Introdução a Ciências de Dados

Aula 4: Técnicas de Agrupamento de dados

Francisco A. Rodrigues
ICMC/USP
francisco@icmc.usp.br

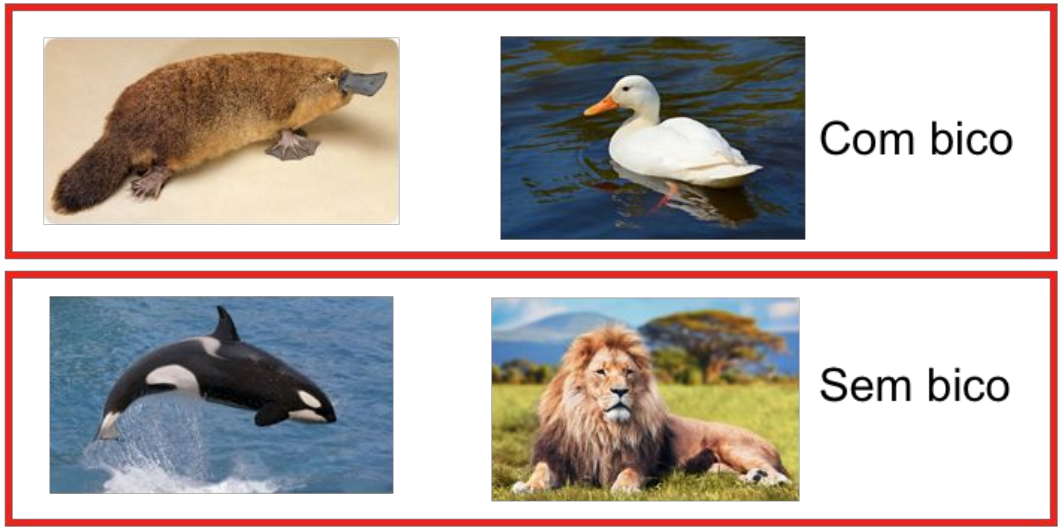
Aula 4 - Agrupamento de dados

- K-means,
- Agrupamento Hierárquico,
- Avaliando Agrupamentos

Agrupamento de dados

Dado um conjunto de objetos, agrupar os objetos em grupos (clusters) baseados na similaridade entre eles.

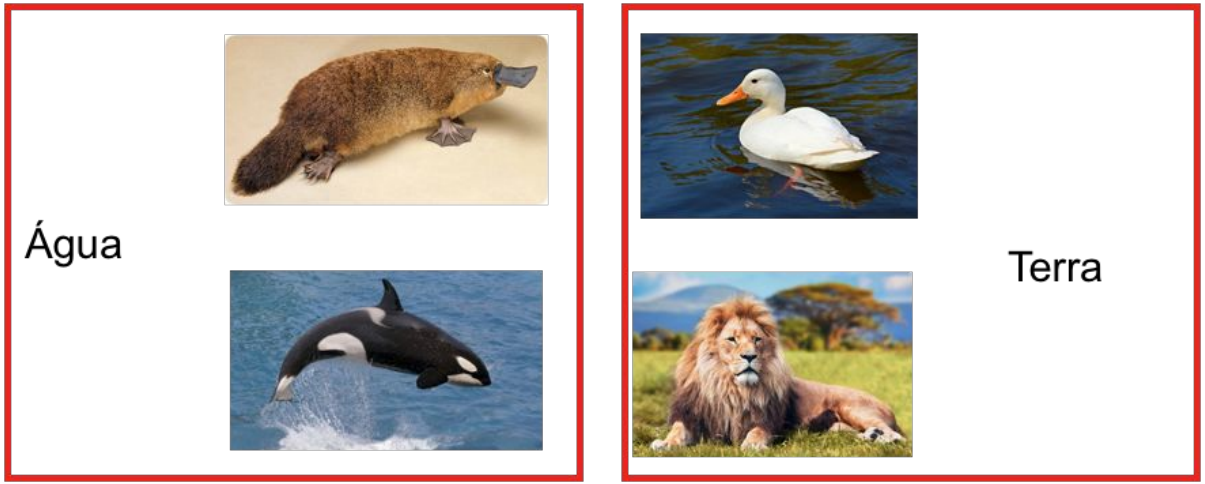
Exemplo: Como agrupar esses animais?



Agrupamento de dados

Dado um conjunto de objetos, agrupar os objetos em grupos (clusters) baseados na similaridade entre eles.

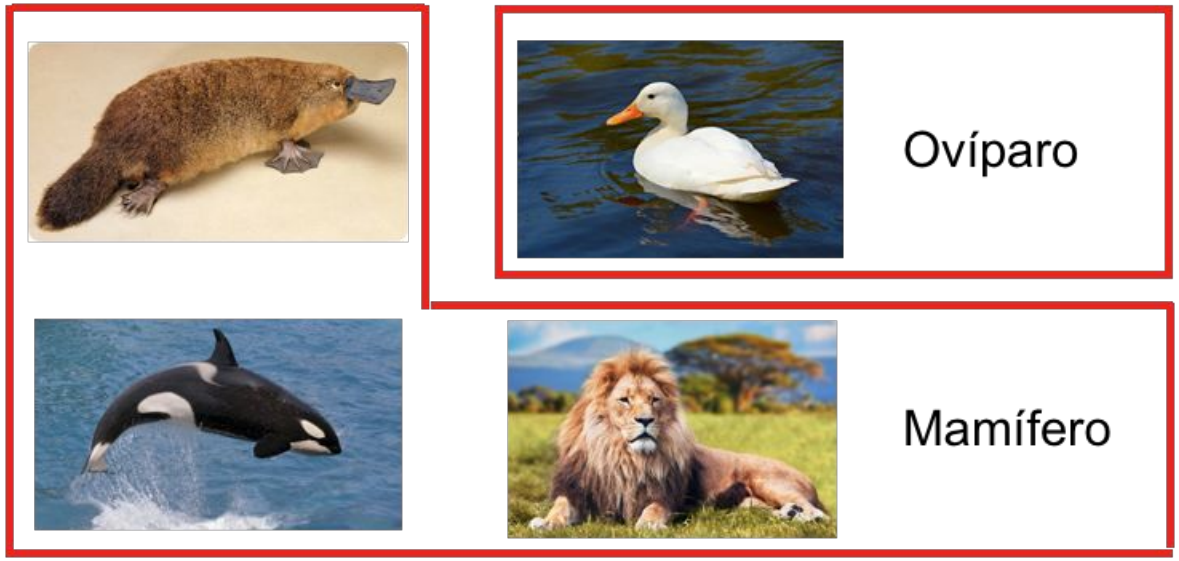
Exemplo: Como agrupar esses animais?



Agrupamento de dados

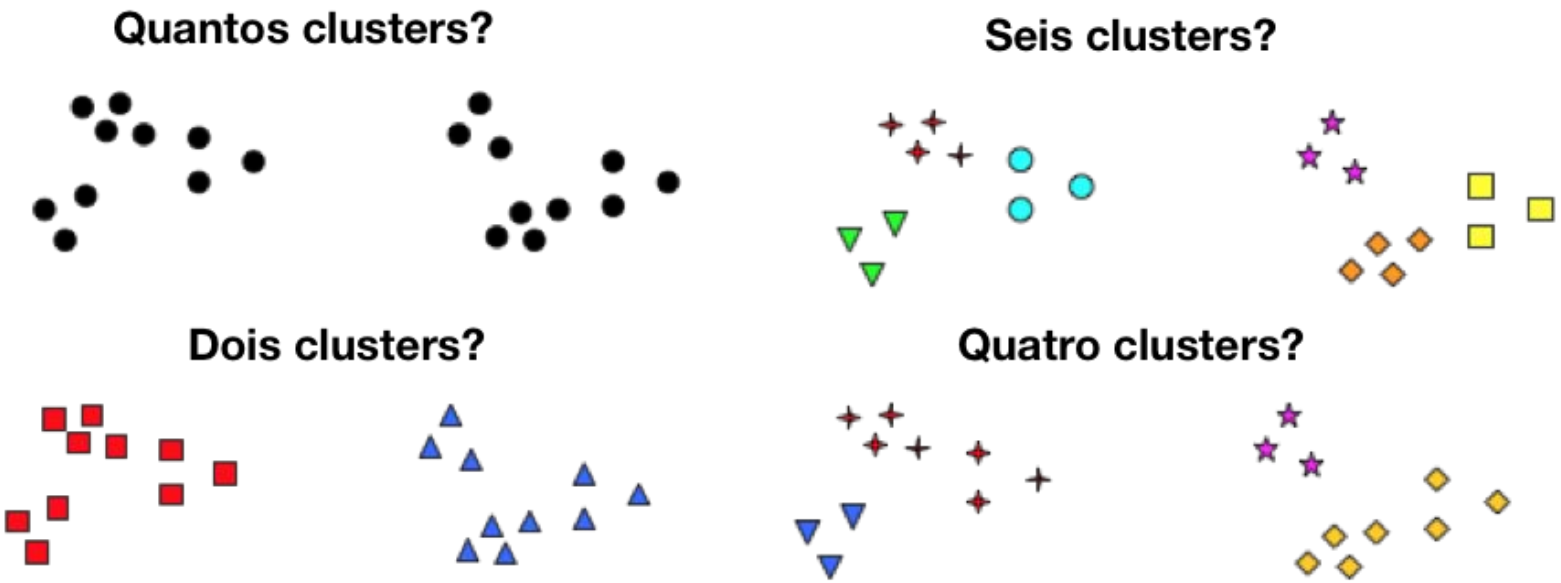
Dado um conjunto de objetos, agrupar os objetos em grupos (clusters) baseados na similaridade entre eles.

Exemplo: Como agrupar esses animais?



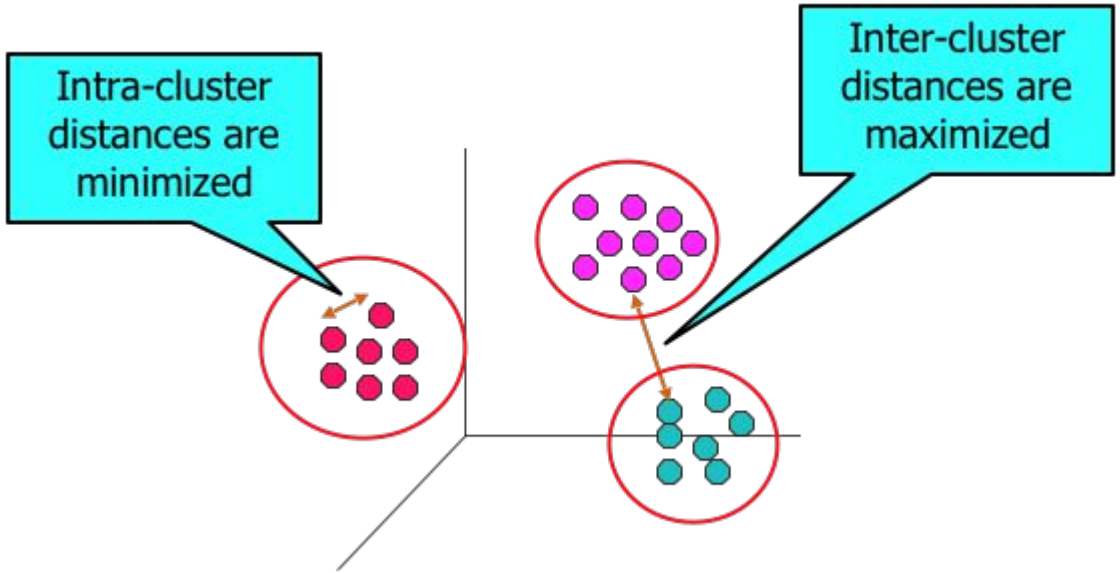
Agrupamento de dados

Limitação: Não há uma definição clara sobre o significado de “cluster” e como encontrá-los.



Agrupamento de dados

Encontrar os grupos de objetos tal que objetos no mesmo grupo serão similares (ou relacionados) um ao outro e diferentes (ou não relacionados) a objetos nos outros grupos.



Agrupamento de dados

- Como definir a similaridade entre os objetos?
- Precisamos definir uma medida de proximidade.
- Medida de similaridade: d(Xi, Xi) é máxima.
  - Exemplo: Número de amigos compartilhados em uma rede social.
- Medida de dissimilaridade: d(Xi,Xi) = 0.
  - Exemplo: distância entre cidades (distância Euclidiana).

Agrupamento de dados

Medida de dissimilaridade:

- $d(p, q) \geq 0$ para todo p e q, e $d(p, q) = 0$ se, e somente se, p = q,
- $d(p, q) = d(q,p)$ para todo p e q,
- $d(p, r) \leq d(p, q) + d(q, r)$ para todo p, q e r, onde $d(p, q)$ é a distância de dissimilaridade entre os pontos (objetos) p e q.

Medida de similaridade:

- $s(p, q) = 1$ (ou máximo de similaridade) se p = q,
- $s(p, q) = s(q, p)$ para todo p e q, onde $s(p, q)$ é a similaridade entre os objetos p e q.

Agrupamento de dados

- **Métricas de distância:**
  - **Euclidiana** $D(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$ $[0,\infty)$
  - **Minkowski** $D(X,Y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}}$ $[0,\infty)$
  - **Cosseno** $D(X,Y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} (x_i)^2} \sqrt{\sum_{i=1}^{n} (y_i)^2}}$ $[0,1]$
  - **Pearson** $D(X,Y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$ $[-1,1]$

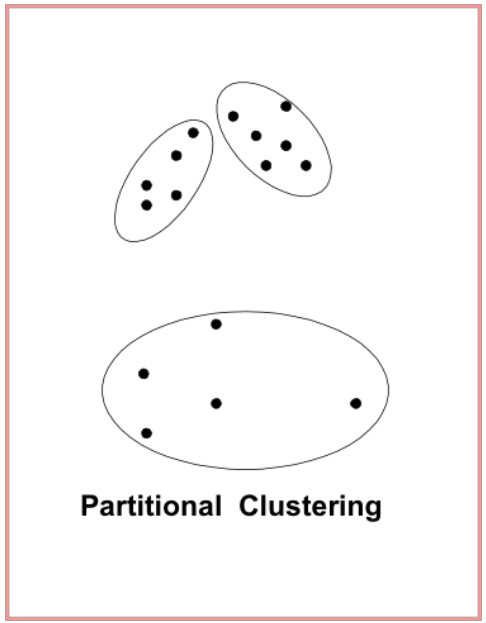
Agrupamento de dados

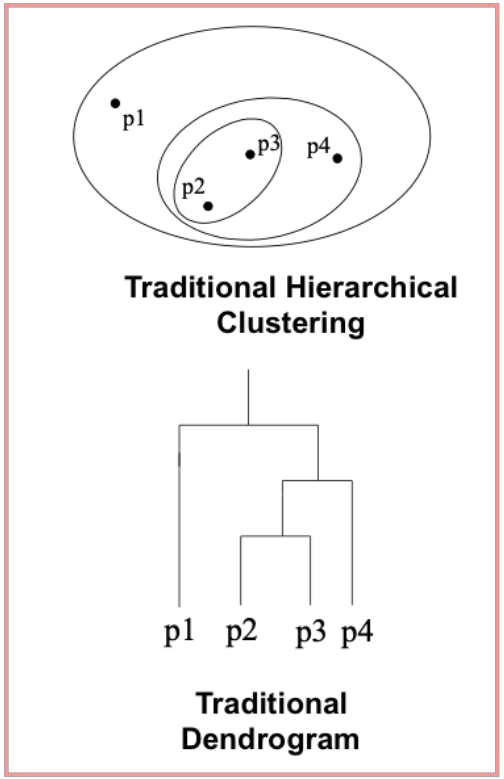
- **Métricas de distância:**

	Similaridade	Dissimilaridade
• **Dados nominais**	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$
• **Dados ordinais**	$s = 1 - \frac{\|p - q\|}{n - 1}$	$d = \frac{\|p - q\|}{n - 1}$

Agrupamento de dados

Tipos:

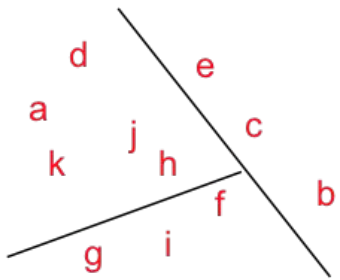




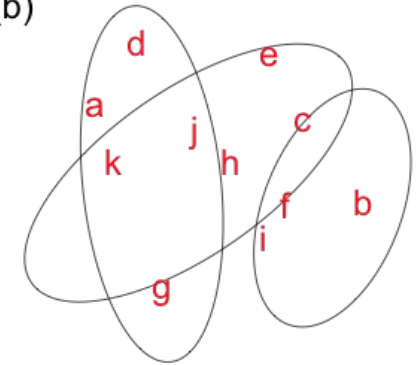
Agrupamento de dados

Representação dos grupos:

(a)



(b)

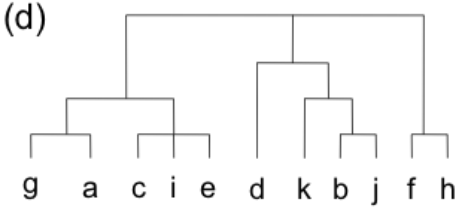


(c)

	1	2	3
a	0.4	0.1	0.5
b	0.1	0.8	0.1
c	0.3	0.3	0.4
...			

Matriz de probabilidades

(d)



Dendograma

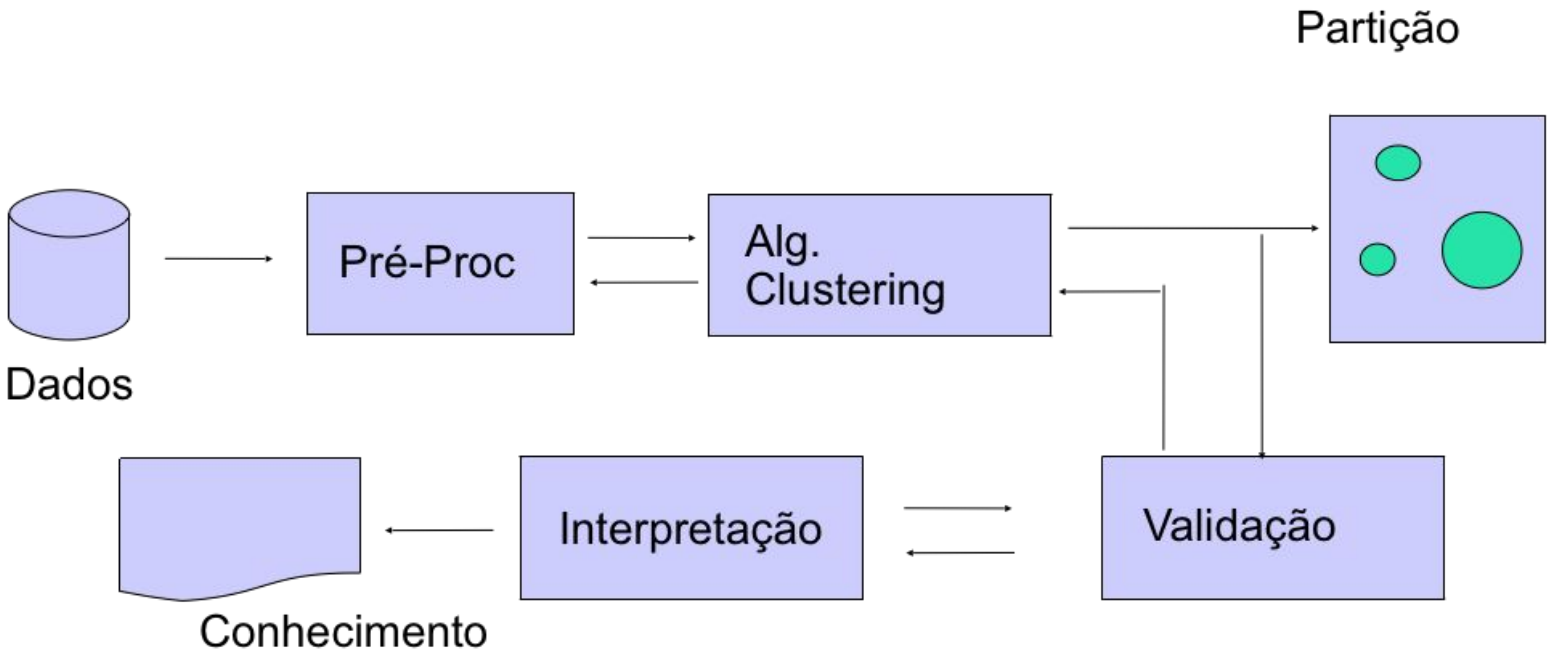
Agrupamento de dados

Estágios:

1. **Seleção de atributos**: Os atributos devem ser selecionados de modo que ocorra o mínimo de redundância entre eles.
2. **Medida de proximidade**: Esta medida deve quantificar o quão similar o dissimilar são os objetos.
3. **Critério de clusterização**: Consiste de uma função custo ou algum tipo de regra.
4. **Algoritmo de clusterização**: Consiste de um conjunto de passos para revelar a estrutura dos dados, baseados na medida de similaridade e no critério adotado.
5. **Validação dos resultados.**
6. **Interpretação dos resultados.**

Agrupamento de dados

Passos:



k-means

k-means

- Amplamente usado na prática:
  - Simplicidade;
  - Interpretabilidade;
  - Eficiência computacional.

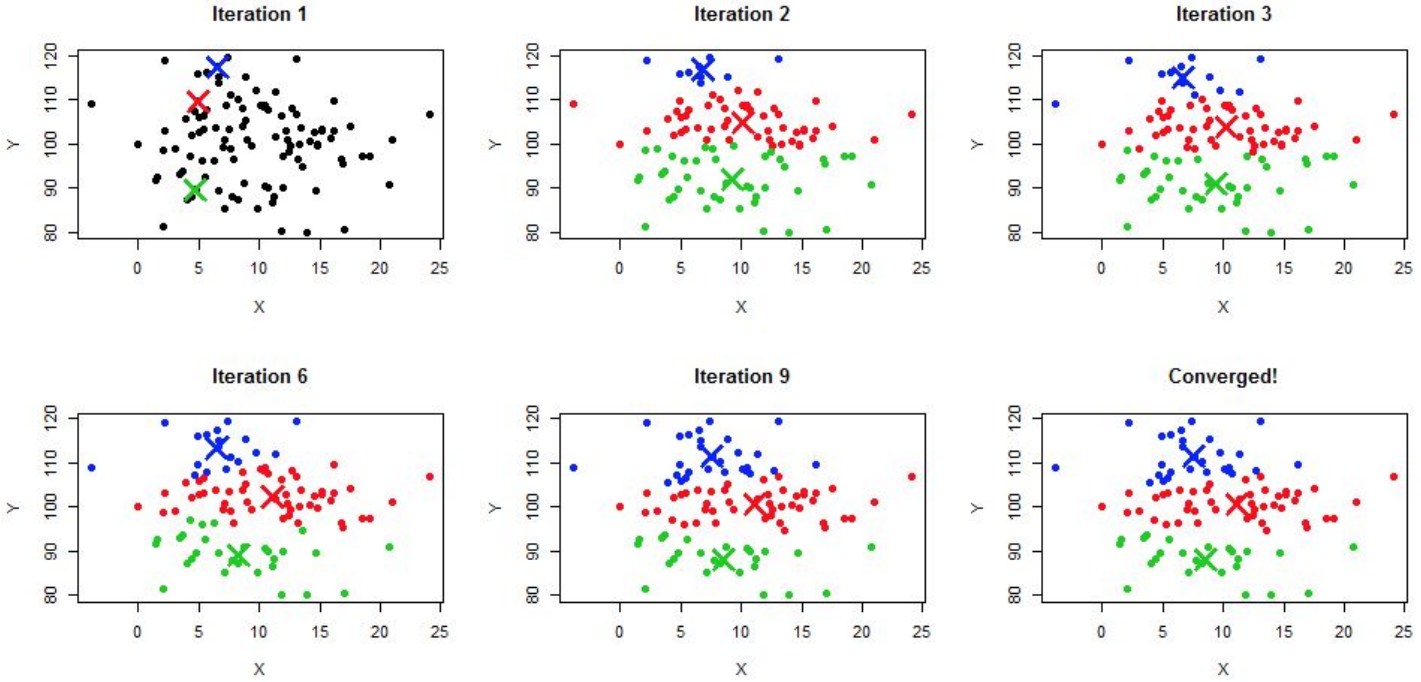
k-means

Algoritmo:

1. Selecione k pontos como centróides iniciais.
2. Repita até que os centróides não mudem.
   a. Forme k grupos associado todos os pontos aos centróide mais próximos,
   b. Calcule o centróide de cada grupo obtido.

k-means

Exemplo:



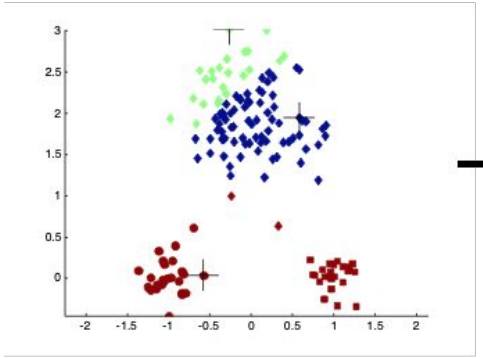
k-means

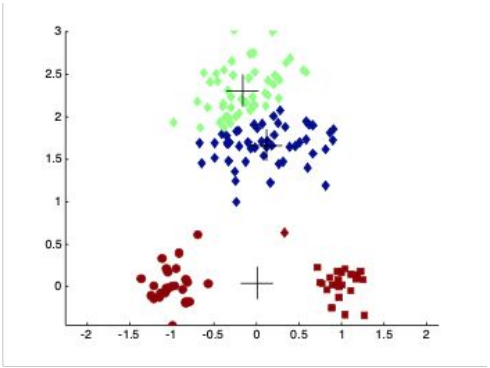
Inicialização:

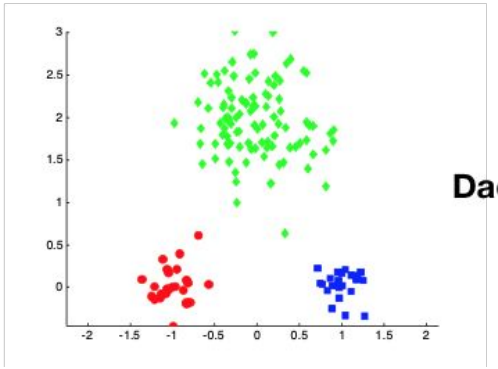
- O algoritmo é sensível à posição inicial das sementes.
- Importante rodar o algoritmo diversas vezes para obter resultados significativos.

k-means

Inicialização:



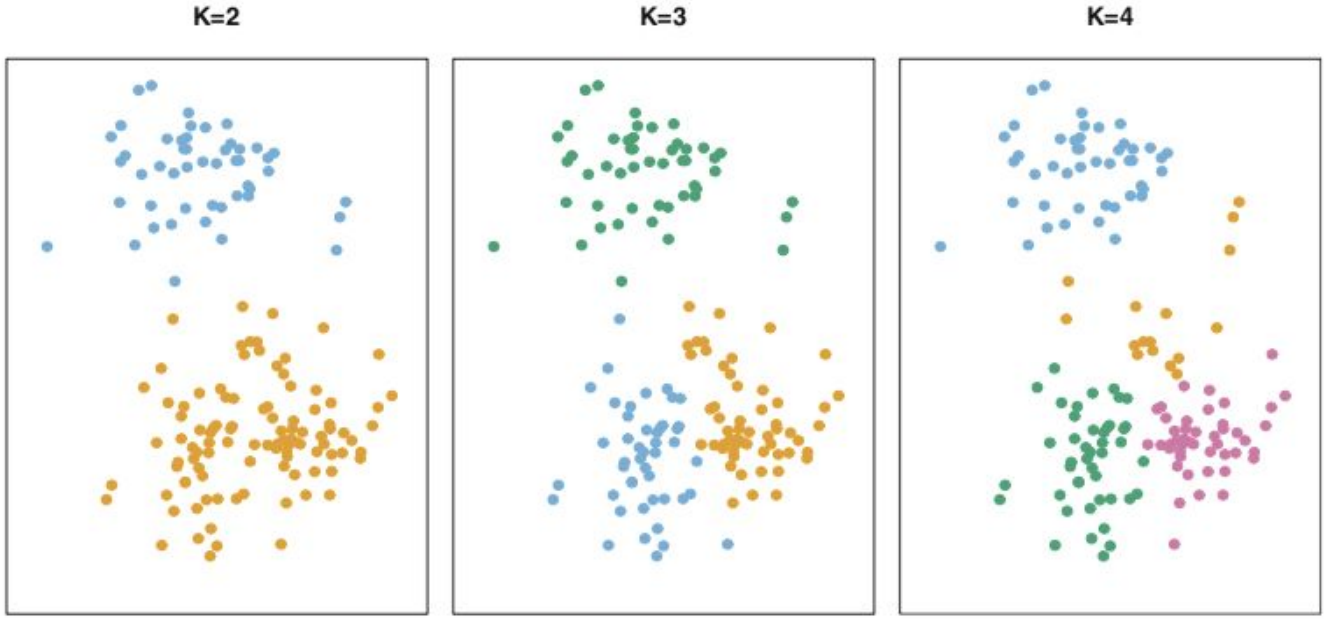




Dados originais

k-means

Qual o número de clusters?



k-means

Elbow method (método do cotovelo):

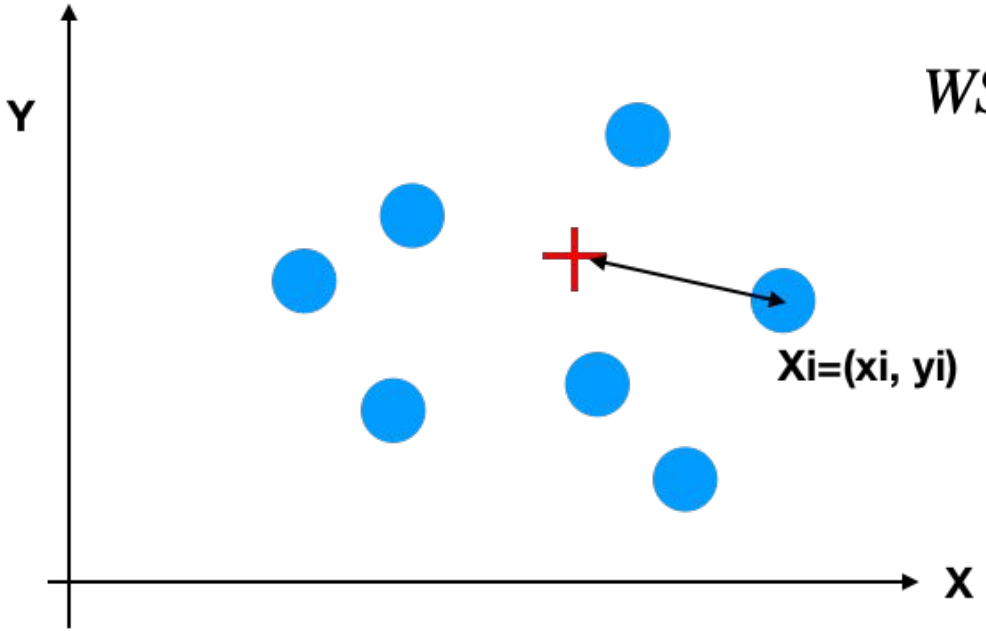
- Usado para encontrar o melhor valor de k, podemos usar a distância média dos pontos dentro de um cluster até o seu centróide (within-cluster sum of squares) para diferentes valores de k.

$$WSS = \sum_{i=1}^{N_c} \sum_{\mathbf{x} \in C_i} d(\mathbf{x}, \bar{x}_{C_i})$$

onde Ci é um grupo e Nc é o número de grupos.

- WSS pode ser entendida como uma medida de compactação.

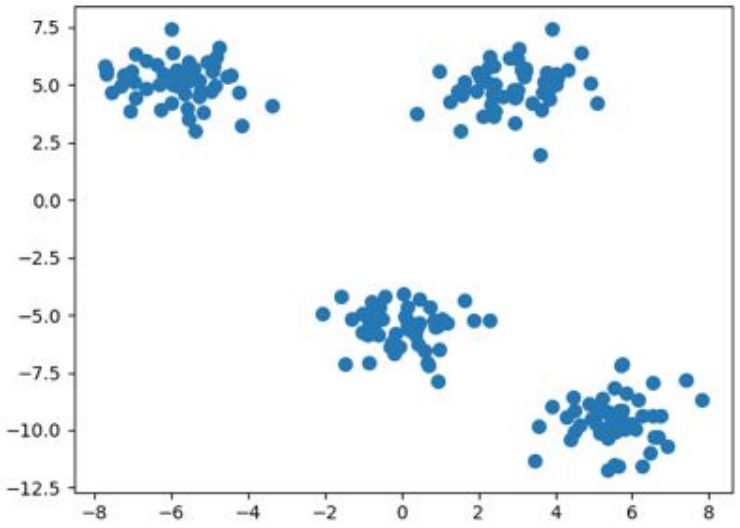
k-means

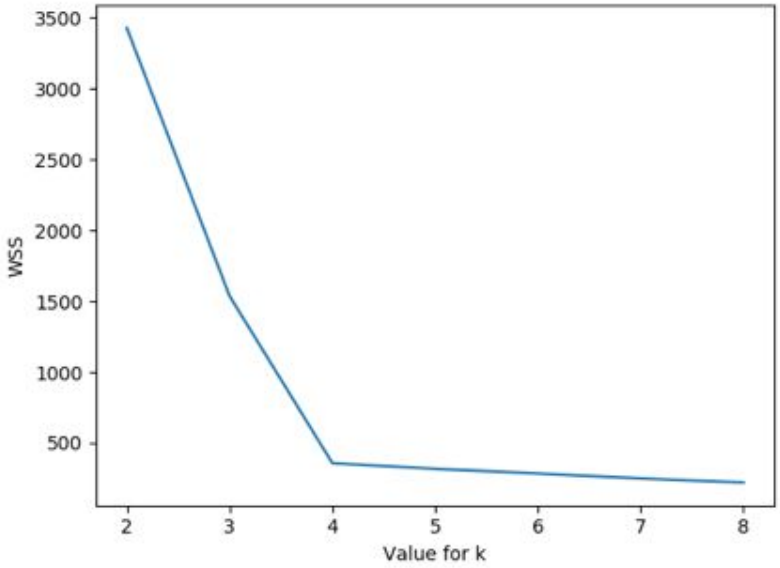


$$WSS = \sum_{i=1}^{N_c} \sum_{\mathbf{x} \in C_i} d(\mathbf{x}, \bar{x}_{C_i})$$

k-means

Elbow method (método do cotovelo):

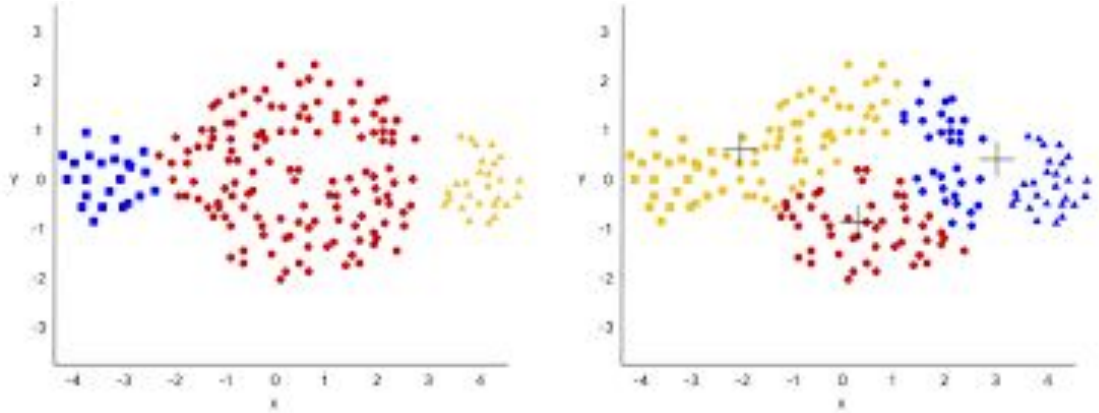




k-means

Limitações:

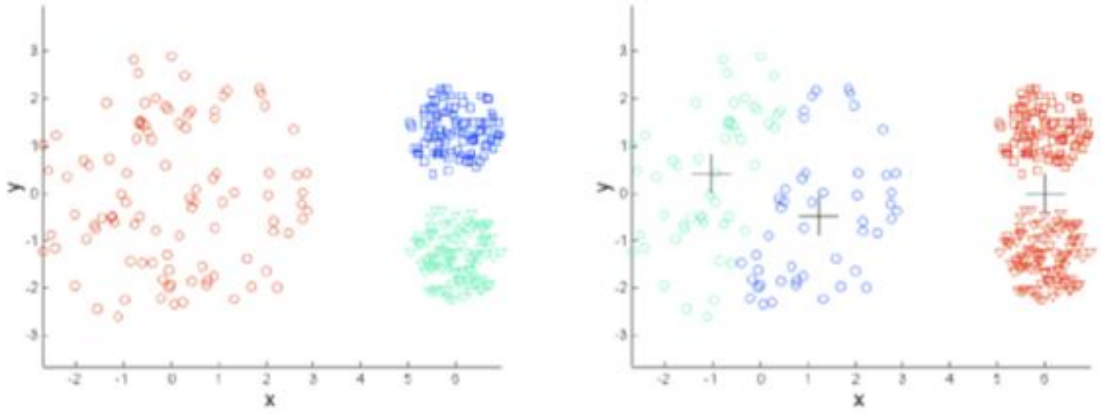
- É bastante susceptível a problemas quando clusters são de diferentes tamanhos.



k-means

Limitações:

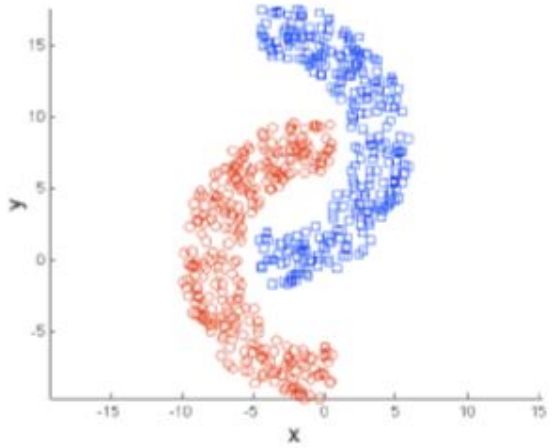
- É bastante susceptível a problemas quando clusters são de diferentes densidades.

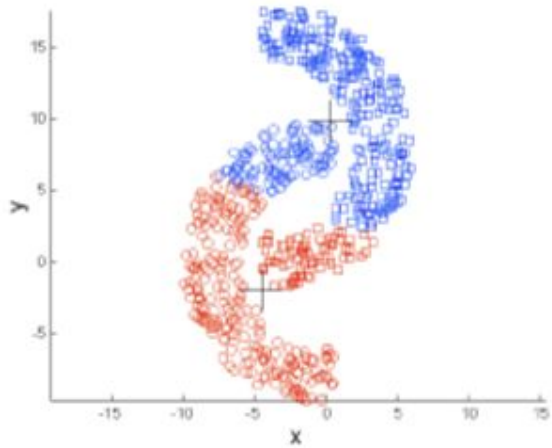


k-means

Limitações:

- É bastante susceptível a problemas quando clusters são de diferentes formatos (em geral não globulares).

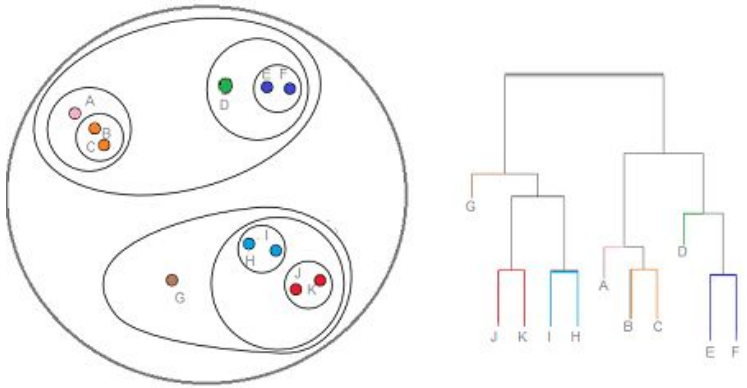




Agrupamento Hierárquico

Agrupamento Hierárquico

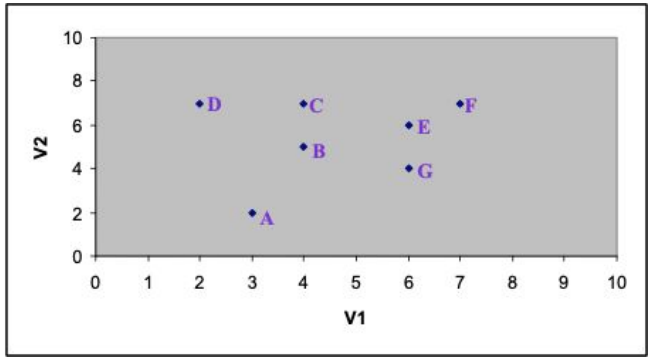
- Um algoritmo de agrupamento hierárquico gera uma estrutura aninhada (árvore) de X, H={H1, H2, ..., HQ} (K ≤ N), tal que:
  - Ci ∈ Hm e Cj ∈ Hl (m > l), implica que
    - Ci ⊆ Cj ou
    - Ci ∩ Cj = ⌀ (i,j,m,l=1,...,Q e i ≠ j)



Agrupamento Hierárquico

- Suponha que um biólogo queira identificar subtipos de um determinado câncer (tumor) com base na expressão gênica do tecido extraído do tumor
- Uma pequena amostra de sete pacientes é selecionada
- A expressão gênica de dois genes (V1 e V2) foi medida para o tumor de cada paciente

Paciente	V1	V2
A	3	2
B	4	5
C	4	7
D	2	7
E	6	6
F	7	7
G	6	4



Introduction to Data Mining, Tan,Steinbach, Kumar, 2004

Agrupamento Hierárquico

Paciente	V1	V2
A	3	2
B	4	5
C	4	7
D	2	7
E	6	6
F	7	7
G	6	4

$$d(A,B) = \sqrt{(3-4)^2 + (2-5)^2} = 3{,}162$$

$$d(C,F) = \sqrt{(4-7)^2 + (7-7)^2} = 3{,}00$$

...

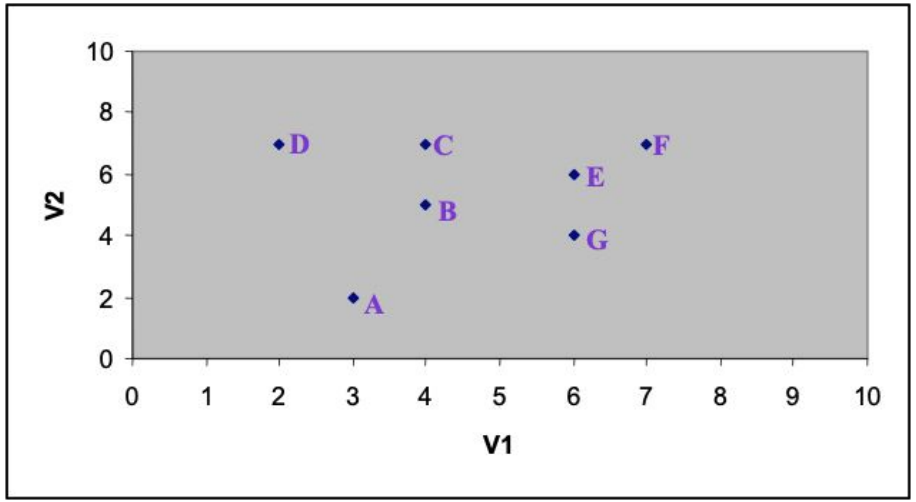
	A	B	C	D	E	F	G
A	0,000						
B	3,162	0,000					
C	5,099	2,000	0,000				
D	5,099	2,828	2,000	0,000			
E	5,000	2,236	2,236	4,123	0,000		
F	6,403	3,606	3,000	5,000	1,414	0,000	
G	3,606	2,236	3,606	5,000	2,000	3,162	0,000

Agrupamento Hierárquico

- Como já temos a medida de similaridade, devemos desenvolver um procedimento para formar grupos.
- Para nosso propósito, usaremos uma regra simples:
  - Identifique as duas observações mais semelhantes (mais próximas) que ainda não estão no mesmo grupo e combine seus grupos.
  - Aplicamos essa regra repetidamente, começando com cada observação em seu próprio grupo e combinando dois grupos por vez, até que todas as observações estejam em um único grupo.
- Este é um procedimento Hierárquico e Aglomerativo.

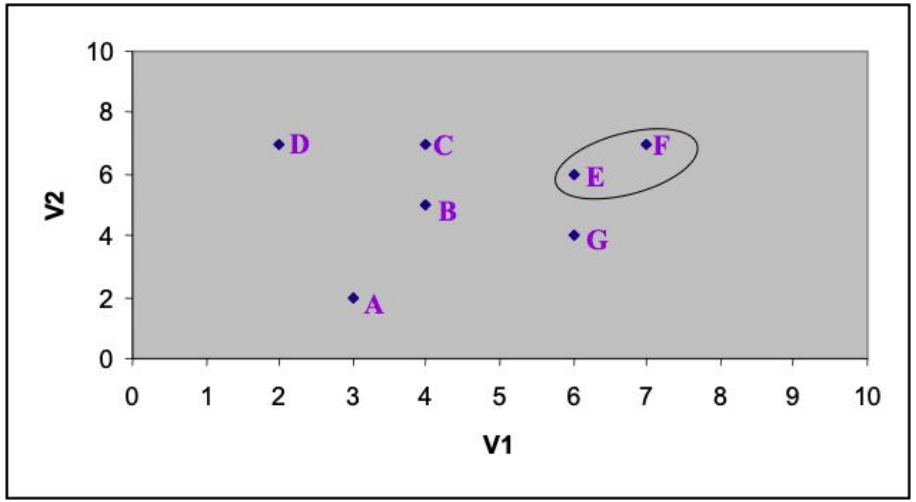
Agrupamento Hierárquico

	A	B	C	D	E	F	G
A	0,000						
B	3,162	0,000					
C	5,099	2,000	0,000				
D	5,099	2,828	2,000	0,000			
E	5,000	2,236	2,236	4,123	0,000		
F	6,403	3,606	3,000	5,000	1,414	0,000	
G	3,606	2,236	3,606	5,000	2,000	3,162	0,000



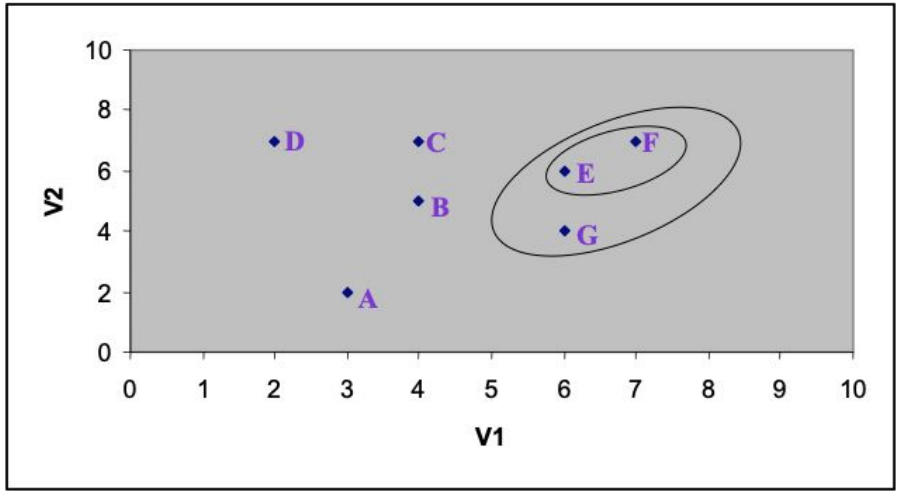
Agrupamento Hierárquico

	A	B	C	D	E	F	G
A	0,000						
B	3,162	0,000					
C	5,099	2,000	0,000				
D	5,099	2,828	2,000	0,000			
E	5,000	2,236	2,236	4,123	0,000		
F	6,403	3,606	3,000	5,000	1,414	0,000	
G	3,606	2,236	3,606	5,000	2,000	3,162	0,000



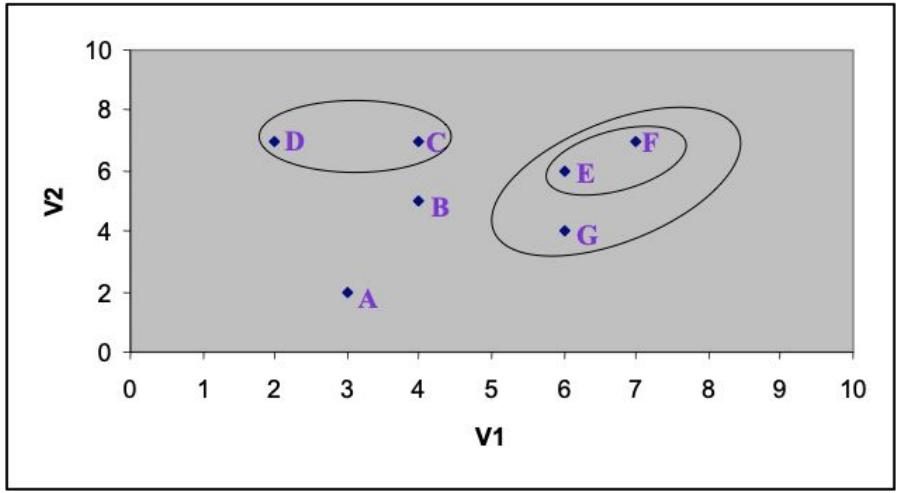
Agrupamento Hierárquico

	A	B	C	D	E	F	G
A	0,000						
B	3,162	0,000					
C	5,099	2,000	0,000				
D	5,099	2,828	2,000	0,000			
E	5,000	2,236	2,236	4,123	0,000		
F	6,403	3,606	3,000	5,000	1,414	0,000	
G	3,606	2,236	3,606	5,000	2,000	3,162	0,000



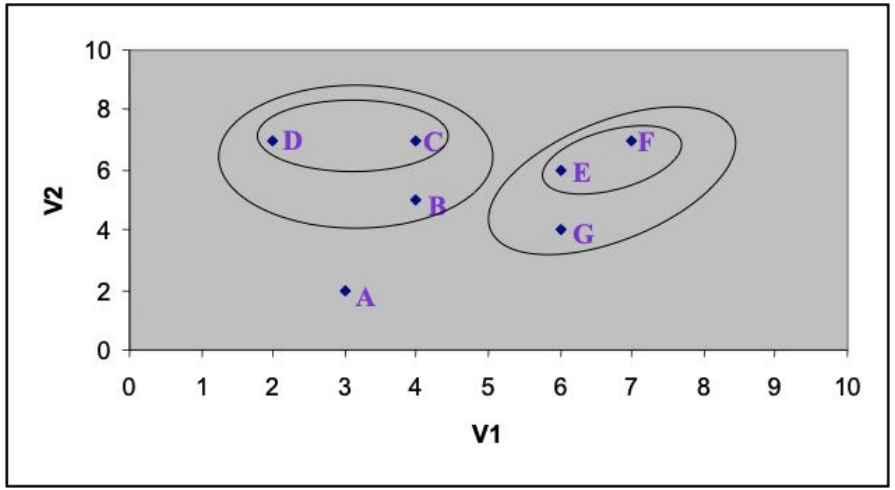
Agrupamento Hierárquico

	A	B	C	D	E	F	G
A	0,000						
B	3,162	0,000					
C	5,099	2,000	0,000				
D	5,099	2,828	2,009	0,000			
E	5,000	2,236	2,236	4,123	0,000		
F	6,403	3,606	3,000	5,000	1,414	0,000	
G	3,606	2,236	3,606	5,000	2,000	3,162	0,000



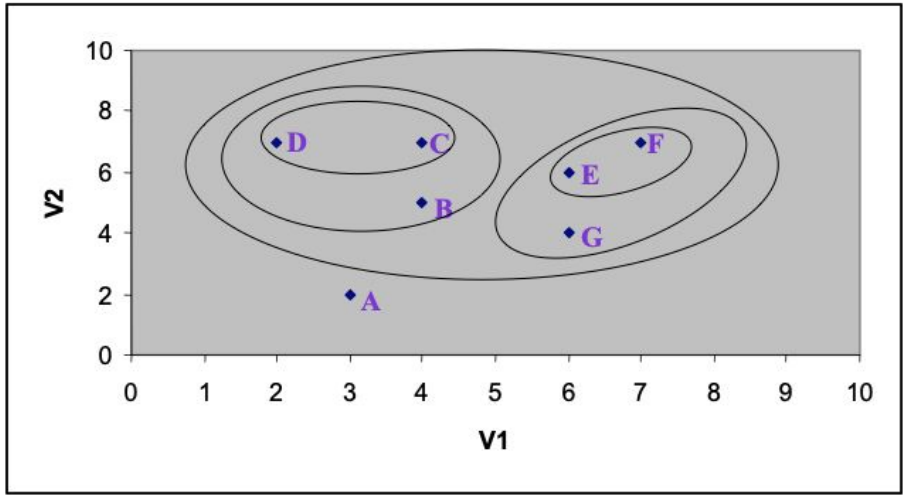
Agrupamento Hierárquico

	A	B	C	D	E	F	G
A	0,000						
B	3,162	0,000					
C	5,099	2,000	0,000				
D	5,099	2,828	2,000	0,000			
E	5,000	2,236	2,236	4,123	0,000		
F	6,403	3,606	3,000	5,000	1,414	0,000	
G	3,606	2,236	3,606	5,000	2,000	3,162	0,000



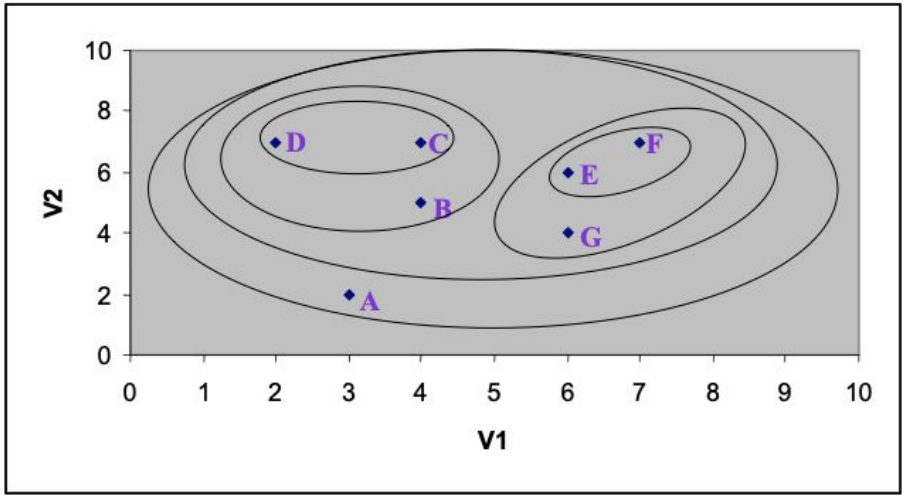
Agrupamento Hierárquico

	A	B	C	D	E	F	G
A	0,000						
B	3,162	0,000					
C	5,099	2,000	0,000				
D	5,099	2,828	2,000	0,000			
E	5,000	2,236	2,236	4,123	0,000		
F	6,403	3,606	3,000	5,000	1,414	0,000	
G	3,606	2,236	3,606	5,000	2,000	3,162	0,000

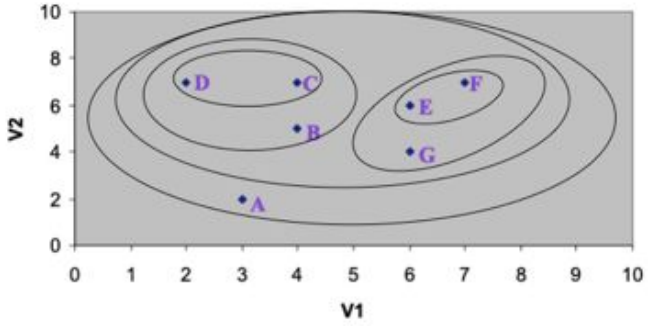


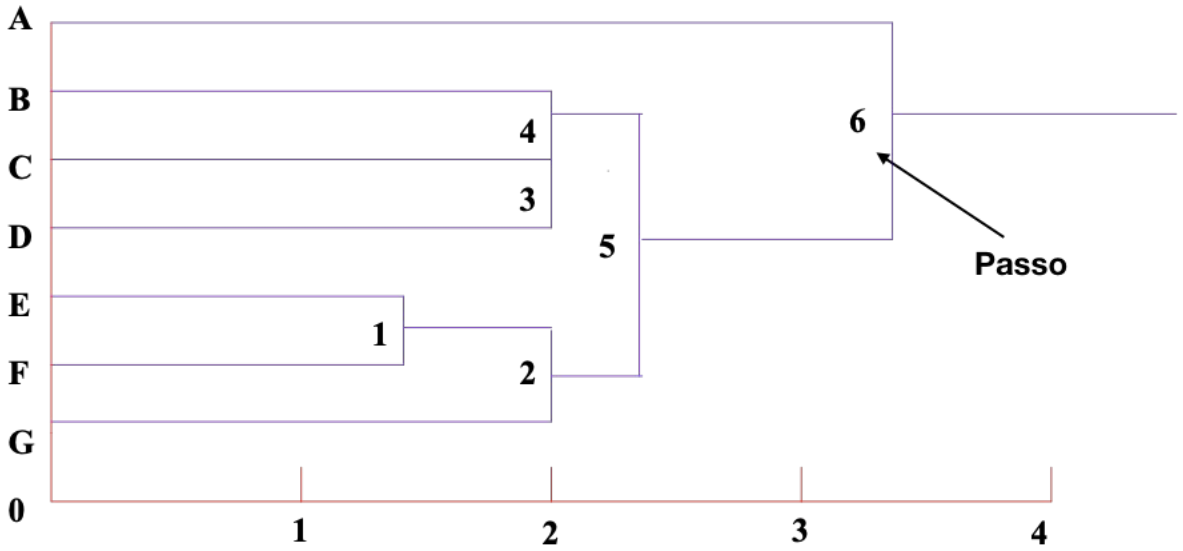
Agrupamento Hierárquico

	A	B	C	D	E	F	G
A	0,000						
B	3,162	0,000					
C	5,099	2,000	0,000				
D	5,099	2,828	2,000	0,000			
E	5,000	2,236	2,236	4,123	0,000		
F	6,403	3,606	3,000	5,000	1,414	0,000	
G	3,606	2,236	3,606	5,000	2,000	3,162	0,000

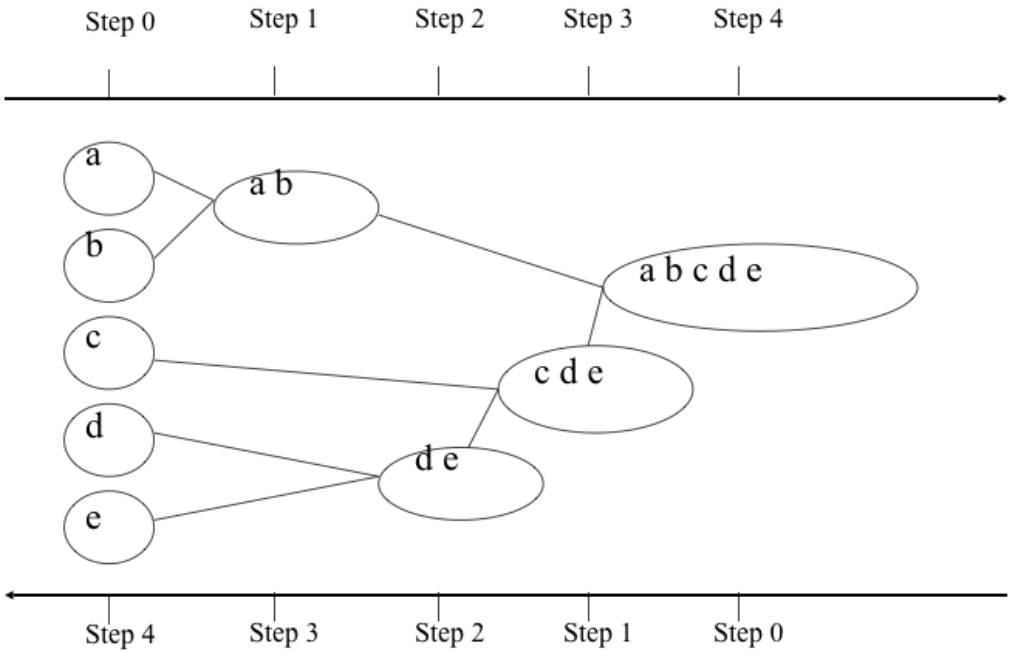


Agrupamento Hierárquico





Agrupamento Hierárquico

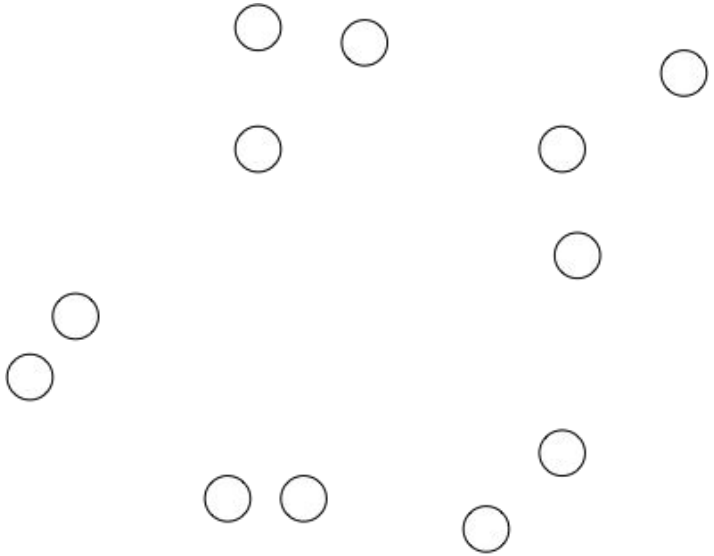


Aglomerativo

Divisivo

Agrupamento Hierárquico

Iniciando com clusters individuais, definimos a matriz de proximidades.

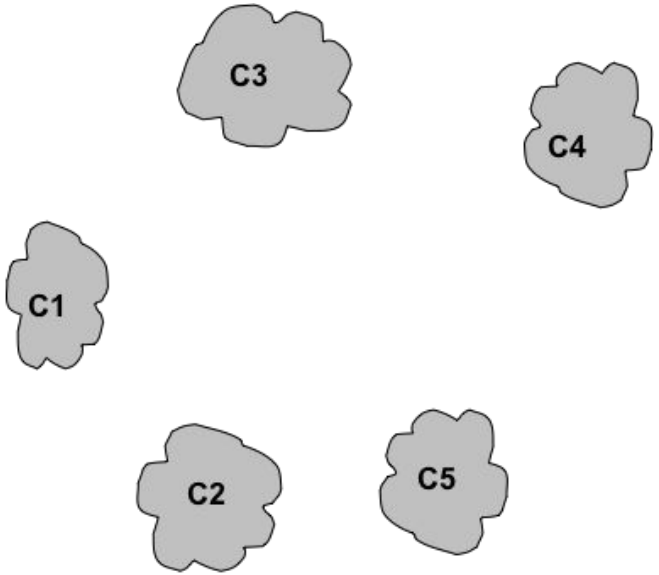


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Matriz de Proximidade

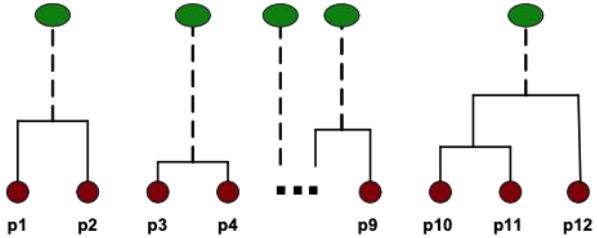
Agrupamento Hierárquico

Depois de alguns passos, temos alguns clusters.



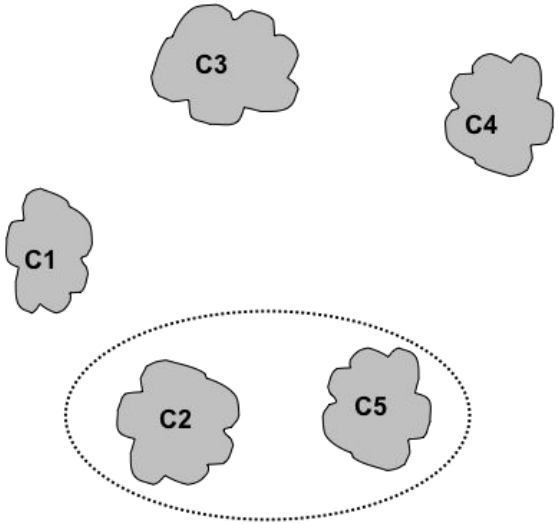
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Matriz de proximidade



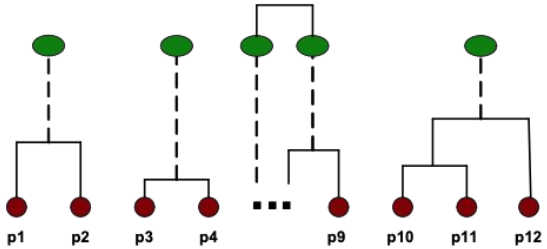
Agrupamento Hierárquico

Nós queremos agrupar os clusters mais próximos (C2 e C5) e atualizar a matriz de proximidades.



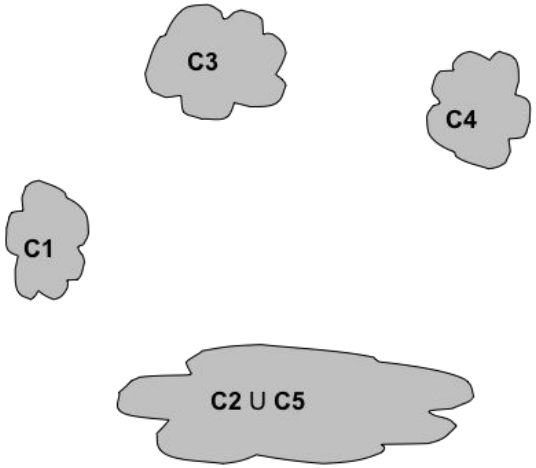
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Matriz de proximidades.



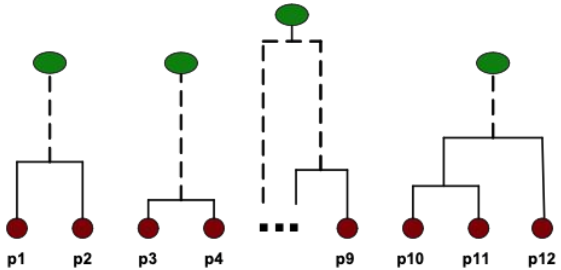
Agrupamento Hierárquico

Uma questão fundamental é: Como realizar a atualização da matriz?



	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Matriz de proximidade

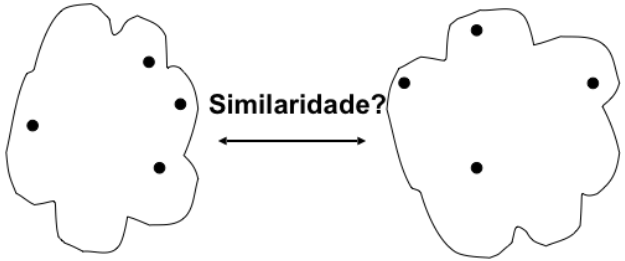


Agrupamento Hierárquico

- MIN (single linkage)
- MAX (complete linkage)
- Média dos grupos
- Distância entre centróides
- Outros métodos que usam uma função objetivo.
- Método de Weard's usa erro quadrático médio.

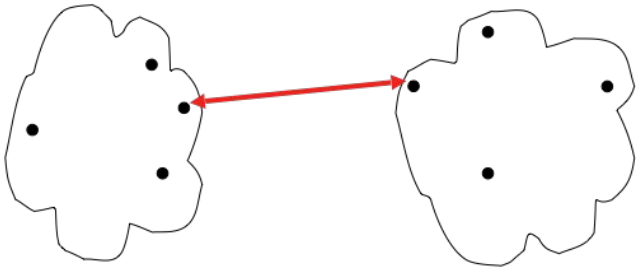
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Matriz de proximidade



Agrupamento Hierárquico

- **MIN (single linkage).**
- MAX (complete linkage).
- Média dos grupos.
- Distância entre centróides.
- Outros métodos que usam uma função objetivo.
- Método de Weard's usa erro quadrático médio.

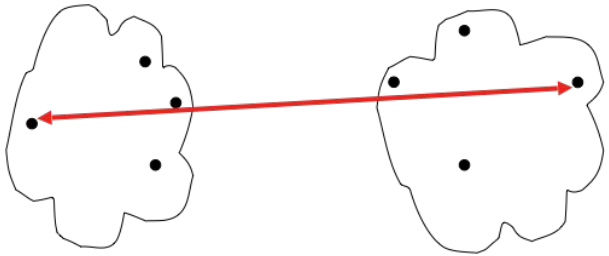


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Matriz de proximidade

Agrupamento Hierárquico

- MIN (single linkage).
- **MAX (complete linkage).**
- Média dos grupos.
- Distância entre centróides.
- Outros métodos que usam uma função objetivo.
- Método de Weard's usa erro quadrático médio.

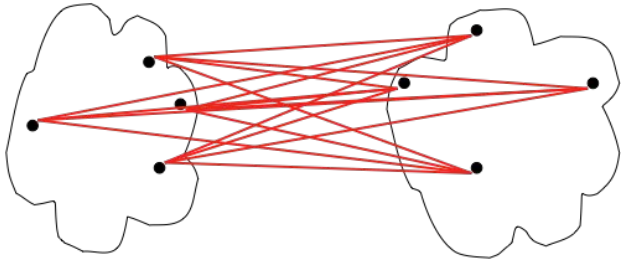


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Matriz de proximidade

Agrupamento Hierárquico

- MIN (single linkage).
- MAX (complete linkage).
- **Média dos grupos.**
- Distância entre centróides.
- Outros métodos que usam uma função objetivo.
- Método de Weard's usa erro quadrático médio.

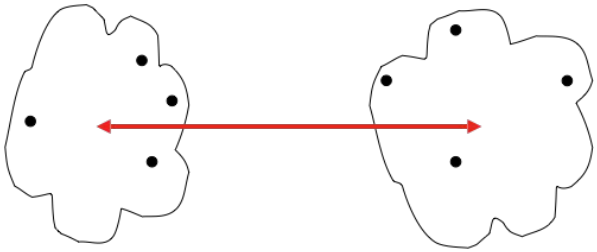


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Matriz de proximidade

Agrupamento Hierárquico

- MIN (single linkage)
- MAX (complete linkage)
- Média dos grupos
- **Distância entre centróides.**
- Outros métodos que usam uma função objetivo.
- Método de Weard's usa erro quadrático médio.

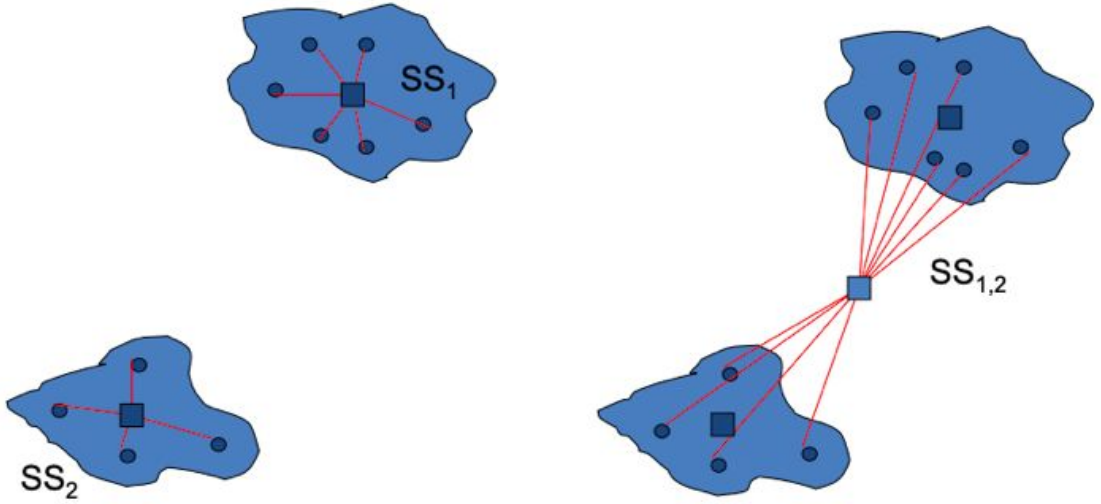


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Matriz de proximidade

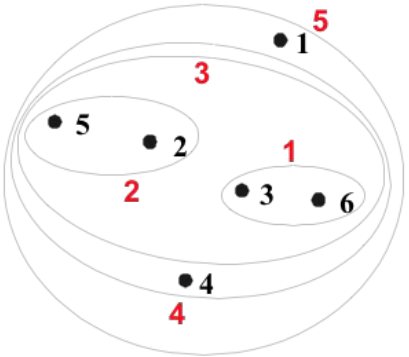
Agrupamento Hierárquico

Método de Ward: Minimizar a perda de informação ao juntar 2 grupos.



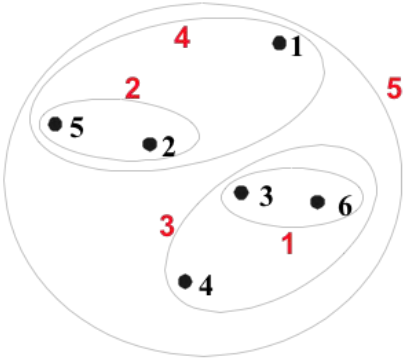
$$d(C_l, C_i) = SS_{l,i} - (SS_l + SS_i)$$

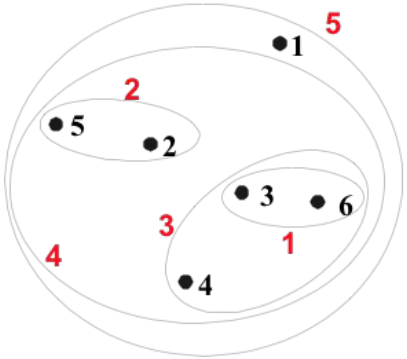
Agrupamento Hierárquico



MIN

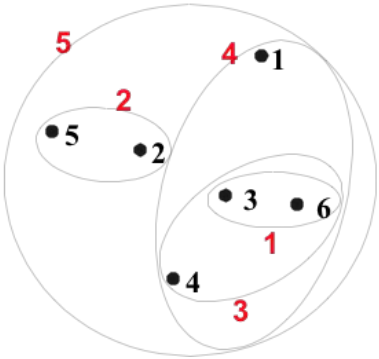
MAX





Média

Método de Ward's

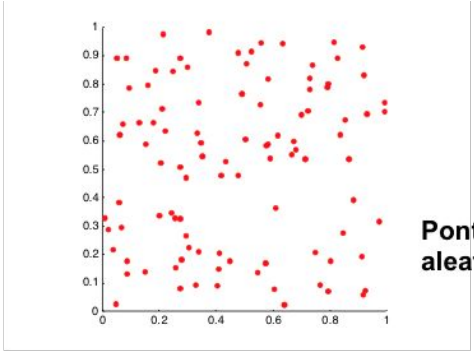


Avaliando Agrupamentos

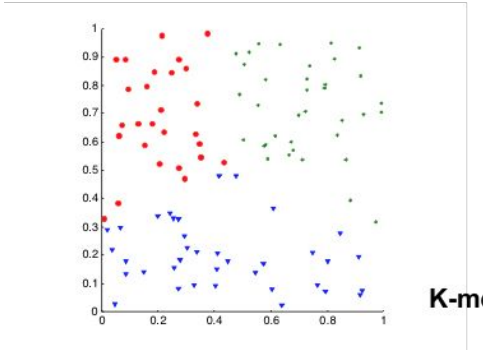
Avaliando Agrupamentos

- Quando encontramos os clusters, uma questão básica é:
- **Quão significativo é o agrupamento?**
- A avaliação de agrupamentos pode ser usada:
  - Para evitar encontrar padrões em ruídos.
  - Para comparar diferentes métodos de agrupamento.
  - Para comparar clusters.

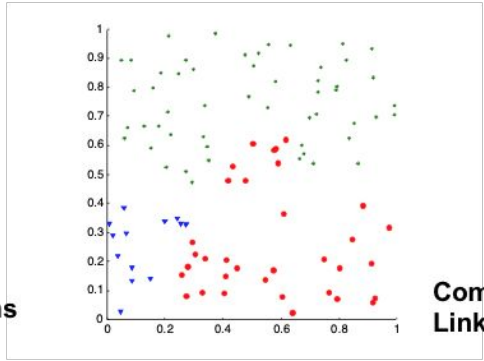
Avaliando Agrupamentos



Pontos aleatórios



K-means



Complete Linkage

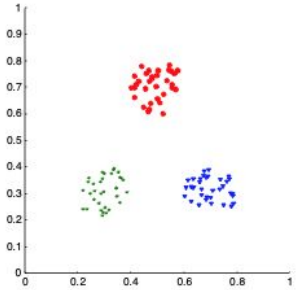
- Podemos encontrar partições em dados aleatórios.

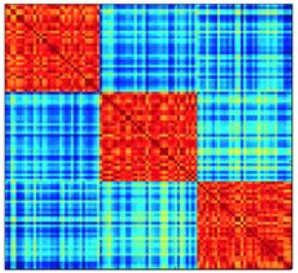
Avaliando Agrupamentos

- Medidas para avaliar agrupamentos são usadas em três casos:
  1. **Índice externo**: Usado quando os rótulos dos objetos são conhecidos e queremos avaliar se os clusters correspondem aos grupos originais.
     a. Exemplo: Medidas de entropia.
  2. **Índice interno**: Usado para avaliar um agrupamento sem usar informações externas.
     a. Exemplo: Soma do erro quadrático
  3. **Índice relativo**: Usado para comprar agrupamentos ou grupos.
     a. Exemplo: Índices internos ou externos são usados para esse fim.

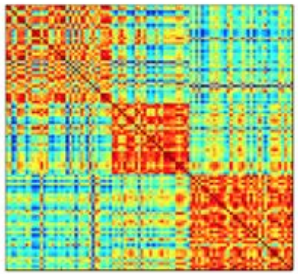
Avaliando Agrupamentos

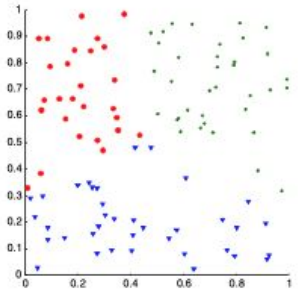
- **Matriz de similaridade**: ordena-se os objetos de acordo com os grupos e inspeciona-se visualmente. Cada elemento da matriz define a similaridade entre dois objetos (por exemplo, distância euclidiana entre dois vetores de atributos):





k-means

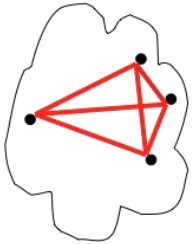




- Limitação: nem sempre essa estrutura é clara.

Avaliando Agrupamentos

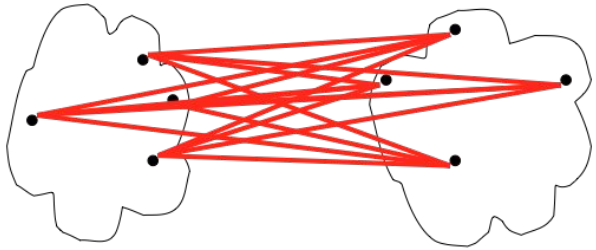
Índice interno: Usado para avaliar um agrupamento sem usar informações externas.



Coesão

- a distância média dos pontos dentro de um cluster até o seu centróide (*within-cluster sum of squares)* (SSE):

$$WSS = \sum_{i=1}^{N_c} \sum_{\mathbf{x} \in C_i} d(\mathbf{x}, \bar{x}_{C_i})$$



Separação

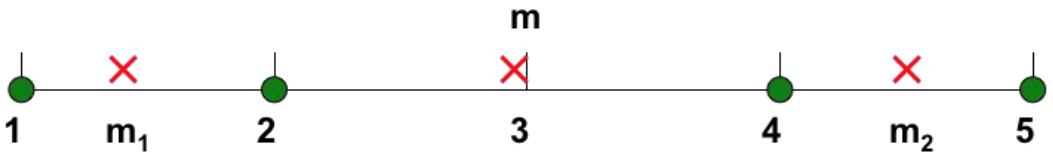
- Soma dos quadrados das distâncias entre os pontos em diferentes clusters:

$$BSS = \sum_{i} |C_i| (m - m_i)^2$$

Avaliando Agrupamentos

Exemplo: BSS + WSS = constante

$$WSS = \sum_{i=1}^{N_c} \sum_{\mathbf{x} \in C_i} d(\mathbf{x}, \bar{x}_{C_i})$$



$$BSS = \sum_{i} |C_i| (m - m_i)^2$$

K=1 cluster:

$$WSS = (1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2 = 10$$

$$BSS = 4 \times (3-3)^2 = 0$$

$$Total = 10 + 0 = 10$$

K=2 clusters:

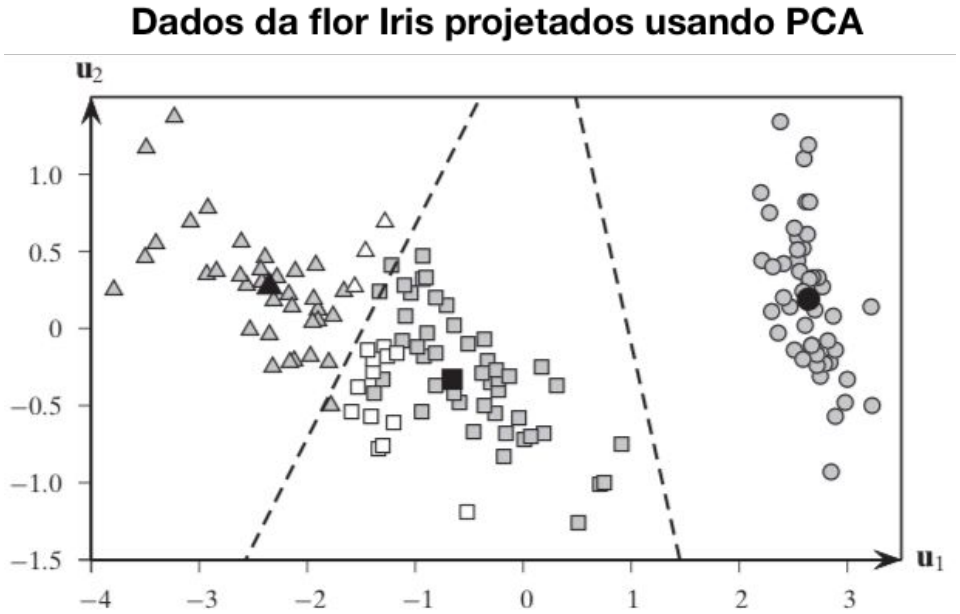
$$WSS = (1-1.5)^2 + (2-1.5)^2 + (4-4.5)^2 + (5-4.5)^2 = 1$$

$$BSS = 2 \times (3-1.5)^2 + 2 \times (4.5-3)^2 = 9$$

$$Total = 1 + 9 = 10$$

Avaliando Agrupamentos

Índice externo: Usado para avaliar um agrupamento quando se tem um ground truth (conhecemos as classes).



Avaliando Agrupamentos

- **Purity (Pureza):** mede o quão “puro" é cada cluster:

$$Purity(i) = \frac{1}{n_i} \max_{j=1}^{k} (|C_i \cap T_j|)$$

onde Ci representa a classe obtida e Tj é a partição esperada.

	iris-setosa	iris-versicolor	iris-virginica	
	T_1	T_2	T_3	n_i
C_1(squares)	0	47	14	61
C_2(circles)	50	0	0	50
C_3(triangles)	0	3	36	39
m_j	50	50	50	$n = 100$

Tabela de contigência

$$Purity(i) = \frac{1}{150}(47 + 50 + 36) = 0{,}887$$

Avaliando Agrupamentos

Índice externo: Outras medidas:

- Maximum matching,
- f-Measure
- Normalized mutual information (NMI).
- ...

Avaliando Agrupamentos

"The validation of clustering structures is the most difficult and frustrating part of cluster analysis.

Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage."

Algorithms for Clustering Data, Jain and Dubes

Sumário

- K-means,
- Agrupamento Hierárquico,
- Avaliando Agrupamentos

Leitura complementar

- Capítulos 11 e 12:
  Inteligência Artificial: Uma abordagem por aprendizado de máquina, Facelli, Lorena, Gamma e Carvalho, LTC.

- Capítulo 8:
  Introduction to Data Mining, Tan,Steinbach, Kumar, 2004