

ANÁLISE DE CLUSTERS (conglomerados ou agrupamentos)

Ana Amélia Benedito-Silva

aamelia@usp.br

<https://www.youtube.com/watch?v=WqMnQuC19Rg>

Análise de Conglomerados

- Análise de conglomerados, análise de agrupamentos ou análise de clusters são técnicas de interdependência
- Permite agrupar casos em um grupo homogêneo, em função de suas similaridades ou semelhanças
- Os objetos (indivíduos) em cada grupo tendem a ser semelhantes entre si e diferentes dos demais objetos (indivíduos) contidos em outros grupos.

Análise de Conglomerados

- é uma técnica exploratória
- permite estudar a estrutura de grupos
- permite identificar *outliers*
- permite levantar hipóteses sobre as associações dos objetos
- é uma técnica não-inferencial, ou seja, não possibilita inferências sobre a população com base na amostra

Análise de Conglomerados

- A Análise de Cluster é aplicada na maioria das vezes em pesquisas de caráter exploratório.
- É uma técnica para analisar interdependência entre casos/indivíduos segundo determinadas variáveis.
- Não é possível determinar antecipadamente as variáveis dependentes e independentes
- Ao contrário, examina relações de interdependência contidas na estrutura dos dados.

Peculiaridades da análise de conglomerados

- possui forte base matemática, mas não estatística.
- suposições como normalidade, linearidade e homoscedasticidade, importantes em outras técnicas multivariadas, possuem pouco impacto

Análise de agrupamentos

Cluster Analysis

Identificação dos *clusters* a partir dos dados

Cada ponto representa uma empresa

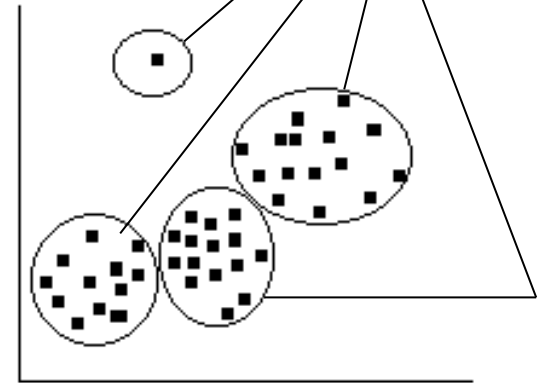
FATURAMENTO



NÚMERO DE
EMPREGADOS

Algoritmo de
análise de
agrupamentos

FATURAMENTO



NÚMERO DE
EMPREGADOS

Clusters

O que é possível fazer com a análise de agrupamentos ?

- **Formar uma taxonomia das observações**
Identificar grupos naturais no interior dos dados, i.e., uma classificação empírica dos objetos.
- **Simplificar os dados (compressão de dados)**
Descrever de forma compacta um razoável volume de dados por meio dos elementos típicos dos clusters (médias ou medianas).
- **Identificar relações entre objetos**
 - Revelar similaridades e diferenças entre objetos não reveladas de outra maneira.
 - Identificar observações aberrantes.

Exemplos

1. Classificar setores censitários de acordo com as diferentes dimensões de justiça/injustiça ambiental
2. Classificar os municípios de SP em função das diferentes dimensões de violência contra a mulher
3. Classificar os bairros do ABC de acordo com a quantidade/perfil dos lançamentos residenciais

Exemplos

4. Classificar os distritos de SP de acordo com as variáveis de infraestrutura e entorno de domicílios
5. Classificar consumidores em relação aos seus hábitos de compra em uma rede de supermercados
6. Um arqueólogo tem dados sobre a localização de restos de cerâmica encontrados em um sítio arqueológico. Para conhecer como era a organização espacial da tribo que lá habitava, ele necessita ter uma ideia mais precisa da dispersão dessas peças.
 - Há locais com alta concentração de peças? Quantos?

Exemplo 4 - Classificar os distritos de SP de acordo com as variáveis de infraestrutura e entorno de domicílios

distrito	População	IDH	Número de hospitais	Número de escolas
Santana				
Butantã				
Freguesia do Ó				
Pinheiros				
..				
...				
...				
...				
...				
Santo Amaro				
Vila Mariana				

Exemplo: Agrupar alunos segundo notas de avaliação



CONCEITO A – 8,8 a 10

CONCEITO B – 7,0 a 8,7

CONCEITO C – 5,0 a 6,9

CONCEITO D – abaixo de 4,9

Identificação_aluno	nota
1	5,0
2	5,5
3	4,3
4	3,0
5	3,3
6	1,2
7	4,4
8	5,4
9	3,0
10	2,2
11	4,1
12	7,0
13	1,0
14	3,9
15	4,2
16	2,6
17	5,9
18	6,2
19	3,2
20	1,8

1 CONCEITO B

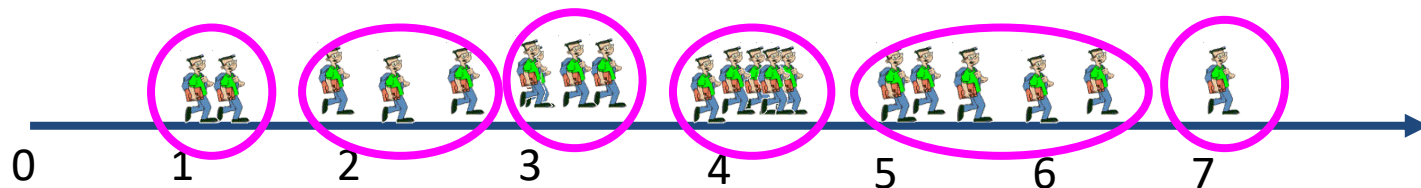
5 CONCEITOS C

14 CONCEITOS D

ANÁLISE DE CLUSTER

ID_Aluno	NOTA
1	5
2	5,5
3	4,3
4	3
5	3,3
6	1,2
7	4,4
8	5,4
9	3
10	2,2
11	4,1
12	7
13	1
14	3,9
15	4,2
16	2,6
17	5,9
18	6,2
19	3,2
20	1,8

SOLUÇÃO = 6
GRUPOS

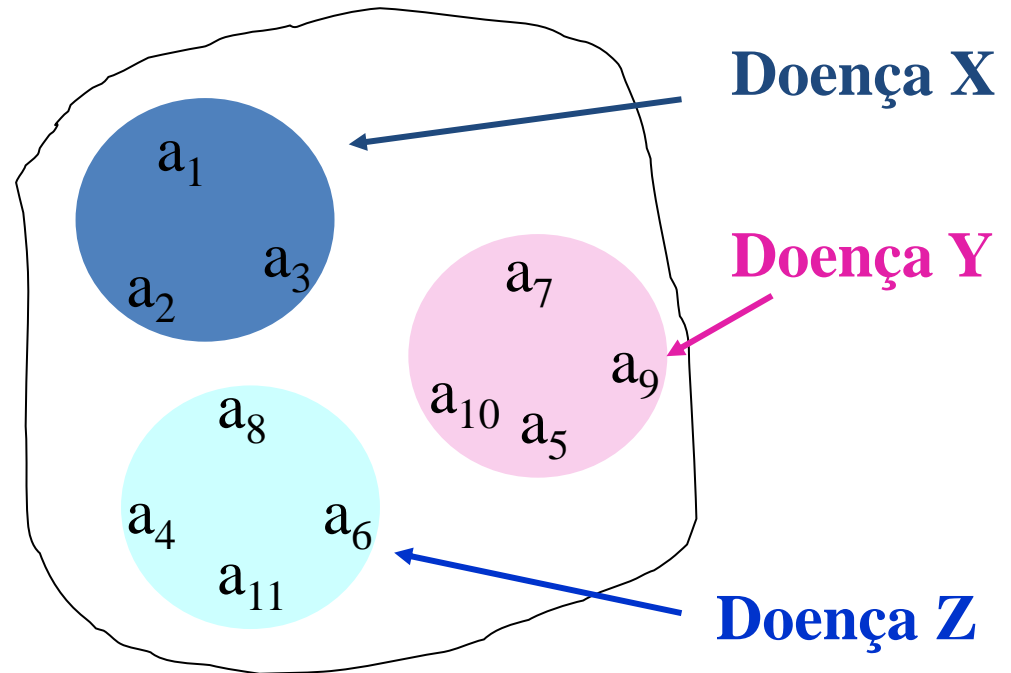


Notas avaliação

Agrupamento -Análise de Clusters

a_1	a	F	1	0	1	1
a_2	b	M	0	0	1	1
•	c	F	1	1	1	0
•	d	F	1	0	0	0
•	e	M	1	1	0	1

Nome Sexo Sintomas



Conceito = Doença

Número de Clusters = 3

ANÁLISE DE CLUSTER

ETAPAS

1. Análise das variáveis e dos objetos a serem agrupados (seleção de variáveis, identificação de *outliers* e padronização)
2. Seleção da medida de distância ou semelhança entre cada par de objetos
3. Seleção do algoritmo de agrupamento: método hierárquico e método não hierárquico
4. Escolha da quantidade de agrupamentos formados
5. Interpretação e validação dos agrupamentos

Etapa 1 - Análise das variáveis e dos objetos

Seleção de variáveis, identificação de *outliers*

- Cabe ao pesquisador selecionar as variáveis relevantes
- A técnica é muito sensível a *outliers*
 - Deve-se localizar os outliers de cada variável
 - Cabe analisar se devem ou não ser retirados
- É comum que os *outliers* formem grupos isolados

Etapa 1 - Análise das variáveis e dos objetos

Padronização das variáveis

- Medidas e/ou escalas diferentes distorcem a estrutura do agrupamento
- Padronização resolve problema de diferentes escalas ou magnitudes das variáveis
- Padronização faz com que seja atribuído o mesmo peso para cada variável

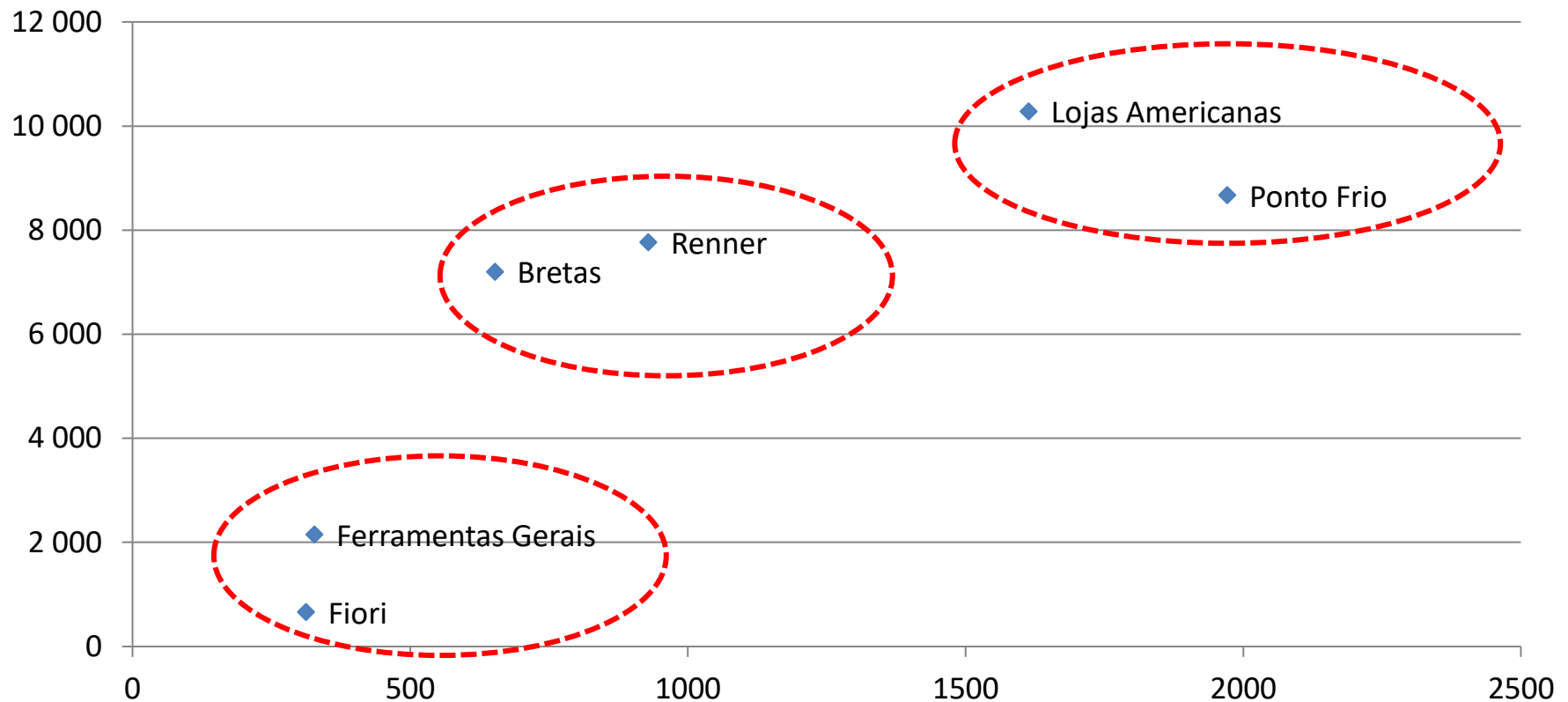
Etapa 1 - Análise das variáveis e dos objetos

Empresas	Vendas (US\$ milhões)	Número empregados
Ferramentas Gerais	327,5	2.150
Fiori	312,2	661
Bretas Supermercados	652,6	7.200
Renner	929	7.764
Lojas Americanas	1.613,5	10.281
Ponto Frio	1.971	8.672

Exemplo: Seria possível separar a amostra de empresas em grupos similares em termos de porte, representado pelas variáveis faturamento e número de empregados?

Etapa 1 - Análise das variáveis e dos objetos

- Exemplo



Etapa 1 - Análise das variáveis e dos objetos

Tipos de padronização

1. z-score
2. Método range: -1 a +1
3. Método range: 0 a 1
4. Método da máxima amplitude
5. Método da média =1
6. Método do desvio-padrão =1

Etapa 1 - Análise das variáveis e dos objetos

1. *Z-score*

É a forma mais utilizada de padronização, com média zero e desvio padrão 1

$$Z = \frac{x - \text{média}}{\text{desvio padrão}}$$

Etapa 1 - Análise das variáveis e dos objetos

Padronização das variáveis

2. Método range: -1 a +1

$$\frac{x}{\text{amplitude}}$$

3. Método range: 0 a +1

$$\frac{x - \text{mínimo}}{\text{amplitude}}$$

4. Método de máxima amplitude

$$\frac{x}{\text{máximo}}$$

Etapa 1 - Análise das variáveis e dos objetos

Padronização de Variáveis

5. Método da média = 1 $\frac{x}{\text{média}}$

6. Método de desvio padrão = 1 $\frac{x}{\text{desvio padrão}}$

Etapa 1 - Análise das variáveis e dos objetos

Padronização dos Dados com z-score

Empresas	Vendas (US\$ milhões)	Número empregados
Ferramentas Gerais	327,5	2.150
Fiori	312,2	661
Bretas Supermercados	652,6	7.200
Renner	929	7.764
Lojas Americanas	1.613,5	10.281
Ponto Frio	1.971	8.672
Média	555,3	6121,3
Desvio-padrão	294,5	3827,7

Etapa 1 - Análise das variáveis e dos objetos

Padronização dos Dados com z-score

Empresas	Vendas (US\$ milhões)	Vendas (z-score)
Ferramentas Gerais	327,5	$(327,5 - 555,3)/294,5 = -0,931$
Fiori	312,2	$(312,2 - 555,3)/294,5 = -0,953$
Bretas Supermercados	652,6	$(652,6 - 555,3)/294,5 = -0,458$
Renner	929	$(929 - 555,3)/294,5 = -0,056$
Lojas Americanas	1.613,5	$(1.613,5 - 555,3)/294,5 = 0,939$
Ponto Frio	1.971	$(1.971 - 555,3)/294,5 = 1,459$
Média	555,3	
Desvio-padrão	294,5	

Etapa 1 - Análise das variáveis e dos objetos

Padronização dos Dados com z-score

Empresas	Nº empregados	Nº empregados (z-score)
Ferramentas Gerais	2.150	$(2150-6121)/3827,3 = -1,038$
Fiori	661	$(661-6121)/3827,3 = -1,427$
Bretas Supermercados	7.200	$(7200-6121)/3827,3 = 0,282$
Renner	7.764	$(7764-6121)/3827,3 = 0,429$
Lojas Americanas	10.281	$(10281-6121)/3827,3 = 1,087$
Ponto Frio	8.672	$(8672-6121)/3827,3 = 0,666$
Média	6121,3	
Desvio-padrão	3827,7	

Etapa 1 - Análise das variáveis e dos objetos

Padronização dos Dados com z-score

Empresas	Vendas (US\$ milhões)	Número empregados	Vendas(z-score)	Nº empregados (z-score)
Ferramentas Gerais	327,5	2.150	-0,931	-1,038
Fiori	312,2	661	-0,953	-1,427
Bretas Supermercados	652,6	7.200	-0,458	0,282
Renner	929	7.764	-0,056	0,429
Lojas Americanas	1.613,5	10.281	0,939	1,087
Ponto Frio	1.971	8.672	1,459	0,666
Média	555,3	6121,3		
Desvio-padrão	294,5	3827,7		

ANÁLISE DE CLUSTER

Etapas

1. Análise das variáveis e dos objetos a serem agrupados (seleção de variáveis, identificação de *outliers* e padronização)
2. Seleção da medida de distância ou semelhança entre cada par de objetos
3. Seleção do algoritmo de agrupamento: método hierárquico e método não hierárquico
4. Escolha da quantidade de agrupamentos formados
5. Interpretação e validação dos agrupamentos

Etapa 2 - Seleção da medida de distância (ou de semelhança) entre cada par de objetos

- As observações são agrupadas segundo algum tipo de métrica de distância.
- Observações com menor distância entre si são mais semelhantes, logo são aglomerados em um mesmo conglomerado.
- Objetos mais distantes participam de conglomerados distintos.

Etapa 2 - Seleção da medida de distância (ou de semelhança) entre cada par de objetos

- São classificadas em 3 tipos:
 - Medidas de distância
 - Medidas correlacionais
 - Medidas de associação
- Escolha da medida depende do tipo de variável (qualitativa ou quantitativa) e da escala de medida

Etapa 2 - Seleção da medida de distância (ou de semelhança) entre cada par de objetos

Medidas de distância:

- Distância euclidiana – mais utilizada
- Distância quadrática euclidiana
- Distância de Minkovski
- Distância absoluta, bloco ou Manhattan
- Distância de Mahalanobis
- Distância de Chebychev

Etapa 2 - Seleção da medida de distância (ou de semelhança) entre cada par de objetos

Medidas de Distância

- Distância Euclidiana: a distância entre duas observações (i e j) correspondente à raiz quadrada da soma dos quadrados das diferenças entre os pares de observações (i e j) para todas as p variáveis

$$d_{ij} = \sqrt{\sum_{k=1}^p (x_{ik} - x_{jk})^2}$$

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{ip} - x_{jp})^2}$$

- Em que x_{ik} é o valor da variável k referente à observação i e x_{jk} representa a variável k para a observação j
- Quanto menor a distância, mais similares são as observações

Etapa 2 - Seleção da medida de distância (ou de semelhança) entre cada par de objetos

Medidas de Distância

- Distância Quadrática Euclidiana: a distância entre duas observações (i e j) correspondente à soma dos quadrados das diferenças entre os pares de observações (i e j) para todas as p variáveis

$$d_{ij}^2 = \sum_{k=1}^p (x_{ik} - x_{jk})^2$$

- Mais comum
- Quanto menor a distância, mais similares são as observações

Etapa 2 - Seleção da medida de distância (ou de semelhança) entre cada par de objetos

Medidas de Distância

- Distância de Minkowski: a distância euclidiana é um caso particular de uma distância mais geral, chamada de Minkowski

$$d_{ij} = \left(\sum_{k=1}^p (|x_{ik} - x_{jk}|)^n \right)^{1/n}$$

- Se aplicarmos $n = 2$, chegamos a distância euclidiana
- Para $n = 1$ temos a Distância City-Block, ou *Manhattan Distance*

Etapa 2 - Seleção da medida de distância (ou de semelhança) entre cada par de objetos

Medidas de Distância

- Distância Absoluta, Bloco, City-Block ou Manhattan:
representa a soma das diferenças absolutas entre os valores das p variáveis para os dois casos

$$d_{ij} = \sum_{k=1}^p |x_{ik} - x_{jk}|$$

Etapa 2 - Seleção da medida de distância (ou de semelhança) entre cada par de objetos

Medidas de Distância

- Distância Mahalanobis: é a distância estatística entre dois indivíduos i e j , considerando a matriz de covariância para o cálculo das distâncias

$$d_{ij} = \sqrt{(x_i - x_j)' S^{-1} (x_i - x_j)}$$

- Em que S é a estimativa amostral da matriz de variância-covariância Σ dentro dos agrupamentos

Etapa 2 - Seleção da medida de distância (ou de semelhança) entre cada par de objetos

Medidas de Distância

- Distância Chebychev: diferença absoluta máxima entre todas as p variáveis entre duas observações

$$d_{ij} = \max |x_{ik} - x_{jk}|$$

- Em que S é a estimativa amostral da matriz de variância-covariância Σ dentro dos agrupamentos

Etapa 2 - Seleção da medida de distância (ou de semelhança) entre cada par de objetos

Medidas Correlacionadas

- Representam similaridade pela correspondência de padrões ao longo das características (X variáveis)
- Correlação de Pearson é a mais utilizada

$$r_{ij} = \frac{\sum_{k=1}^p (x_{1k} - \bar{x}_i)(x_{1j} - \bar{x}_j)}{\sqrt{\sum_{k=1}^p (x_{1k} - \bar{x}_i)^2 \sum_{k=1}^p (x_{1j} - \bar{x}_j)^2}}$$

Etapa 2 - Seleção da medida de distância (ou de semelhança) entre cada par de objetos

Medidas de Associação

- Utilizado com variáveis binárias
- Tabela de Contingência

		Indivíduo <i>j</i>		
		1	0	Total
Indivíduo <i>i</i>	1	a	b	a+b
	0	c	d	c+d
Total		a+c	b+d	$p = a+b+c+d$

Etapa 2 - Seleção da medida de distância (ou de semelhança) entre cada par de objetos

Medidas de Associação

- Medida de Semelhança (S_{ij}) $S_{ij} = \frac{a}{a+b+c}$
- Medida de Distância (d_{ij}) $d_{ij} = \frac{b+c}{a+b+c}$

Exemplo com 6 empresas: Seria possível separar a amostra de empresas em grupos similares em termos de porte, representado pelas variáveis faturamento e número de empregados?

Empresas	Vendas (US\$ milhões)	Número de empregados
Ferramentas Gerais (1)	327,5	2150
Fiori (2)	312,2	661
Bretas Supermercados (3)	652,6	7200
Renner (4)	929,0	7764
Lojas Americanas (5)	1613,5	10281
Ponto Frio (6)	1971,0	8672

Exemplo com 6 empresas

Padronização

Empresas	x= Vendas (z-score)	y =Nº de empregados (z-score)
Ferramentas Gerais (1)	-0,931	-1,038
Fiori (2)	-0,953	-1,427
Bretas Supermercados (3)	-0,458	0,282
Renner (4)	-0,056	0,429
Lojas Americanas (5)	0,939	1,087
Ponto Frio (6)	1,459	0,666

Exemplo com 6 empresas

Cálculo das distâncias

Empresas	x= Vendas (z-score)	y =Nº de empregados (z-score)
Ferramentas Gerais (1)	-0,931	-1,038
Fiori (2)	-0,953	-1,427
Bretas Supermercados (3)	-0,458	0,282
Renner (4)	-0,056	0,429
Lojas Americanas (5)	0,939	1,087
Ponto Frio (6)	1,459	0,666

Distância quadrática euclidiana = $((x_1 - x_2)^2 + (y_1 - y_2)^2)$

Distância (empresa₁-empresa₂) = $(-0,931 - (-0,953))^2 + (-1,038 - (-1,427))^2 = \mathbf{0,152}$

Exemplo com 6 empresas

Matriz de similaridade pela Distância Quadrática Euclidiana

	Ferramentas Gerais (1)	Fiori (2)	Bretas Supermercado s (3)	Renner (4)	Lojas Americana s (5)	Ponto Frio (6)
Ferramentas Gerais (1)	0,000					
Fiori (2)	0,152	0,000				
Bretas Super- mercados (3)	1,964	3,163	0,000			
Renner (4)	2,916	4,248	0,183	0,000		
Lojas Americanas (5)	8,010	9,898	2,601	1,423	0,000	
Ponto Frio (6)	8,616	10,200	3,824	2,353	0,447	0,000

ANÁLISE DE CLUSTER

Etapas

1. Análise das variáveis e dos objetos a serem agrupados (seleção de variáveis, identificação de *outliers* e padronização)
2. Seleção da medida de distância ou semelhança entre cada par de objetos
3. Seleção do algoritmo de agrupamento: método hierárquico e método não hierárquico
4. Escolha da quantidade de agrupamentos formados
5. Interpretação e validação dos agrupamentos

Etapa 3 - Seleção do algoritmo de agrupamento: método hierárquico e método não hierárquico

- Envolve a escolha do **algoritmo de agrupamento** e a decisão quanto ao número de grupos
- **Algoritmo de agrupamento:**
 - procedimento para colocar objetos similares dentro de grupos
 - hierárquicos e os não-hierárquicos
- Todo **algoritmo** visa maximizar as diferenças entre os grupos em confronto com a variação dentro dos grupos (*between-cluster* x *within-cluster*).

Etapa 3 - Seleção do algoritmo de agrupamento: método hierárquico e método não hierárquico

- **Hierárquicos:** identificam agrupamentos e o provável o n° g de grupos, por:
 - a) Uma série de fusões sucessivas (técnicas *aglomerativas*);
 - b) Ou uma série de sucessivas divisões (técnicas *divisas*).

Os resultados de ambos, aglomerativos e divisivos, são observados no *dendrograma*, que ilustra as fusões ou divisões feitas em níveis sucessivos.

- **Não hierárquicos:** o n° g de grupos é pré-especificado.

Métodos de agrupamento

	Hierárquicos		Não-hierárquicos
Processo de aglomeração		Single linkage	K-means
		Complete linkage	Sequential threshold
	Aglomerativos	Average linkage	Parallel threshold
		Ward	optimization
		Centroid method	Selecting seed points
	Divisivos		

Métodos hierárquicos

Métodos hierárquicos

Aglomerativo (método mais comum)

- No início cada objeto forma um *cluster* que sucessivamente sofre uma série de fusões com outros *clusters* até que no final todos os objetos estejam em um único agrupamento.
- Um *cluster* formado em uma dada interação corresponde a união de *clusters* formados em passos anteriores.

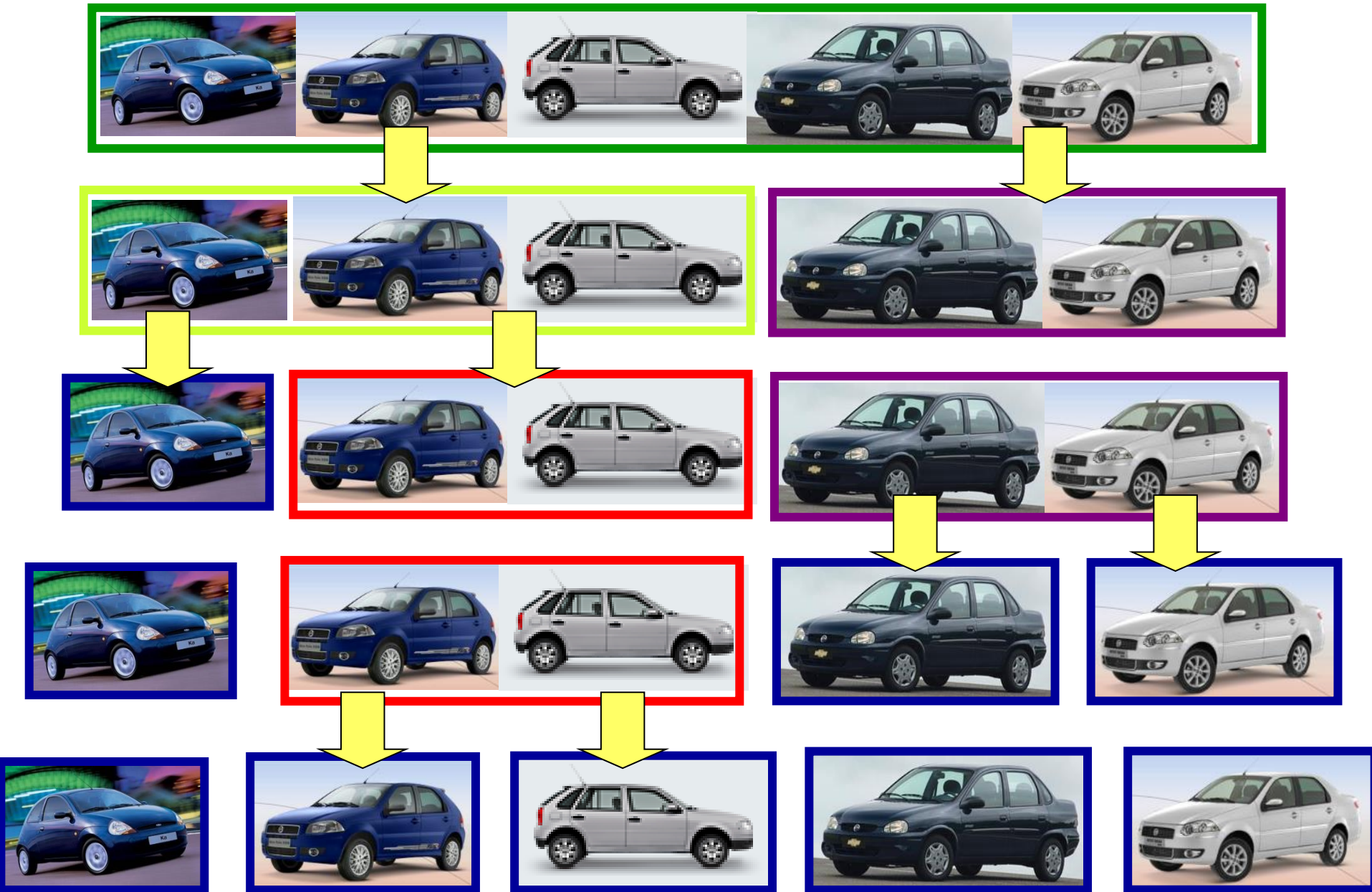
Divisivo

- No início há apenas um *cluster* formado pelo conjunto de objetos que é dividido sucessivamente até que no final cada *cluster* contenha apenas um objeto.
- *Clusters* formados em uma dada interação correspondem a fragmentação de um *cluster* formado no passo anterior.

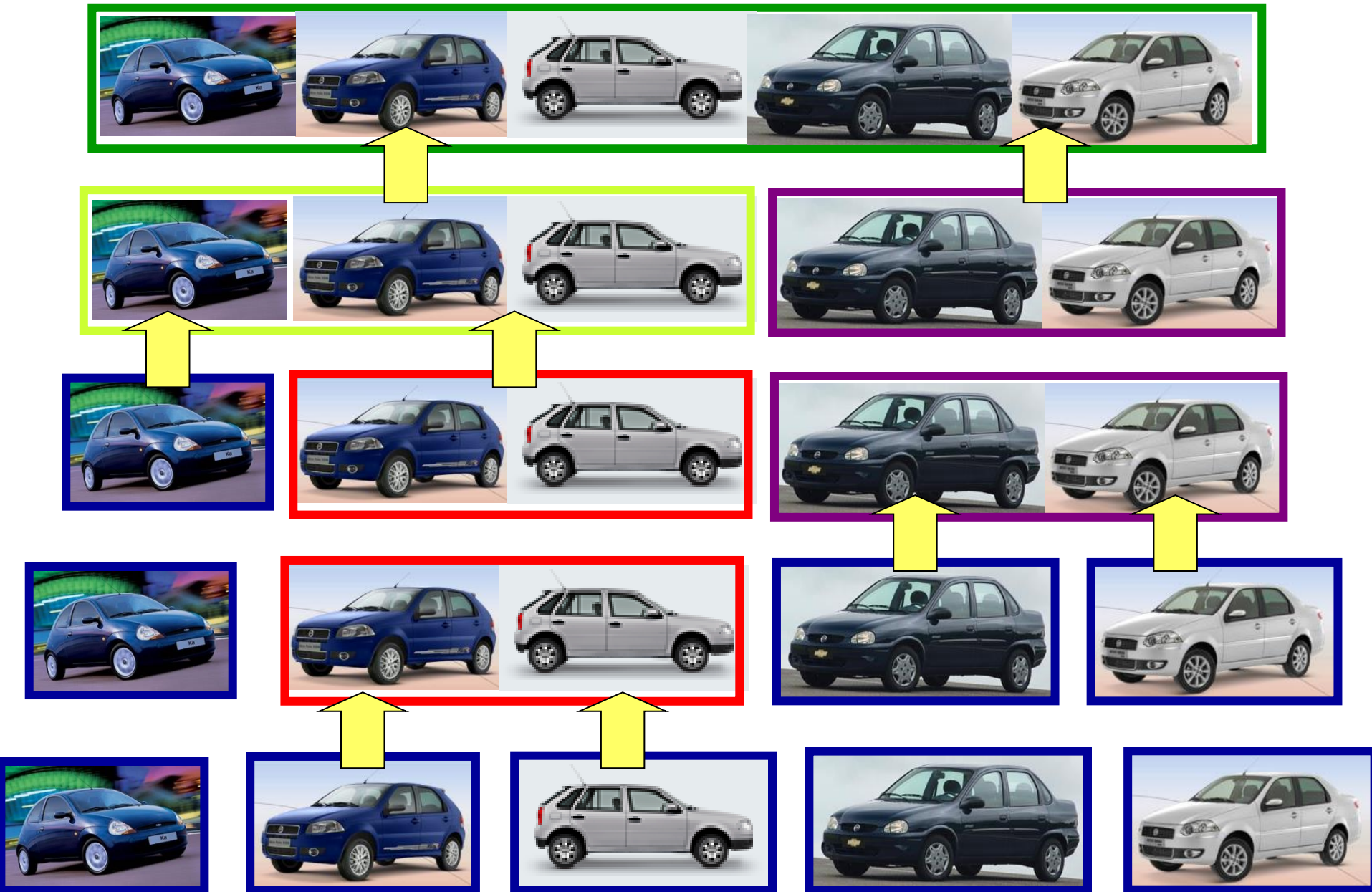
Exemplo:

Considere 5 veículos caracterizados pelos seguintes atributos: nº de portas, autonomia (km/l), preço, cilindrada, conforto, tamanho.

Métodos hierárquicos divisivos



Métodos hierárquicos aglomerativos



Procedimentos hierárquicos de agrupamento

- envolvem a construção de uma hierarquia semelhante a uma árvore.
- São de dois tipos: aglomerativos e divisivos.
- Algoritmos mais populares:
 - *single linkage*
 - *complete linkage*
 - *average linkage*
 - *Ward's method*
 - *centroid method*

Procedimentos hierárquicos de agrupamento

Single linkage: baseado na distância mínima entre dois objetos. É também chamado de *nearest neighbor*

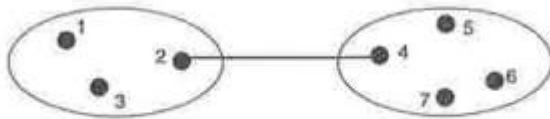


Figura 6.2: Esquematização do método da Ligação Individual.

Complete linkage: é baseado na distância máxima, razão pela qual é conhecido como a abordagem do vizinho mais longe (*furthest neighbor*)

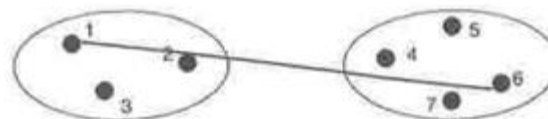


Figura 6.3: Esquematização do método da Ligação Completa.

Procedimentos hierárquicos de agrupamento

Average linkage: ou ligação média, onde o critério é a distância de todos os indivíduos de um grupo em relação a todos de outro. Tende a produzir grupos com aproximadamente a mesma variância

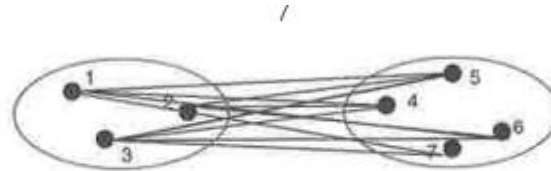


Figura 6.4: Esquemática do método da Ligação Média.

Ward's method: minimiza a soma dos quadrados entre dois grupos em relação a todas as variáveis. Tende a produzir grupos com mesmo número de observações

Procedimentos hierárquicos de agrupamento

Centroid method

- a distância entre os grupos é a distância entre seus centróides, que são os valores médios das observações em relação às variáveis.
- Cada vez que indivíduos são agrupados, um novo centróide é calculado.
- Tanto este método quanto o de *Ward* exigem a distância euclidiana.

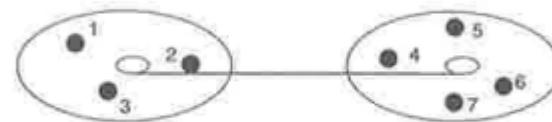


Figura 6.5: Esquematização do método Centróide.

Métodos não-hierárquicos

Procedimentos não-hierárquicos de agrupamento

- número inicial de clusters é definido pelo pesquisador
- Objetivo: partição de n elementos em k grupos de modo que a partição atenda 2 requisitos:
 - coesão interna (ou semelhança interna)
 - isolamento (ou separação)
- não requerem cálculo e armazenamento de uma nova matriz de distância a cada processo
- reduzem tempo computacional
- produzem apenas uma solução ao contrário do hierárquico que fornece uma série de soluções correspondentes a diferentes número de agrupamentos

Procedimentos não-hierárquicos de agrupamento

- K-means
 - o mais popular
 - objetivo: minimizar a variância interna dos grupos e maximizar a variância entre os grupos
 - pode-se fornecer informações sobre os centróides ou sementes iniciais
 - se os centróides forem desconhecidos todas as observações são consideradas centróides
 - verificação é feita pela ANOVA (neste caso espera-se rejeitar H_0 , ou seja, as médias não são iguais)

Procedimentos não-hierárquicos de agrupamento

1. escolhe-se o número k de conglomerados
2. assume-se inicialmente um centro de aglomeração (centróide) para cada grupo
3. calcula-se a distância euclidiana de cada sujeito aos centróides e coloca-se cada sujeito no cluster com centro de aglomeração mais próximo
4. quando todos os sujeitos estiverem alocados recalculam-se os novos centróides para cada grupo
5. repetem-se os passos 3 e 4 até os valores dos centróides dos grupos não mais variarem

Combinação dos métodos de agrupamento hierárquico e não-hierárquico

- inicialmente, através de um método hierárquico, estabelece-se o número de grupos, traça-se o perfil dos núcleos centrais e identificam-se possíveis *outliers*;
- após a eliminação de eventuais *outliers*, aplica-se um método não-hierárquico (k-means), tendo como grupos sementes os núcleos centrais definidos através do método hierárquico.

ANÁLISE DE CLUSTER

Etapas

1. Análise das variáveis e dos objetos a serem agrupados (seleção de variáveis, identificação de *outliers* e padronização)
2. Seleção da medida de distância ou semelhança entre cada par de objetos
3. Seleção do algoritmo de agrupamento: método hierárquico e método não hierárquico
4. Escolha da quantidade de agrupamentos formados
5. Interpretação e validação dos agrupamentos

Etapa 4 - Escolha da quantidade de agrupamentos formados

- Quantos grupos devem ser formados?
 - Não existe um critério categórico
 - Uma regra de parada (*stopping rule*) simples é examinar a distância entre os grupos a cada passo sucessivo;
 - Outra regra seria adaptar um teste estatístico de significância;
 - Além disso, o pesquisador deve confrontar com o referencial teórico, que pode sugerir um número natural de grupos;
 - Deve-se, ao final, buscar a melhor solução dentre as possíveis.

Etapa 4 - Escolha da quantidade de agrupamentos formados

- A análise *cluster* deve ser estruturada novamente?
 - Analisar se existe um disparate acentuado entre o tamanho dos grupos, ou se existem grupos com uma ou duas observações (possíveis *outliers*);
 - Comparar a solução final com as expectativas do pesquisador;

ANÁLISE DE CLUSTER

Etapas

1. Análise das variáveis e dos objetos a serem agrupados (seleção de variáveis, identificação de *outliers* e padronização)
2. Seleção da medida de distância ou semelhança entre cada par de objetos
3. Seleção do algoritmo de agrupamento: método hierárquico e método não hierárquico
4. Escolha da quantidade de agrupamentos formados
5. Interpretação e validação dos agrupamentos

Etapa 5 - Interpretação e validação dos agrupamentos

Alguns procedimentos de validação da solução:

- 1) dividir a amostra em dois grupos;
- 2) usar outras variáveis conhecidas por discriminar entre os grupos, ou refazer a análise excluindo algumas variáveis;
- 3) refazer a análise utilizando outros métodos de agrupamento e outras medidas de similaridade.

Etapas 5 - Interpretação e validação dos agrupamentos

- consiste na descrição das características de cada grupo para explicar como elas podem diferir em dimensões relevantes.
- Utilizam-se dados não previamente incluídos no procedimento de agrupamento (demográficos, psicográficos etc.).
- O enfoque é na descrição, não do que determinou diretamente os grupos, mas das características dos grupos depois de que eles foram identificados.

Exemplos

Exemplo 1 - hierárquico

Exemplo 1: Seria possível separar a amostra de empresas em grupos similares em termos de porte, representado pelas variáveis faturamento e número de empregados?

Empresas	Vendas (US\$ milhões)	Número empregados
Ferramentas Gerais	327,5	2.150
Fiori	312,2	661
Bretas Supermercados	652,6	7.200
Renner	929	7.764
Lojas Americanas	1.613,5	10.281
Ponto Frio	1.971	8.672

Etapa 1 - Análise das variáveis e dos objetos

Empresas	x= Vendas (z-score)	y =Nº de empregados (z-score)
Ferramentas Gerais (1)	-0,931	-1,038
Fiori (2)	-0,953	-1,427
Bretas Supermercados (3)	-0,458	0,282
Renner (4)	-0,056	0,429
Lojas Americanas (5)	0,939	1,087
Ponto Frio (6)	1,459	0,666

Etapas 1 - Análise das variáveis e dos objetos

Padronização dos dados com z-score

Etapa 2 - cálculo das distâncias

Empresas	x= Vendas (z-score)	y =Nº de empregados (z-score)
Ferramentas Gerais (1)	-0,931	-1,038
Fiori (2)	-0,953	-1,427
Bretas Supermercados (3)	-0,458	0,282
Renner (4)	-0,056	0,429
Lojas Americanas (5)	0,939	1,087
Ponto Frio (6)	1,459	0,666

Distância quadrática euclidiana = $d(\text{empresa}_1 - \text{empresa}_2)^2 = ((x_1 - x_2)^2 + (y_1 - y_2)^2)^2$

Distância $(\text{empresa}_1 - \text{empresa}_2)^2 = (-0,931 - (-0,953))^2 + (-1,038 - (-1,427))^2 = 0,152$

Etapa 2 - cálculo das distâncias

matriz de similaridade pela Distância Quadrática Euclidiana

	Ferramentas Gerais (1)	Fiori (2)	Bretas Supermercados (3)	Renner (4)	Lojas Americanas (5)	Ponto Frio (6)
Ferramentas Gerais (1)	0,000					
Fiori (2)	0,152	0,000				
Bretas Super- mercados (3)	1,964	3,163	0,000			
Renner (4)	2,916	4,248	0,183	0,000		
Lojas Americanas (5)	8,010	9,898	2,601	1,423	0,000	
Ponto Frio (6)	8,616	10,200	3,824	2,353	0,447	0,000

Etapa 3 - agrupamento

	Ferrament as Gerais (1)	Fiori (2)	Bretas Supermercados (3)	Renner (4)	Lojas Americanas (5)	Ponto Frio (6)
Ferramentas Gerais (1)	0,000					
Fiori (2)	0,152	0,000				
Bretas Super- mercados (3)	1,964	3,163	0,000			
Renner (4)	2,916	4,248	0,183	0,000		
Lojas Americanas (5)	8,010	9,898	2,601	1,423	0,000	
Ponto Frio (6)	8,616	10,200	3,824	2,353	0,447	0,000

Menor distância = 0,152 ; logo reúno 1 com 2

Após formar o cluster 12, calcula-se a distancia entre o cluster 12 e as outras empresas:

$$D_{(12),3} = \min (d_{13}, d_{23}) = \min (1,964; 3,163) = 1,964$$

$$D_{(12)4} = \min (d_{14}, d_{24}) = \min (2,916; 4,248) = 2,916$$

$$D_{(12),5} = \min (d_{15}, d_{25}) = \min (8,010; 9,898) = 8,010$$

$$D_{(12)6} = \min (d_{16}, d_{26}) = \min (8,616; 10,200) = 8,616$$

Etapa 3 - agrupamento

	1+2	Bretas Supermercados (3)	Renner (4)	Lojas Americanas (5)	Ponto Frio (6)
1+2	0,000				
Bretas Super-mercados (3)	1,964	0,000			
Renner (4)	2,916	0,183	0,000		
Lojas Americanas (5)	8,010	2,601	1,423	0,000	
Ponto Frio (6)	8,616	3,824	2,353	0,447	0,000

Menor distância = 0,183 ; logo reúno 3 com 4

Após formar o cluster 34, calcula-se a distancia entre o cluster 34 e as outras empresas:

$$D_{(34),(12)} = \min (d_{3(12)}, d_{4(12)}) = \min (1,964; 2,916) = 1,964$$

$$D_{(34)5} = \min (d_{35}, d_{45}) = \min (2,601; 1,423) = 1,423$$

$$D_{(34),6} = \min (d_{36}, d_{46}) = \min (3,824; 2,353) = 2,353$$

Etapa 3 - agrupamento

single linkage – menor distância

	1+2	3+4	Lojas Americanas (5)	Ponto Frio (6)
1+2	0,000			
3+4	1,964	0,000		
Lojas Americanas (5)	8,010	1,423	0,000	
Ponto Frio (6)	8,616	2,353	0,447	0,000

Menor distância = 0,447 ; logo reúno 5 com 6

Após formar o cluster 56, calcula-se a distancia entre o cluster 56 e as outras empresas:

$$D_{(56),(12)} = \min (d_{5(12)}, d_{6(12)}) = \min (8,010; 8,616) = 8,010$$

$$D_{(56)(34)} = \min (d_{5(34)}, d_{6(34)}) = \min (1,423; 2,353) = 1,423$$

Etapa 3 - agrupamento

single linkage – menor distância

	1+2	3+4	5+6
1+2	0,000		
3+4	1,964	0,000	
5+6	8,010	1,423	0,000

Menor distância = 1,423 ; logo reúno 34 com 56

Após formar o cluster 3456, calcula-se a distancia entre o cluster 3456 e as outras empresas:

$$D_{(56),(12)} = \min (d_{5(12)}, d_{6(12)}) = \min (8,010; 8,616) = 8,010$$

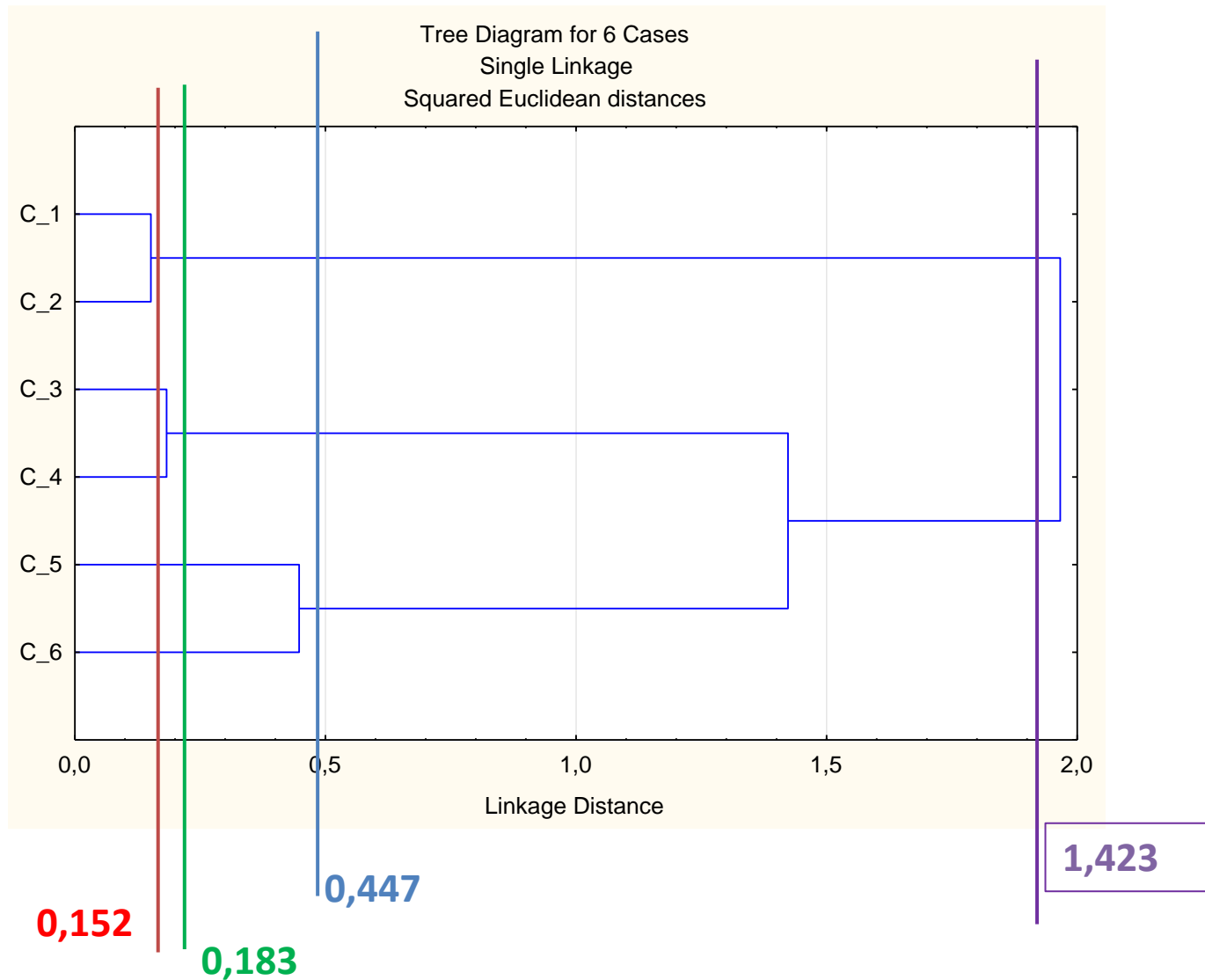
$$D_{(56)(34)} = \min (d_{5(34)}, d_{6(34)}) = \min (1,423; 2,353) = 1,423$$

Etapa 3 - agrupamento

single linkage – menor distância

	(1+2)	3+4+5+6
1+2	0,000	
3+4+5+6	1,964	0,000

dendrograma



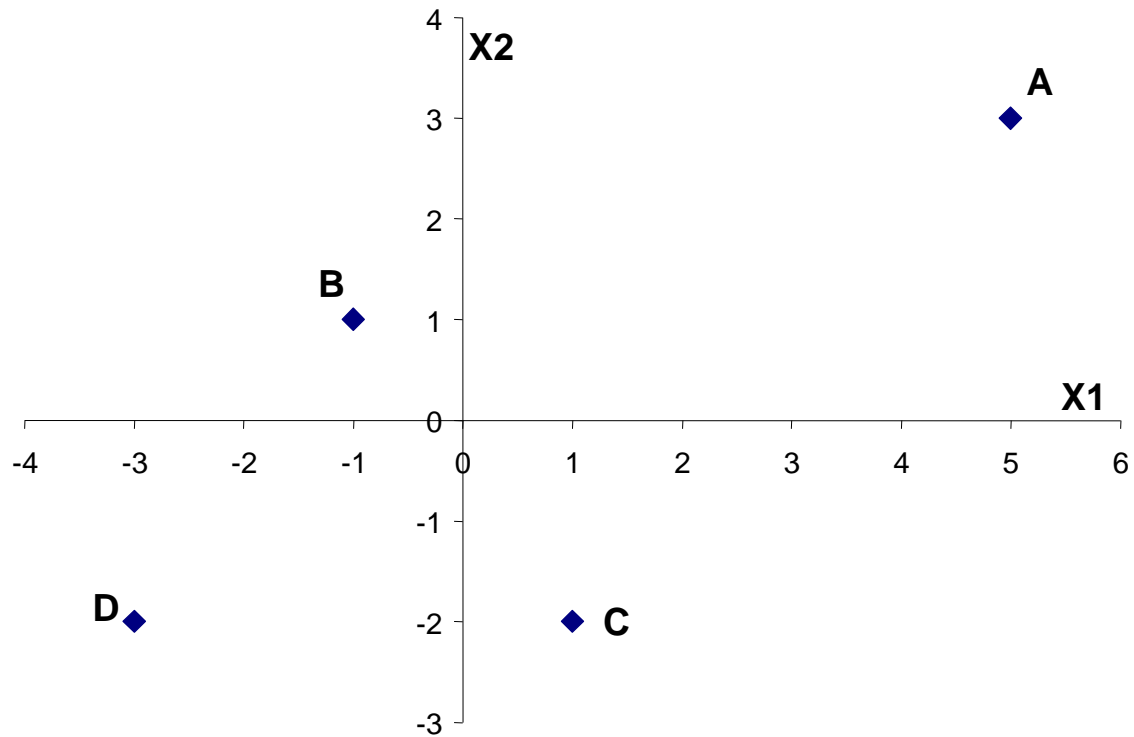
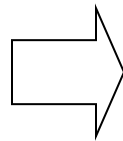
Exemplo 2 – não-hierárquico

K-Means

Exemplo

Dado o conjunto de 4 objetos ($n=4$), use o algoritmo k-Means para identificar 2 clusters ($k=2$)

Objetos	Coordenadas	
	X1	X2
A	5	3
B	-1	1
C	1	-2
D	-3	-2



K-Means

Exemplo

Centróides iniciais (seleção aleatória)

<i>Centróide</i>	<i>Coordenadas dos centróides dos clusters</i>	
	<i>X1</i>	<i>X2</i>
<i>C1</i>	2	2
<i>C2</i>	-1	-2

Lista de objetos

<i>Objetos</i>	<i>Coordenadas</i>	
	<i>X1</i>	<i>X2</i>
<i>A</i>	5	3
<i>B</i>	-1	1
<i>C</i>	1	-2
<i>D</i>	-3	-2

K-Means

Exemplo

Centróides iniciais (seleção aleatória)

Centróide	Coordenadas dos centróides dos clusters	
	X1	X2
C1	2	2
C2	-1	-2


Lista de objetos

Objetos	Coordenadas	
	X1	X2
A	5	3
B	-1	1
C	1	-2
D	-3	-2

Alocação dos objetos aos clusters


$$||A - C1||^2 = (5-2)^2 + (3-2)^2 = \mathbf{10}$$

$$||A - C2||^2 = (5+1)^2 + (3+2)^2 = 61$$

A  Cluster com centróide C1


$$||B - C1||^2 = (-1-2)^2 + (1-2)^2 = 10$$

$$||B - C2||^2 = (-1+1)^2 + (1+2)^2 = \mathbf{9}$$

B  Cluster com centróide C2


$$||C - C1||^2 = (1-2)^2 + (-2-2)^2 = 5$$

$$||C - C2||^2 = (1+1)^2 + (-2+2)^2 = \mathbf{4}$$

C  Cluster com centróide C2

$$||D - C1||^2 = (-3-2)^2 + (-2-2)^2 = 39$$

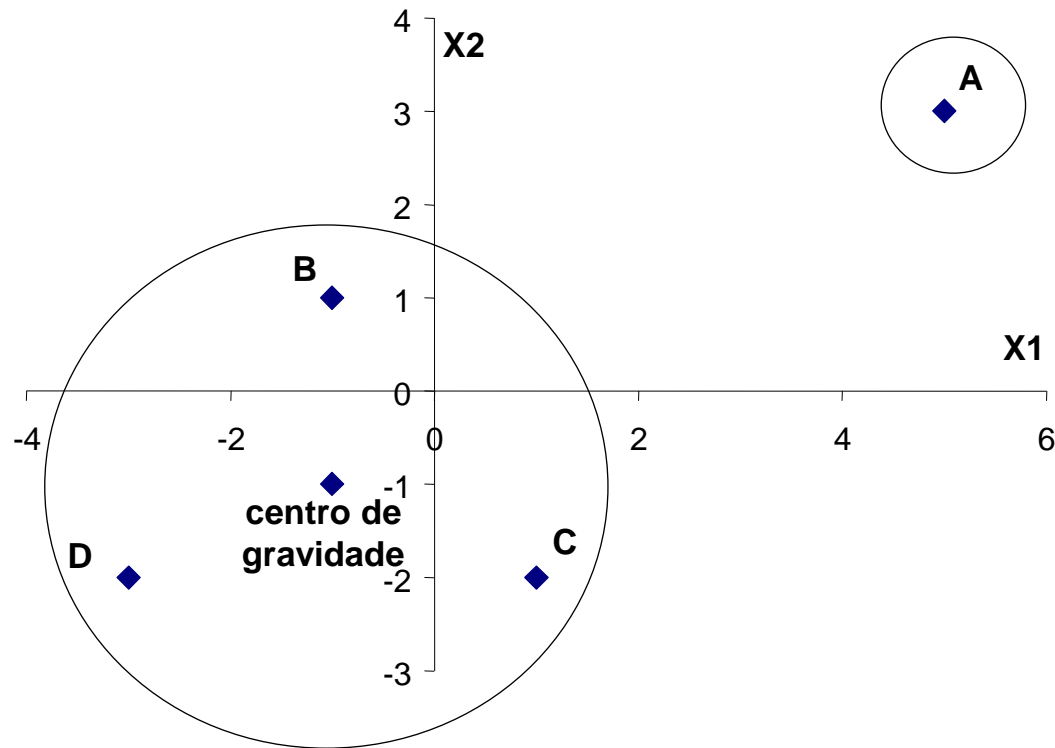
$$||D - C2||^2 = (-3+1)^2 + (-2+2)^2 = \mathbf{4}$$

D  Cluster com centróide C2

K-Means

Exemplo

Atualiza os
centróides



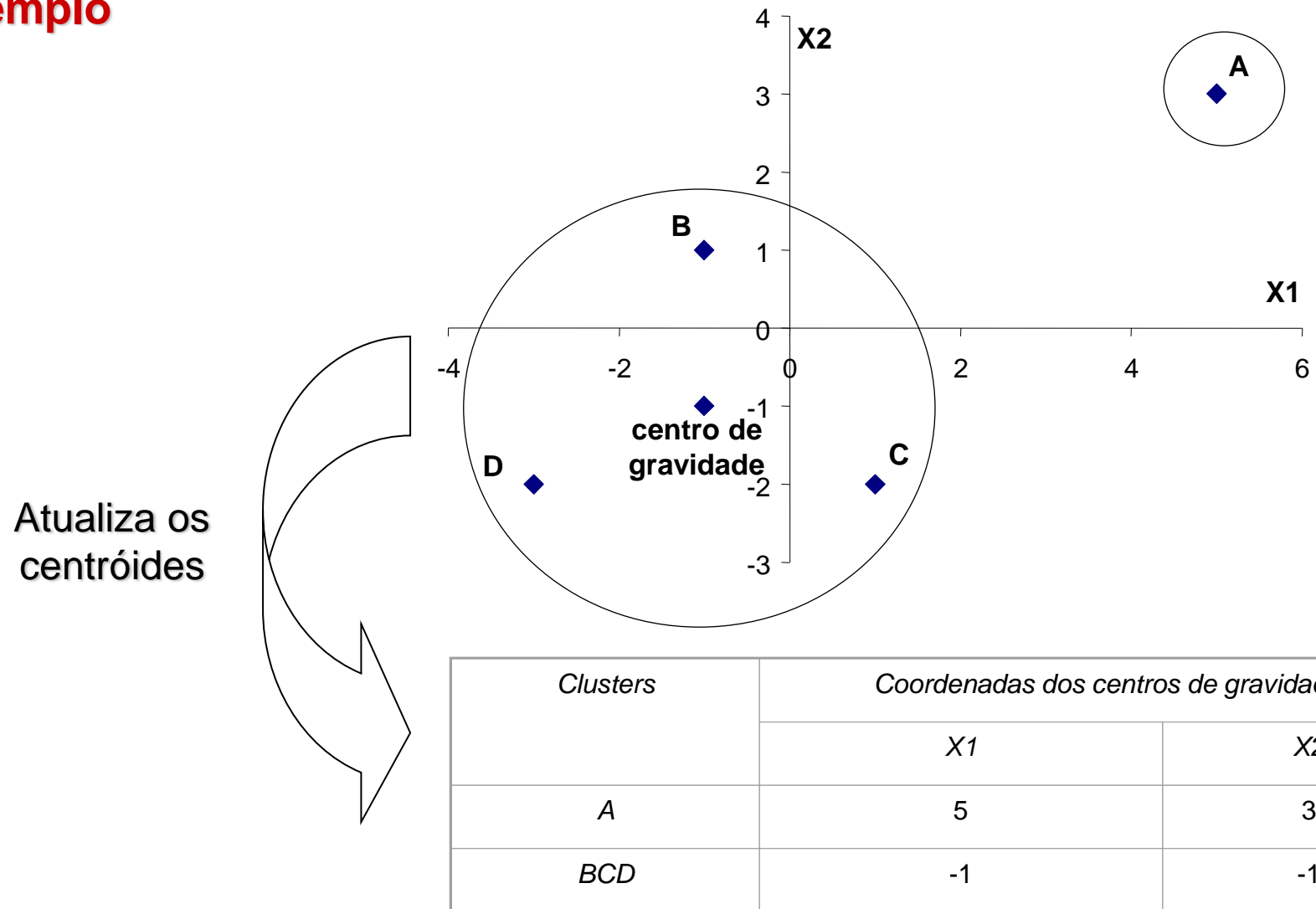
O centro de gravidade de um cluster é a média dos seus objetos:

Abcissa do centróide BCD = $(-1 + 1 - 3) / 3 = -1$

Ordenada do centróide BCD = $(1 - 2 - 2) / 3 = -1$

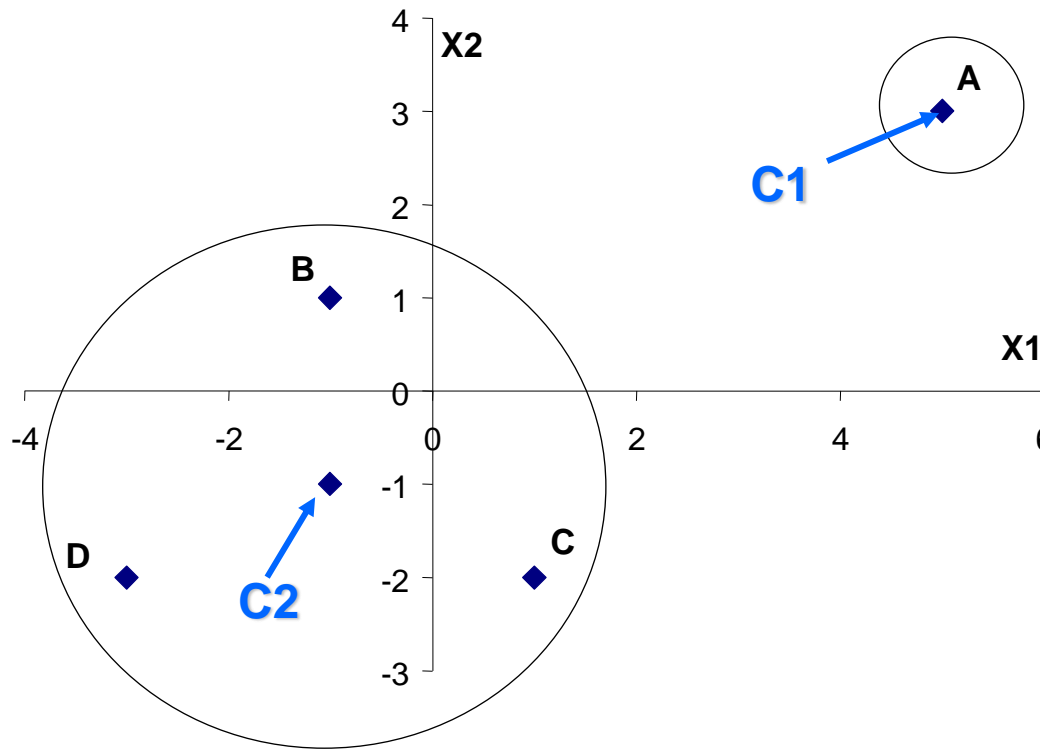
K-Means

Exemplo



K-Means

Exemplo



Não houve realocação de objetos, portanto, o algoritmo convergiu e dois clusters foram identificados: A e B,C,D

Alocação dos objetos aos clusters

$$||A - C1||^2 = (5-5)^2 + (3-3)^2 = \mathbf{0}$$

$$||A - C2||^2 = (5+1)^2 + (3+1)^2 = 61$$

A ➡ Cluster com centróide C1

$$||B - C1||^2 = (-1-5)^2 + (1-3)^2 = 40$$

$$||B - C2||^2 = (-1+1)^2 + (1+1)^2 = \mathbf{4}$$

B ➡ Cluster com centróide C2

$$||C - C1||^2 = (1-5)^2 + (-2-3)^2 = 39$$

$$||C - C2||^2 = (1+1)^2 + (-2+1)^2 = \mathbf{5}$$

C ➡ Cluster com centróide C2

$$||D - C1||^2 = (-3-5)^2 + (-2-2)^2 = 41$$

$$||D - C2||^2 = (-3+1)^2 + (-2+1)^2 = \mathbf{5}$$

D ➡ Cluster com centróide C2

Kmeans

<https://www.youtube.com/watch?v=WqMnQuC19Rg>

A partir do minuto 5

<https://www.youtube.com/watch?v=WqMnQuC19Rg>