Curso 2 – CD, AM e DM

Mineração de Dados

Parte 5

Pós-processamento

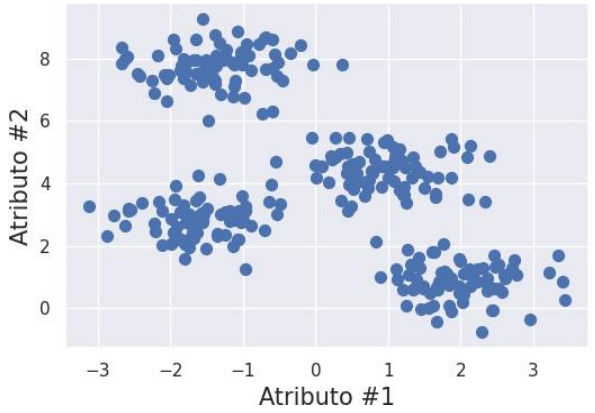
Validação de Agrupamentos

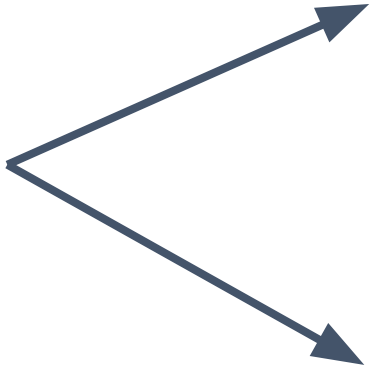
Prof. Ricardo M. Marcacini
ricardo.marcacini@icmc.usp.br

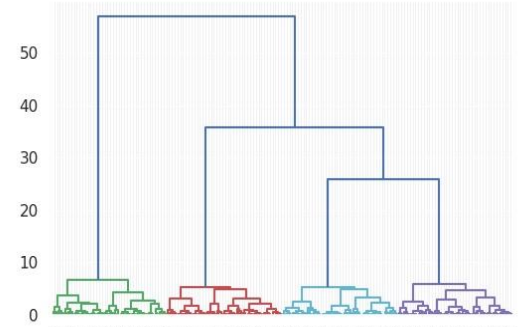
Métodos para Agrupamento de Dados

- **Hierárquicos:** organizar dados em uma decomposição hierárquica de *clusters* e *subclusters*
- **Particionais:** organizar dados em uma partição de *k clusters*

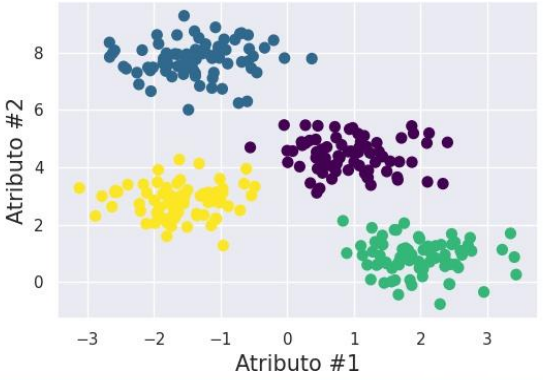
Conjunto de Dados







Agrupamento Hierárquico

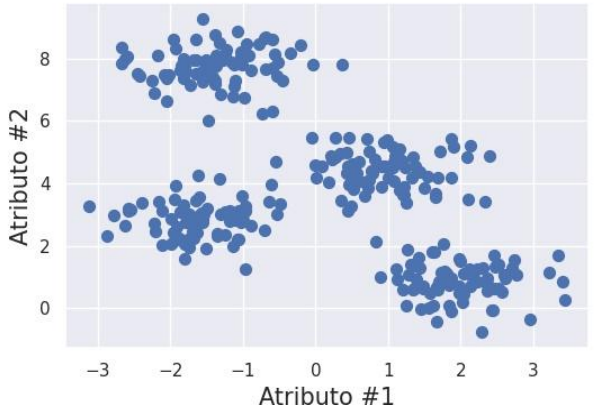


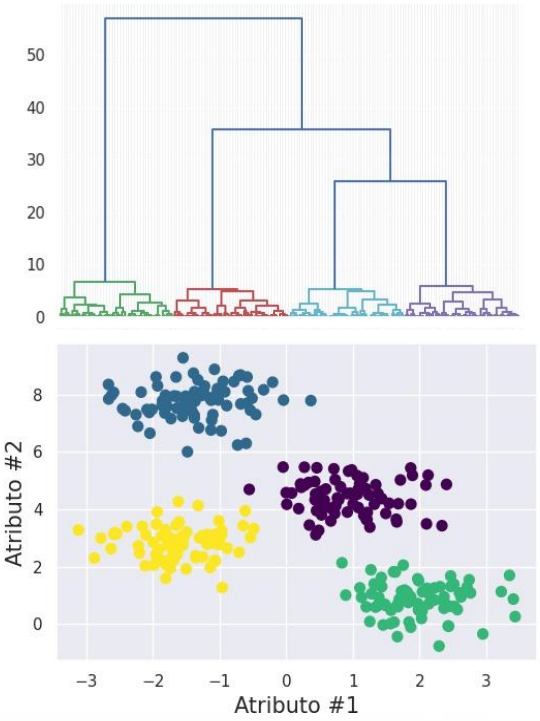
Agrupamento Particional

Métodos para Agrupamento de Dados

- **Hierárquicos:** *Single-Link, Complete-Link, Average-Link e Bisecting K-Means*
- **Particionais:** *k-Means e k-Medoides*

Conjunto de Dados





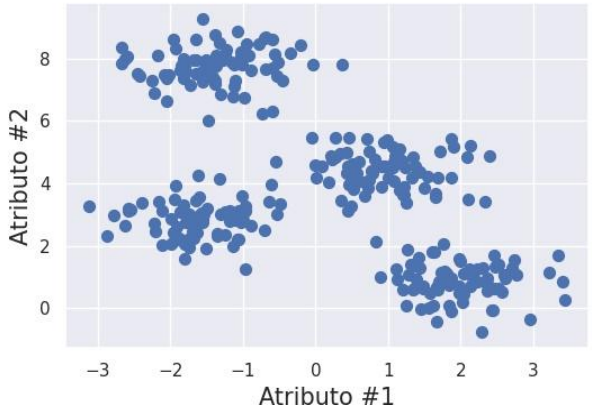
Agrupamento Hierárquico

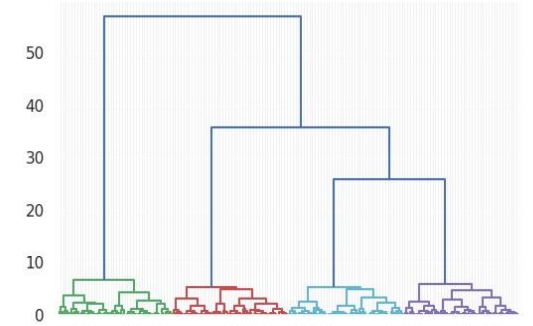
Agrupamento Particional

Métodos para Agrupamento de Dados

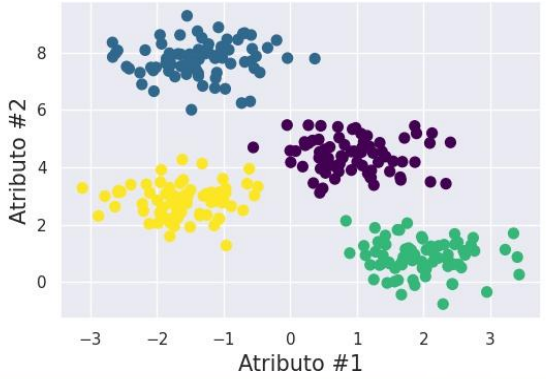
Qual método e algoritmo de agrupamento escolher?
Qual é o número (k) apropriado de *clusters*?

Conjunto de Dados





Agrupamento Hierárquico



Agrupamento Particional

Pós-Processamento

Quais os critérios de avaliação?

Validação de Agrupamentos

- Analisar o “mérito” e qualidade dos clusters
- Validação por inspeção visual
- Índices de validação de agrupamentos
  - Índices internos
  - Índices relativos
  - Índices externos

Validação de Agrupamentos

- Analisar o “mérito” e qualidade dos clusters

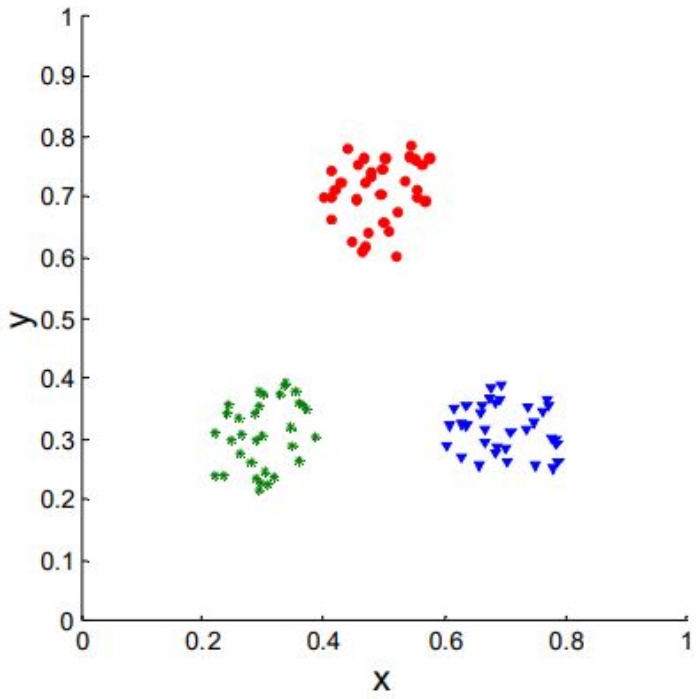
- Validação por inspeção visual

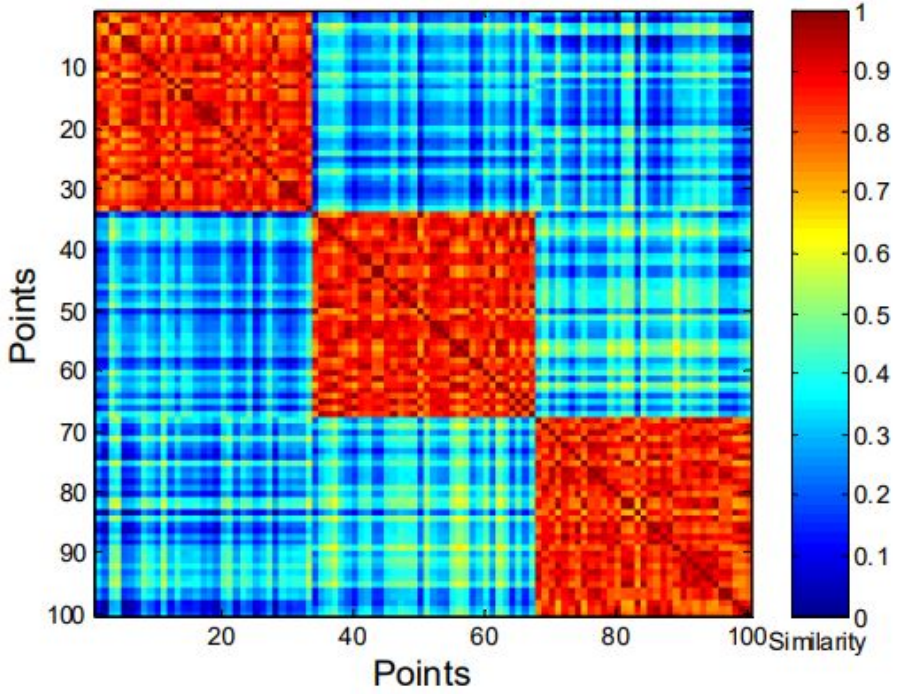
- Índices de validação de agrupamentos
  - Índices internos
  - Índices relativos
  - Índices externos

Validação de Agrupamentos

- Validação por inspeção visual

Após encontrar os clusters, construir a matriz de dissimilaridades (ou similaridade) e ordená-la de acordo com os clusters. Em seguida, colorir a matriz conforme a medida de proximidade.

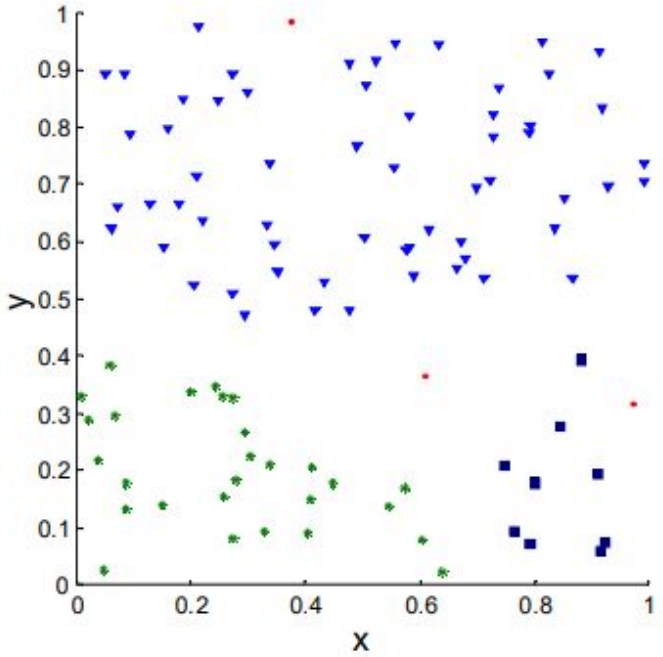


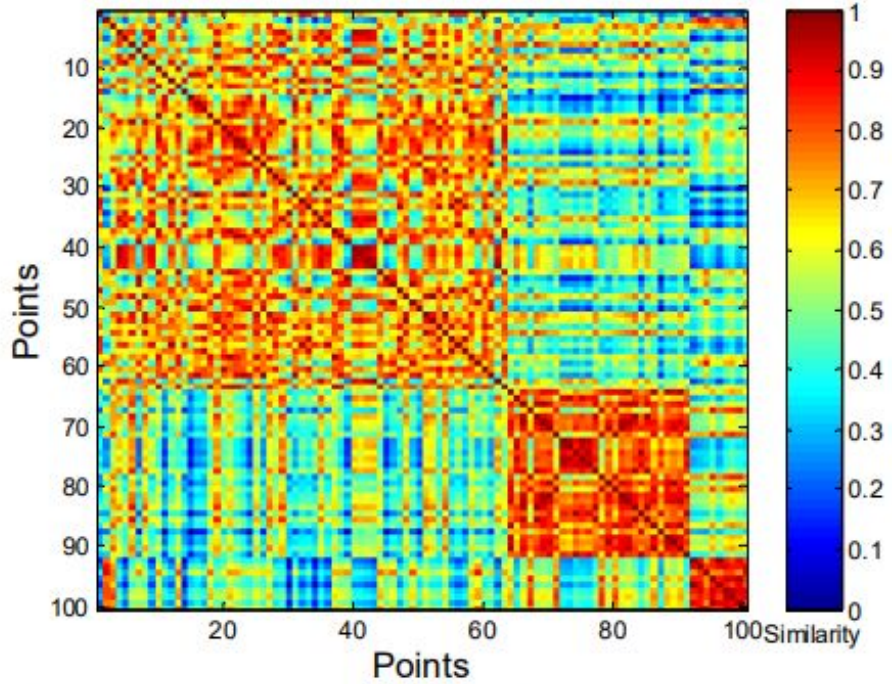


Validação de Agrupamentos

- Validação por inspeção visual

Observe que dados sem uma estrutura de cluster bem definida se destacam menos na inspeção visual via matriz de (dis)similaridades.



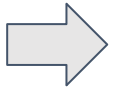


Validação de Agrupamentos

- Validação por inspeção visual

Verificar visualmente os clusters em um espaço de baixa dimensionalidade, como uma projeção bidimensional.

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{bmatrix}$$



Medida de Proximidade

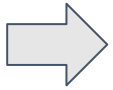
Algoritmo de Agrupamento

Validação de Agrupamentos

- Validação por inspeção visual

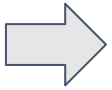
Verificar visualmente os clusters em um espaço de baixa dimensionalidade, como uma projeção bidimensional.

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{bmatrix}$$



Medida de Proximidade

Algoritmo de Agrupamento



Resultado: Partição (Clusters)

Validação de Agrupamentos

- Validação por inspeção visual

Verificar visualmente os clusters em um espaço de baixa dimensionalidade, como uma projeção bidimensional.

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{bmatrix}$$

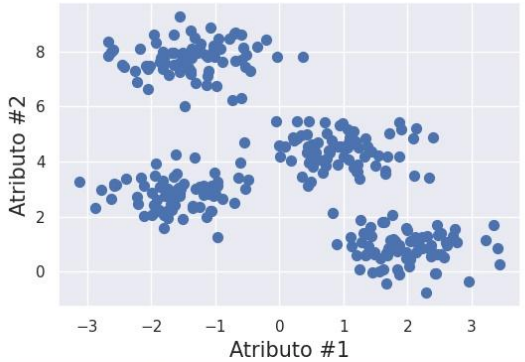
Medida de Proximidade

Algoritmo de Agrupamento

Resultado: Partição (Clusters)

Projeção dos dados em duas dimensões

Ex: PCA



Validação de Agrupamentos

- Validação por inspeção visual

Verificar visualmente os clusters em um espaço de baixa dimensionalidade, como uma projeção bidimensional.

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{bmatrix}$$

Medida de Proximidade

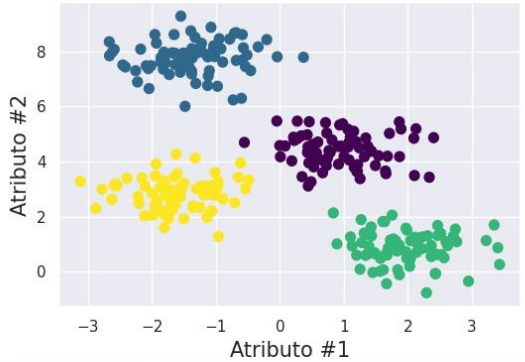
Algoritmo de Agrupamento

Resultado: Partição (Clusters)

Plotar cores nos objetos conforme os *clusters*

Projeção dos dados em duas dimensões

Ex: PCA



Validação de Agrupamentos

- Validação por inspeção visual
  - São opções importantes para explorar os clusters
  - Evitar o uso como única forma de validação

Validação de Agrupamentos

- Validação por inspeção visual
  - São opções importantes para explorar os clusters
  - Evitar o uso como única forma de validação

- Limitações
  - Subjetividade na validação
  - Projeção bi-dimensionais dos dados podem perder informações relevantes

Validação de Agrupamentos

- Índices de validação de agrupamento
  - São critérios que analisam de forma quantitativa e objetiva a “qualidade” dos *clusters* obtidos

Validação de Agrupamentos

- Índices de validação de agrupamento
  - São critérios que analisam de forma quantitativa e objetiva a “qualidade” dos *clusters* obtidos

  - **Índices de validade interna**

    Analisam a estrutura de *clusters* sem uso de informação externa

Validação de Agrupamentos

- Índices de validação de agrupamento
  - São critérios que analisam de forma quantitativa e objetiva a “qualidade” dos *clusters* obtidos

  - **Índices de validade interna**

    Analisam a estrutura de *clusters* sem uso de informação externa
  - **Índices de validade relativa**

    Visam comparar diferentes partições, geralmente para identificar o número apropriado de clusters (diferentes algoritmos de *clustering*)

Validação de Agrupamentos

- Índices de validação de agrupamento
  - São critérios que analisam de forma quantitativa e objetiva a “qualidade” dos *clusters* obtidos

  - **Índices de validade interna**

    Analisam a estrutura de *clusters* sem uso de informação externa
  - **Índices de validade relativa**

    Visam comparar diferentes partições, geralmente para identificar o número apropriado de clusters (diferentes algoritmos de *clustering*)
  - **Índices de validade externa**

    Medem o quanto a estrutura de cluster representa uma estrutura de organização previamente estabelecida

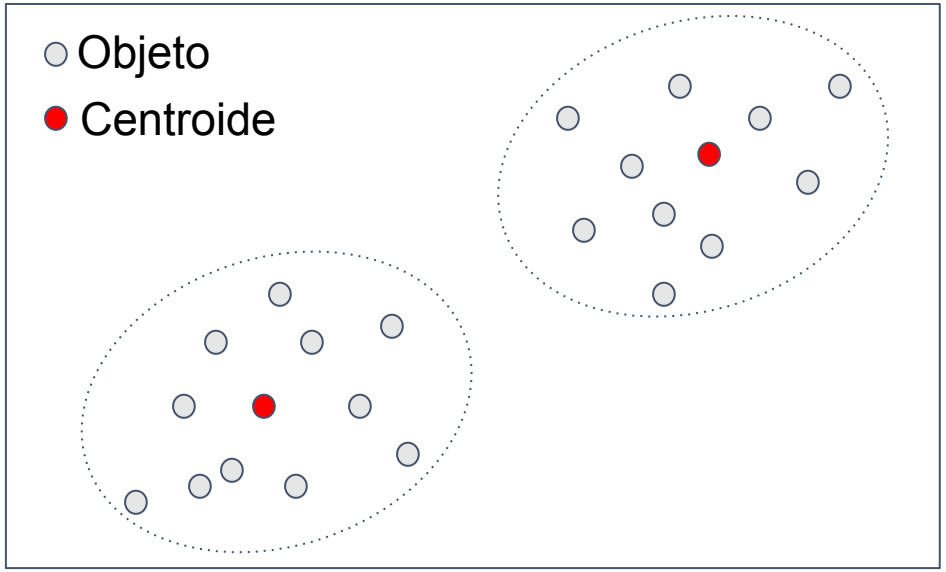
Validação de Agrupamentos

- Índices de validade interna: Erro Quadrático
  - Já estudamos indiretamente essa medida
  - É utilizada pelo *k-means* para escolher a melhor execução após diferentes inicializações de centroides

$$E = \sum_{i=1}^{k} \sum_{\mathbf{x} \in \mathbf{C_i}} d^2(\mu_i, \mathbf{x})$$

Centroide do cluster i

Objeto alocado no cluster i

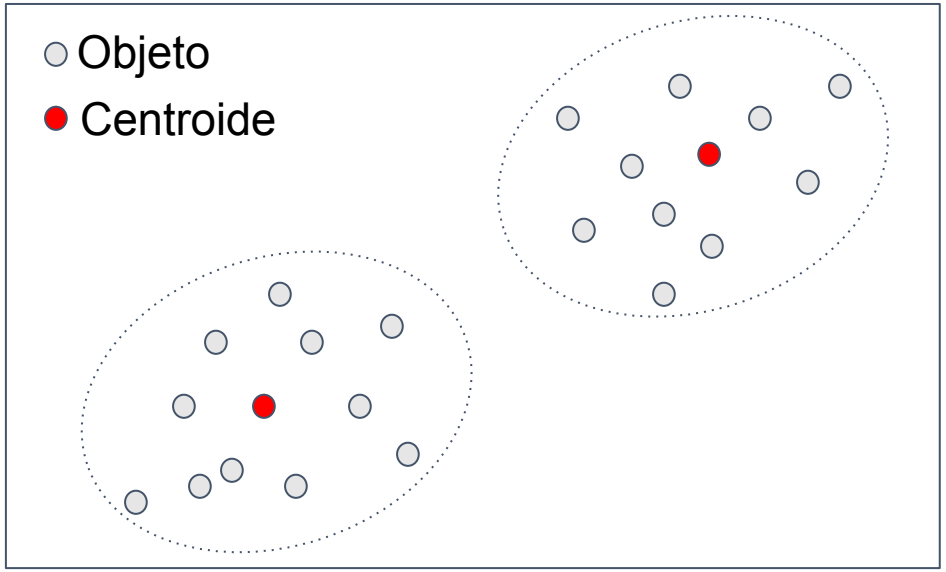


Validação de Agrupamentos

- Índices de validade interna: Erro Quadrático
  - Já estudamos indiretamente essa medida
  - É utilizada pelo *k-means* para escolher a melhor execução após diferentes inicializações de centroides

$$E = \sum_{i=1}^{k} \sum_{\mathbf{x} \in \mathbf{C_i}} d^2(\mu_i, \mathbf{x})$$

Quanto menor o valor de erro *E*, melhor a solução.

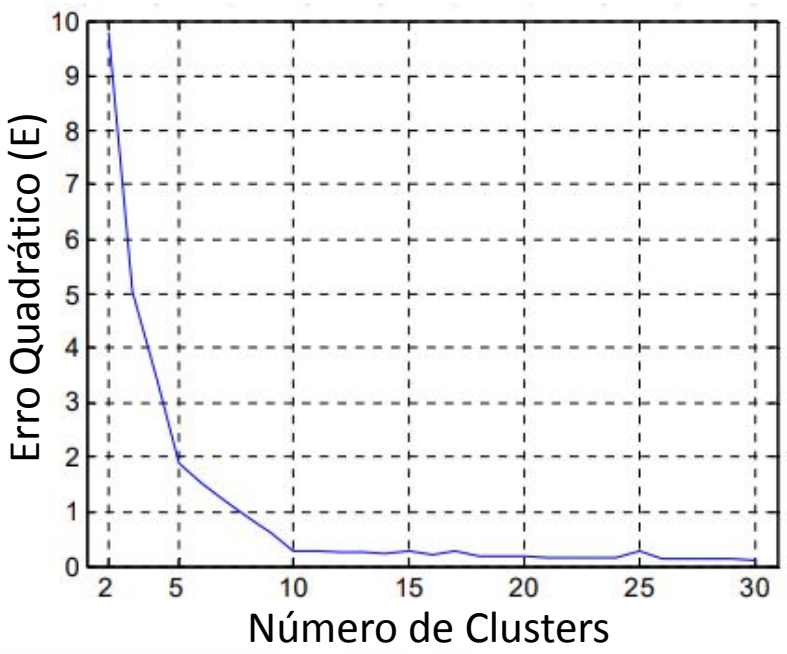


Validação de Agrupamentos

- Índices de validade interna: Erro Quadrático
  - Já estudamos indiretamente essa medida
  - É utilizada pelo *k-means* para escolher a melhor execução após diferentes inicializações de centroides

$$E = \sum_{i=1}^{k} \sum_{\mathbf{x} \in \mathbf{C_i}} d^2(\mu_i, \mathbf{x})$$

O erro quadrático (E) naturalmente é reduzido ao aumentar o número de *clusters*

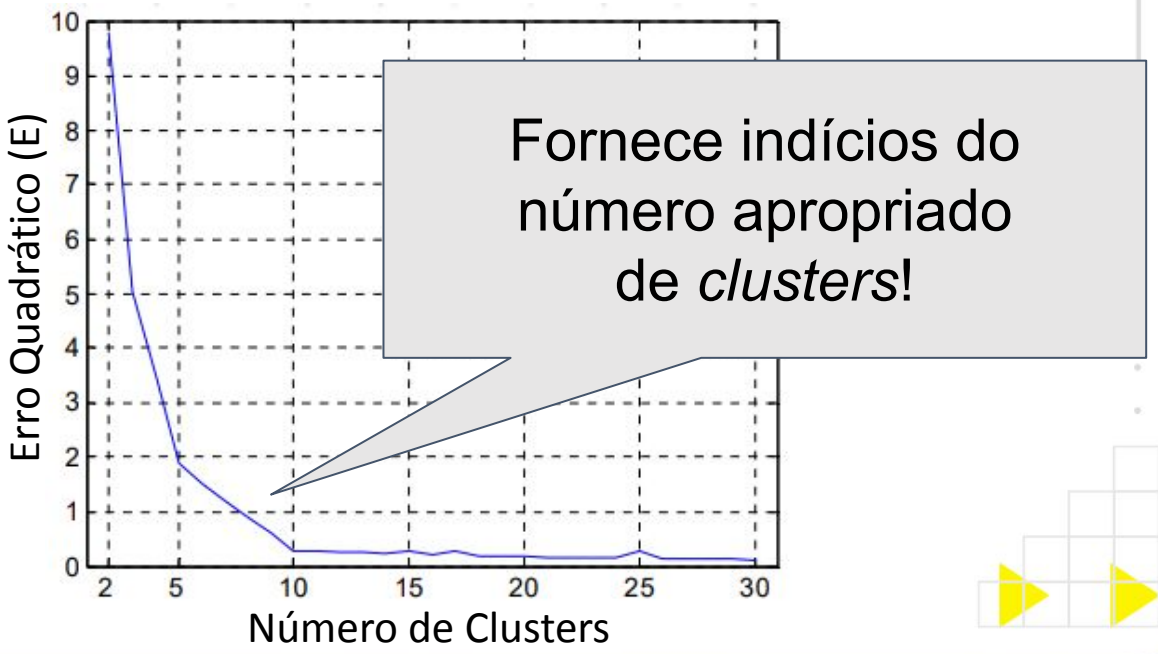


Validação de Agrupamentos

- Índices de validade interna: <u>Erro Quadrático</u>
  - Já estudamos indiretamente essa medida
  - É utilizada pelo *k-means* para escolher a melhor execução após diferentes inicializações de centroides

$$E = \sum_{i=1}^{k} \sum_{\mathbf{x} \in \mathbf{C_i}} d^2(\mu_i, \mathbf{x})$$

O erro quadrático (E) naturalmente é reduzido ao aumentar o número de *clusters*



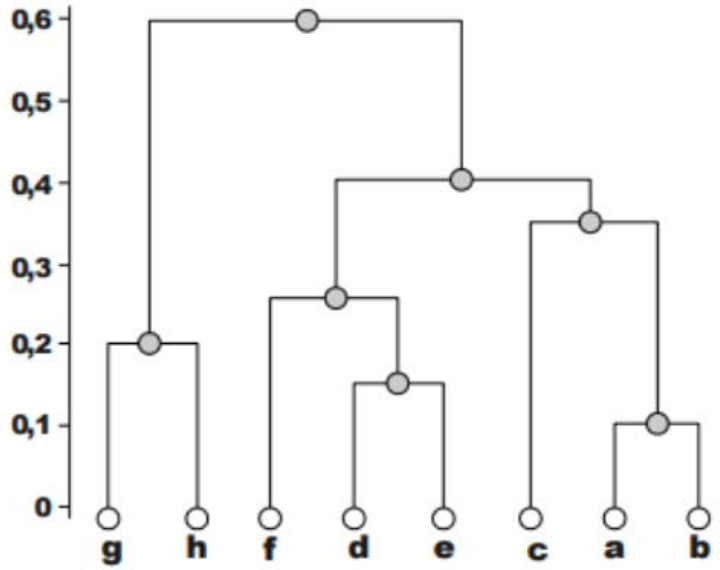
Validação de Agrupamentos

- Índices de validade interna: Correlação Cofenética
  - Avaliação de dendrogramas

	a	b	c	d	e	f	g	h
a	0	0,1	0,35	0,75	0,7	0,65	0,9	1
b	0,1	0	0,8	0,95	0,85	0,7	0,6	0,95
c	0,35	0,8	0	0,4	0,7	1	0,8	0,75
d	0,75	0,95	0,4	0	0,15	0,95	0,85	1
e	0,7	0,85	0,7	0,15	0	0,25	1	0,7
f	0,65	0,7	1	0,95	0,25	0	0,85	1
g	0,9	0,6	0,8	0,85	1	0,85	0	0,2
h	1	0,95	0,75	1	0,7	1	0,2	0

Matriz de Distâncias

Agrupamento Hierárquico



Dendrograma

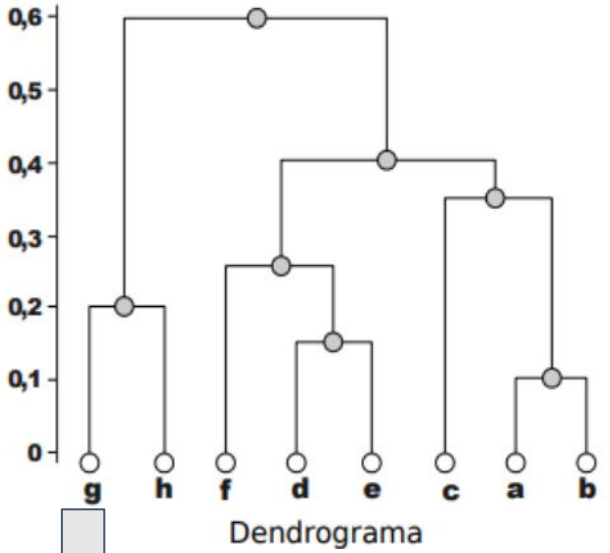
Validação de Agrupamentos

- Índices de validade interna: Correlação Cofenética
  - Avaliação de dendrogramas

	a	b	c	d	e	f	g	h
a	0	0,1	0,35	0,75	0,7	0,65	0,9	1
b	0,1	0	0,8	0,95	0,85	0,7	0,6	0,95
c	0,35	0,8	0	0,4	0,7	1	0,8	0,75
d	0,75	0,95	0,4	0	0,15	0,95	0,85	1
e	0,7	0,85	0,7	0,15	0	0,25	1	0,7
f	0,65	0,7	1	0,95	0,25	0	0,85	1
g	0,9	0,6	0,8	0,85	1	0,85	0	0,2
h	1	0,95	0,75	1	0,7	1	0,2	0

Matriz de Distâncias

Agrupamento Hierárquico



Dendrograma

	a	b	c	d	e	f	g	h
a	0	0,1	0,35	0,4	0,4	0,4	0,6	0,6
b	0,1	0	0,35	0,4	0,4	0,4	0,6	0,6
c	0,35	0,35	0	0,4	0,4	0,4	0,6	0,6
d	0,4	0,4	0,4	0	0,15	0,25	0,6	0,6
e	0,4	0,4	0,4	0,15	0	0,25	0,6	0,6
f	0,4	0,4	0,4	0,25	0,25	0	0,6	0,6
g	0,6	0,6	0,6	0,6	0,6	0,6	0	0,2
h	0,6	0,6	0,6	0,6	0,6	0,6	0,2	0

Cophenetic Difference

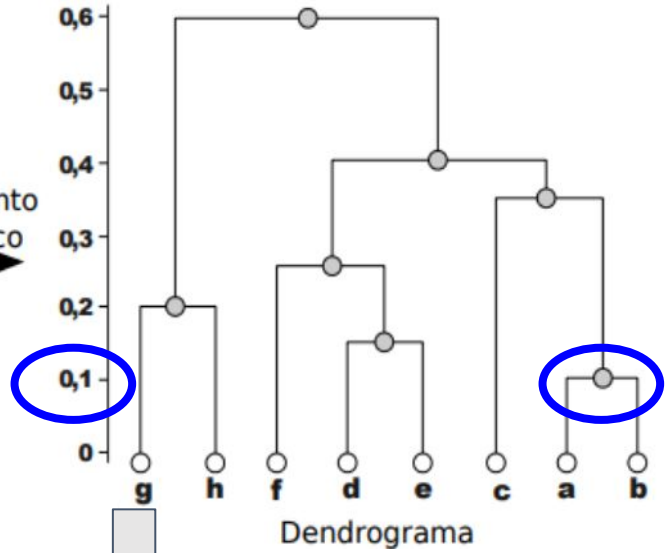
Validação de Agrupamentos

- Índices de validade interna: <u>Correlação Cofenética</u>
  - Avaliação de dendrogramas

	a	b	c	d	e	f	g	h
a	0	0,1	0,35	0,75	0,7	0,65	0,9	1
b	0,1	0	0,8	0,95	0,85	0,7	0,6	0,95
c	0,35	0,8	0	0,4	0,7	1	0,8	0,75
d	0,75	0,95	0,4	0	0,15	0,95	0,85	1
e	0,7	0,85	0,7	0,15	0	0,25	1	0,7
f	0,65	0,7	1	0,95	0,25	0	0,85	1
g	0,9	0,6	0,8	0,85	1	0,85	0	0,2
h	1	0,95	0,75	1	0,7	1	0,2	0

Matriz de Distâncias

Agrupamento Hierárquico



Dendrograma

	a	b	c	d	e	f	g	h
a	0	0,1	0,35	0,4	0,4	0,4	0,6	0,6
b	0,1	0	0,35	0,4	0,4	0,4	0,6	0,6
c	0,35	0,35	0	0,4	0,4	0,4	0,6	0,6
d	0,4	0,4	0,4	0	0,15	0,25	0,6	0,6
e	0,4	0,4	0,4	0,15	0	0,25	0,6	0,6
f	0,4	0,4	0,4	0,25	0,25	0	0,6	0,6
g	0,6	0,6	0,6	0,6	0,6	0,6	0	0,2
h	0,6	0,6	0,6	0,6	0,6	0,6	0,2	0

Cophenetic Difference

A matriz cofenética é construída a partir do dendrograma!

Validação de Agrupamentos

- Índices de validade interna: <u>Correlação Cofenética</u>
  - Avaliação de dendrogramas

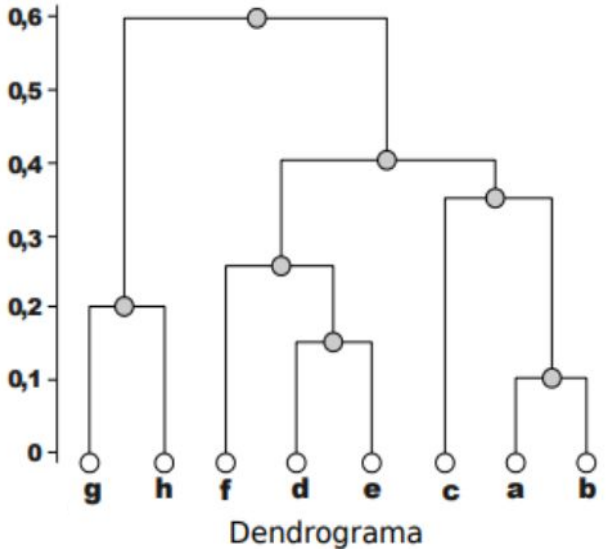
	a	b	c	d	e	f	g	h
a	0	0,1	0,35	0,75	0,7	0,65	0,9	1
b	0,1	0	0,8	0,95	0,85	0,7	0,6	0,95
c	0,35	0,8	0	0,4	0,7	1	0,8	0,75
d	0,75	0,95	0,4	0	0,15	0,95	0,85	1
e	0,7	0,85	0,7	0,15	0	0,25	1	0,7
f	0,65	0,7	1	0,95	0,25	0	0,85	1
g	0,9	0,6	0,8	0,85	1	0,85	0	0,2
h	1	0,95	0,75	1	0,7	1	0,2	0

Matriz de Distâncias

	a	b	c	d	e	f	g	h
a	0	0,1	0,35	0,4	0,4	0,4	0,6	0,6
b	0,1	0	0,35	0,4	0,4	0,4	0,6	0,6
c	0,35	0,35	0	0,4	0,4	0,4	0,6	0,6
d	0,4	0,4	0,4	0	0,15	0,25	0,6	0,6
e	0,4	0,4	0,4	0,15	0	0,25	0,6	0,6
f	0,4	0,4	0,4	0,25	0,25	0	0,6	0,6
g	0,6	0,6	0,6	0,6	0,6	0,6	0	0,2
h	0,6	0,6	0,6	0,6	0,6	0,6	0,2	0

Cophenetic Difference

Agrupamento Hierárquico



Dendrograma

Calcular a correlação entre as duas matrizes! (Correlação de Pearson)

Validação de Agrupamentos

- Índices de validade interna: Correlação Cofenética
  - Avaliação de dendrogramas

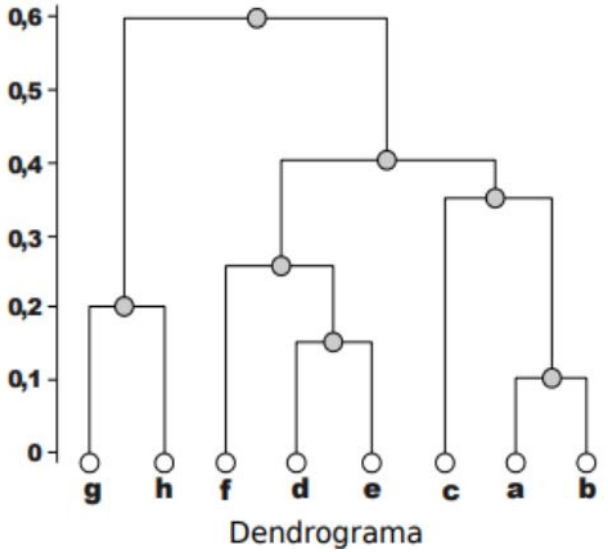
	a	b	c	d	e	f	g	h
a	0	0,1	0,35	0,75	0,7	0,65	0,9	1
b	0,1	0	0,8	0,95	0,85	0,7	0,6	0,95
c	0,35	0,8	0	0,4	0,7	1	0,8	0,75
d	0,75	0,95	0,4	0	0,15	0,95	0,85	1
e	0,7	0,85	0,7	0,15	0	0,25	1	0,7
f	0,65	0,7	1	0,95	0,25	0	0,85	1
g	0,9	0,6	0,8	0,85	1	0,85	0	0,2
h	1	0,95	0,75	1	0,7	1	0,2	0

Matriz de Distâncias

	a	b	c	d	e	f	g	h
a	0	0,1	0,35	0,4	0,4	0,4	0,6	0,6
b	0,1	0	0,35	0,4	0,4	0,4	0,6	0,6
c	0,35	0,35	0	0,4	0,4	0,4	0,6	0,6
d	0,4	0,4	0,4	0	0,15	0,25	0,6	0,6
e	0,4	0,4	0,4	0,15	0	0,25	0,6	0,6
f	0,4	0,4	0,4	0,25	0,25	0	0,6	0,6
g	0,6	0,6	0,6	0,6	0,6	0,6	0	0,2
h	0,6	0,6	0,6	0,6	0,6	0,6	0,2	0

Cophenetic Difference

Agrupamento Hierárquico



Dendrograma

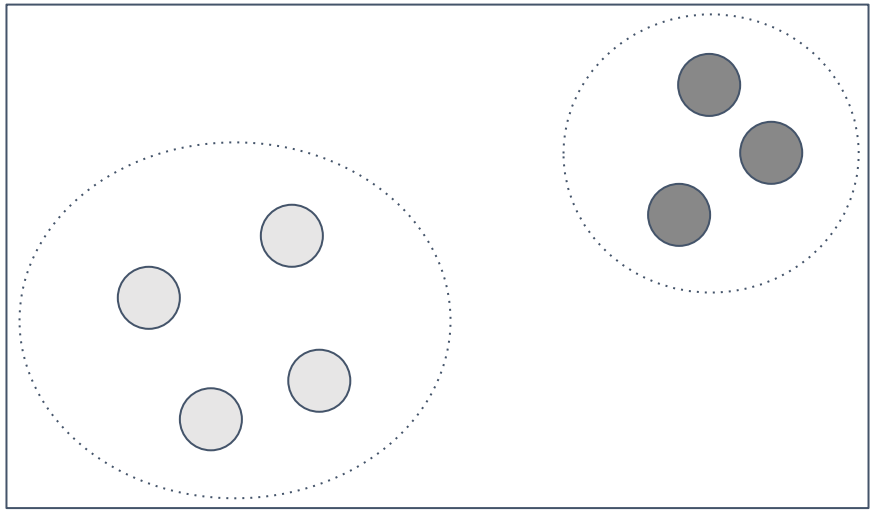
Quanto mais próximo de 1, melhor a qualidade do dendrograma

Validação de Agrupamentos

- Índices de validade relativa: Silhueta
  - Avaliar a qualidade de uma partição (*clusters*)
  - Comparar partições obtidas por diferentes algoritmos
  - Determinar o número apropriado de *clusters*
  - Verificar se um objeto está bem alocado no seu *cluster*
  - Visualização (diagrama de silhueta)

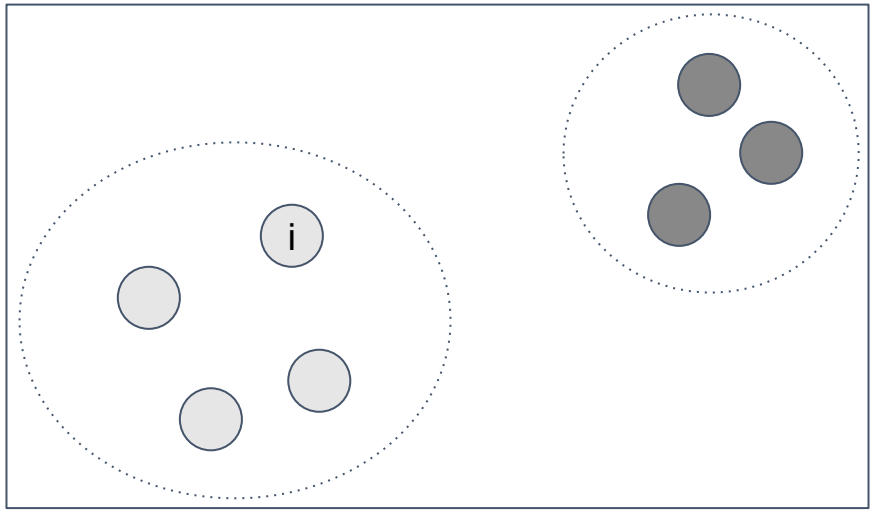
Validação de Agrupamentos

- Índices de validade relativa: Silhueta



Validação de Agrupamentos

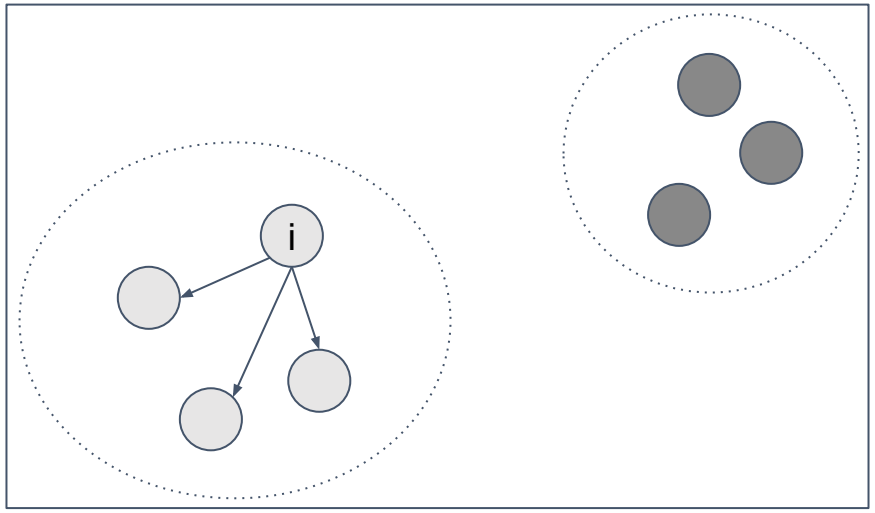
- Índices de validade relativa: Silhueta



1. O quão bem o objeto *i* está alocado em seu próprio *cluster*?

Validação de Agrupamentos

- Índices de validade relativa: Silhueta

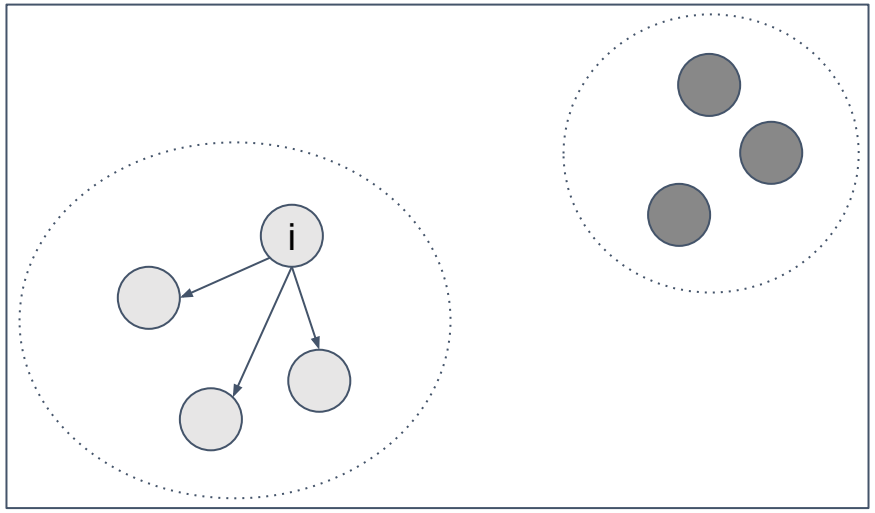


1. O quão bem o objeto *i* está alocado em seu próprio *cluster*?

a(i) = distância média entre o objeto *i* e todos os outros objetos do seu *cluster.*

Validação de Agrupamentos

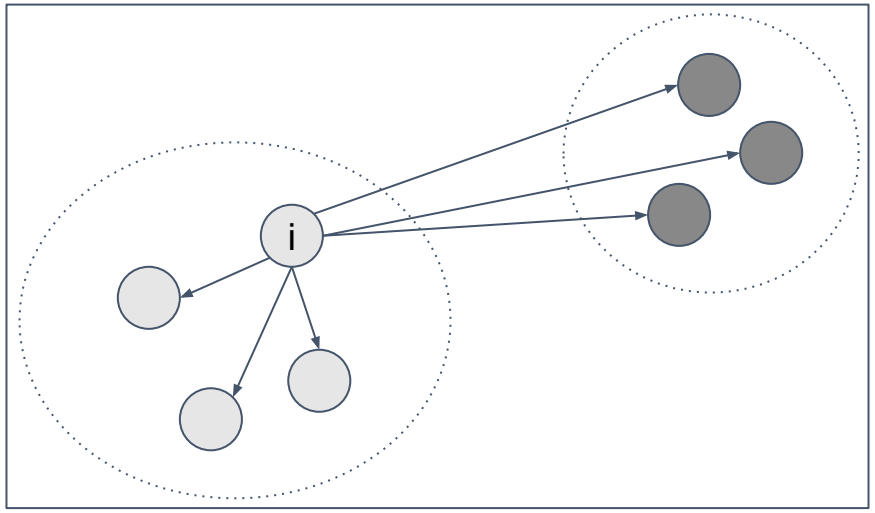
- Índices de validade relativa: Silhueta



2. O quão próximo o objeto i está do seu *cluster* vizinho?

Validação de Agrupamentos

- Índices de validade relativa: Silhueta

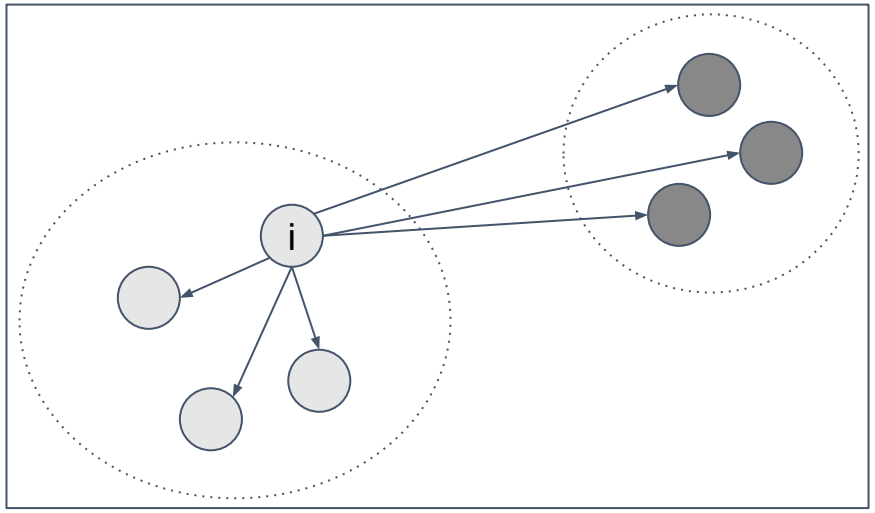


2. O quão próximo o objeto i está do seu *cluster* vizinho?

b(i) = distância média entre o objeto *i* e todos os outros objetos do *cluster* vizinho.

Validação de Agrupamentos

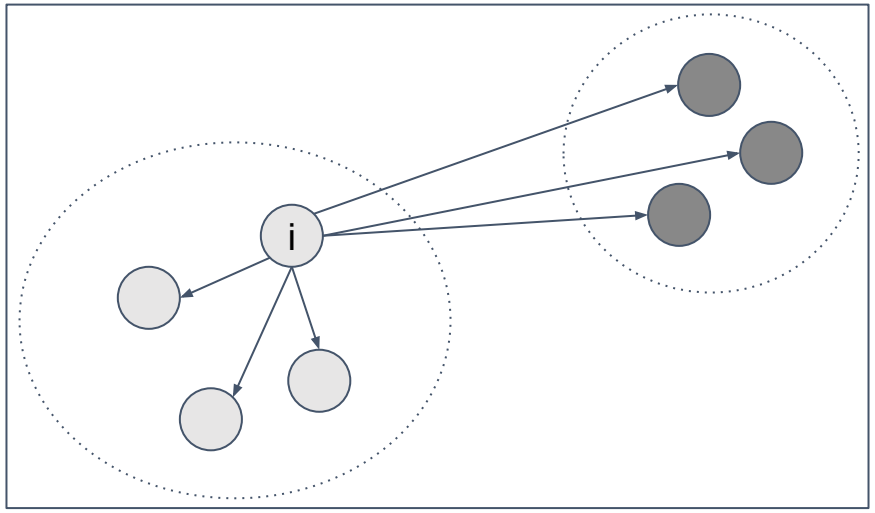
- Índices de validade relativa: Silhueta



3. Qual é o valor do índice de silhueta do objeto *i*?

Validação de Agrupamentos

- Índices de validade relativa: Silhueta



3. Qual é o valor do índice de silhueta do objeto *i*?

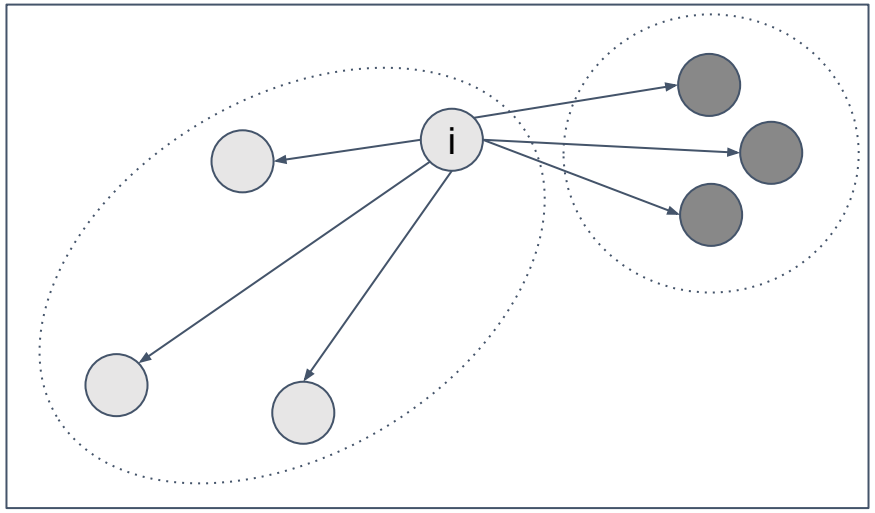
$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

a(i) = distância média entre o objeto *i* e todos os outros objetos do seu *cluster.*

b(i) = distância média entre o objeto *i* e todos os outros objetos do *cluster* vizinho.

Validação de Agrupamentos

- Índices de validade relativa: Silhueta



3. Qual é o valor do índice de silhueta do objeto *i*?

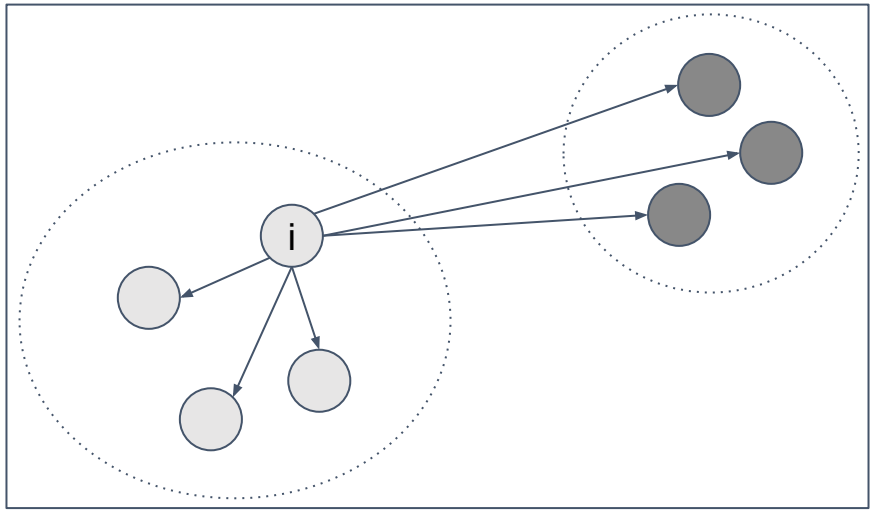
$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

a(i) = distância média entre o objeto *i* e todos os outros objetos do seu *cluster.*

b(i) = distância média entre o objeto *i* e todos os outros objetos do *cluster* vizinho.

Validação de Agrupamentos

- Índices de validade relativa: Silhueta

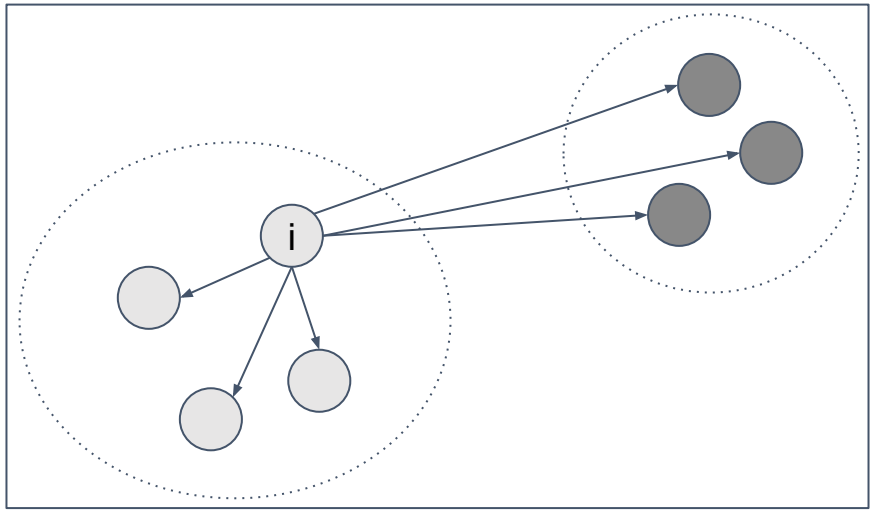


4. Calcular a silhueta de todos os objetos e computar a silhueta média

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

Validação de Agrupamentos

- Índices de validade relativa: Silhueta



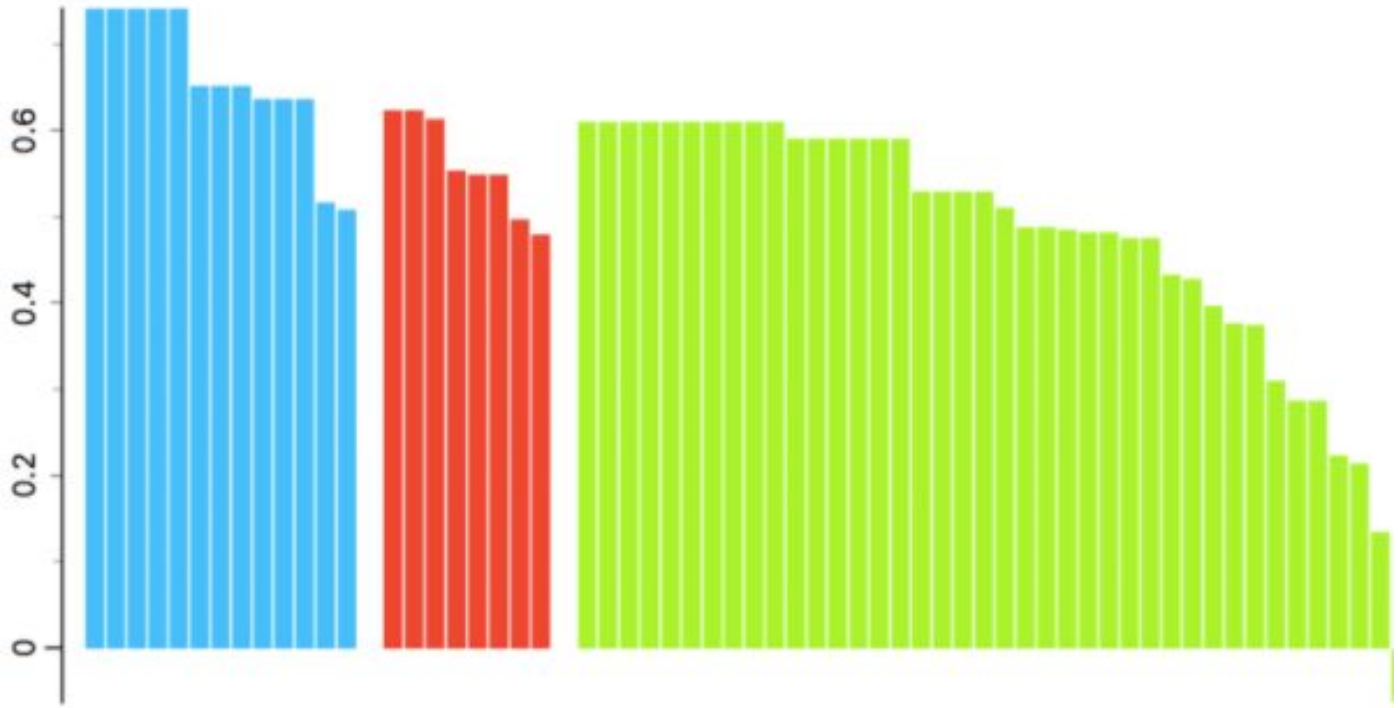
4. Calcular a silhueta de todos os objetos e computar a silhueta média

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

A silhueta média S, $-1 \leq S \leq 1$, indica a qualidade geral do agrupamento.

Validação de Agrupamentos

- Índices de validade relativa: Silhueta
  - Ordenar os objetos por cluster e por valor de silhueta, fornece o diagrama de silhueta.



Validação de Agrupamentos

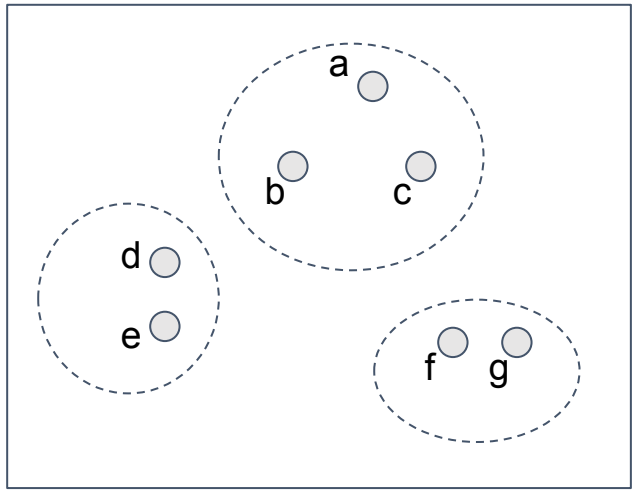
- Índices de validade externa
  - Comparar nossa partição com uma partição de referência
  - Exige uma rotulacão prévia de uma parcela dos dados
- Uso prático?
  - Cenários e experimentos controlados

Validação de Agrupamentos

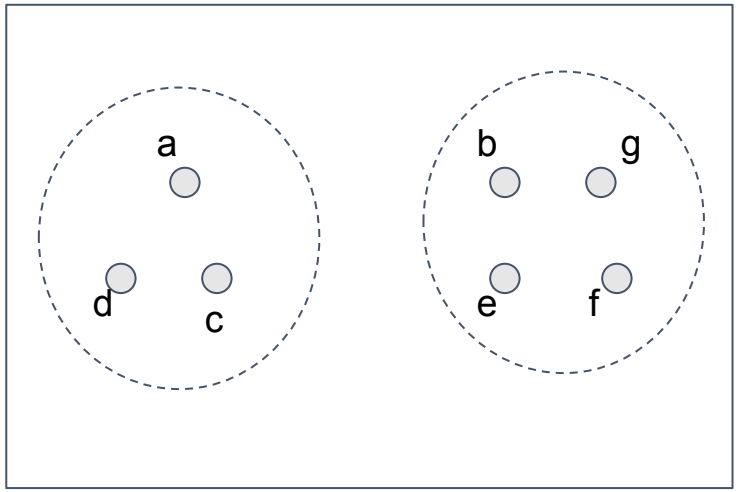
- Índices de validade externa
  - Considere:
    - P = clusters obtidos pelo seu algoritmo
    - R = clusters de referência (e.g. anotado por humanos)

Validação de Agrupamentos

- Índices de validade externa: RAND Index



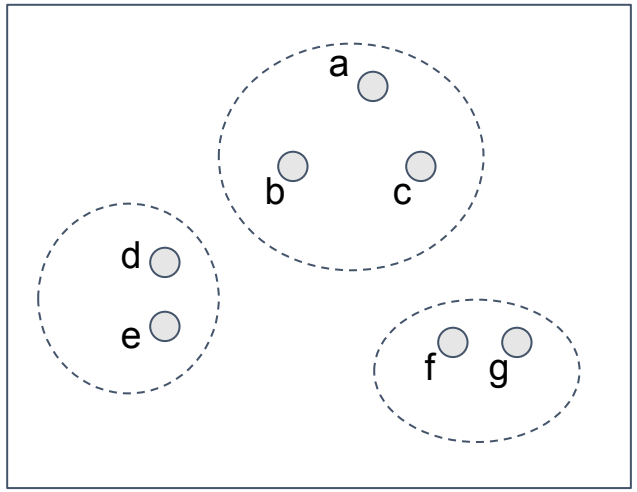
Clusters obtidos pelo algoritmo (P)



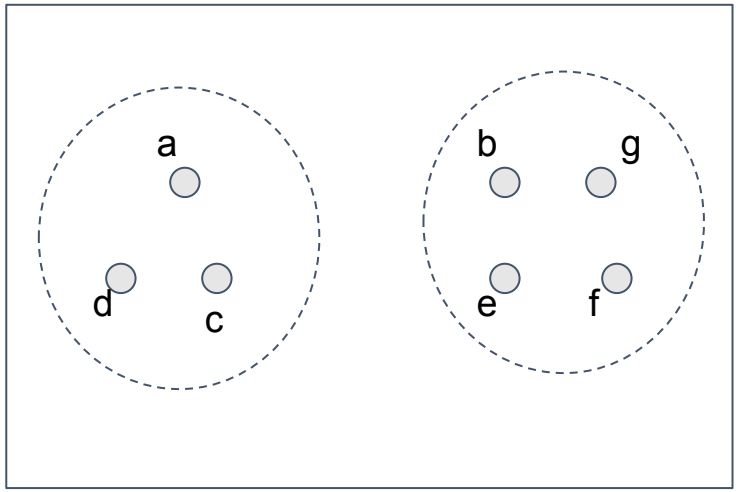
Clusters de Referência (R)

Validação de Agrupamentos

- Índices de validade externa: RAND Index



Clusters obtidos pelo algoritmo (P)



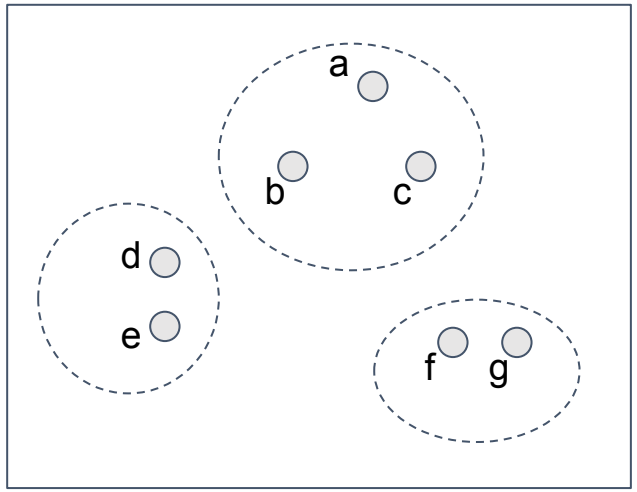
Clusters de Referência (R)

A	B
C	D

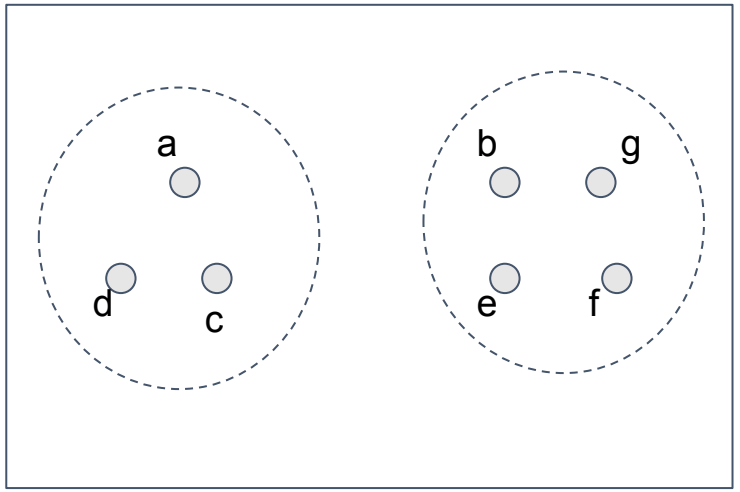
A = quantidade de vezes em que objetos *i* e *j* estão no mesmo cluster em P e no mesmo cluster em R

Validação de Agrupamentos

- Índices de validade externa: RAND Index



Clusters obtidos pelo algoritmo (P)



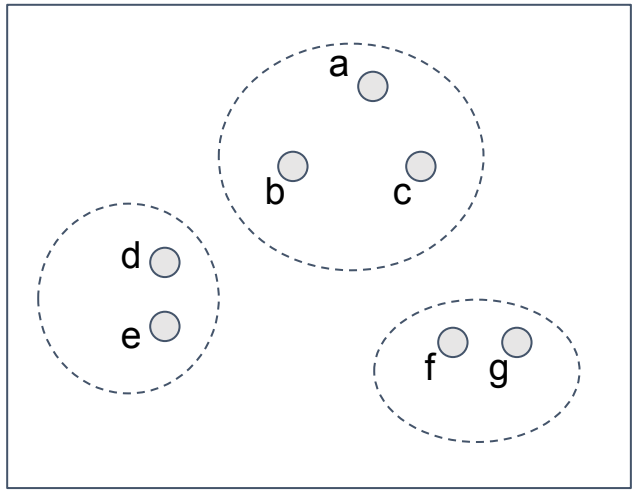
Clusters de Referência (R)

A=2	B
C	D

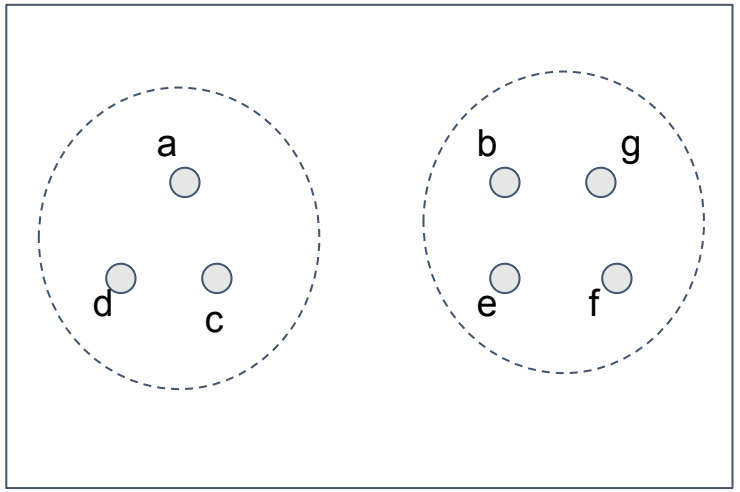
A = quantidade de vezes em que objetos *i* e *j* estão no mesmo cluster em P e no mesmo cluster em R

Validação de Agrupamentos

- Índices de validade externa: RAND Index



Clusters obtidos pelo algoritmo (P)



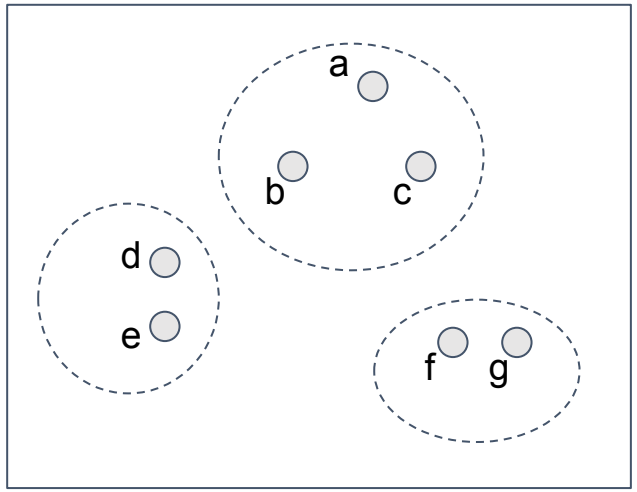
Clusters de Referência (R)

A=2	B=3
C	D

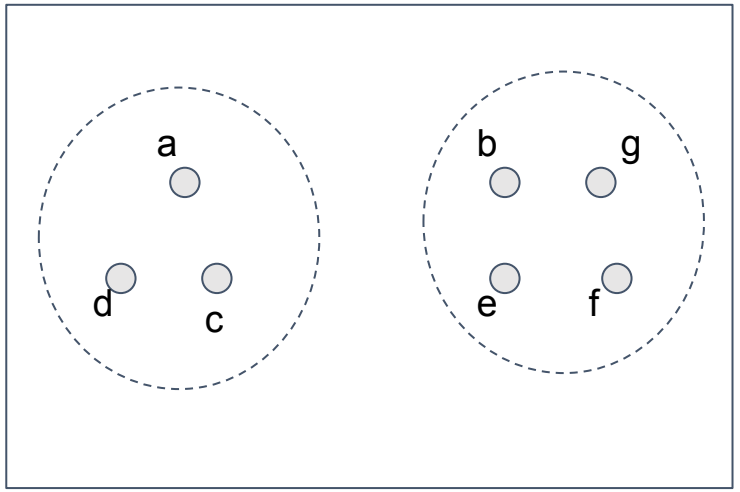
B = quantidade de vezes em que objetos *i* e *j* estão no mesmo cluster em P e em clusters diferentes em R

Validação de Agrupamentos

- Índices de validade externa: RAND Index



Clusters obtidos pelo algoritmo (P)



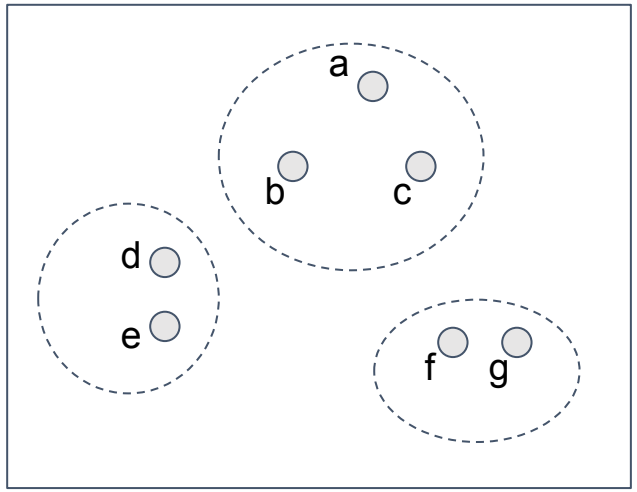
Clusters de Referência (R)

A=2	B=3
C=7	D=9

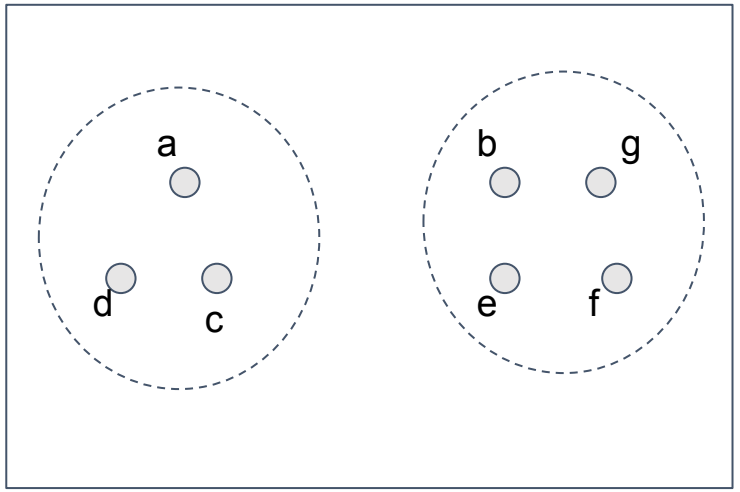
C = quantidade de vezes em que objetos *i* e *j* estão clusters diferentes em P e no mesmo cluster em R

Validação de Agrupamentos

- Índices de validade externa: RAND Index



Clusters obtidos pelo algoritmo (P)



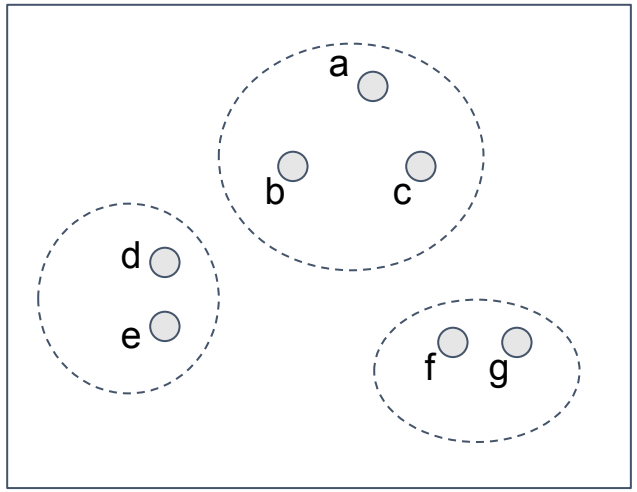
Clusters de Referência (R)

A=2	B=3
C=7	D=9

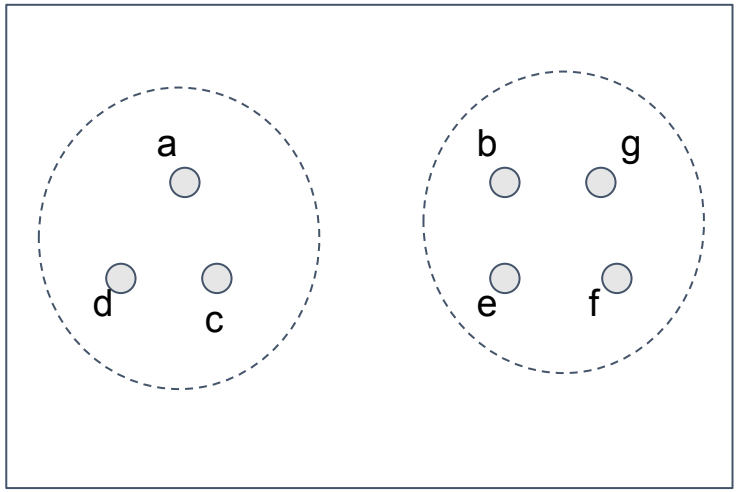
D = quantidade de vezes em que objetos *i* e *j* estão clusters diferentes em P e em clusters diferentes em R

Validação de Agrupamentos

- Índices de validade externa: RAND Index



Clusters obtidos pelo algoritmo (P)



Clusters de Referência (R)

A=2	B=3
C=7	D=9

$$RAND = \frac{A+D}{A+B+C+D}$$

Validação de Agrupamentos

- Analisar o “mérito” e qualidade dos clusters
- Validação por inspeção visual
- Índices de validação de agrupamentos
  - Índices internos
  - Índices relativos
  - Índices externos

Eficiência computacional?

Bibliografia

Rezende, S. O. (2003). *Sistemas inteligentes: fundamentos e aplicações*. Editora Manole Ltda.

Tan, P.N.; Steinbach, M.; Karpatne, A.; Kumar, V. (2016). *Introduction to Data Mining (2nd Edition)*. Pearson.