



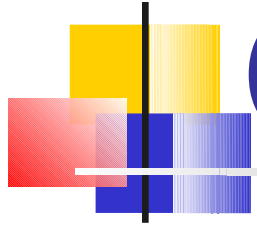
Inteligência Artificial

Profa. Patrícia R. Oliveira
EACH / USP

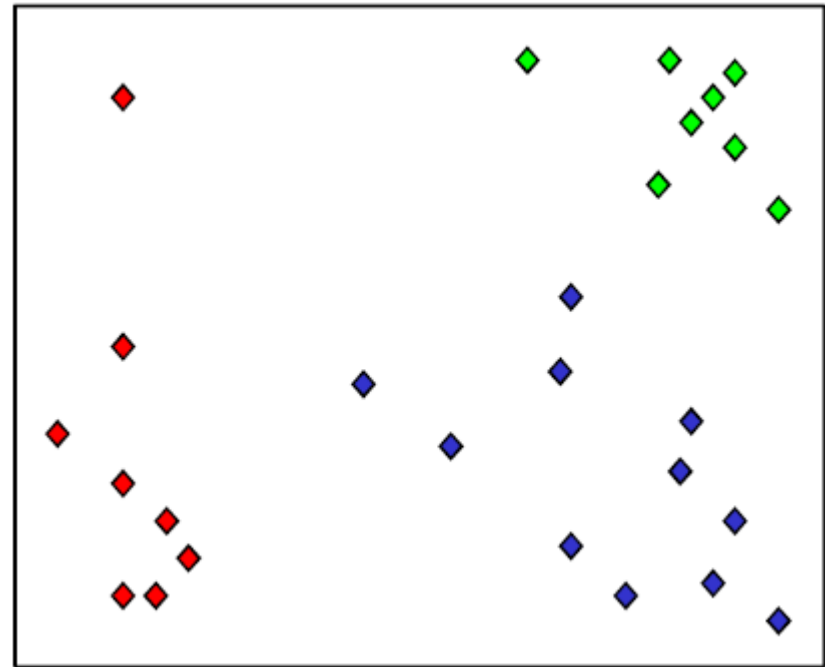
Parte 7 – Aprendizado Não Supervisionado: Clustering

Este material é parcialmente baseado em slides
do Prof. Eduardo Raul Hruschka (ICMC/USP)

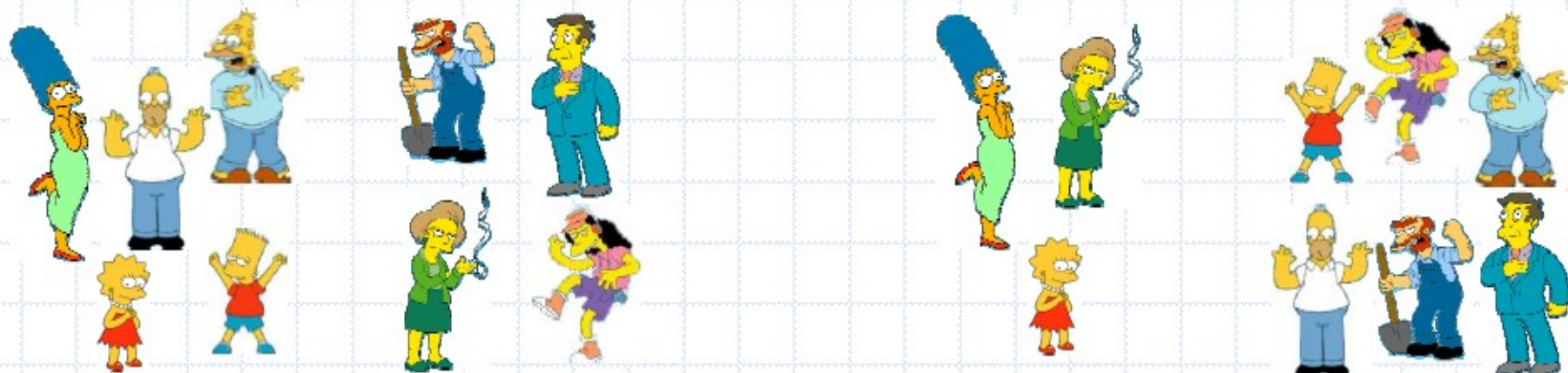
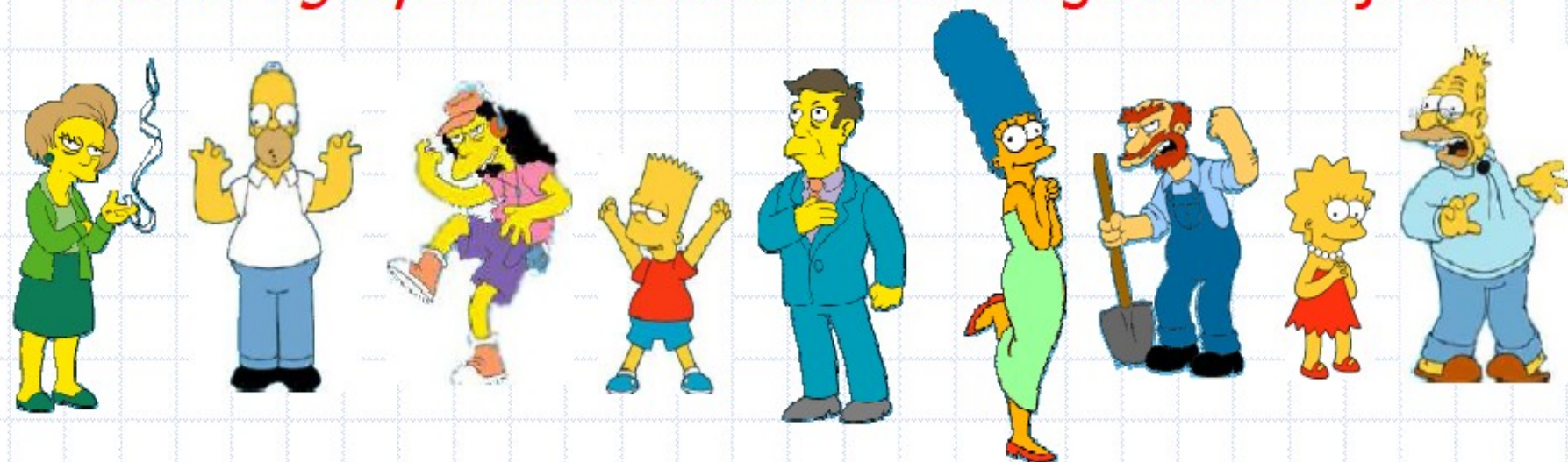
O que é Clustering (Agrupamento de dados)?



- Organização dos dados em clusters (grupos) tal que:
 - O grau de similaridade entre objetos dentro do mesmo cluster seja alta;
 - O grau de similaridade entre objetos de clusters diferentes seja baixo.
- Um método de clustering deve encontrar agrupamentos naturais entre os objetos.



Como *agrupar naturalmente* os seguintes objetos?



Família

Empregados

Mulheres

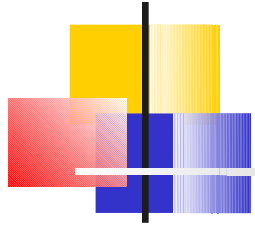
Homens

→ *Cluster* é um conceito subjetivo!

O que é um cluster (grupo)?

- Definições subjetivas:
 - “Semelhanças entre objetos”.
 - Quais atributos devemos considerar para computar similaridades?



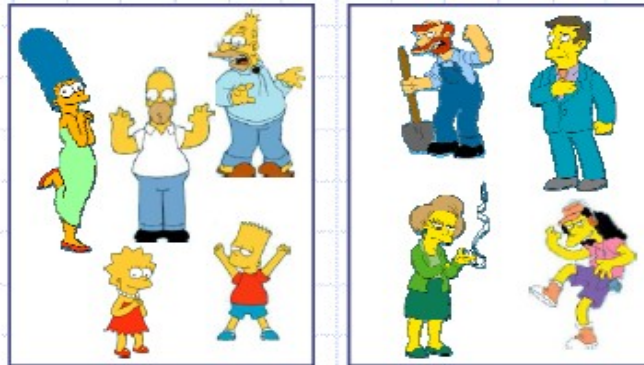


Tipos de Clustering

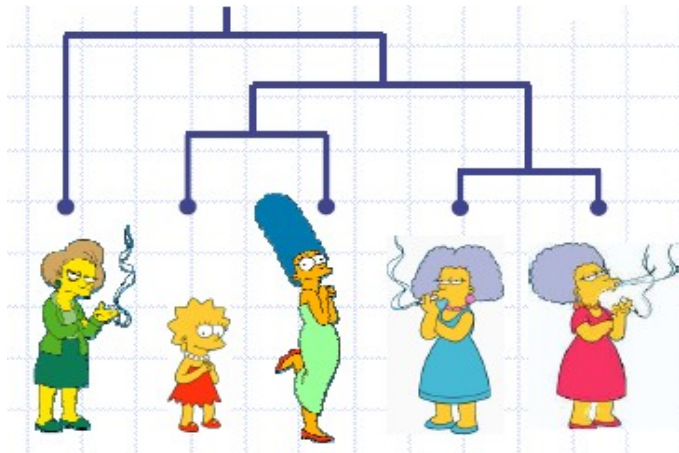
- Métodos de Clustering têm como principal objetivo organizar dados em:
 - 1) Partições: o método constrói partições dos dados, de modo a otimizar algum critério.
 - Ex: Algoritmo k-médias
 - 2) Hierarquias: cria decomposições hierárquicas dos objetos utilizando um determinado critério.
 - Ex: Técnica de clustering hierárquico

Tipos de Clustering

- Algoritmos de particionamento: constróem partições.

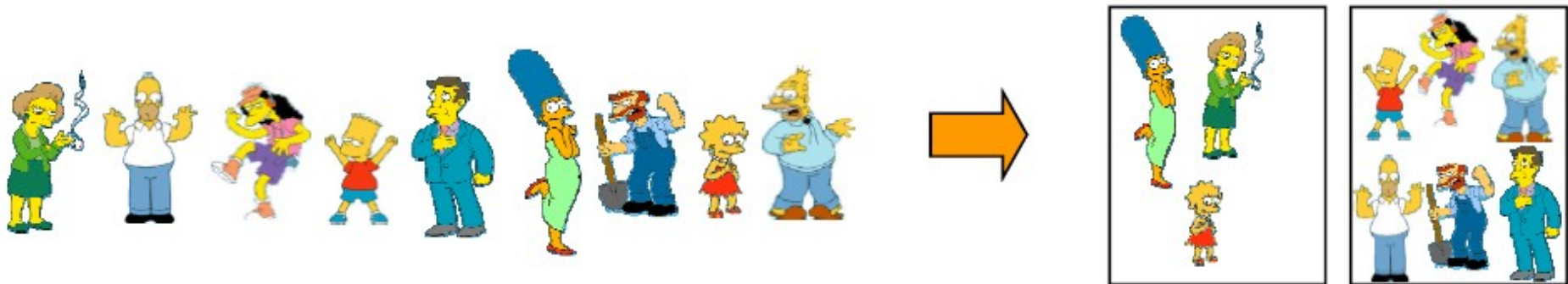


- Algoritmos hierárquicos: criam uma decomposição hierárquica.



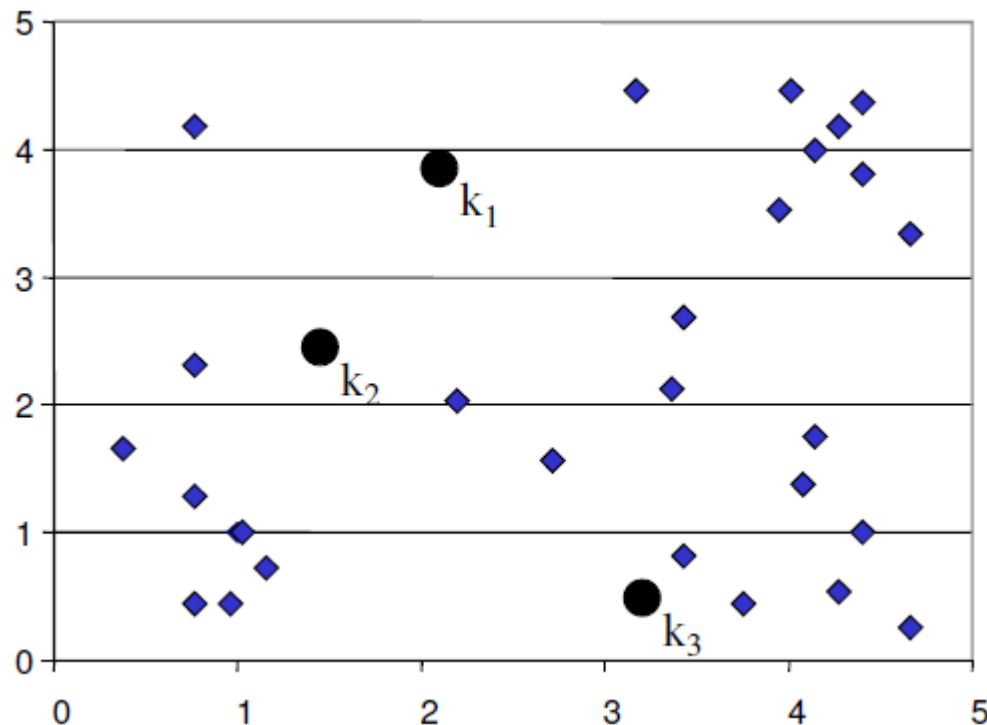
Algoritmos de Particionamento

- Na abordagem clássica, cada objeto é alocado em exatamente um dos K clusters disjuntos.
- O usuário deve estipular o número desejado de clusters (K).



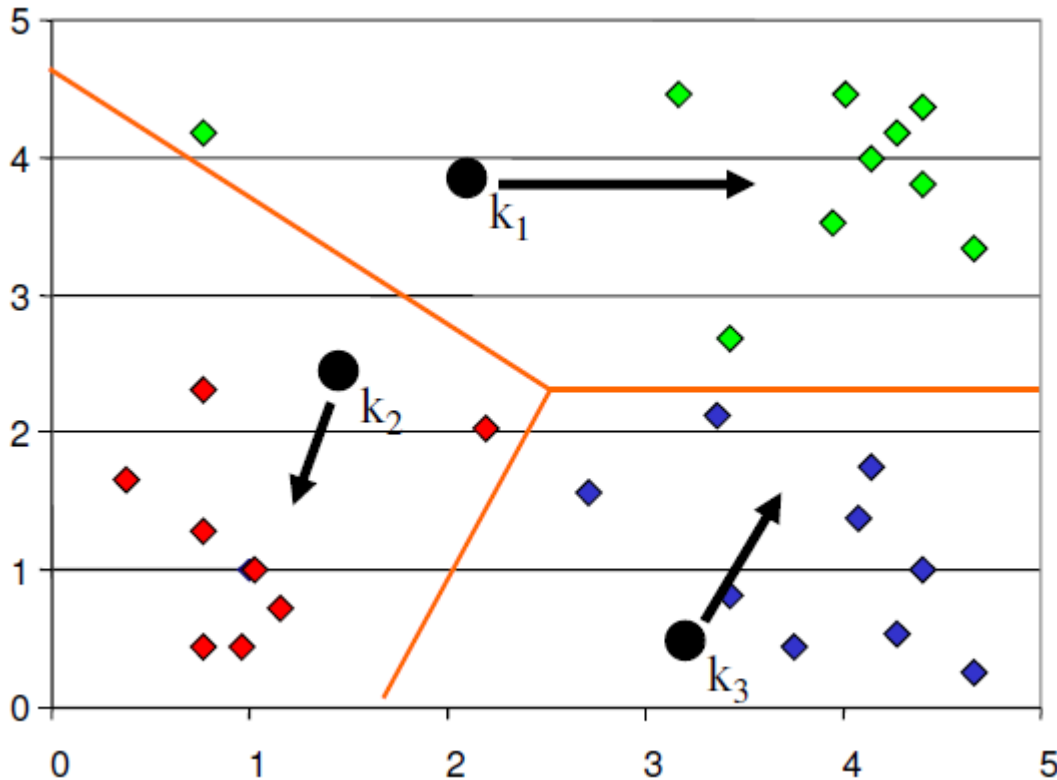
Algoritmo K-médias: inicialização

- Estabeleça o número de clusters (K) e escolha os protótipos (centróides) de cada cluster aleatoriamente.



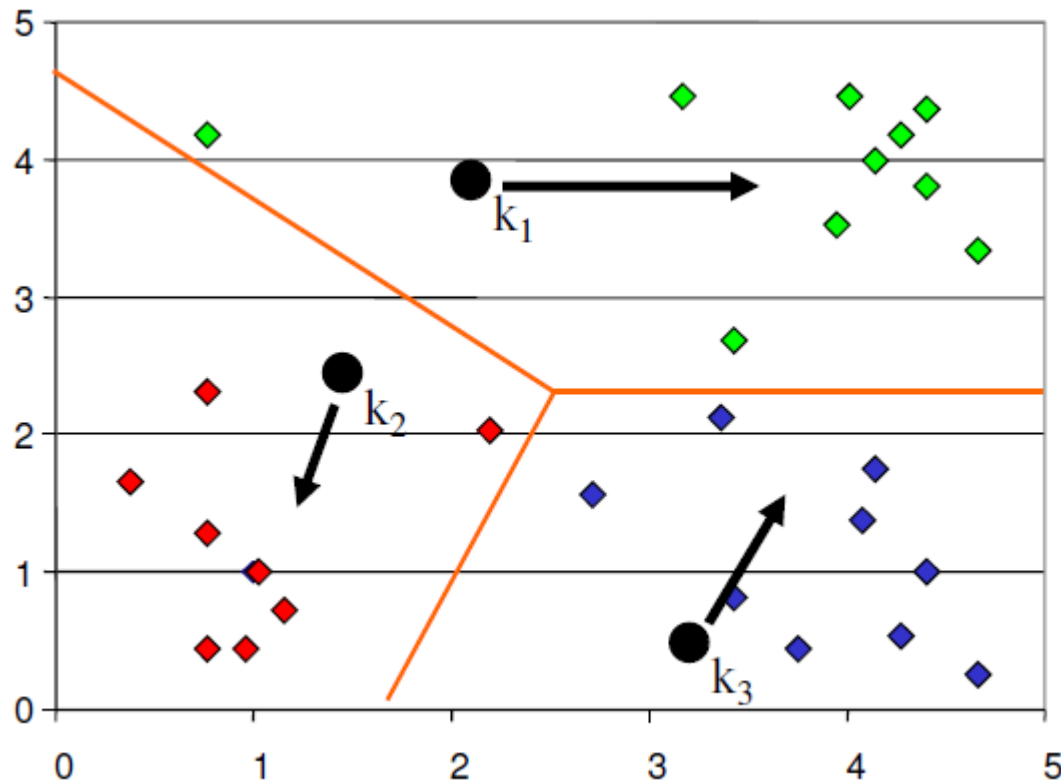
Algoritmo K-médias: iteração 1

- Calcule as distâncias entre objetos e os protótipos (k_1, k_2, k_3), encontrando clusters iniciais pela regra do vizinho mais próximo:



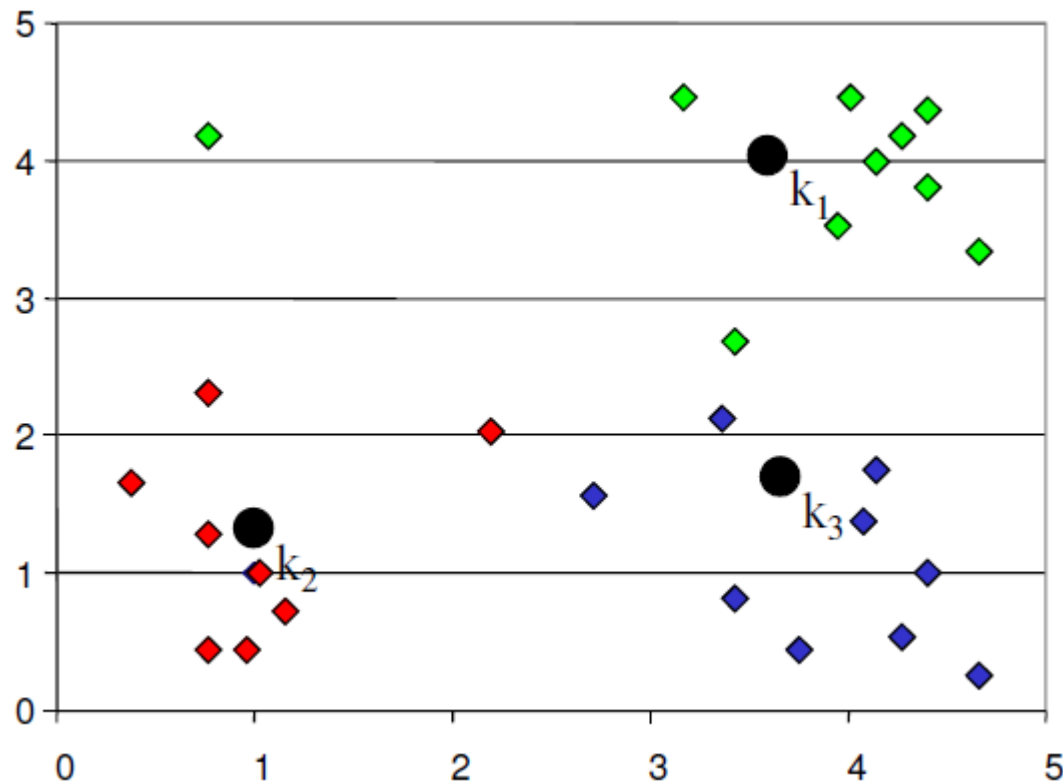
Algoritmo K-médias: iteração 1

- Depois, atualize os protótipos (k_1, k_2, k_3), calculando o vetor de médias para os objetos alocados em cada cluster.



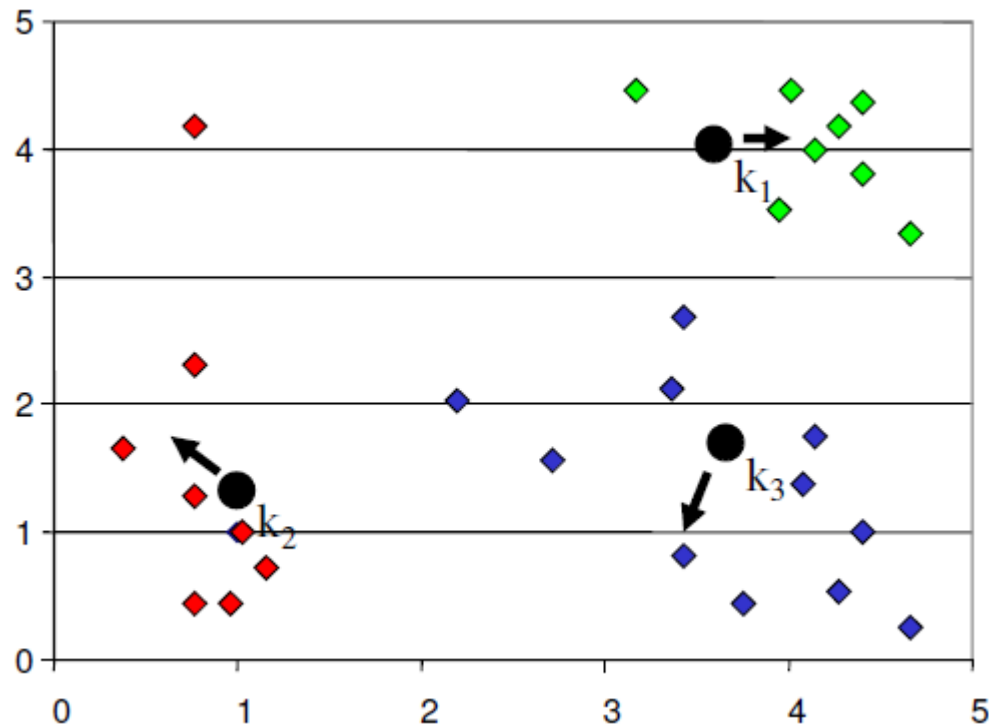
Algoritmo K-médias: iteração 2

- Realoque os objetos de acordo com a menor distância aos protótipos atualizados.



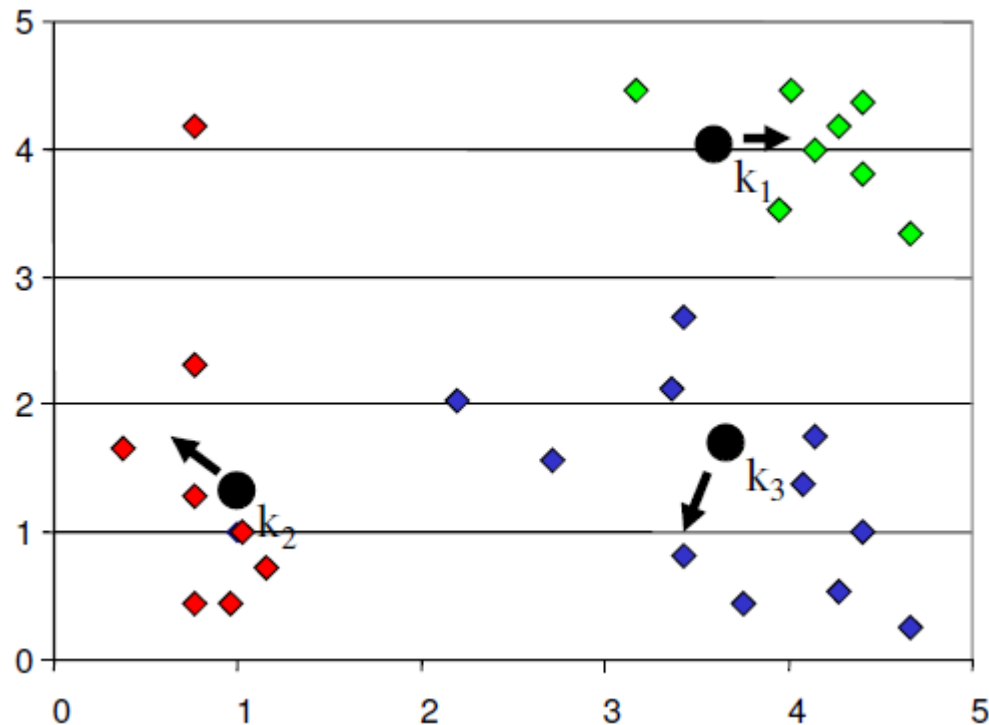
Algoritmo K-médias: iteração 2

- Depois, atualize os protótipos, calculando o vetor de médias para os objetos alocados em cada novo cluster.



Algoritmo K-médias: término

- Prossiga, realocando os objetos e atualizando os protótipos até que não haja mais mudança na configuração dos clusters.





Algoritmo K-médias

- 1. Estabeleça um valor para K , o número de clusters.
- 2. Escolha, aleatoriamente, K pontos como protótipos iniciais (centróides) dos clusters.
- 3. Aloque cada um dos N objetos, associando-os ao cluster mais próximo.
- 4. Atualize os protótipos dos clusters, calculando o vetor de médias para os objetos em cada cluster.
- 5. Repita os passos 3 e 4 até que nenhum dos objetos mude de cluster.

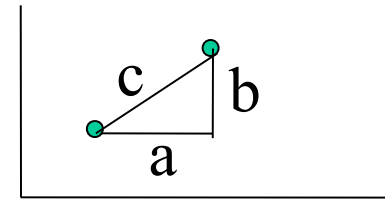
Medida de dis(similaridade)

- A mais comum é a medida de distância euclidiana entre dois pontos.
- Baseada no Teorema de Pitágoras.
- Para dois pontos X e Y:

$$X=(x_1, x_2, \dots, x_n) \text{ e } Y=(y_1, y_2, \dots, y_n)$$

- A distância euclidiana entre X e Y é dada por:

$$d(X, Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2}$$



$$a^2 + b^2 = c^2$$



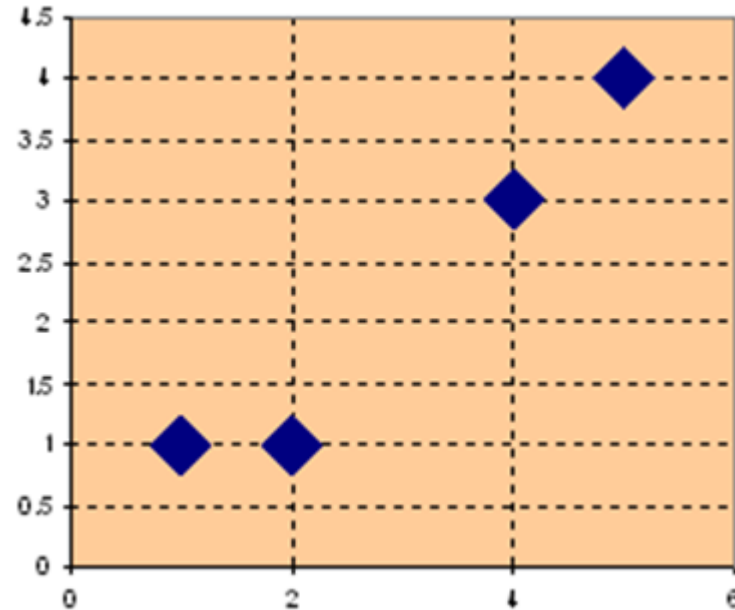
Exemplo

- Utilize o K-médias para agrupar 4 objetos, que representam 4 medicamentos diferentes, em 2 clusters. Cada um desses objetos está caracterizado por dois atributos: PH e Nível de Concentração (NC).
- Suponha que os medicamentos A e B foram escolhidos pelo algoritmo como protótipos iniciais dos clusters.

Objeto	PH	NC
Medicamento A	1	1
Medicamento B	2	1
Medicamento C	4	3
Medicamento D	5	4

Exemplo

- Cada medicamento representa um ponto com dois atributos (X, Y) que podem ser representados como coordenadas no espaço bidimensional.

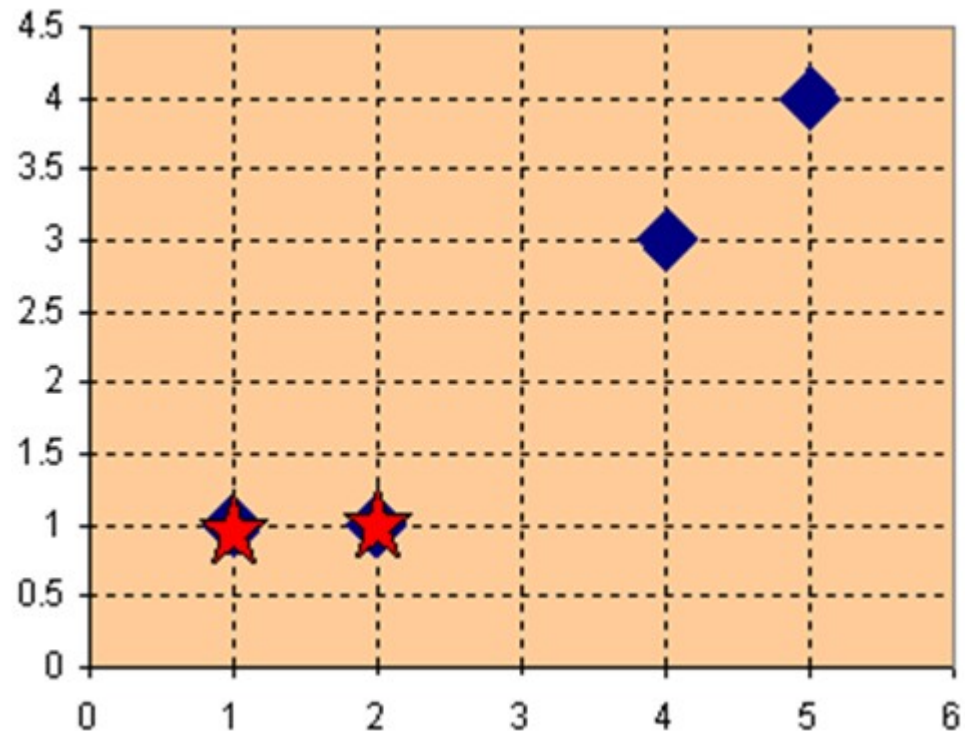


Exemplo

- Estamos supondo os medicamentos A e B como os primeiros centróides.
- Neste caso:

$$K_1 = (1, 1)$$

$$K_2 = (2, 1)$$





Exemplo

- Calcula-se as distâncias entre os objetos e cada centróide de cluster.
- Usando a distância euclidiana, encontra-se a seguinte matriz de distâncias:

$$Dist = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} \quad \begin{matrix} K_1 = (1, 1) \\ K_2 = (2, 1) \end{matrix}$$

A B C D

- Por exemplo, para o medicamento C:

$$Dist(C, k_1) = \sqrt{(4-1)^2 + (3-1)^2} = 3.61$$

$$Dist(C, k_2) = \sqrt{(4-2)^2 + (3-1)^2} = 2.83$$



Exemplo

- Associa-se cada objeto a um cluster, baseando-se no critério da distância mínima.

$$\textit{Clustering} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \quad \begin{array}{l} \text{Cluster 1} \\ \text{Cluster 2} \end{array}$$

A B C D

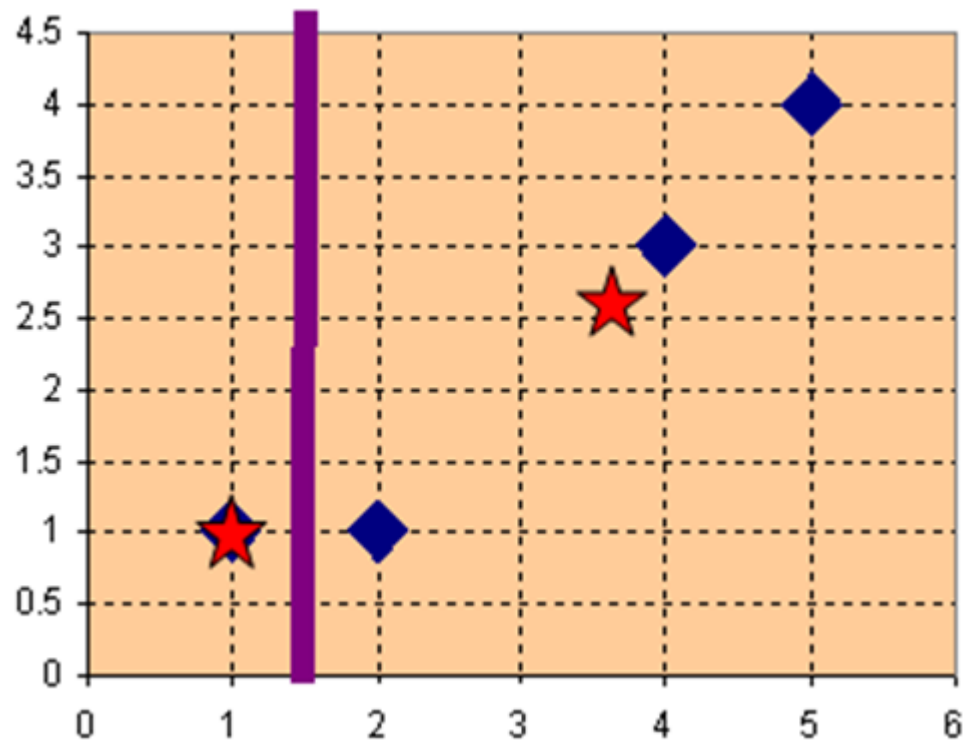
- Atualiza-se os centróides dos clusters:

$$k_1 = (1, 1)$$

$$k_2 = \left(\frac{2+4+5}{3}, \frac{1+3+4}{3} \right)$$

Exemplo

- Após a primeira iteração, os clusters apresentam a configuração mostrada ao lado.





Exemplo

- O próximo passo é recalcular as distâncias entre os objetos e cada novo centróide de cluster.
- Encontra-se, então, a seguinte matriz de distâncias:

$$Dist = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{bmatrix} \quad \begin{matrix} k_1 = (1, 1) \\ k_2 = \left(\frac{11}{3}, \frac{8}{3}\right) \end{matrix}$$

A B C D

- Por exemplo, para o medicamento C:

$$Dist(C, k_1) = \sqrt{(4-1)^2 + (3-1)^2} = 3.61$$

$$Dist(C, k_2) = \sqrt{\left(4 - \frac{11}{3}\right)^2 + \left(3 - \frac{8}{3}\right)^2} = 0.47$$



Exemplo

- Novamente, associa-se cada objeto a um cluster, baseando-se no critério da distância mínima.

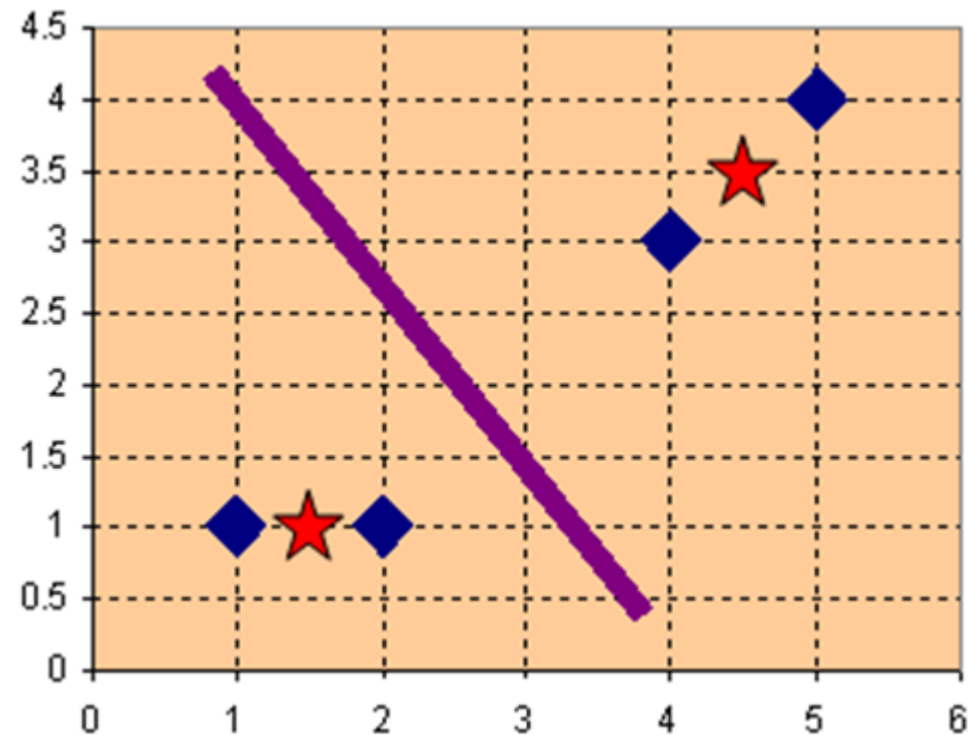
$$\textit{Clustering} = \begin{matrix} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} & \begin{matrix} \text{Cluster 1} \\ \text{Cluster 2} \end{matrix} \\ \begin{matrix} A & B & C & D \end{matrix} & \end{matrix}$$

- Atualiza-se os centróides dos clusters:

$$k_1 = \left(\frac{1+2}{2}, \frac{1+1}{2} \right)$$
$$k_2 = \left(\frac{4+5}{2}, \frac{3+4}{2} \right)$$

Exemplo

- Após a segunda iteração, os clusters apresentam a configuração mostrada ao lado.





Exemplo

- Calcula-se, novamente as distâncias entre os objetos e cada novo centróide de cluster.
- Encontra-se, então, a seguinte matriz de distâncias:

$$Dist = \begin{bmatrix} 0.5 & 0.5 & 3.20 & 4.61 \\ 4.30 & 3.54 & 0.71 & 0.71 \end{bmatrix} \quad \begin{matrix} k_1 = (1.5, 1) \\ k_2 = (4.5, 3.5) \end{matrix}$$

A B C D

- Por exemplo, para o medicamento C:

$$Dist(C, k_1) = \sqrt{(4 - 1.5)^2 + (3 - 1)^2} = 3.20$$

$$Dist(C, k_2) = \sqrt{(4 - 4.5)^2 + (3 - 3.5)^2} = 0.71$$



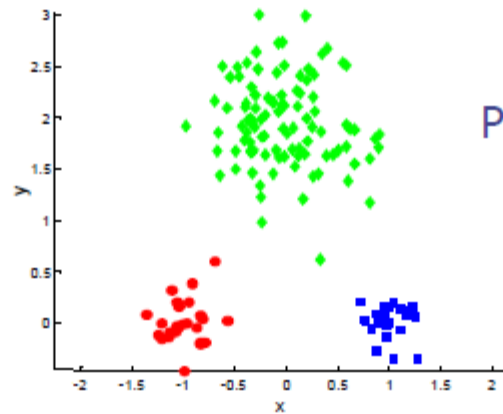
Exemplo

- Novamente, associa-se cada objeto a um cluster, baseando-se no critério da distância mínima.

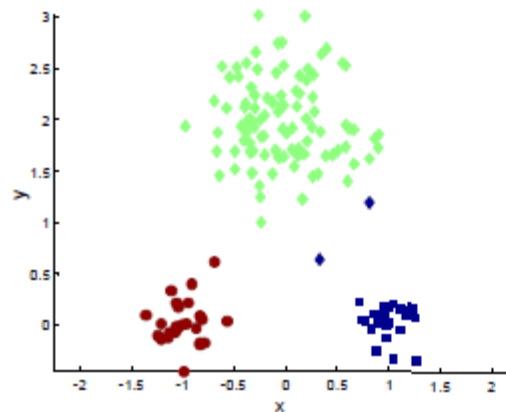
$$\textit{Clustering} = \begin{array}{cccc} \left[\begin{array}{cccc} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{array} \right] & \begin{array}{l} \text{Cluster 1} \\ \text{Cluster 2} \end{array} \\ \begin{array}{cccc} A & B & C & D \end{array} \end{array}$$

- Comparando-se esse agrupamento com aquele encontrado na última iteração, observa-se que não houve mudança nas configurações dos clusters.
 - Término do algoritmo!

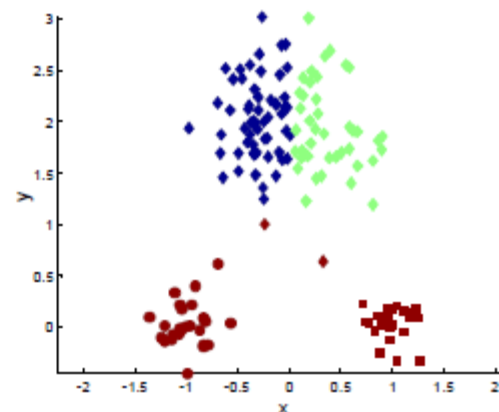
Importância da escolha dos centróides iniciais



Pontos originais

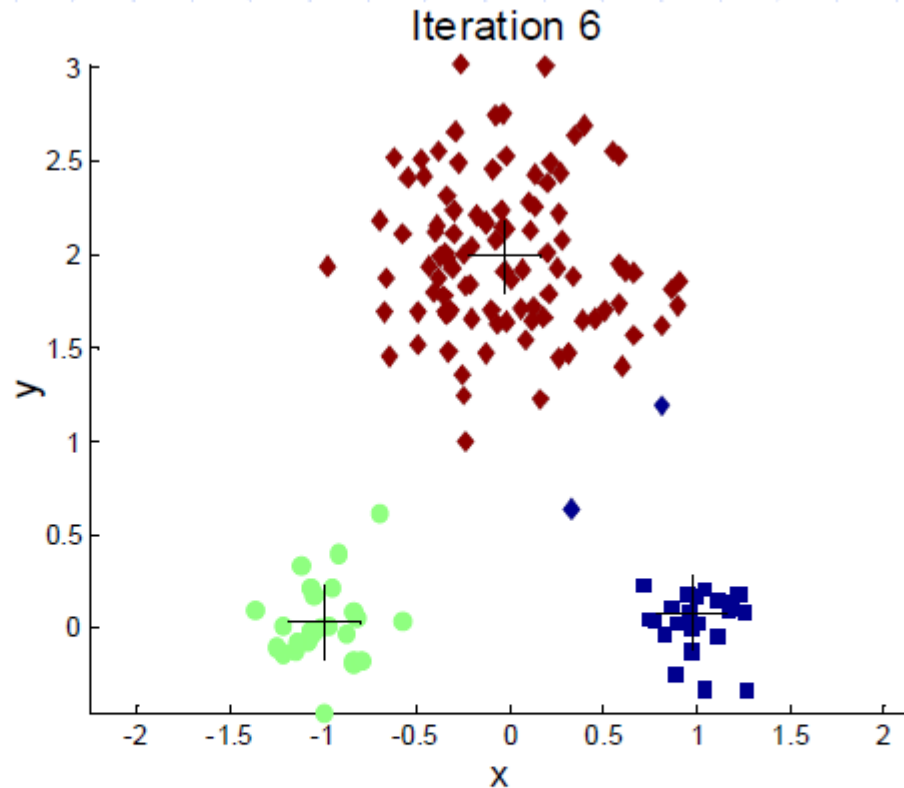


Partição ótima

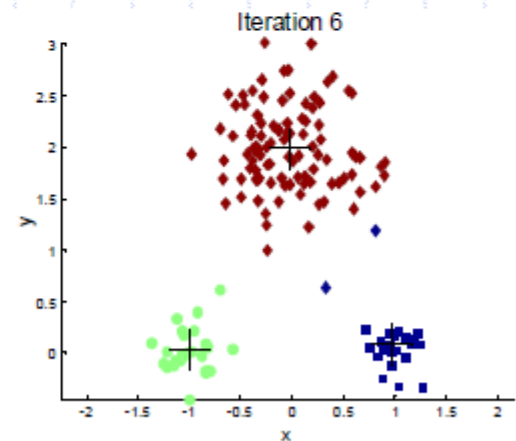
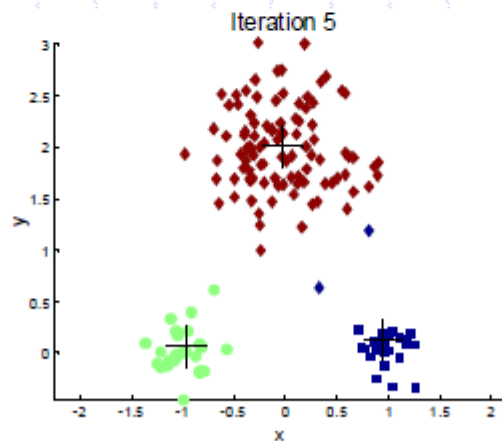
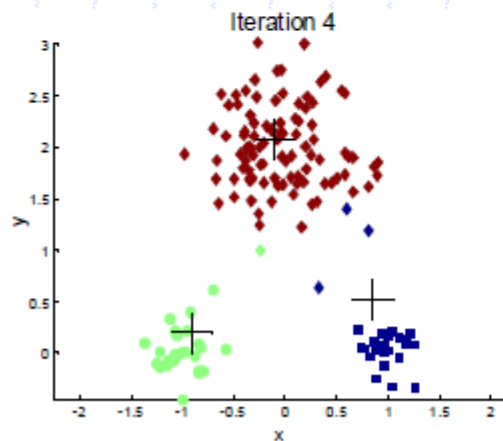
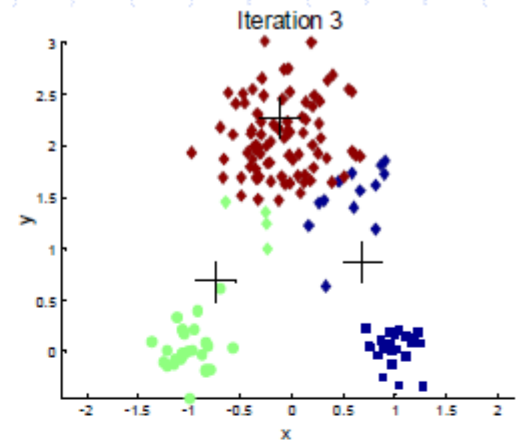
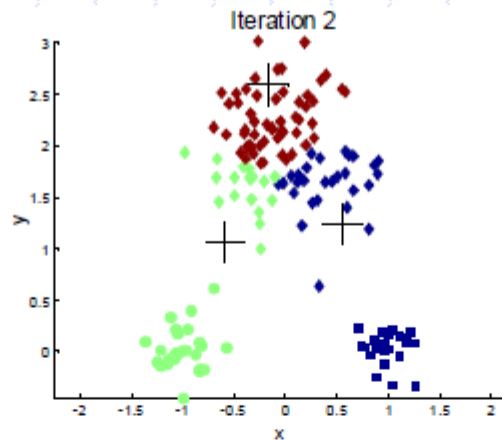
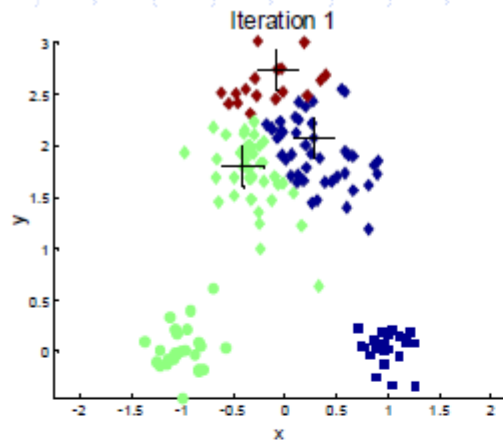


Partição Sub-ótima

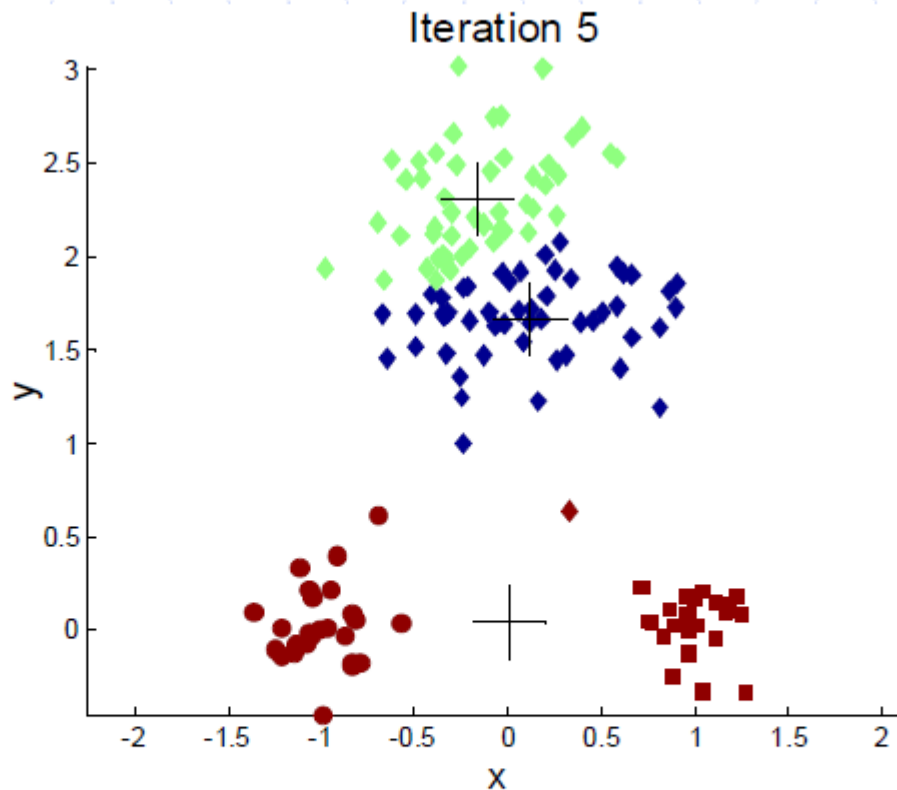
Importância da escolha dos centróides iniciais



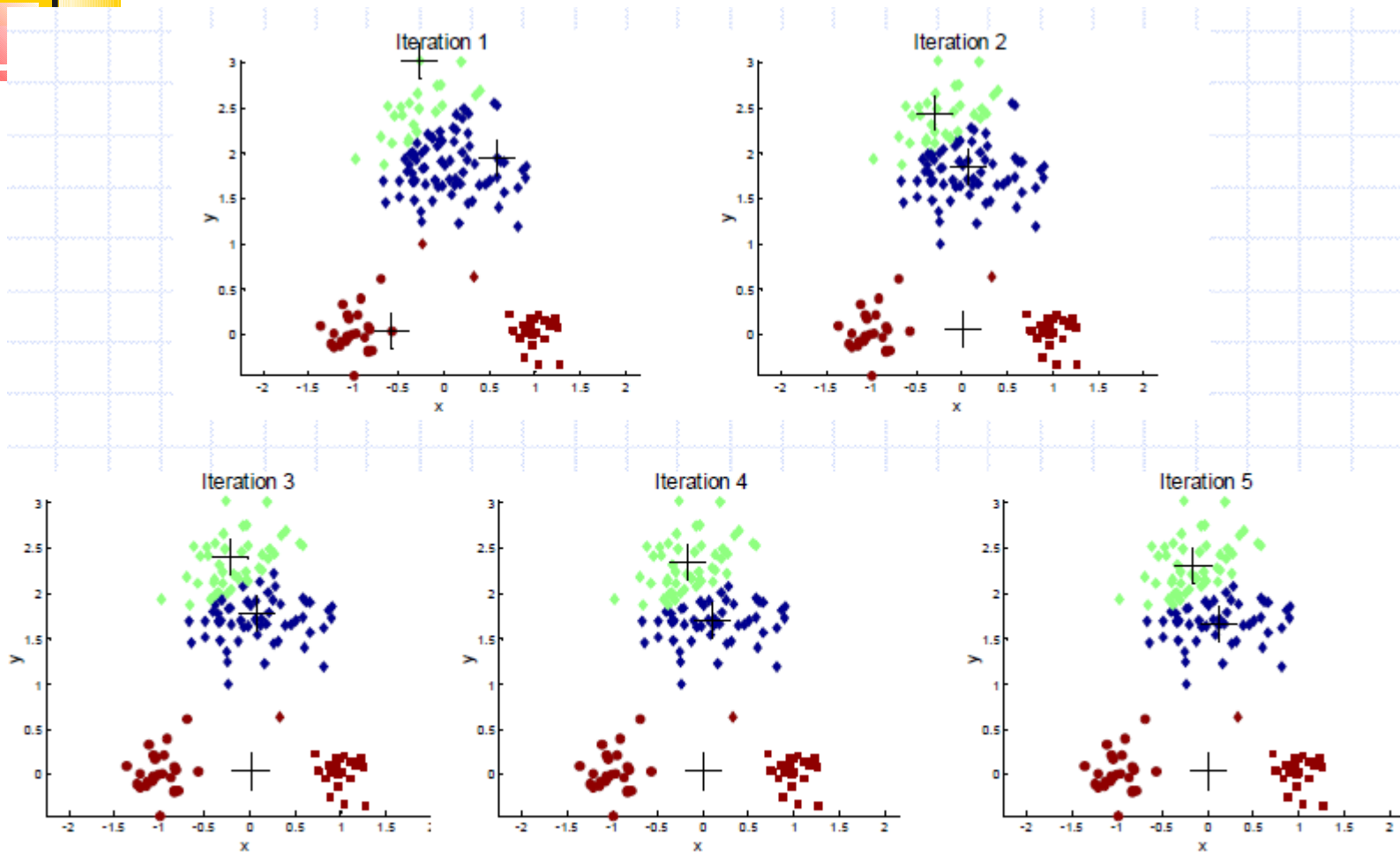
Importância da escolha dos centróides iniciais



Importância da escolha dos centróides iniciais



Importância da escolha dos centróides iniciais





Importância da escolha dos centróides iniciais

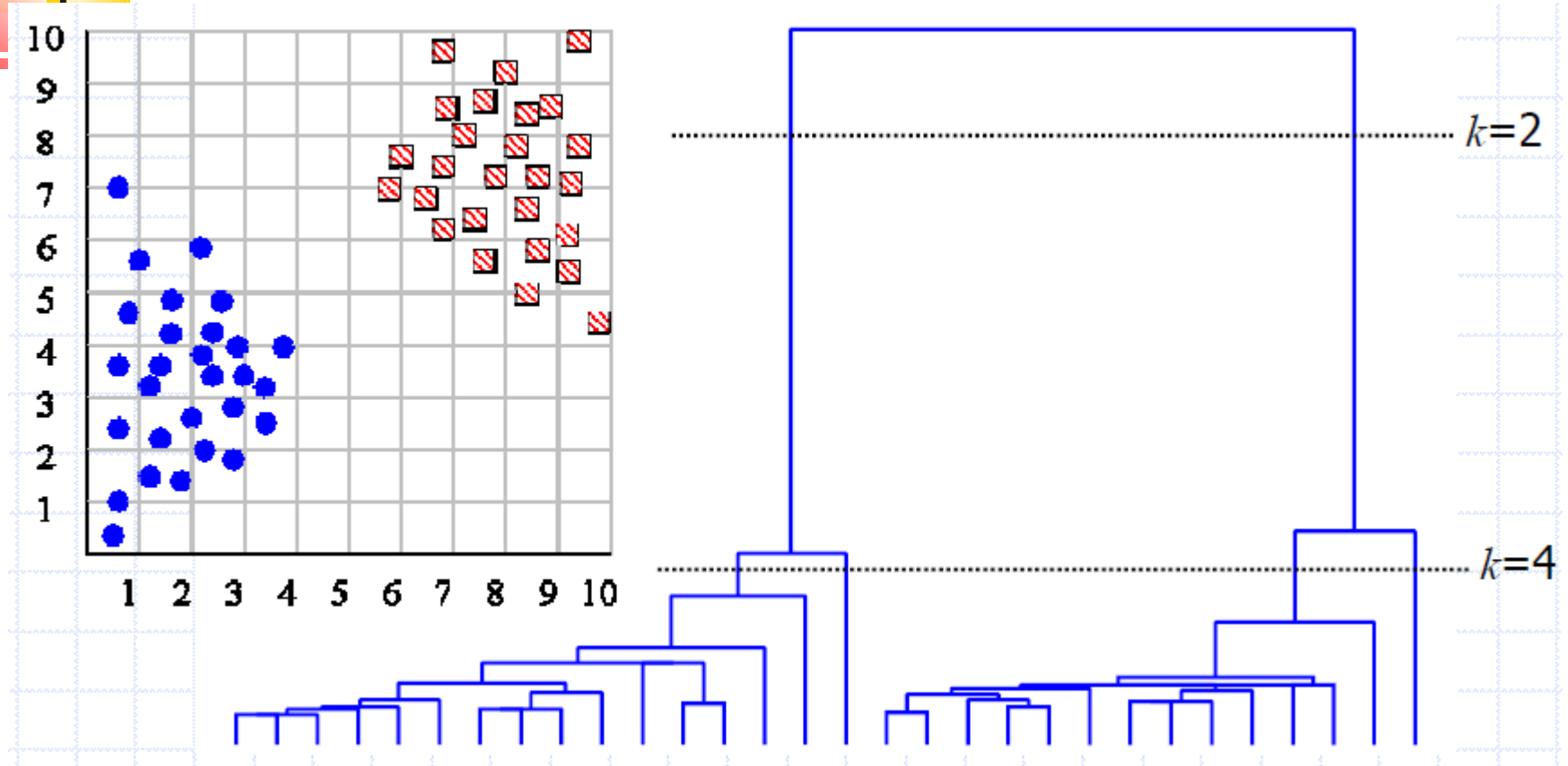
- Para tentar evitar que o k-médias encontre uma solução sub-ótima, é possível realizar várias execuções do algoritmo e escolher aquela que minimiza a Soma dos Erros Quadráticos (SEQ):

$$SEQ = \sum_{i=1}^k \sum_{\mathbf{x} \in C_i} dist^2(\mathbf{m}_i, \mathbf{x})$$

em que \mathbf{x} é um objeto e \mathbf{m}_i é o centróide do cluster \mathbf{C}_i .

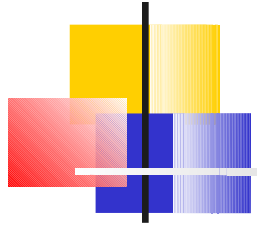
- Por exemplo, dadas duas partições com o mesmo número k de clusters, escolher aquela que apresenta a menor SEQ.
 - O aumento de k , por si só, já diminui a SEQ.

Métodos de Clustering Hierárquicos



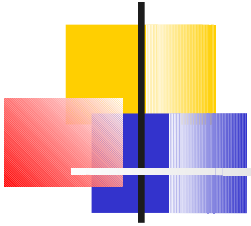
- Conjunto de “partições aninhadas”.
- Estimativa do número de clusters.

Principais Métodos de Clustering Hierárquicos

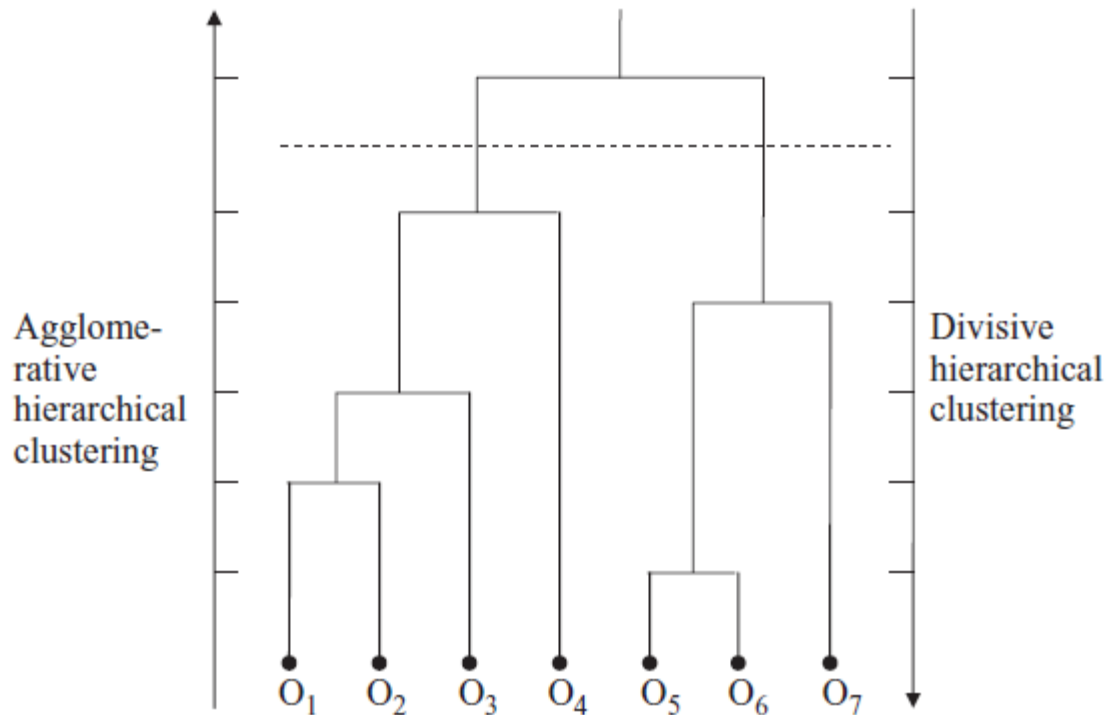


- Bottom-Up (aglomerativos):
 - iniciar colocando cada objeto em um cluster;
 - encontrar o melhor par de clusters para uni-los;
 - repetir até que todos os clusters sejam reunidos em um só cluster.
- Top-Down (divisivos):
 - iniciar com todos os objetos num único cluster;
 - considerar possíveis divisões do cluster em dois;
 - escolher a melhor divisão e recursivamente operar em ambos os lados até que cada objeto forme um cluster.

Métodos de Clustering Hierárquicos

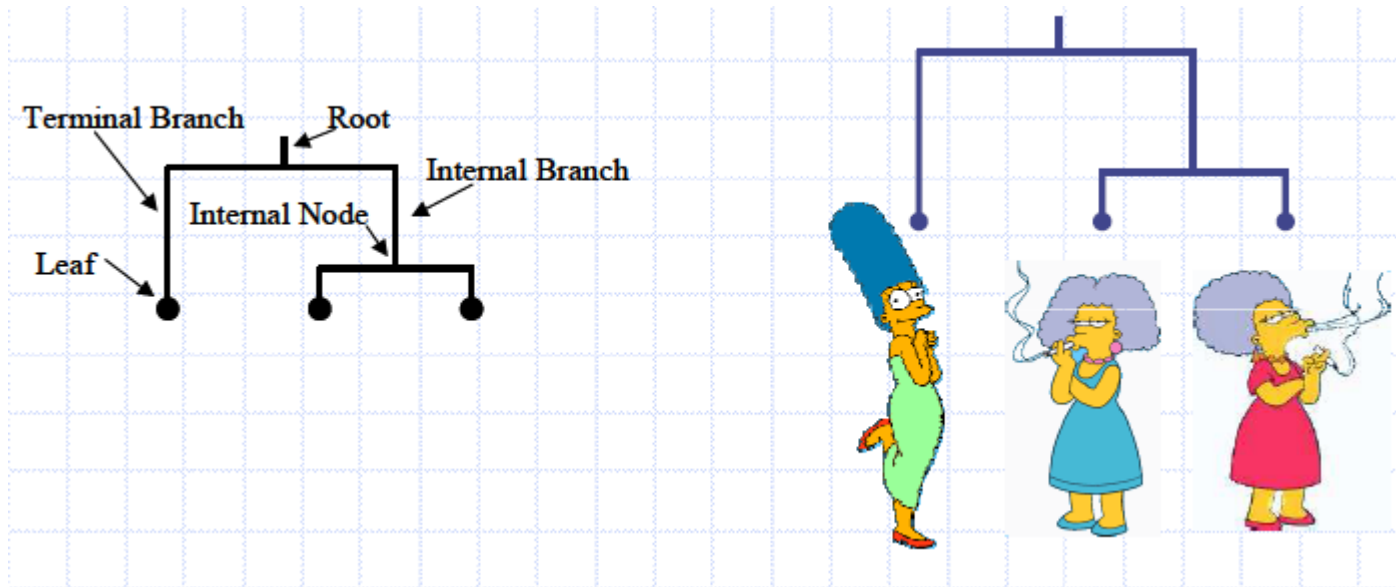


- Métodos aglomerativos e divisivos seguem direções opostas na estrutura de árvore (denominada de dendrograma).

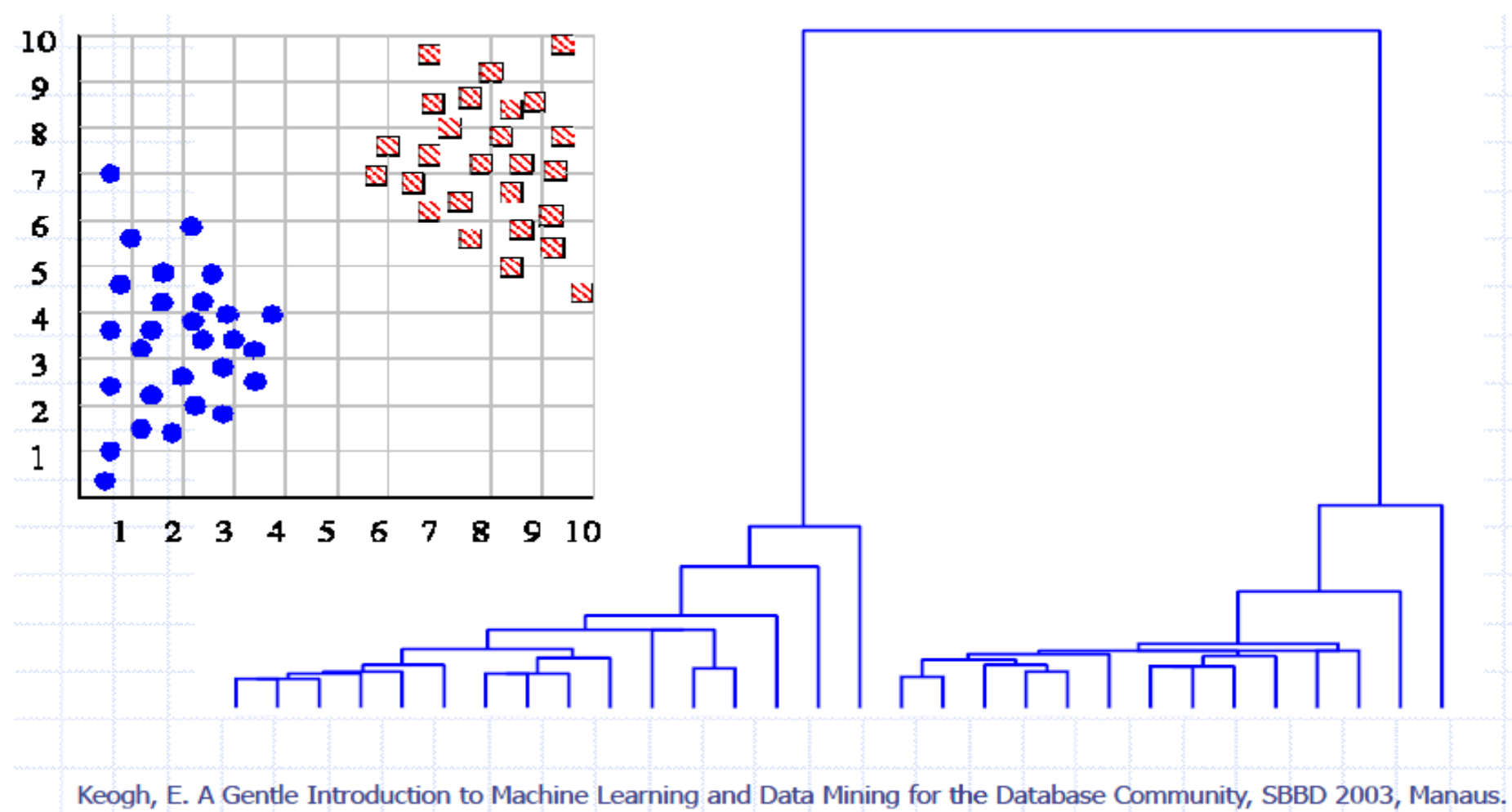


Dendrograma

- Dendrograma: é uma ferramenta útil para sumarizar medidas de (dis)similaridade.



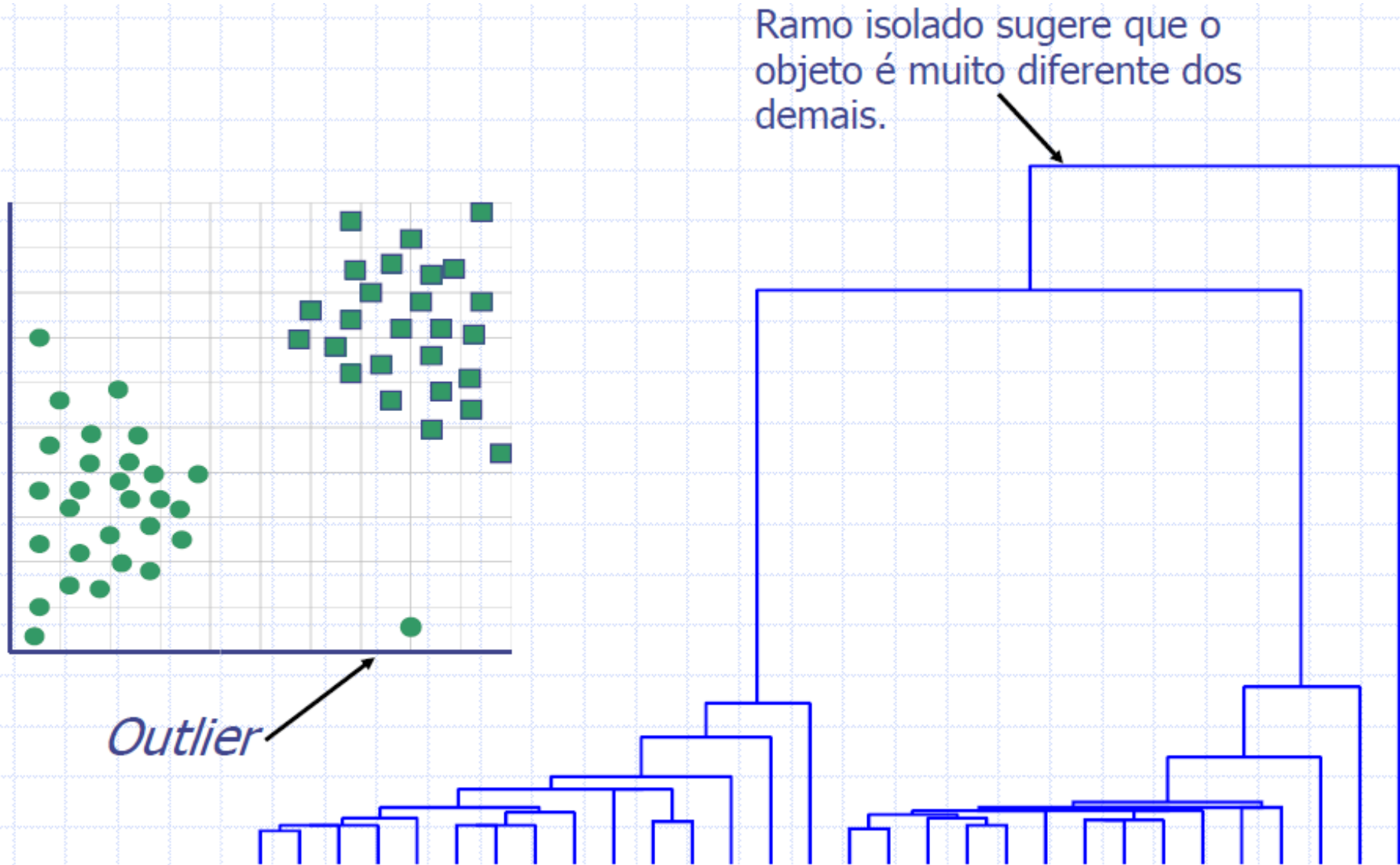
- A dissimilaridade entre dois clusters é representada como a altura do nó interno mais baixo compartilhado.



Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBB 2003, Manaus.

- Pode-se examinar o dendrograma para estimar o número correto de clusters.
- No caso acima, existem duas sub-árvores bem separadas, sugerindo dois clusters. Infelizmente na prática as distinções não são tão simples...

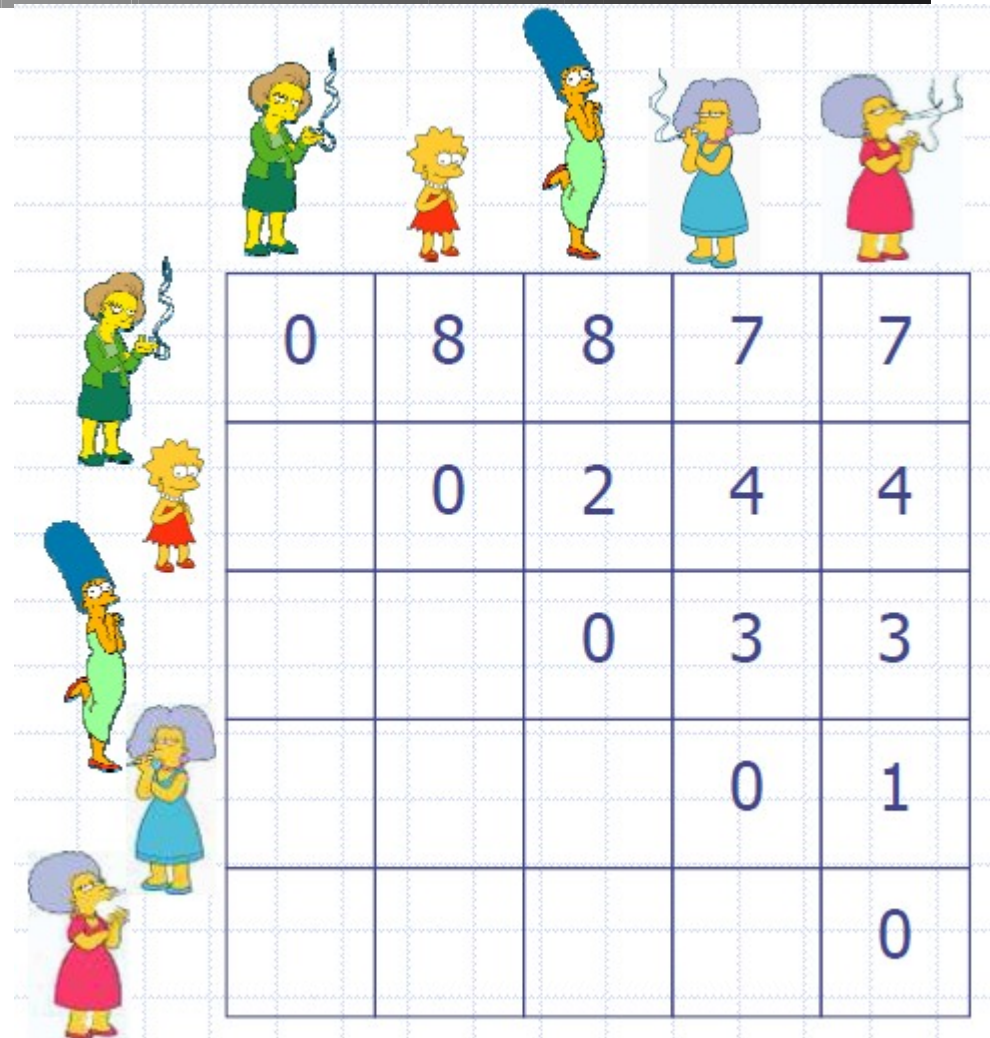
- O dendrograma pode ser útil para detecção de outliers.



Métodos de Clustering Hierárquicos

- Inicialmente é calculada uma matriz de distâncias.
- Esta contém as distâncias entre cada par de objetos.

$$D(\text{Mrs. Muntz}, \text{Lisa Simpson}) = 8$$
$$D(\text{Marge Simpson}, \text{Bart Simpson}) = 1$$

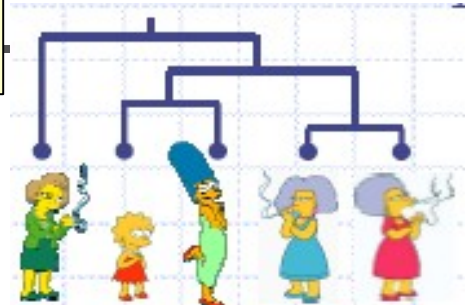
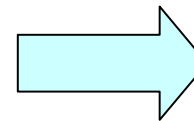


	Mrs. Muntz	Lisa Simpson	Marge Simpson	Bart Simpson	Marge Simpson
Mrs. Muntz	0	8	8	7	7
Lisa Simpson		0	2	4	4
Marge Simpson			0	3	3
Bart Simpson				0	1
Marge Simpson					0

Método Bottom-Up (Aglomerativo)

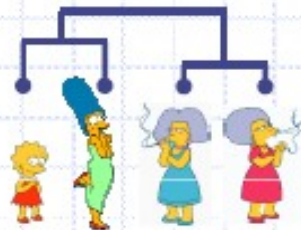
Iniciando com cada objeto em seu próprio cluster, encontrar o melhor par de clusters para unir. Repetir até que todos os clusters sejam fundidos em um único cluster.

Nível 4:



Nível 3:

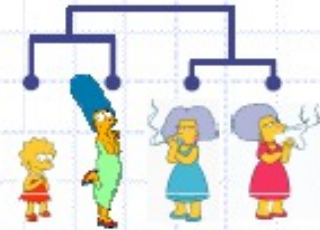
Considerar todas as uniões possíveis ...



...

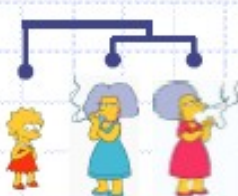


Escolher a melhor



Nível 2:

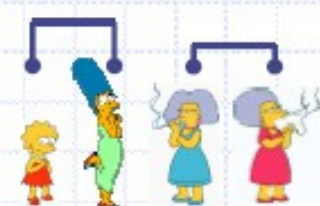
Considerar todas as uniões possíveis ...



...



Escolher a melhor



Nível 1:

Considerar todas as uniões possíveis ...



...



...



Escolher a melhor





Algoritmo de Clustering Aglomerativo

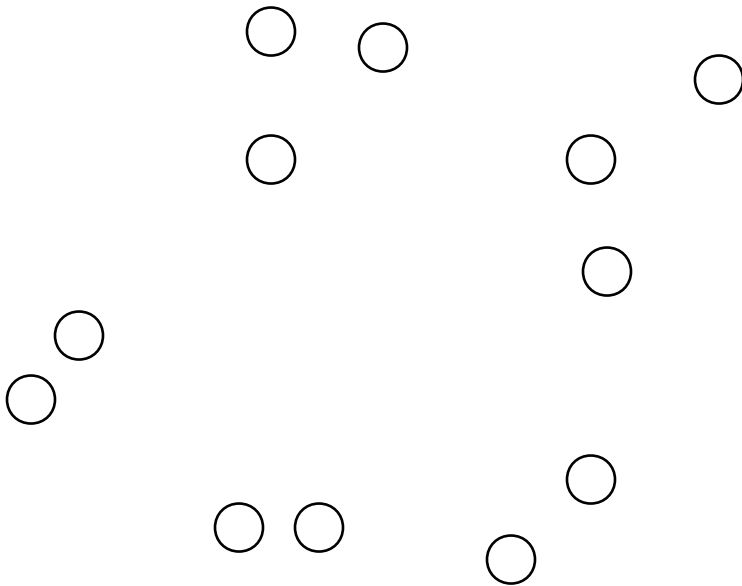
1. Compute a matriz de distâncias entre os objetos.
2. Considere cada objeto como sendo um cluster.
3. **Repita**
4. Una os dois clusters mais próximos.
5. Atualize a matriz de distâncias.
6. **Até que** haja somente um cluster.

- Essa é a técnica de clustering hierárquico mais popular.
- A operação chave é o cálculo da matriz de distâncias entre dois clusters.
 - Abordagens diferentes para definir distâncias entre clusters geram algoritmos diferentes.



Situação Inicial

- Considere cada objeto como um cluster individual e calcule a matriz de distâncias entre esses clusters.



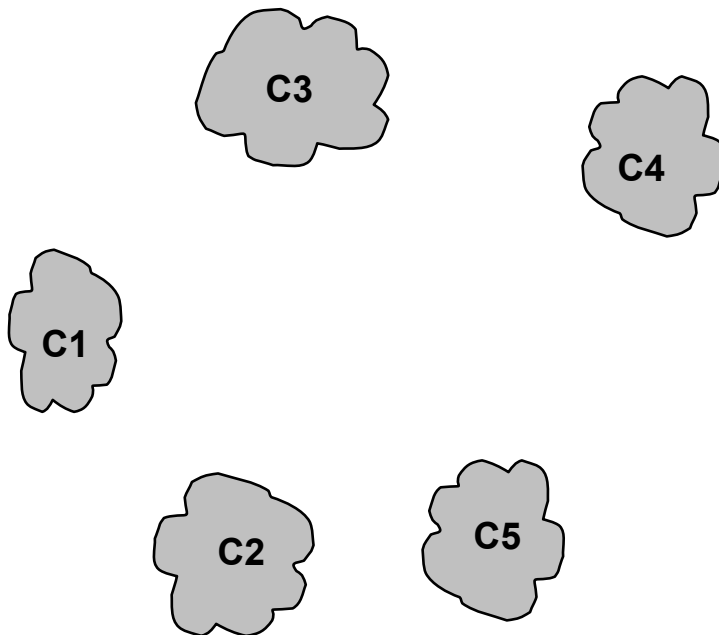
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						

Matriz de distâncias



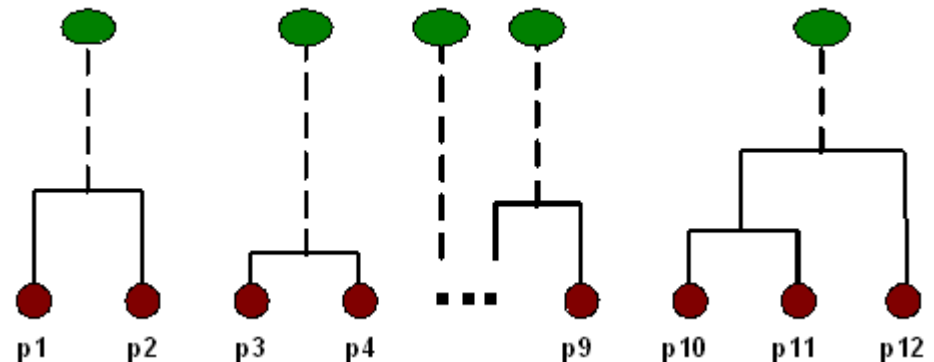
Situação Intermediária

- Após algumas fusões, encontramos os seguintes clusters...



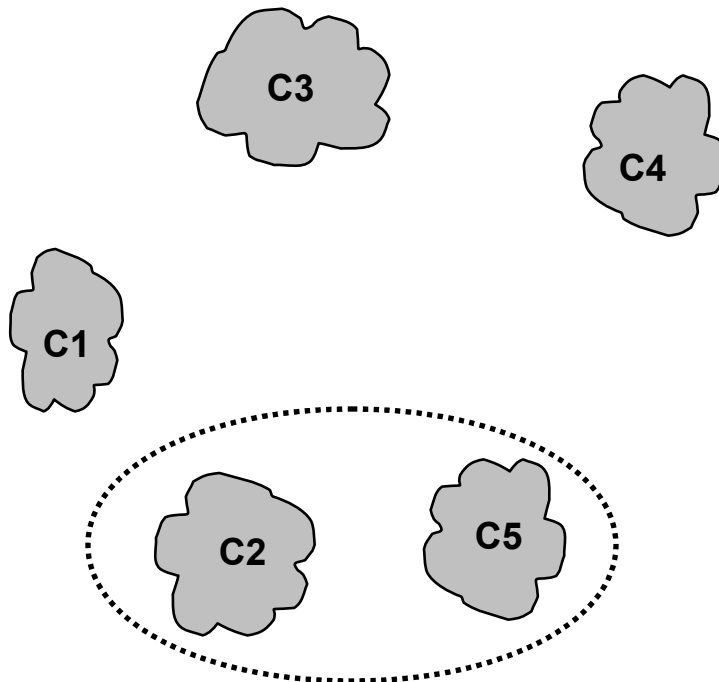
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Matriz de distâncias



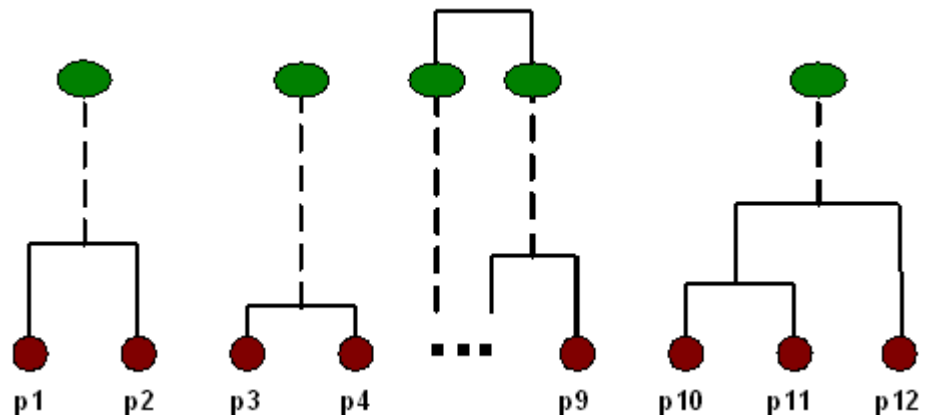
Situação Intermediária

- Une-se os clusters mais próximos (C2 e C5) e atualiza-se a matriz de distâncias.



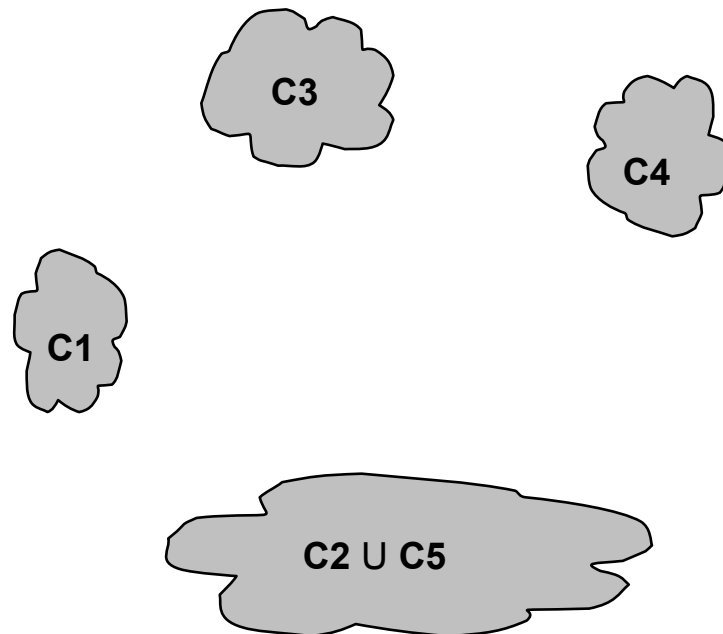
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Matriz de distâncias



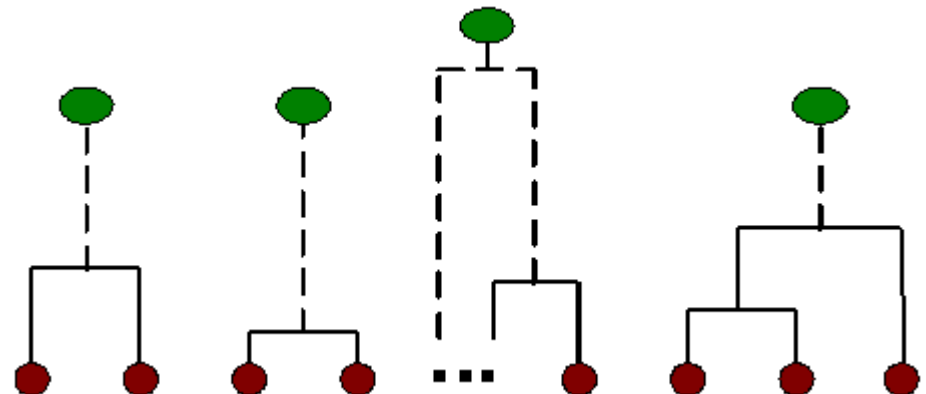
Depois de uma operação de fusão...

- A questão é: como atualizar a matriz de distâncias?

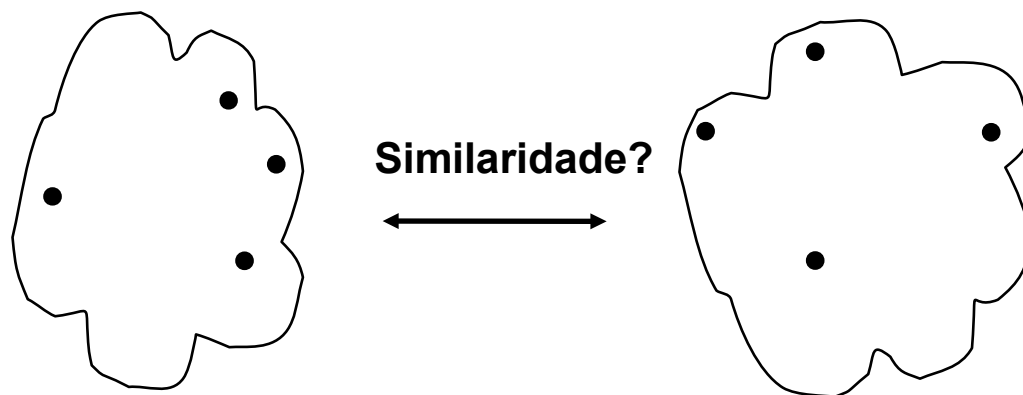
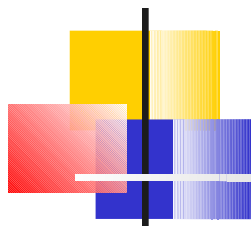


	C1	$\begin{matrix} C2 \\ \cup \\ C5 \end{matrix}$	C3	C4
C1		?		
$C2 \cup C5$?	?	?	?
C3		?		
C4		?		

Matriz de distâncias

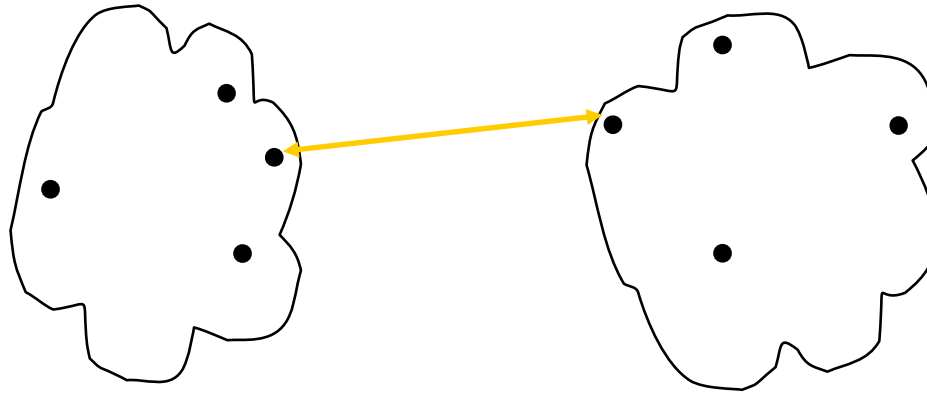


Medidas de similaridades entre clusters



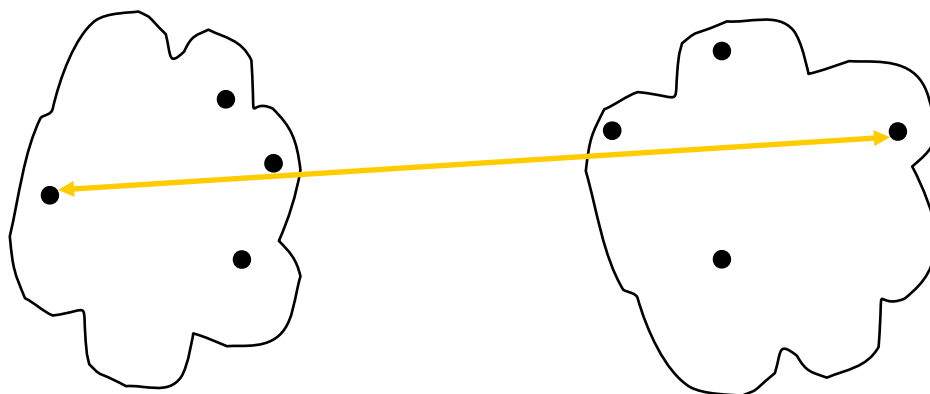
- Dada uma matriz de dis(similaridades):
 - MIN: distância entre os 2 pontos mais próximos em diferentes clusters;
 - MAX: distância entre os 2 pontos mais distantes em diferentes clusters;
 - Média das distâncias entre todos os pontos em diferentes clusters;
 - Distância entre os centróides dos clusters;

Vizinho mais Próximo (MIN)



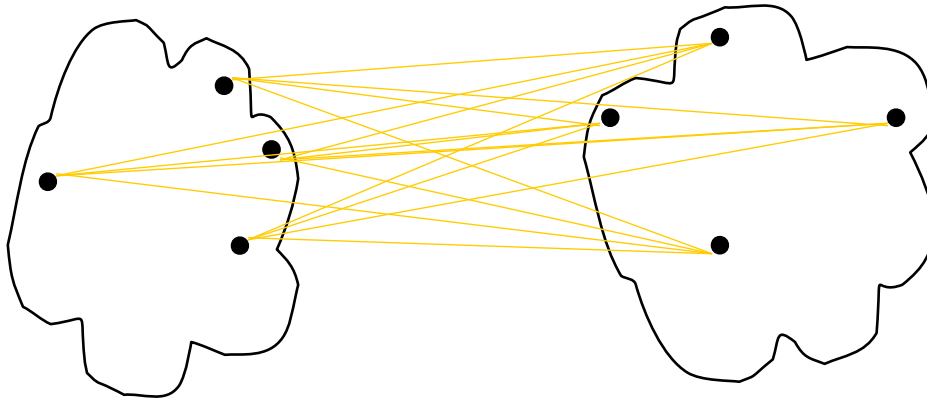
- Conhecido também como método de vinculação única (single link).
- Grafos: menor aresta entre dois nós de subconjuntos distintos.

Vizinho mais distante (MAX)



- Conhecido também como método de vinculação completa (complete link).
- Grafos: maior aresta entre dois nós de subconjuntos distintos.

Média de distâncias



- A distância entre dois clusters (C_i e C_j) é dada pela média das distâncias entre objetos de clusters distintos:

$$d(C_i, C_j) = \frac{\sum_{\substack{\mathbf{x}_i \in C_i \\ \mathbf{x}_j \in C_j}} d(\mathbf{x}_i, \mathbf{x}_j)}{|C_i| |C_j|}$$



Exemplo

- Dada a seguinte matriz de distâncias entre 6 objetos, utilize o algoritmo de clustering hierárquico para produzir dendrogramas, considerando os critérios de distância entre clusters apresentados anteriormente, ou seja:

- MIN
- MAX
- Média de distâncias

	I1	I2	I3	I4	I5	I6
I1	0,00	0,24	0,22	0,37	0,34	0,23
I2		0,00	0,15	0,20	0,14	0,25
I3			0,00	0,15	0,28	0,11
I4				0,00	0,29	0,22
I5					0,00	0,39
I6						0,00

Exemplo - MIN

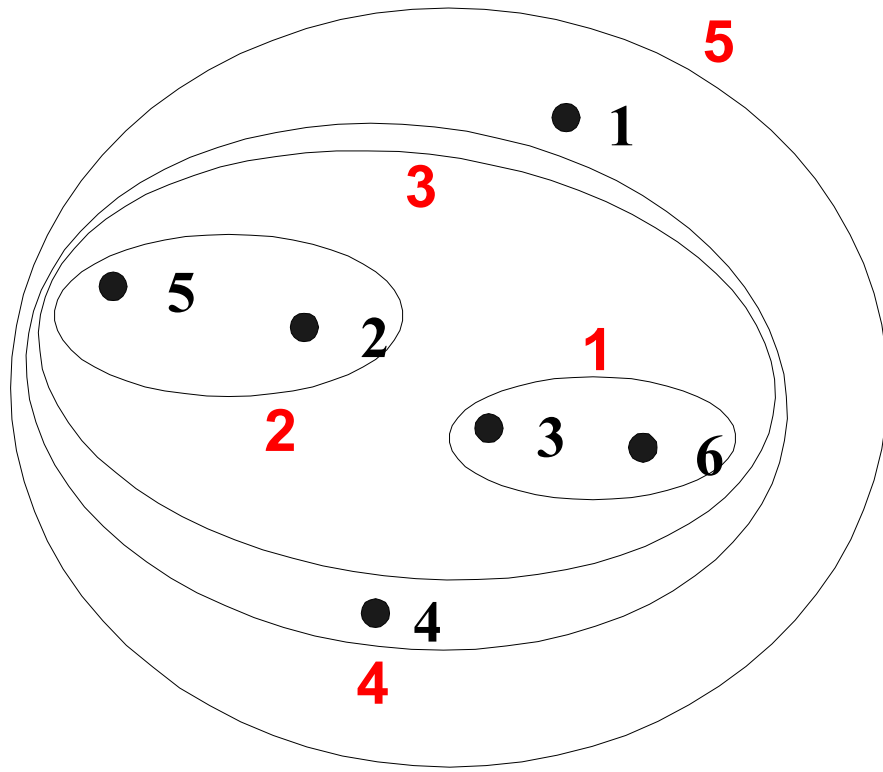
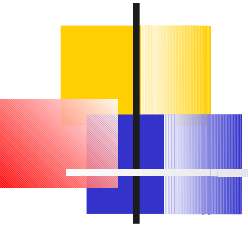
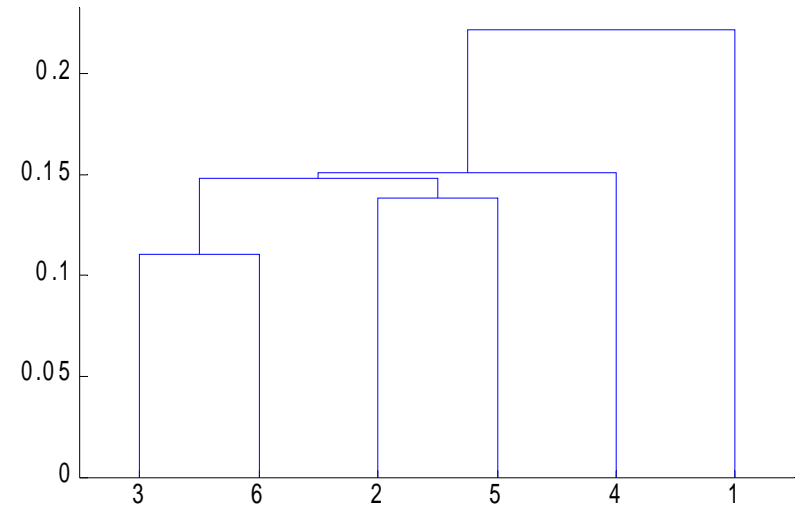


Diagrama de Venn



Dendrograma

Exemplo - MAX

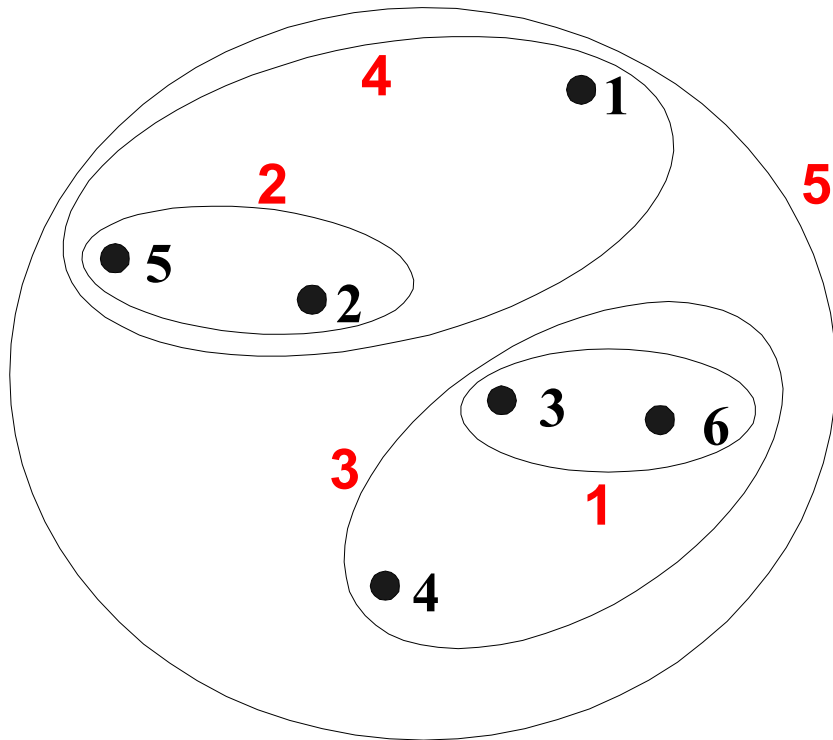
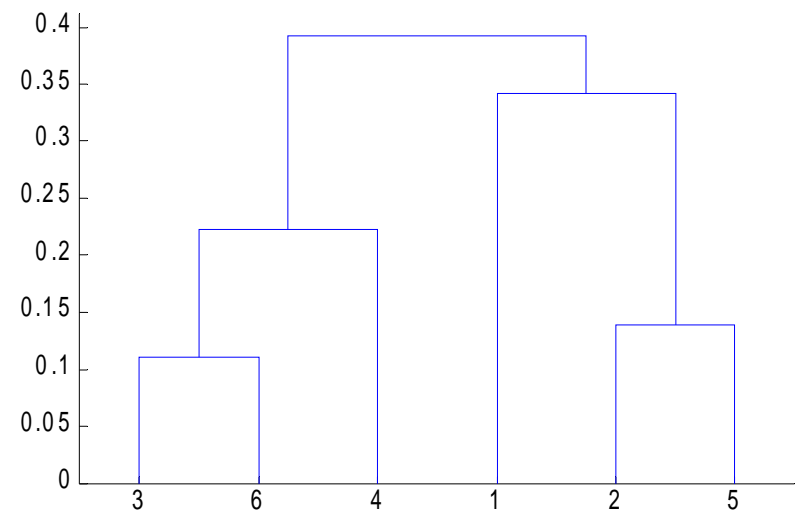


Diagrama de Venn



Dendrograma

Exemplo – Média de Distâncias

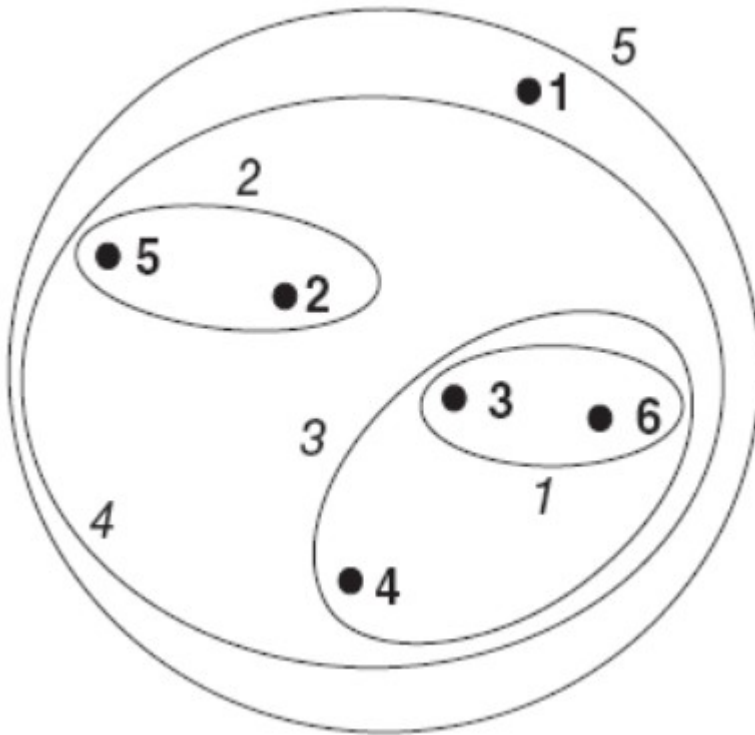
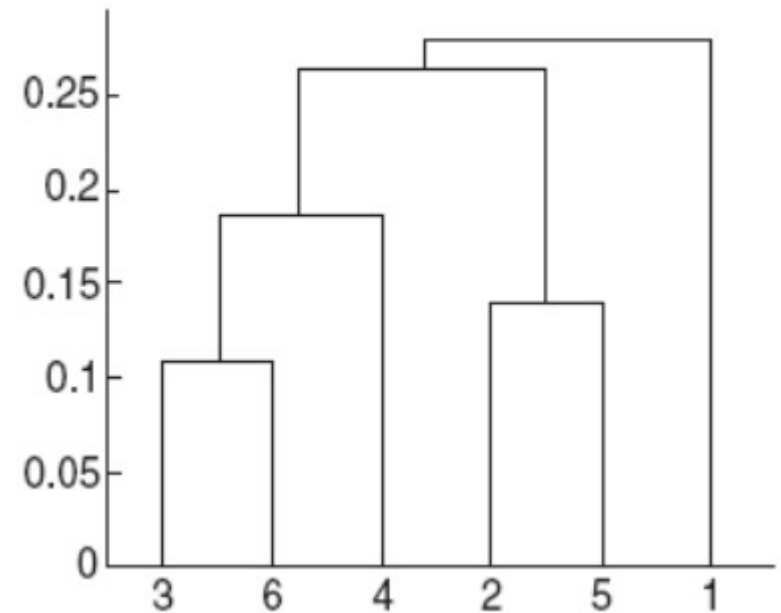


Diagrama de Venn



Dendrograma



Leituras

- Tan, P., Steinback, M., Kumar, V., Introduction to Data Mining, Addison Wesley, 2005.
 - Seção 8.2: K-means, pp. 496-506.
 - Seção 8.3: Agglomerative Hierarchical Clustering, pp. 515-522.
- Xu, R., Wunsch, D., Clustering, IEEE Press, 2009.
 - Capítulo 3: Hierarchical Clustering, pp. 31-37.
 - Capítulo 4: Partitional Clustering, pp. 63-73.
- Jain, A. K., Dubes, R. C., Algorithms for Clustering Data, Prentice Hall, 1988.
 - Capítulo 3: Clustering Methods and Algorithms, pp. 55-101.