



Escola Politécnica de Pernambuco
Especialização em Ciência de Dados e Analytics

Estatística Computacional

Aula 4.1 – Análise de Conglomerados e Fatorial – PARTE I

Prof. Dr. Rodrigo Lins Rodrigues

rodrigo.linsrodrigues@ufrpe.br

O que veremos nesta aula ?

- Introdução a análise de conglomerados;
- Análise das variáveis e objetos a serem agrupados;
- Medidas de similaridade ou distância;
- Método de agrupamento hierárquico;
- Método de agrupamento não-hierárquico;
- Quantidades de cluster;
- Validação de agrupamentos;
- Entendimento de scripts em R.



Ao final da aula será capaz...

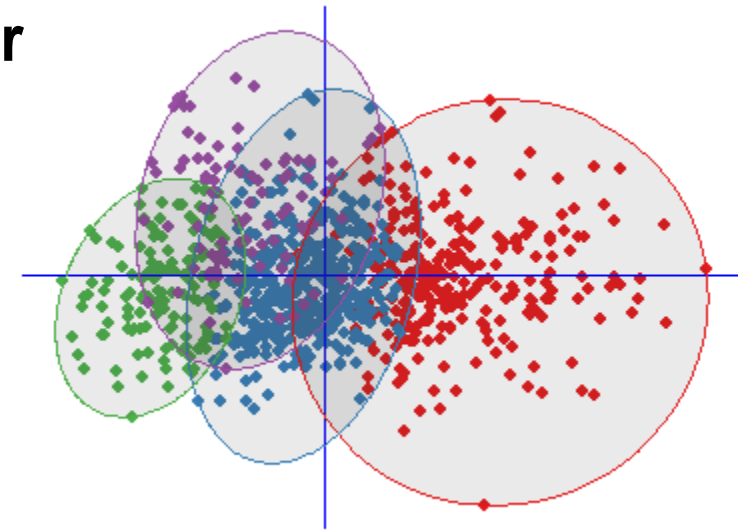
- **Identificar situações** que sejam apropriadas à utilização de análise de conglomerados;
- Explicar as **principais medidas** de similaridade ou distância;
- Entender as principais diferenças entre os procedimentos **hierárquicos e não-hierárquicos**;
- Interpretar os resultados apresentados pela técnica;
- Determinar o número de grupos mais adequado para o problema;
- Praticar computacionalmente.





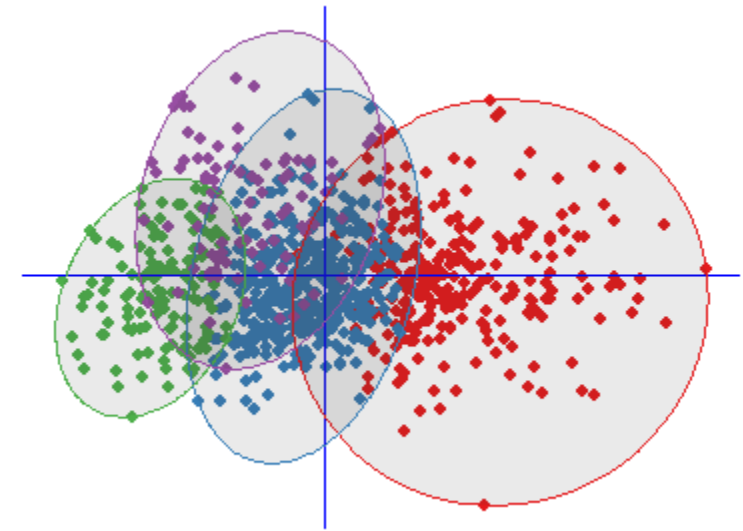
Análise de Conglomerados

- Conhecida na área de Data Mining como **Análise de Cluster**;
- É uma técnica de **interdependência** que busca **agrupar os elementos** conforme sua estrutura;
- Agrupar elementos em grupos **homogêneos internamente**;
- Busca que os grupos sejam **heterogêneos entre si**;



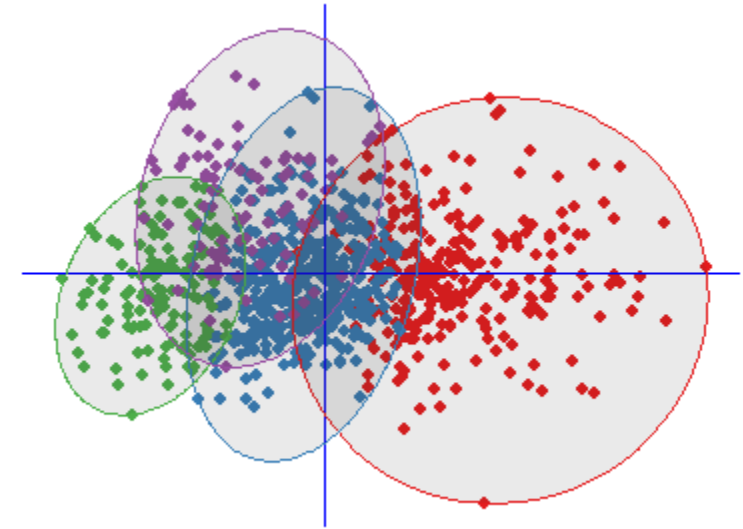
Análise de Conglomerados

- Busca alocar as observações mais **similares no mesmo grupo**;
- Com base em suas características (*variáveis*);
- Buscando assim maximizar a homogeneidade dos objetos em cada grupo;
- Esta técnica teve suas origens na psicologia, em busca de agrupar comportamentos de pacientes.



Análise de Conglomerados

- Cabe ressaltar que a análise de conglomerados é uma **técnica descritiva** ou **exploratória**;
- Não possui o **rigor da inferência estatística**;
- É dividida em duas categorias:
 - ✓ Hierárquica (abordagem estatística);
 - ✓ Não-hierárquica (abordagem de machine learning).



Análise de Conglomerados

Em quais situações poderíamos utilizar **análise de conglomerados** ?



Análise de Conglomerados

- **Exemplo 1:**

- ✓ Um gestor está interessado em identificar grupos de investimentos de acordo com o perfil comportamental;
- ✓ O objetivo é lançar pacotes de serviços de acordo com o perfil de cada grupo (mais conservador ou grupo de risco).



Análise de Conglomerados

- **Exemplo 2:**
 - ✓ Um diretor de marketing busca **identificar segmentos** homogêneos de consumidores, a fim de estabelecer programas de marketing específicos para cada público;
 - ✓ Propaganda focada nos **hábitos de consumo** de cada grupo.



Análise de Conglomerados

- **Exemplo 3:**

- ✓ Um educador visa identificar grupos de alunos mais **propensos à evasão** escolar;
- ✓ A partir da identificação destes grupos, gestores podem tomar **decisões pedagógicas**.



Análise de Conglomerados

- Exemplo 4:

- ✓ Uma seguradora busca identificar grupos de clientes de **menor risco** para lançar mão de produtos e promoções;
- ✓ Redução em franquias;



Análise de Conglomerados

- **Etapas de uma análise de conglomerados**

1. Análise das variáveis e dos objetos a serem agrupados:

- ✓ Seleção de variáveis;
- ✓ Identificação de *outliers*;
- ✓ Padronização de variáveis;

2. Seleção da medida de distância entre grupos;



Análise de Conglomerados

- **Etapas de uma análise de conglomerados**


3. Seleção do algoritmo de agrupamento:

- ✓ Hierárquico;
- ✓ Não-hierárquico;

4. Escolha da quantidade de grupos formados;

5. Interpretação e validação dos grupos formados.



A young man with dark hair, wearing a red and white striped shirt, is holding a magnifying glass over a chalkboard. The chalkboard is filled with various mathematical diagrams, including bar charts, line graphs, and geometric shapes like triangles and circles. The man has a focused expression, looking through the magnifying glass. A blue semi-transparent banner is overlaid on the bottom half of the image, containing the title in yellow text.

Variáveis e Objetos a serem agrupados

Variáveis e Objetos

- A **seleção de variáveis**, para análise de conglomerados, deve ser feita com **extremo cuidado**;
- Os grupos a serem formados **refletirão a estrutura** inerente das variáveis escolhidas;
- De acordo com a **natureza das variáveis** deve-se escolher a medida de similaridade a qual corresponde o critério de construção dos grupos.



Variáveis e Objetos

- Cabe ressaltar que a técnica **não distingue se as variáveis são ou não relevantes** para o estudo;
- Essa tarefa fica a **cargo do analista** de dados;
- A inclusão de variáveis **não representativas** ou com **multicolinearidade** pode distorcer os resultados;
- A multicolinearidade **interfere na ponderação** das medidas de similaridade.



Variáveis e Objetos

- Antes de realizar a análise de conglomerados é recomendável **verificar a existência de *outliers***;
- Cabe ao analista de dados **decidir se deve retirar os *outliers*** ou **aplicar alguma técnica** de categorização;
- Observações atípicas (*outliers*) **podem formar grupos específicos** e estes grupos podem ser de interesse do analista de dados.



Padronização de Variáveis

- Após a escolha das variáveis representativas do fenômeno o analista de dados deve **verificar as escalas** das mesmas;
- Variáveis com **escalas/medidas diferentes** podem **distorcer a estrutura** do agrupamento;
- A maior parte das medidas de distância **sofrem influência das diferenças de escala** entre as variáveis;



Padronização de Variáveis

- A variável que apresenta **maior dispersão** tende a ter um **peso mais elevado** no cálculo das medidas de distância;
- Para contornar este problema existem as técnicas de **padronização de variáveis**;
- A padronização faz com que seja atribuído o **mesmo peso** para todas as variáveis;



Padronização de Variáveis

- **Método Z Padrão (*Z scores*):**

- ✓ É a forma mais utilizada de padronização de dados;
- ✓ Padroniza cada variável com média zero (0) e desvio padrão um (1);

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

Padronização de Variáveis

- **Método Range -1 a 1:**

- ✓ Faz com que a variável padronizada tenha amplitude 1;

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

- **Método Range 0 a 1:**

- ✓ Faz com que a variável apresente variação de 0 a 1;

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

Padronização de Variáveis

- **Método de máxima amplitude:**

- ✓ Confere à variável o valor máximo de 1;

$$\frac{x}{\text{máximo}}$$

- **Método de média 1:**

- ✓ Transforma a variável de maneira que apresente média 1;

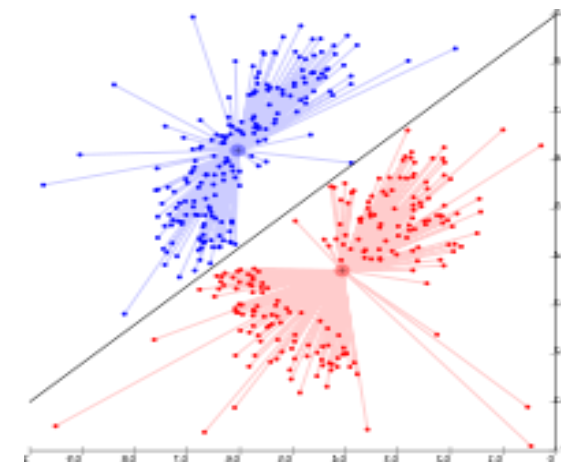
$$\frac{x}{\text{média}}$$



Medidas de Similaridade

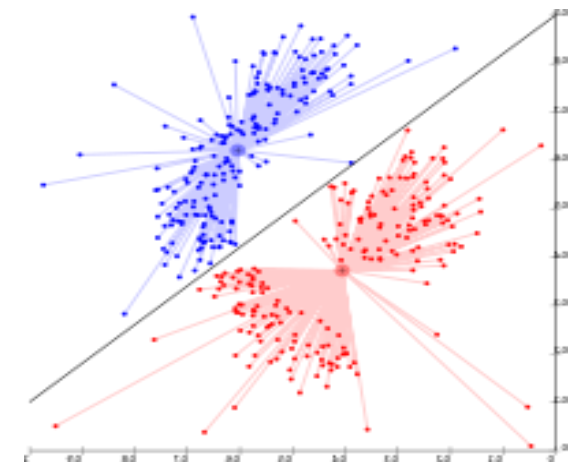
Medidas de Similaridade

- Após a **seleção das variáveis** e verificação da necessidade ou não de **padronizar**;
- A próxima etapa é a escolha da **medida de similaridade**;
- A técnica de agrupamento **só é possível** com a adoção de uma **medida de similaridade**;
- Permite fazer o **comparativo entre as observações** da amostra.



Medidas de Similaridade

- Existem **diversas medidas** de similaridade ou distância;
- A medida depende da **natureza das variáveis**;
- Para ilustrar o conceito de uma medida geométrica vamos observar um **exemplo**;



Medidas de Similaridade

- **Exemplo:**

✓ Um analista de dados pretende agrupar seis empresas no setor de comércio. Para isto utilizou as seguintes informações:

Empresas	Vendas (em US\$ milhões)	Número de 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

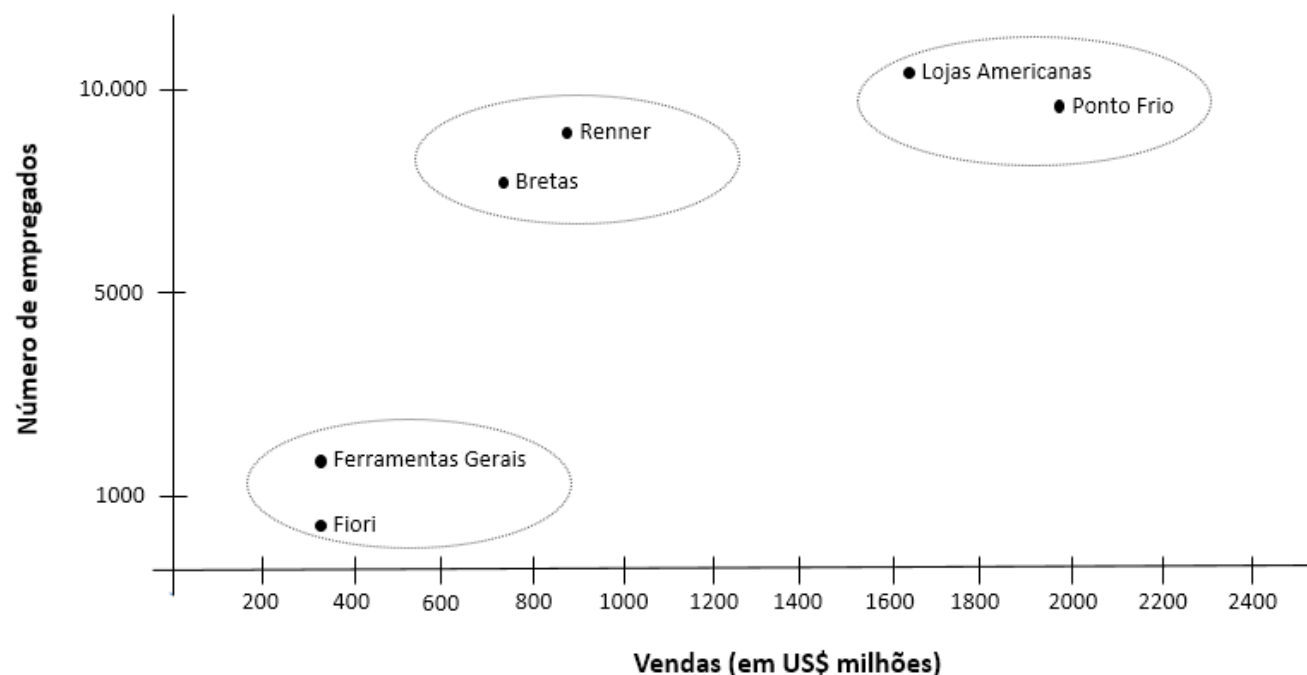
Medidas de Similaridade

- **Exemplo:**

- ✓ Para fins ilustrativos estamos analisando **apenas duas variáveis** (*Vendas e Número de Empregados*);
- ✓ Para este exemplo podemos representar cada observação como um ponto em um **espaço bidimensional**;
- ✓ Foi construído um **gráfico de dispersão** onde é possível visualizar a presença de três grupos;

Medidas de Similaridade

- **Exemplo:**
 - ✓ Representação em um espaço bidimensional;



Medidas de Similaridade

- Para um maior número de variáveis, já **não seria possível** identificar **visualmente** os grupos formados;
- Data a **limitação gráfica** do espaço tridimensional;
- Para isto é utilizado outros **critérios de aglomeração**;
- São utilizadas as **medidas de distância** ou similaridade;

Medidas de Similaridade

- A **escolha das medidas** de similaridade implica o conhecimento da **natureza das variáveis** (discreta, contínua, binária) e da **escala de medida** (nominal, ordinal);
- Dependendo da natureza das variáveis as **medidas podem ser**:
 - ✓ Medidas de distância (variáveis numéricas);
 - ✓ Medidas correlacionais (variáveis numéricas);
 - ✓ Medidas de associação (variáveis qualitativas).

Medidas de Similaridade

- **Distância Euclidiana:**

- ✓ Distância entre duas observações (i e j);
- ✓ Corresponde à raiz quadrada da soma dos quadrados das diferenças entre os pares (i e j) para todas as p variáveis:

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

- ✓ Onde x_{ik} é o valor da variável k referente à observação i e x_{jk} representa a variável k para a observação j .

Medidas de Similaridade

- **Distância Quadrática Euclidiana:**

✓ É a distância entre duas observações (i e j) que corresponde à soma dos quadrados das diferenças entre i e j para todas as p variáveis:


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

Medidas de Similaridade

- **Continuação do Exemplo**

- ✓ Vamos refazer o exemplo agora usando a distância euclidiana;
- ✓ O primeiro passo é realizar a padronização das variáveis por meio da técnica *Z Padrão*.

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



média zero
e desvio
padrão um

Medidas de Similaridade

- **Continuação do Exemplo**

✓ Logo, a tabela com valores padronizados ficará da seguinte forma:

Empresas	Vendas (em US\$ milhões)	Número de empregados
Ferramentas Gerais	-0,931	-1,038
Fiori	-0,953	-1,427
Bretas Supermercados	-0,458	0,282
Renner	-0,056	0,429
Lojas Americanas	0,939	1,087
Ponto Frio	1,459	0,666

Medidas de Similaridade

- **Continuação do Exemplo**

✓ Agora iremos aplicar a **distância quadrática euclidiana** na tabela padronizada:

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

✓ Logo, calculando a distância d_{12} temos:

$$d_{ij}^2 = (-0,931 - (-0,953))^2 + (-1,038 - (-1,427))^2 = 0,152$$

Medidas de Similaridade

- **Continuação do Exemplo**

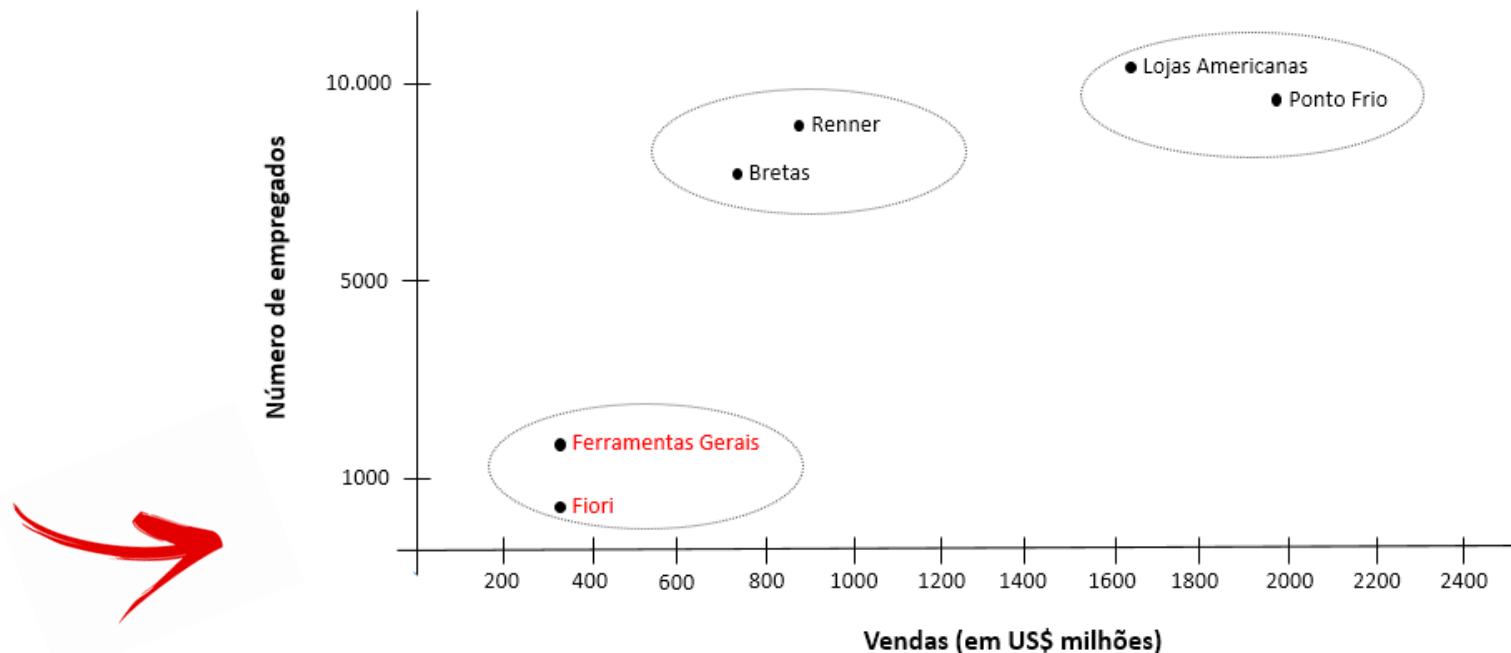
- ✓ A distância d_{12} refere-se a distância entre *Ferramentas Gerais* e *Fiori*;
- ✓ A matriz de similaridade mostra os valores de distâncias:

	Ferramentas Gerais	Fiori	Bretas	Renner	Lojas Americanas	Ponto Frio
Ferramentas Gerais	0,000					
Fiori	0,152	0,000				
Bretas	1,964	3,163	0,000			
Renner	2,916	4,248	0,183	0,000		
Lojas Americanas	8,01	9,898	2,601	1,423	0,000	
Ponto Frio	8,616	10,2	3,824	2,353	0,447	0,000

Medidas de Similaridade

- **Continuação do Exemplo**

- ✓ Comparando com os grupos formados pelo gráfico de dispersão:



Medidas de Similaridade

- **Continuação do Exemplo**

- ✓ A matriz de similaridade **mostra as distâncias** de cada par de empresas;
- ✓ Percebe-se que a menor distância entre empresas *Ferramentas gerais* e *Fiore* **denotará o primeiro grupo** a ser formado;
- ✓ Cabe ressaltar que as **medidas mais utilizadas** são as de similaridade;
- ✓ Especialmente a distância euclidiana **simples** e a **quadrática**.

Medidas de Similaridade

- **Distância de Mahalanobis:**

✓ É a distância estatística entre dois indivíduos i e j , considerando a matriz de covariância para o cálculo das distâncias

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

Medidas de Similaridade

- **Existem diversas outras distâncias, tais como:**
 - ✓ Distância de Minkowski;
 - ✓ Distância absoluta ou Manhattan;
 - ✓ Distância de Chebychev;
 - ✓

Medidas de Similaridade

- **Medidas Correlacionadas:**

- ✓ As medidas correlacionadas baseiam-se na força da relação entre variáveis;
- ✓ A mais utilizada é a medida de correlação de *Pearson*:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \times \sum_{i=1}^n (y_i - \bar{y})^2}}$$

Medidas de Similaridade

- **Medidas de Associação:**

- ✓ São utilizadas para representar similaridade em **variáveis de natureza qualitativas**;
- ✓ É realizada por meio de **tabelas de contingência**;
- ✓ São medidas **pouco utilizadas** na prática de análise de conglomerados;

Agora é com vocês!

- ✓ Porque é importante padronizar as variáveis inicialmente ?
- ✓ O que é uma medida de similaridade e qual seu objetivo?
- ✓ Quais são as medidas de similaridade mais utilizadas ?

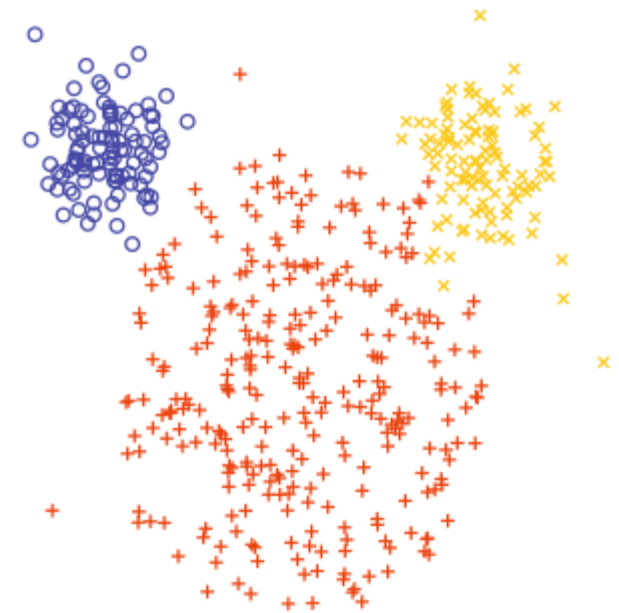


Métodos de Agrupamento



Métodos de Agrupamento

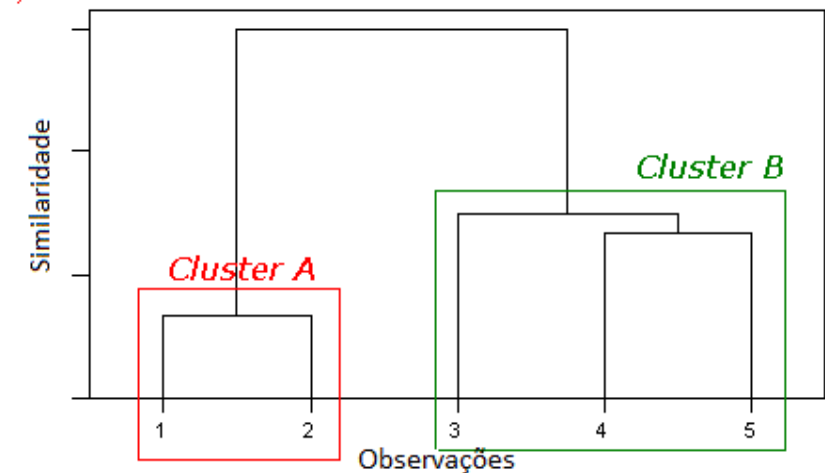
- Uma vez (1) **selecionada as variáveis**, (2) **padronizado** as variáveis, escolhida a (3) **medida de similaridade** e construída a (4) **matriz de similaridade**;
- É necessário determinar o **algoritmo** que fará o processo de **agrupamento**;
- Basicamente há dois métodos: (1) **Hierárquico** e (2) **Não-hierárquico**.



Métodos de Agrupamento

- **Método Hierárquico:**

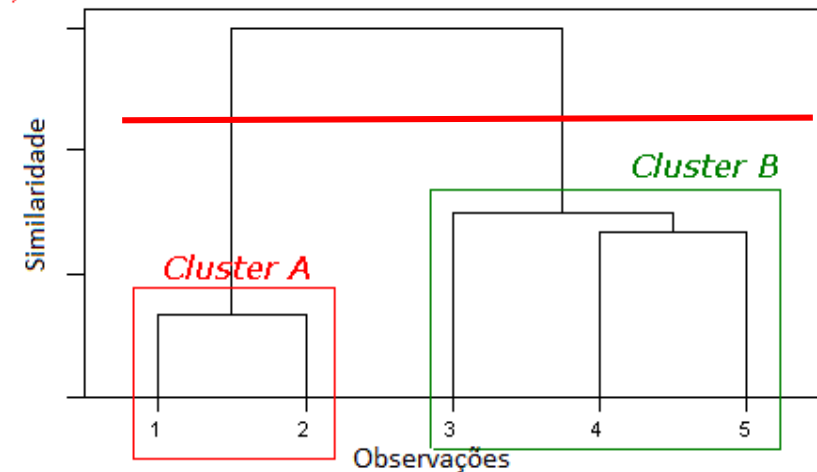
- ✓ Apresentam a estrutura de uma árvore;
- ✓ Inicialmente todos os elementos pertencem a um só grupo;
- ✓ Em seguida os grupos vão sendo formados de acordo com o método para cálculo da distância:
 - Ward;
 - Centroide;
 - Menor distância ou Ligação individual.



Métodos de Agrupamento

- **Método Hierárquico:**

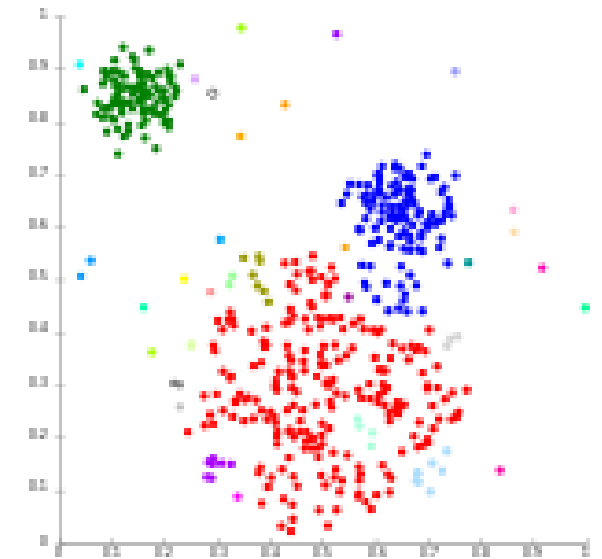
- ✓ Uma maneira de **representar graficamente** é por meio do dendrograma;
- ✓ Mostra as **etapas da aglomeração** dos grupos;
- ✓ É possível **visualizar os elementos** de cada cluster;
- ✓ De acordo com o corte é possível verificar o **número de grupos** formados.



Métodos de Agrupamento

- **Método Não-hierárquico:**

- ✓ É necessário **especificar o número de grupos** anteriormente;
- ✓ É um **processo iterativo** em busca da solução ótima;
- ✓ O processo busca satisfazer duas condições: (1) **coesão interna** e (2) **isolamento entre grupos**;
- ✓ É **menos custoso** computacionalmente;



Métodos de Agrupamento

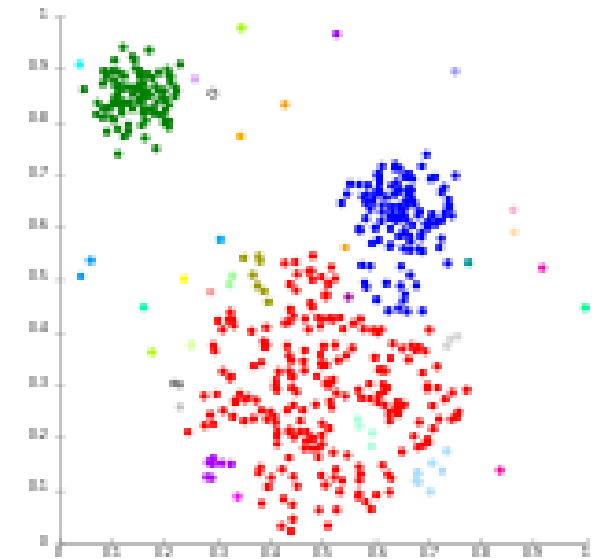
- Método Não-hierárquico:

Como definir o **número de cluster** (k) inicialmente ?



Métodos de Agrupamento

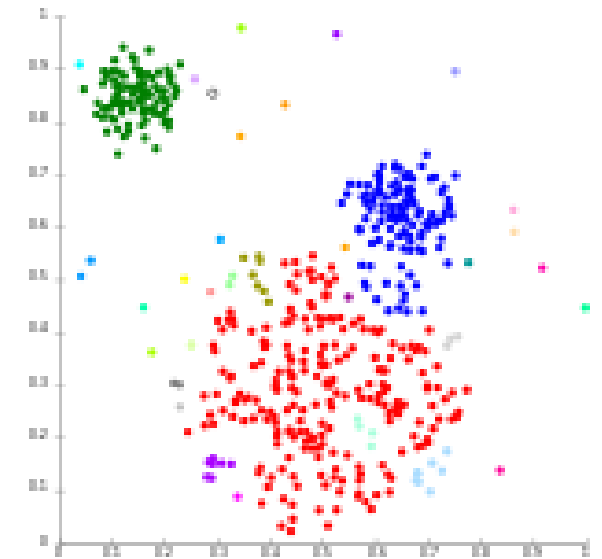
- **Método Não-hierárquico:**
 - ✓ Existe **dificuldade** em estabelecer o **número de clusters** de partida;
 - ✓ Uma alternativa é realizar o método hierárquico como técnica exploratória para **verificar o K**;
 - ✓ Posteriormente utiliza-se o método não-hierárquico **para alocar as observações**;



Métodos de Agrupamento

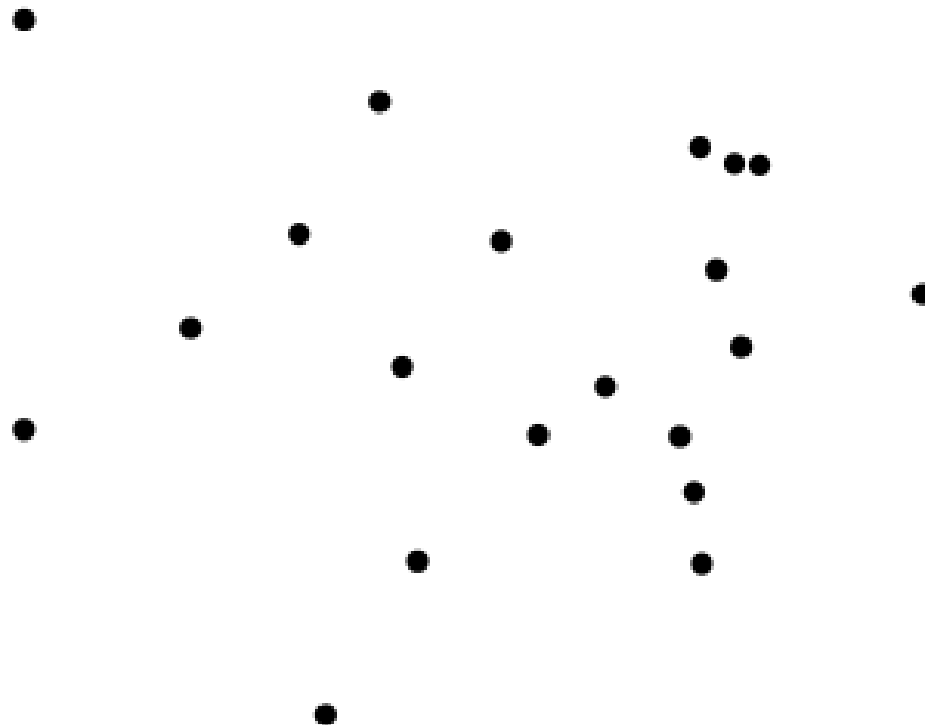
- **Método Não-hierárquico:**

- ✓ O algoritmo mais popular é o *K-means*;
- ✓ Pode ser usado para agrupamento de grandes quantidades de dados;
- ✓ O critério de distância mais utilizado no k-means é a distância euclidiana;



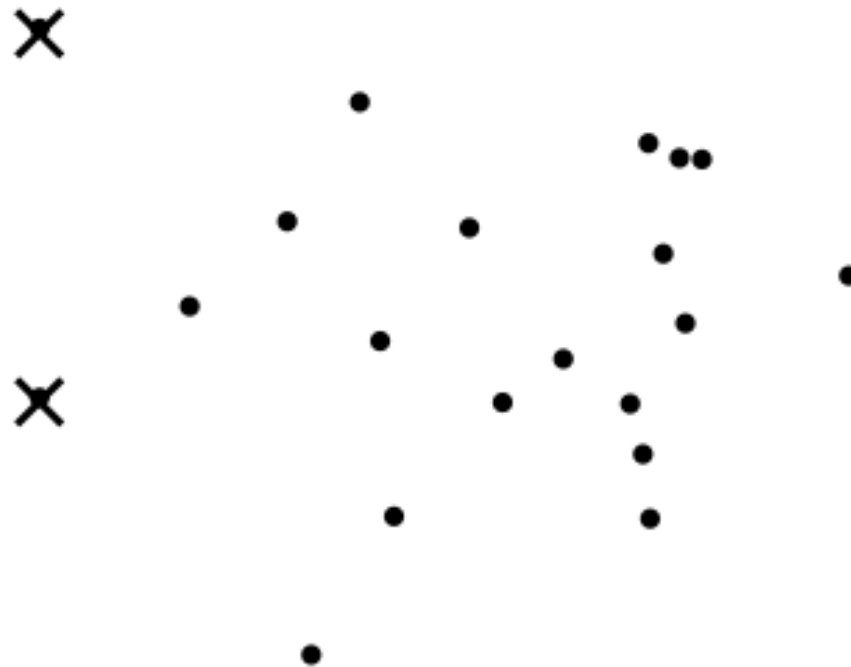
Métodos de Agrupamento

- **Exemplo:** Conjunto de dados a serem agrupados



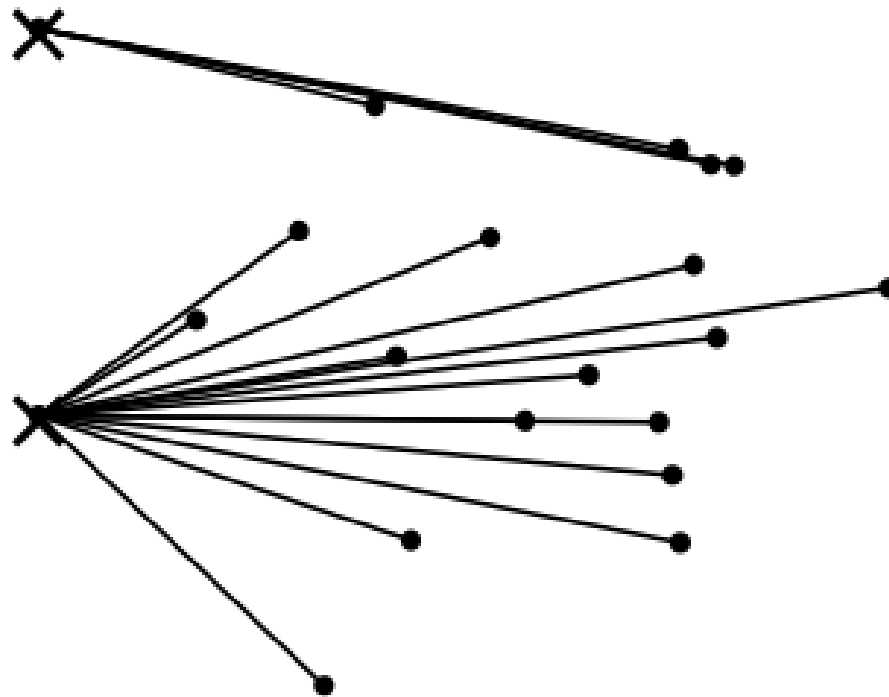
Métodos de Agrupamento

- Seleção inicial dos Centroides



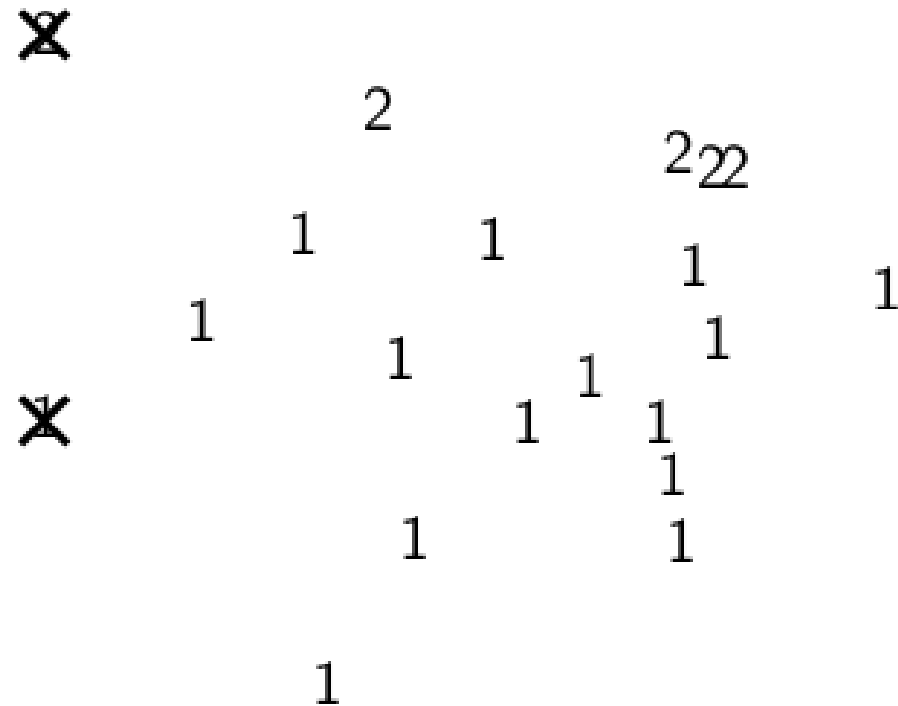
Métodos de Agrupamento

- Atribuir os pontos aos centroides mais próximos



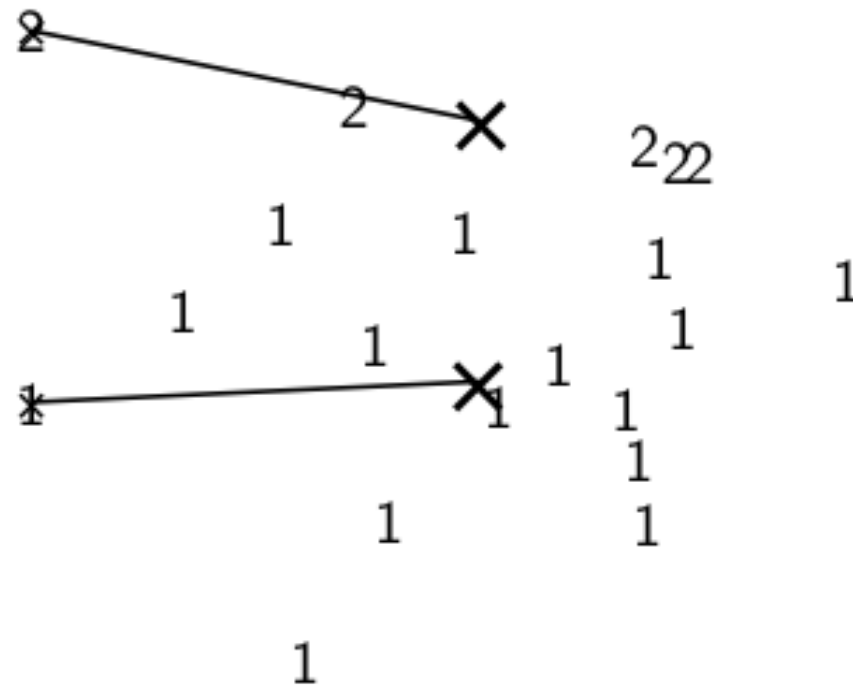
Métodos de Agrupamento

- Atribuição



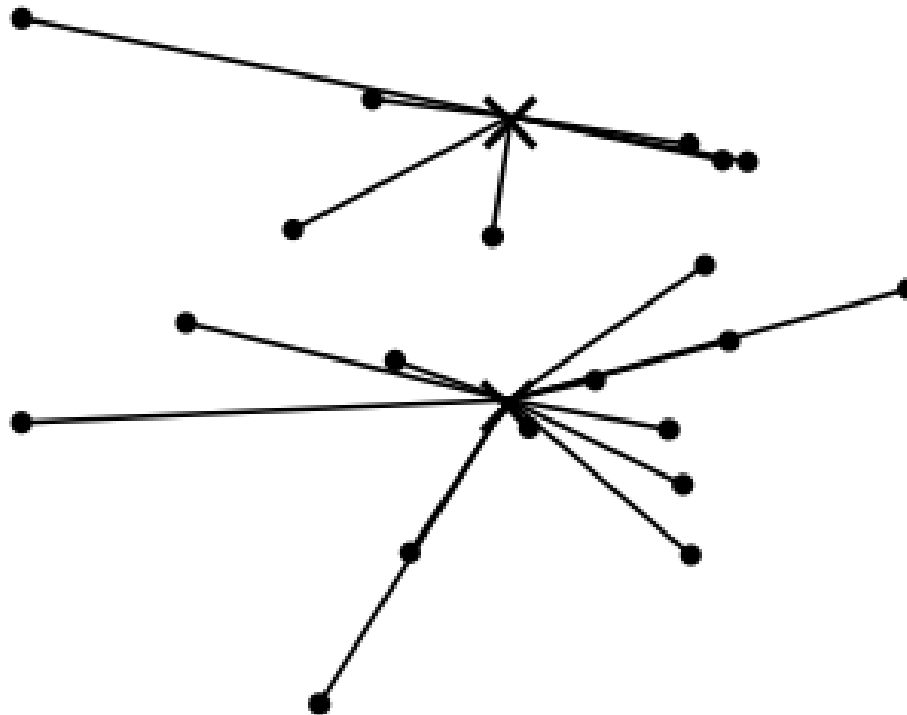
Métodos de Agrupamento

- Recalcular os centroides



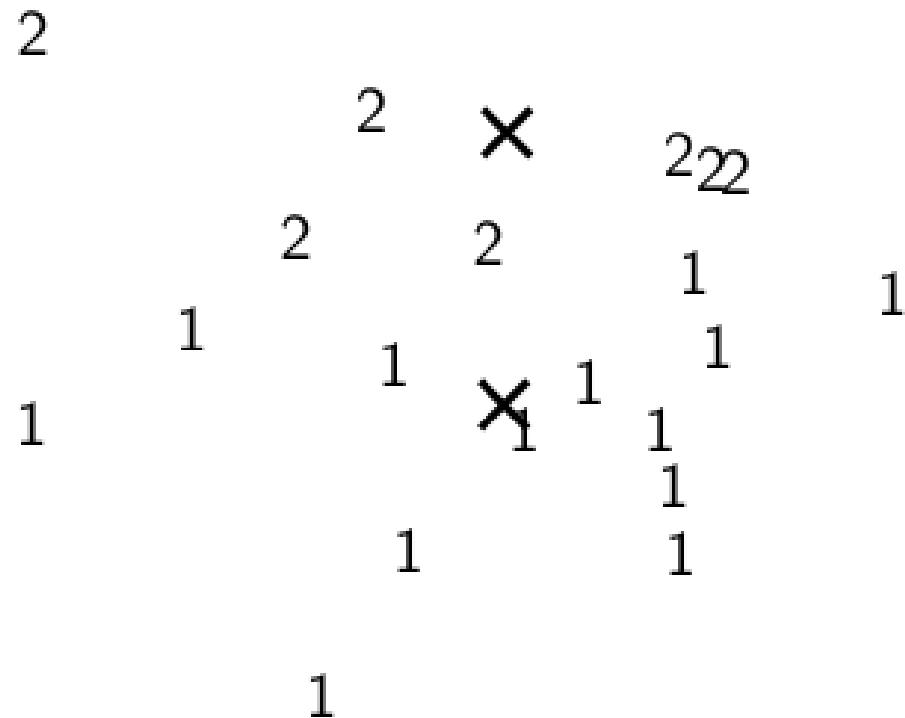
Métodos de Agrupamento

- Atribuir os pontos aos centroides mais próximos



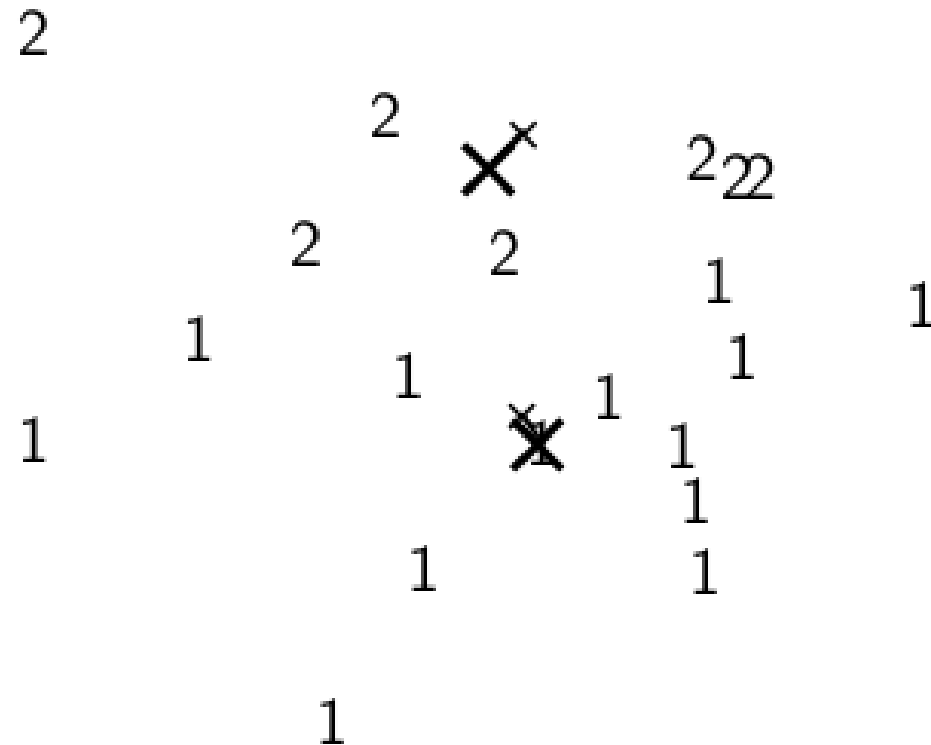
Métodos de Agrupamento

- Atribuição



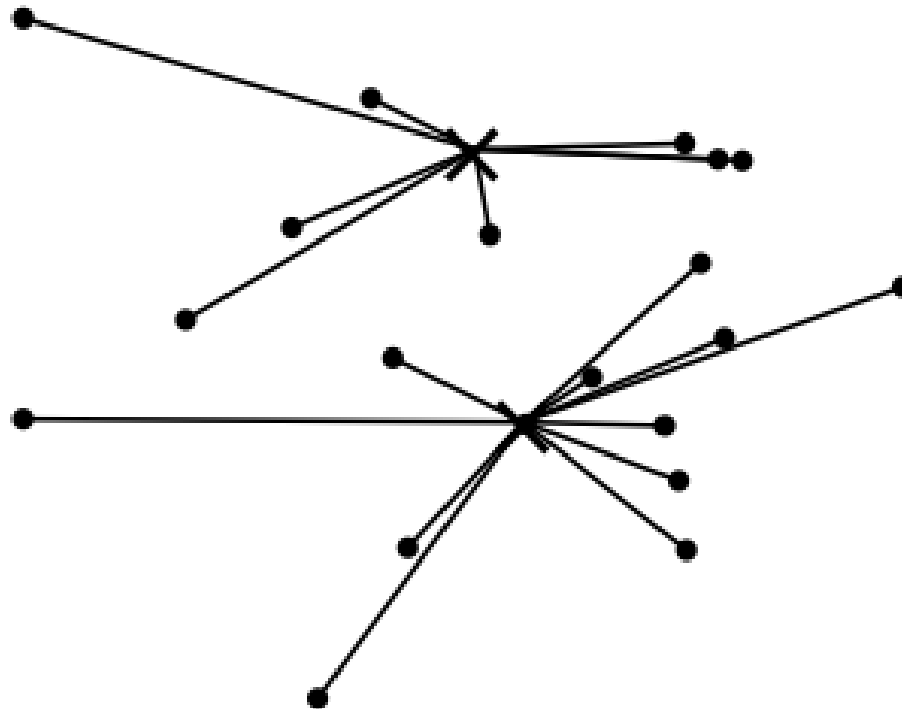
Métodos de Agrupamento

- Recalcular os centroides



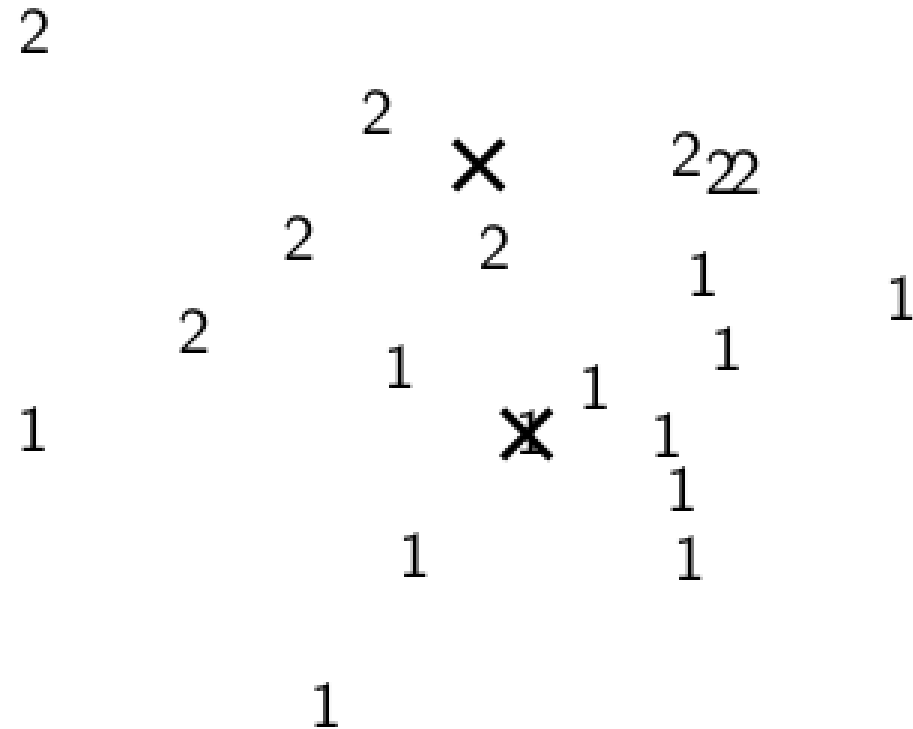
Métodos de Agrupamento

- Atribuir os pontos aos centroides mais próximos



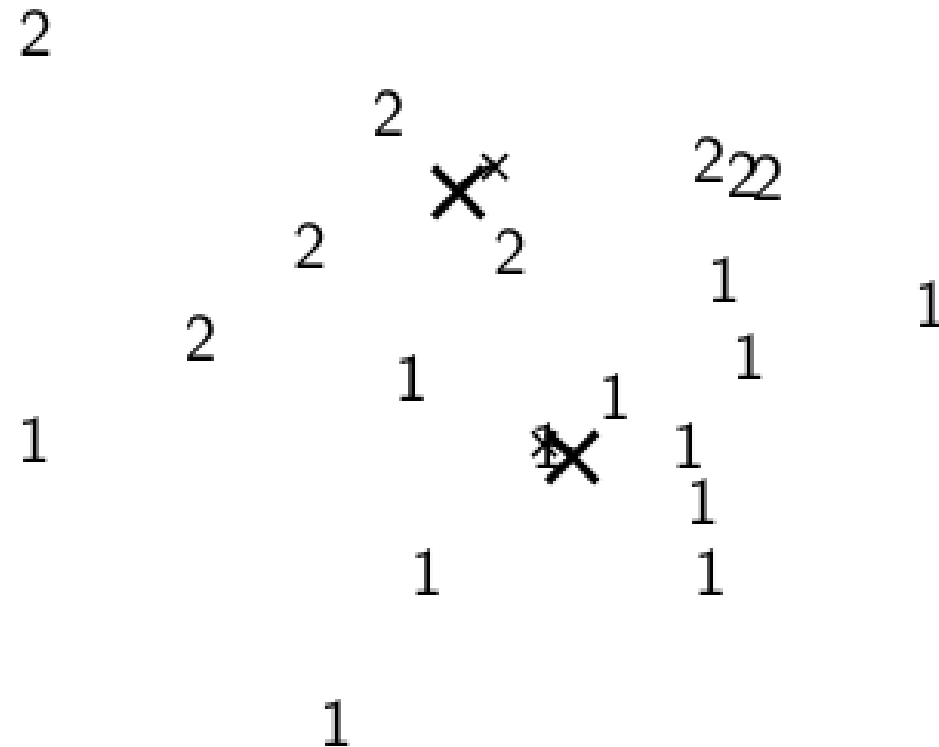
Métodos de Agrupamento

- Atribuição



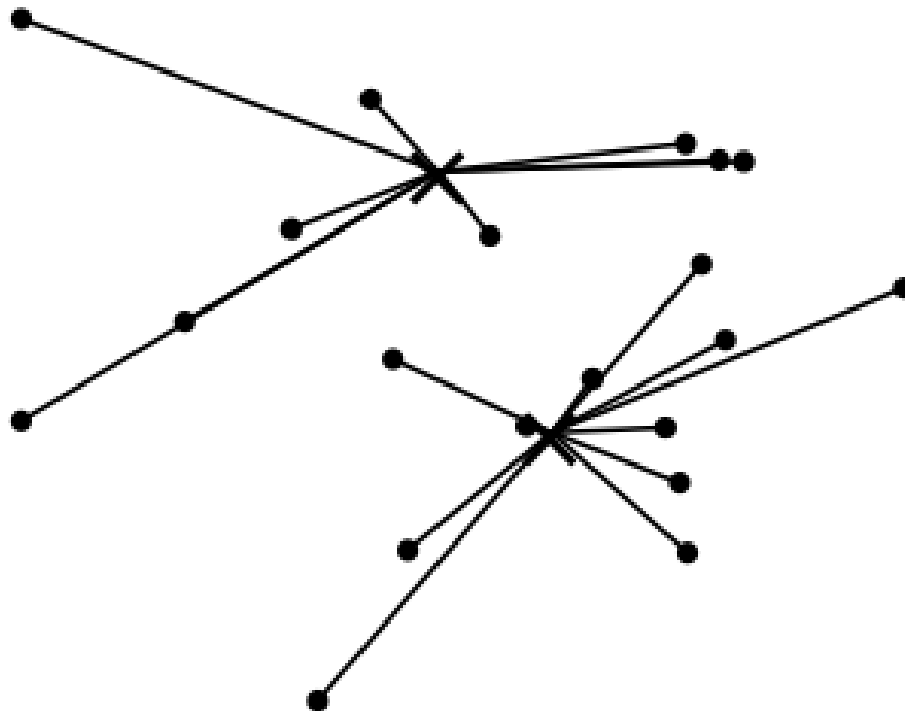
Métodos de Agrupamento

- Recalcular os centroides



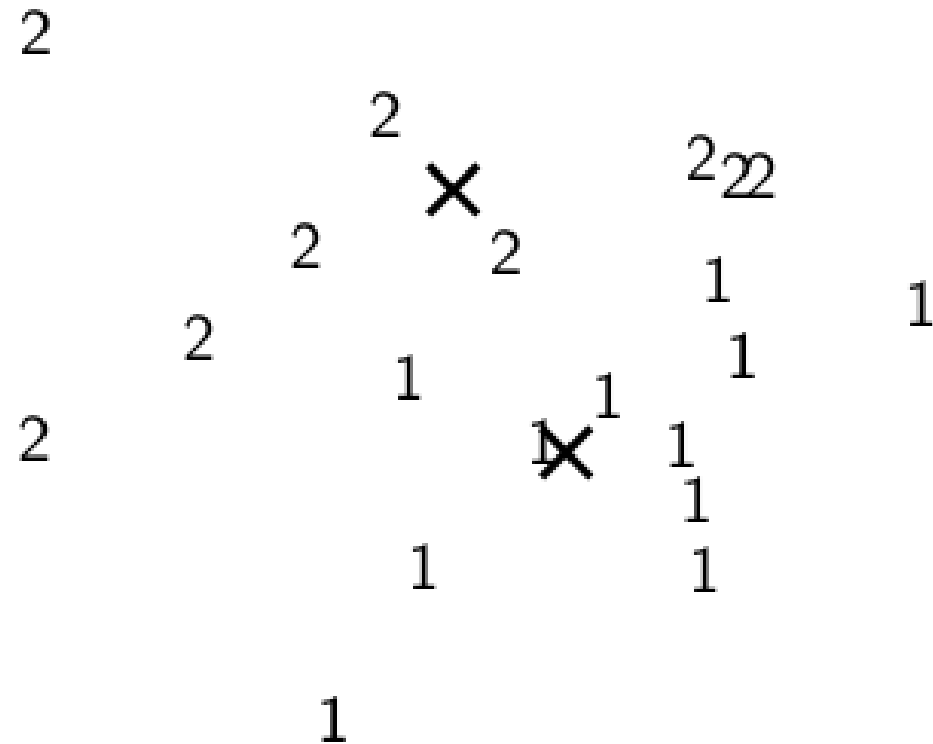
Métodos de Agrupamento

- Atribuir os pontos aos centroides mais próximos



Métodos de Agrupamento

- Atribuição



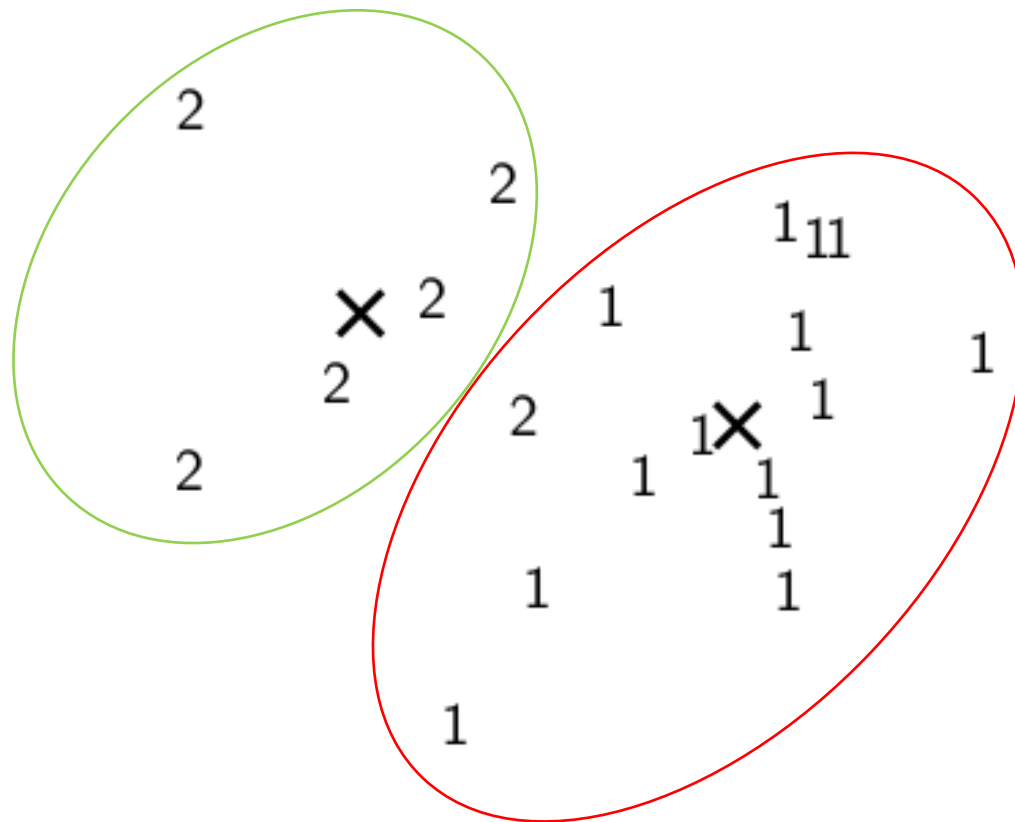
Métodos de Agrupamento

- Depois de muitas interações....



Métodos de Agrupamento

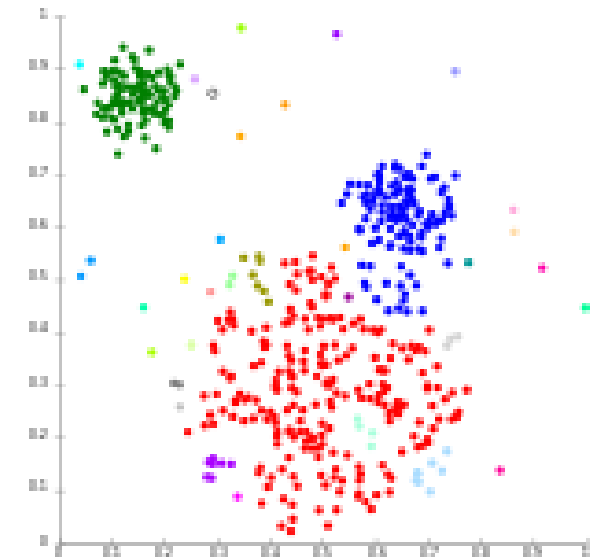
- Centroides e atribuições após a convergência



Métodos de Agrupamento

- **Método Não-hierárquico:**

- ✓ Um dos problemas enfrentados pelo algoritmo k-means é a seleção das ementes;
- ✓ Pode produzir diferentes resultados;
- ✓ É interessante validar o método de agrupamento.



Avaliação de Conglomerados

Avaliação de Conglomerados

- Critério interno:
 - ✓ Um exemplo é a avaliação de cluster pelo critério da ANOVA;
- Porém, um critério interno muitas vezes não avalia a utilidade do agrupamento para uma aplicação;
- Alternativa: Critério externo
 - ✓ Avaliar de acordo com um critério definido por especialistas da área;
 - ✓ Verifica se os grupos formados fazem sentido.

Avaliação de Conglomerados

- Após a especificação do número de clusters, o pesquisador deve ficar atento ao número de observações que compõem cada cluster;
- Verificar se existem grupos com quantidade muito pequena de elementos;
- Formação de grupo com apenas um único elemento pode haver elementos com observações atípicas.



```
script, filename = argv

print "We're going to erase %r." % filename
print "If you don't want that, hit CTRL-C (^C)."
print "If you do want that, hit RETURN."

raw_input("?")

from sys import argv
from os.path import exists

script, from filename = argv

print "Opening the file..."
target = open(filename, 'w')

print_two("Zed", "Shaw")
print_two_again("Zed", "Shaw")
print_one("First!")
print_none()

def br (stuff):
    "function will be"

    lit(' '

:
5. ""
ds)
```

Prática Computacional no R

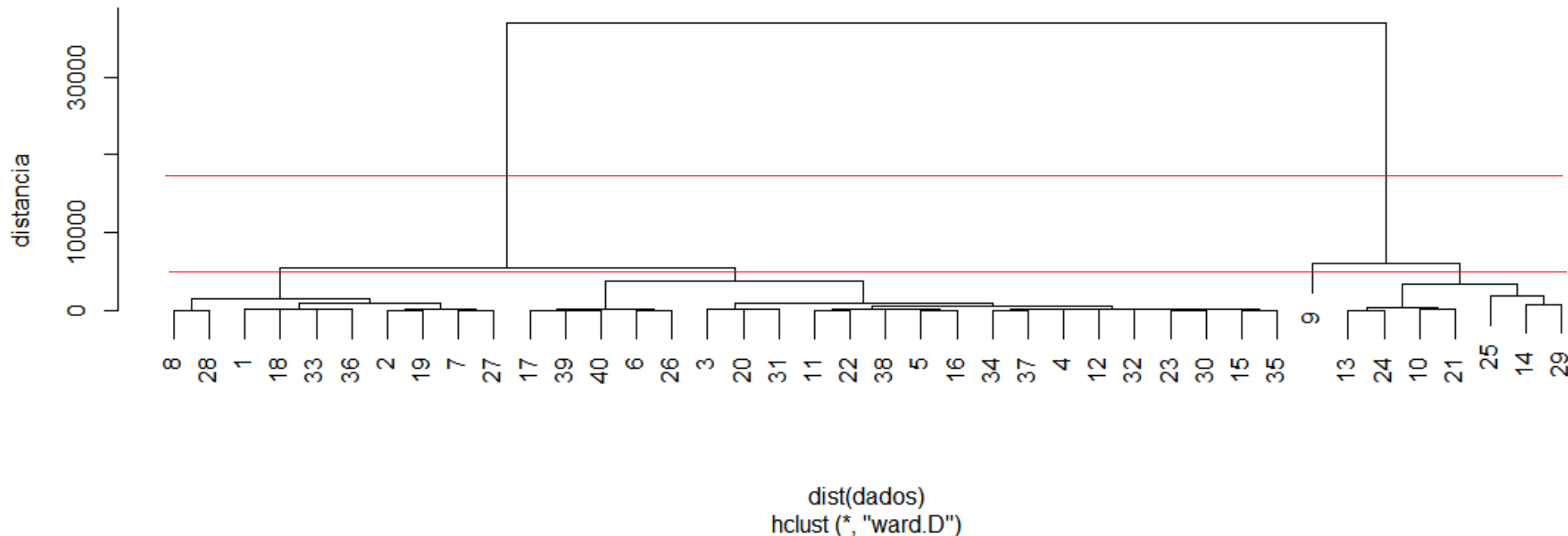
Prática Computacional

- Script em R para Agrupamento **Hierárquico** (*sem padronização*)

```
7  # Lendo a base de dados
8  dados <- read.csv2("dados.csv", head=T)
9
10 #Com isso o R entende que os nomes da primeira linha da base importada são
11 #as variáveis e cada linha refere-se a um registro, sendo a primeira coluna a
12 #identificação de cada registro
13 variaveis<-names(dados)
14 objetos<-rownames(dados)
15
16 # aplicando o metodo de cluster hierarquico
17 output_cluster<-hclust(dist(dados), method='ward.D')
18 dendograma_output_cluster<-plclust(output_cluster, labels=objetos, ylab='distancia')
```

Prática Computacional

- **Dedrograma** do Agrupamento Hierárquico (*sem padronização*)



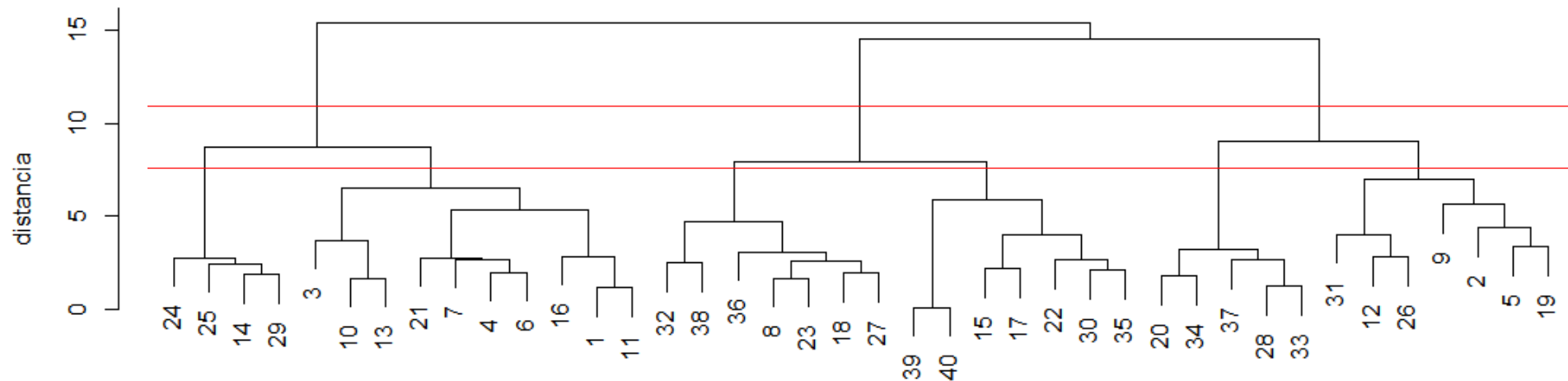
Prática Computacional

- Script em R para Agrupamento **Hierárquico** (*com padronização*)

```
7  # Lendo a base de dados
8  dados <- read.csv2("dados.csv", head=T)
9
10 # Padronizando a base de dados pelo Zscore
11 dados <- data.frame(scale(dados))
12
13 #Com isso o R entende que os nomes da primeira linha da base importada são
14 #as variáveis e cada linha refere-se a um registro, sendo a primeira coluna a
15 #identificação de cada registro
16 variaveis<-names(dados)
17 objetos<-rownames(dados)
18
19 # aplicando o metodo de cluster hierarquico
20 output_cluster<-hclust(dist(dados), method='ward.D')
21 dendograma_output_cluster<-plclust(output_cluster, labels=objetos, ylab='distancia')
```


Prática Computacional

- **Dedrograma** do Agrupamento Hierárquico (*com padronização*)



dist(dados)
hclust (*, "ward.D")

Prática Computacional

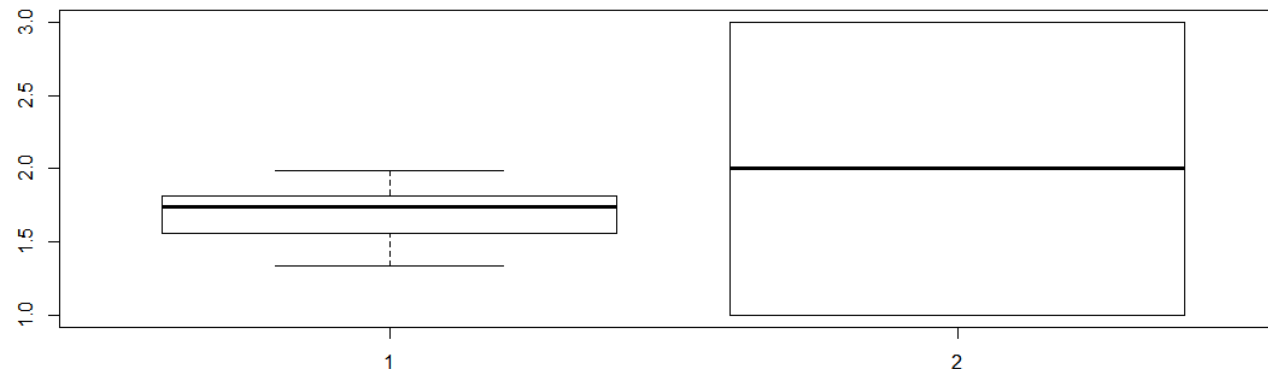
- Script em R para Agrupamento **Não-Hierárquico**

```
3 # Lendo a base de dados
4 dados <- read.csv2("dados.csv", head=T)
5
6 # Padronizando a base de dados pelo Zscore
7 dados <- data.frame(scale(dados))
8
9 # Realizando o agrupamento não-hierárquico por Kmeans
10 output_cluster<-kmeans(dados,3)
11 grupos<-output_cluster$cluster
12 matriz<-cbind(dados,grupos)
13 write.table(file='dadosKMeans.csv',matriz, sep=';',dec=',')
```

Prática Computacional

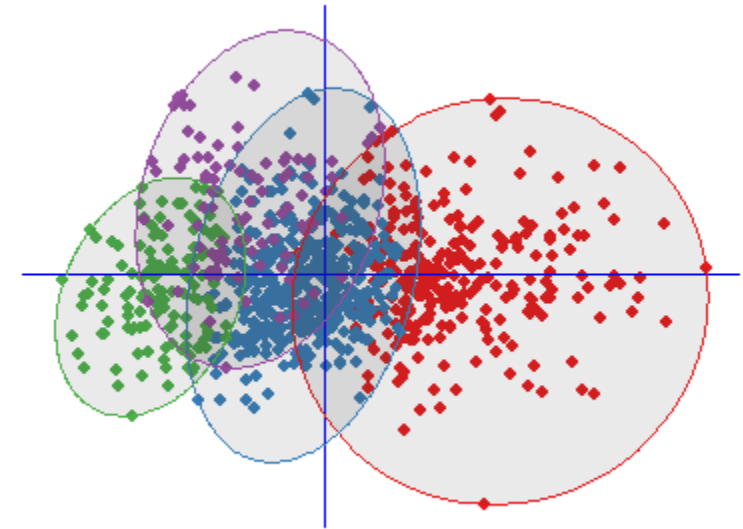
- Construindo boxplot para agrupamento **Não-Hierárquico**

```
19 # Incluindo a coluna de agrupamento nos dados não padronizados
20 dados<-cbind(dados,grupos)
21
22 # Fazendo o boxplot com os dados agrupados e não padronizados
23 boxplot(dados$Altura,dados$grupos)
```



Prática Computacional

- Após a construção dos grupos pode-se aplicar diversas análises descritivas;
- É possível verificar o comportamento de cada grupo;



Dúvidas



- **Contatos:**

- ✓ Email: rodrigo.linsrodrigues@ufrpe.br
- ✓ Facebook: [/rodrigomuribec](https://www.facebook.com/rodrigomuribec)