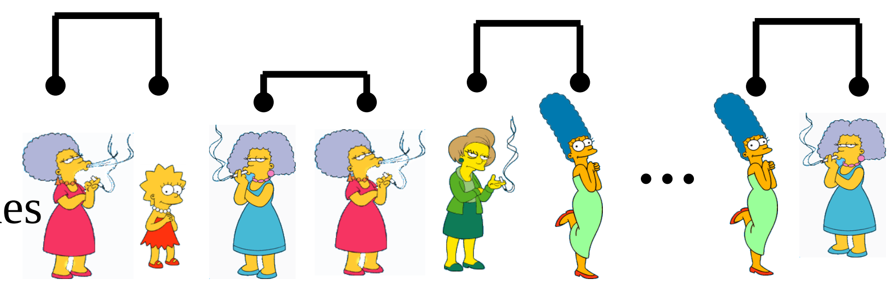
D() = 8

D() = 1

Bottom-Up (aglomerativo):

Começando com cada item em seu próprio *cluster*, encontrar o melhor par para aglomerar em um novo *cluster*. Repetir até que todos os *clusters* tenham sido aglomerados em um único.

Considere todas as possibilidades …

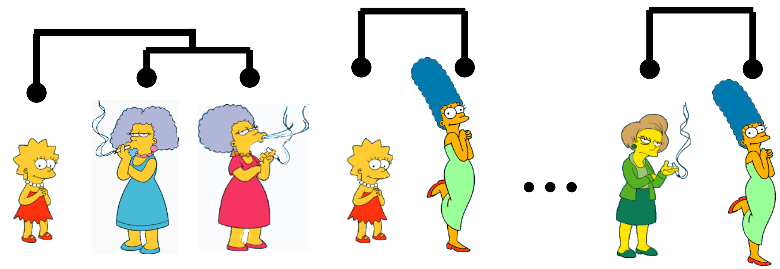


Escolha a melhor

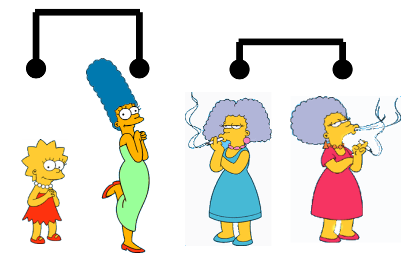
Bottom-Up (aglomerativo):

Começando com cada item em seu próprio *cluster*, encontrar o melhor par para aglomerar em um novo *cluster*. Repetir até que todos os clusters tenham sido aglomerados em um único.

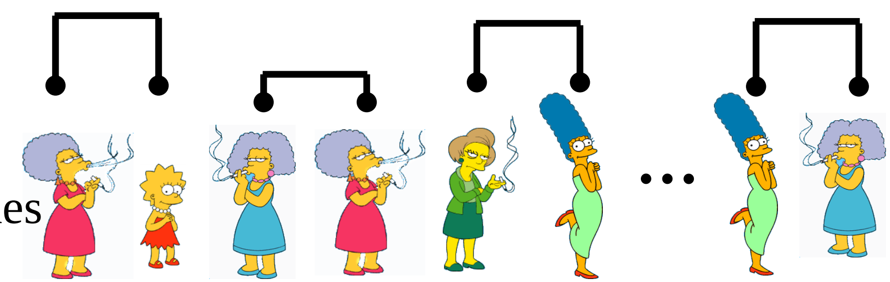
Considere todas as possibilidades …



Escolha a melhor



Considere todas as possibilidades …

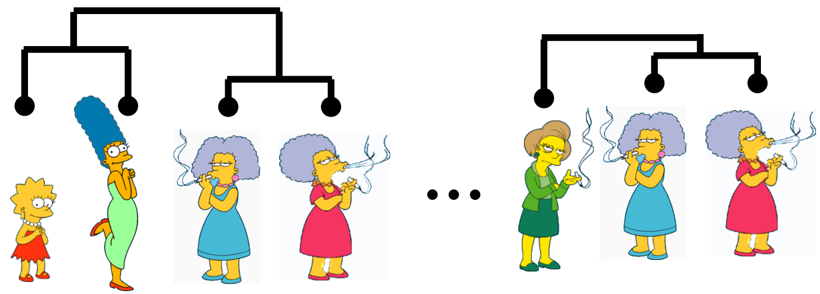


Escolha a melhor

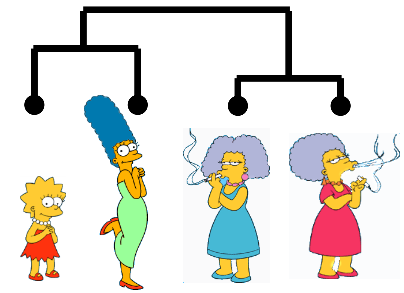
Bottom-Up (aglomerativo):

Começando com cada item em seu próprio *cluster*, encontrar o melhor par para aglomerar em um novo *cluster*. Repetir até que todos os clusters tenham sido aglomerados em um único.

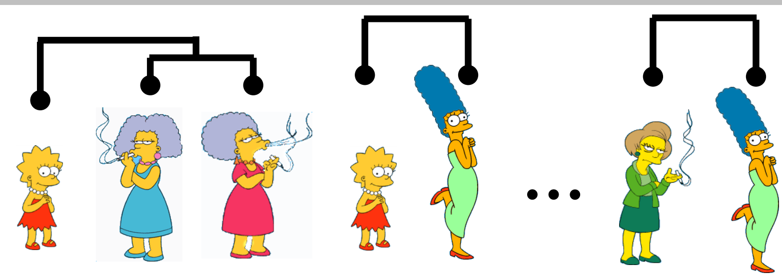
Considere todas as possibilidades …



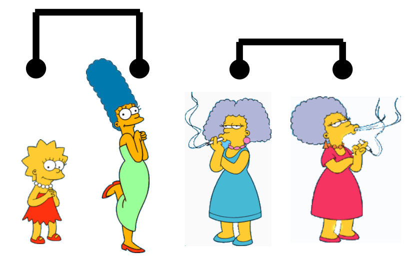
Escolha a melhor



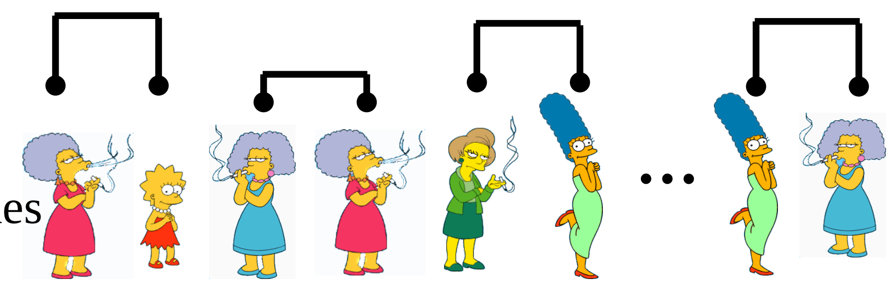
Considere todas as possibilidades …



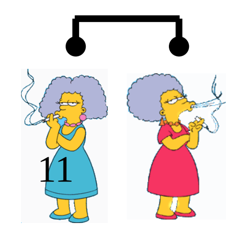
Escolha a melhor



Considere todas as possibilidades …

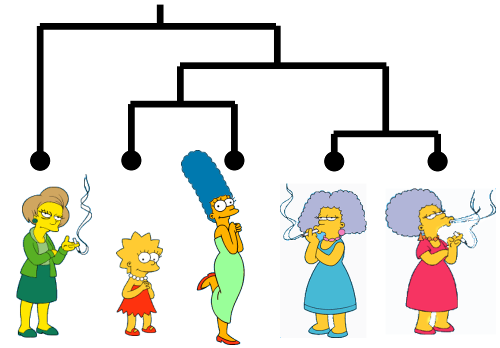


Escolha a melhor

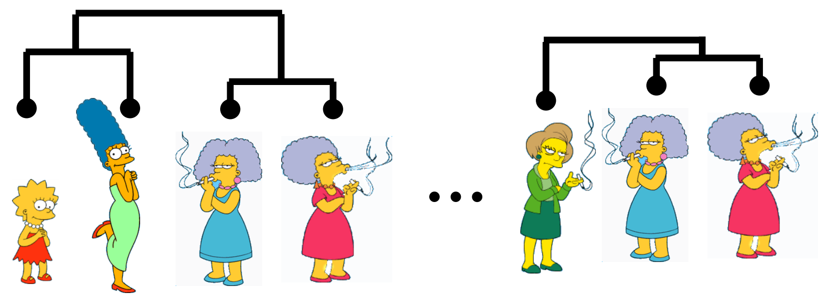


Bottom-Up (aglomerativo):

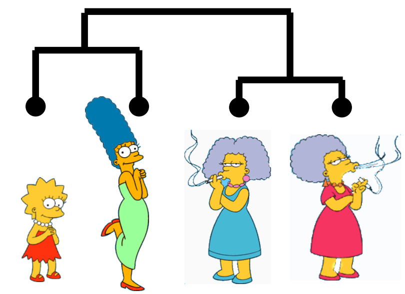
Começando com cada item em seu próprio *cluster*, encontrar o melhor par para aglomerar em um novo *cluster*. Repetir até que todos os clusters tenham sido aglomerados em um único.



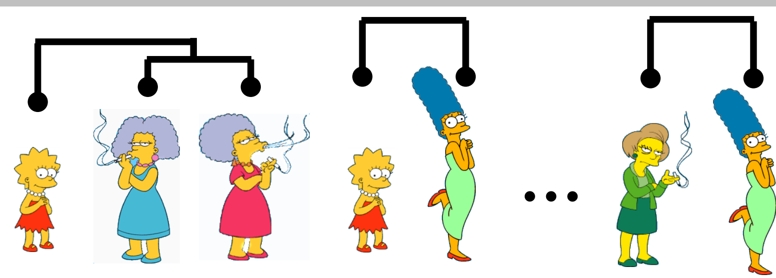
Considere todas as possibilidades …



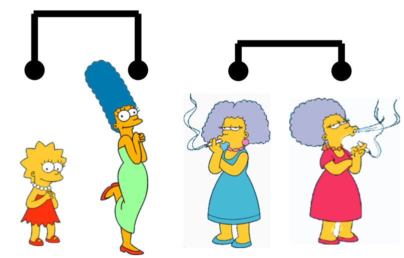
Escolha a melhor



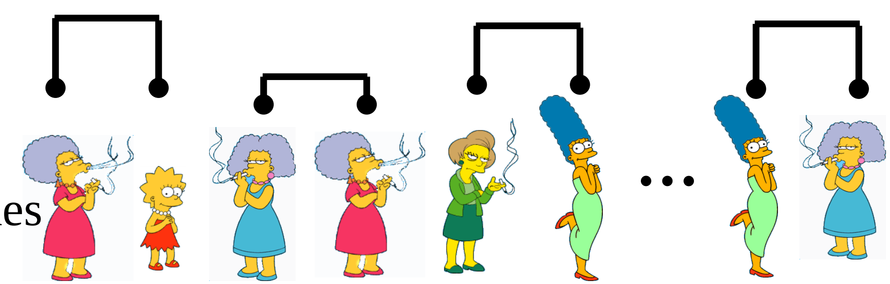
Considere todas as possibilidades …



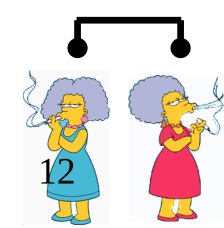
Escolha a melhor



Considere todas as possibilidades …



Escolha a melhor



O algoritmo Single-link

- O algoritmo hierárquico simples (ou *Single Linkage*) é um algoritmo aglomerativo. Ele inicia com um objeto pertencendo a um grupo e os aglomera até que todos pertençam a um único grupo

- O processo se inicia com uma *matriz de distâncias* entre todos os objetos. Na sequência, *um processo iterativo* de união sempre entre os dois grupos mais similares é realizado. Os grupos selecionados são substituídos na matriz de distâncias por um novo grupo, onde a distância desse novo grupo aos demais grupos é definida pela menor distância entre os dois grupos selecionados e os grupos restantes

O algoritmo Single-link

1. Calcular a matriz de distância entre os objetos
2. Enquanto existir mais de um grupo, faça
   a) Encontre e junte os dois grupos mais próximos
   b) Atualize a matriz de distância entre os grupos
3. Defina um ponto de corte para obter o agrupamento

Exemplo: Single Link

1. **Método do vizinho mais próximo (Método da ligação simples- *Single Link*)**

   Para o nosso exemplo suponha a seguinte matriz de distâncias:

$$
\begin{array}{cc}
 & \begin{array}{ccccc} \text{A} & \text{B} & \text{C} & \text{D} & \text{E} \end{array} \\
\begin{array}{c} \text{B} \\ \text{C} \\ \text{D} \\ \text{E} \\ \text{F} \end{array} &
\begin{bmatrix}
0,67 & & & & \\
1,41 & 0,74 & & & \\
2,12 & 1,47 & 0,77 & & \\
0,79 & 0,67 & 1,09 & 1,62 & \\
2,49 & 1,84 & 1,13 & 0,37 & 1,96
\end{bmatrix}
\end{array}
$$

Exemplo: Single Link

- <u>Passo 1:</u> inicialmente, cada caso forma um grupo, isto é, temos 6 grupos iniciais.

- <u>Passo 2:</u> olhando-se a matriz de distâncias, observa-se que as duas observações mais próximas são D e F, corresponde a uma distância de 0,37, assim, esta duas observações são agrupadas, formando o primeiro grupo. Necessita-se, agora, das distâncias deste grupo aos demais. A partir da matriz de distâncias iniciais têm-se**:**

$$d(A,DF) = \min\{d(A,D), d(A,F)\} = \min\{2{,}12\,;\,2{,}49\} = 2{,}12$$

$$d(B,DF) = \min\{d(B,D), d(B,F)\} = \min\{1{,}47\,;\,1{,}84\} = 1{,}47$$

$$d(C,DF) = \min\{d(C,D), d(C,F)\} = \min\{0{,}77\,;\,1{,}13\} = 0{,}77$$

$$d(E,DF) = \min\{d(E,D), d(E,F)\} = \min\{1{,}62\,;\,1{,}96\} = 1{,}62$$

Com isso, temos a seguinte matriz de distâncias:

Exemplo: Single Link

$$
\begin{array}{c|cccc}
 & A & B & C & E \\
\hline
B & 0{,}67 & & & \\
C & 1{,}41 & 0{,}74 & & \\
E & 0{,}79 & 0{,}67 & 1{,}09 & \\
DF & 2{,}12 & 1{,}47 & 0{,}77 & 1{,}62
\end{array}
$$

- **<u>Passo 3:</u> Agrupar A e B ao nível de 0,67, e recalcular:**

$$d(C,AB) = min\{d(C,A),d(C,B)\} = min\{1{,}41;0{,}74\} = 0{,}74$$

$$d(E,AB) = min\{d(E,A),d(E,B)\} = min\{0{,}79;0{,}67\} = 0{,}67$$

$$d(DF,AB) = min\{d(D,A),d(D,B),d(F,A),d(F,B)\} = min\{2{,}12;1{,}47;2{,}49;1{,}84\} = 1{,}47$$

A matriz resultante será:

Exemplo: Single Link

$$\begin{array}{cc} & \begin{array}{ccc} C & E & DF \end{array} \\ \begin{array}{c} E \\ DF \\ AB \end{array} & \begin{bmatrix} 1{,}09 & & \\ 0{,}77 & 1{,}62 & \\ 0{,}74 & 0{,}67 & 1{,}47 \end{bmatrix} \end{array}$$

- **<u>Passo 4:</u> Agrupar AB com E ao nível de 0,67, e recalcular:**

$$d(C,ABE) = min\{d(C,A), d(C,B), d(C,E)\} = min\{1{,}41; 0{,}74; 1{,}09\} = 0{,}74$$

$$d(DF,ABE) = min\{d(D,A), d(D,B), d(D,E), d(F,A), d(F,B), d(F,E)\} =$$
$$min\{2{,}12; 1{,}47; 1{,}62; 2{,}49; 1{,}84; 1{,}96\} = 1{,}47$$

Matriz resultante:

$$\begin{array}{cc} & \begin{array}{cc} C & DF \end{array} \\ \begin{array}{c} DF \\ ABE \end{array} & \begin{bmatrix} 0{,}77 & \\ 0{,}74 & 1{,}47 \end{bmatrix} \end{array}$$

Exemplo: Single Link

- **<u>Passo 5:</u> Agrupar C com ABE ao nível de 0,74, e recalcular:**

$$d(DF, ABCE) =$$
$$\min\{d(D,A), d(D,B), d(D,C), d(D,E), d(F,A), d(F,B), d(F,C), d(F,E)\} =$$
$$\min\{2{,}12; 1{,}47; 0{,}77; 1{,}62; 2{,}49; 1{,}84; 1{,}13; 1{,}96\} = 0{,}77$$

Matriz resultante:

$$\begin{array}{cc} & \textbf{DF} \\ \textbf{ABCE} & \begin{bmatrix} 0{,}77 \end{bmatrix} \end{array}$$

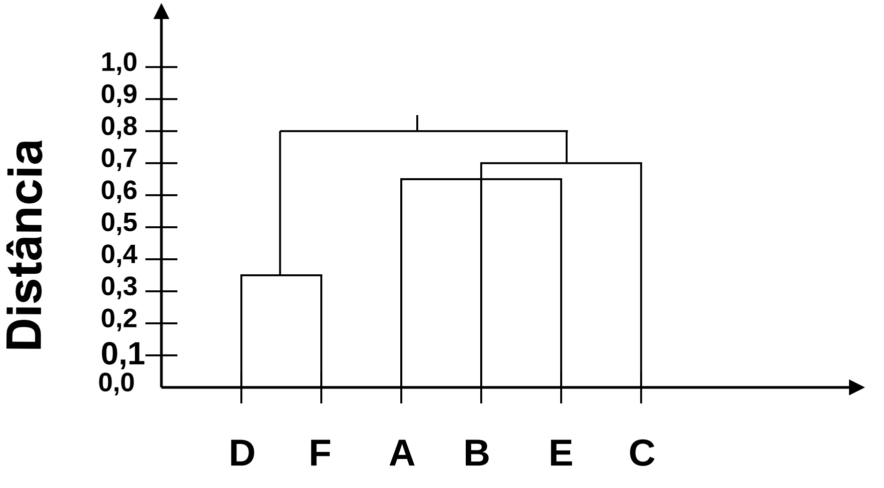
- **<u>Passo 6:</u> O último passo cria um único agrupamento contendo os 6 objetos, que serão similares a um nível de 0,77**

Exemplo: Single Link

Resumindo-se, temos:

Nó	Fusão	Nível
1	D e F	0,37
2	A e B	0,67
3	AB e E	0,67
4	ABE e C	0,74
5	ABCE e DF	0,77

Dendrograma:



Resumo sobre agrupamento hierárquico

• Não existe a necessidade de especificar o número de grupos, *a priori*

• As divisões ou fusões não podem ser desfeitas

• O método opera com uma matriz de proximidade (dissimilaridade) entre os objetos

•O resultado de um algoritmo hierárquico é um dendrograma representando o agrupamento

Métricas de Avaliação de Agrupamento

- A análise de desempenho de algoritmos fornece um grau de confiabilidade dos resultados produzidos pelos mesmos. Isto implica na decisão de escolher o algoritmo mais eficiente e eficaz para uma determinada aplicação

- Medidas de avaliação de algoritmos de agrupamento que consideram uma partição desejada são denominadas de *medidas externas*, enquanto as *medidas internas* consideram apenas distâncias inter- e/ou intragrupos para quantificar o desempenho dos algoritmos

- Em aplicações práticas a qualidade das partições deve ser avaliada por alguma *medida interna*, uma vez que o rótulo dos objetos não é conhecido *a priori*

Algumas Medidas Externas

- Entropia

  - A entropia mede a homogeneidade de um grupo
  - Esta informação mostra como os objetos da base de dados estão distribuídos nos grupos encontrados

$$\boldsymbol{E(S_r) = -\frac{1}{\log k}\sum_{i=1}^{k}\frac{n_r^i}{n_r}\, log\,\frac{n_r^i}{n_r}}$$

  - $\boldsymbol{S}_r$ é o grupo avaliado
  - k é o número total de classes da base de dados
  - n_r^i é o número de objetos da classe i
  - n_r o número de objetos no grupo $\boldsymbol{S}_r$

Algumas Medidas Externas

- Ainda sobre entropia...

  - Baixo valor de entropia indica melhor qualidade do grupo.
  - A entropia global é a soma da entropia de cada grupo, ponderada pelo tamanho de cada grupo:

$$\boldsymbol{E_g = \sum_{r=1}^{k} \frac{n_r}{n} E_r}$$

  - g indica que se trata da entropia global
  - r representa um grupo particular
  - k é o total de grupos na base de dados
  - n_r é o número de objetos no grupo r
  - n é o número de objetos na base de dados

Algumas Medidas Externas

- Pureza

  - A pureza indica quão puro é o grupo avaliado, isto é, a razão da classe dominante de um grupo em relação ao número total de objetos neste grupo

$$\boldsymbol{P(\mathbf{S}_r)} = \frac{\mathbf{max}(\boldsymbol{n_r^i})}{\boldsymbol{n_r}}$$

  - n_r^i é o número de objetos da classe *i* no grupo *r*
  - n_r o número de objetos no grupo $\mathbf{S}_r$

Algumas Medidas Externas

- Ainda sobre pureza...

  - Quanto maior a pureza melhor é a qualidade do grupo. A pureza global segue a mesma ideia do cálculo da entropia global:

$$\boldsymbol{P_g = \sum_{r=1}^{k} \frac{n_r}{n} P_r}$$

  - g indica que se trata da entropia global
  - r representa a um grupo particular
  - k é o número total de grupos na base de dados
  - n_r é o número de objetos no grupo $\boldsymbol{S}_r$
  - n é o número de objetos da base de dados

Algumas Medidas Internas

- Silhueta Simplificada

  - Mostra quais objetos estão bem situados dentro dos seus grupos e quais estão fora do cluster apropriado

$$\boldsymbol{s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}}$$

  - *i* é o objeto da base dados
  - *a*(i) é distância do objeto *i* ao seu respectivo centroide
  - *b*(i) é a menor distância do objeto *i* aos centroides dos demais grupos

Algumas Medidas Internas

- Ainda sobre a Silhueta Simplificada...

$$\boldsymbol{S} = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}} \boldsymbol{s}(\boldsymbol{i})$$

  - S é silhueta média sobre todos os objeto da base dados
  - n é a quantidade de objetos
  - s(i) é a silhueta do i-ésimo objeto
  - $S \in [-1,1]$; quanto maior o S, melhor é o agrupamento

Algumas Medidas Internas

- Xie-Beni

  - Avalia simultaneamente quão compactos e separados são os grupos dentro de uma partição nebulosa (ou *crisp*):

$$\mathbf{x}\boldsymbol{b} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k}\left(\mu_{ij}\right)^{2} * \left\|X_i - C_j\right\|^{2}}{n * min(\left\|C_j - C_k\right\|^{2}), j \neq k}$$

  - n e k são o número de objetos na base de dados e o número de protótipos, respectivamente
  - μ_{ij} representa o grau de pertinência do objeto $\mathbf{X}_i$ ao protótipo $\mathbf{C}_j$
  - $\|.\|^2$ é a distância Euclidiana
  - $xb \in [0,1]$; quanto menor o xb, mais compactos e separados são os grupos encontrados