Agrupamento de Textos e suas Aplicações em Inteligência Analítica

Agrupamento de Textos: k-Means e Bisecting k-Means

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

Cursos de Extensão – Difusão de Conhecimento – Dezembro de 2021

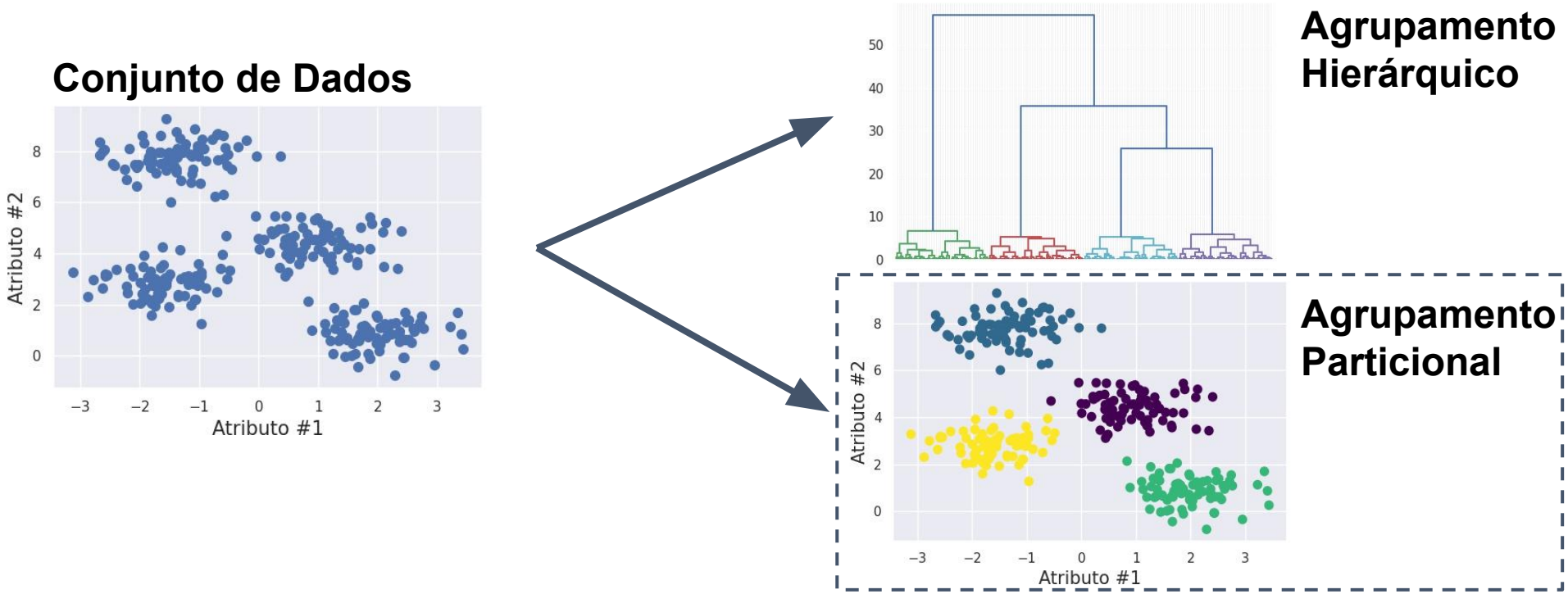
ICMC SÃO CARLOS USP

LABIC Laboratório de Inteligência Computacional ICMC.USP SÃO CARLOS

WebSensors

Agrupamento de Textos

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



Agrupamento Particional

- Falaremos sobre métodos de agrupamento para obter partições rígidas dos dados
- **Partição rígida:** clusters não possuem sobreposição
  - Dado um conjunto de ***n*** documentos

$$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$$

Agrupamento Particional

- Falaremos sobre métodos de agrupamento para obter partições rígidas dos dados
- **Partição rígida:** clusters não possuem sobreposição
  - Dado um conjunto de ***n*** documentos

$$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$$

  - Obter um agrupamento *C* em *k clusters*

$$C = \{C_1, C_2, \cdots, C_k\}$$
$$C_1 \cup C_2 \cup \cdots \cup C_k = \mathbf{X}$$

Agrupamento Particional

- Falaremos sobre métodos de agrupamento para obter partições rígidas dos dados
- **Partição rígida:** clusters não possuem sobreposição
  - Dado um conjunto de ***n*** documentos

$$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$$

  - Obter um agrupamento *C* em *k clusters*

$$C = \{C_1, C_2, \cdots, C_k\}$$
$$C_1 \cup C_2 \cup \cdots \cup C_k = \mathbf{X}$$

  - Sem clusters vazios e sem sobreposição

$$C_i \neq \emptyset \qquad\qquad C_i \cap C_j = \emptyset \text{ para } i \neq j$$

Algoritmo *k*-Médias ou *k-Means*

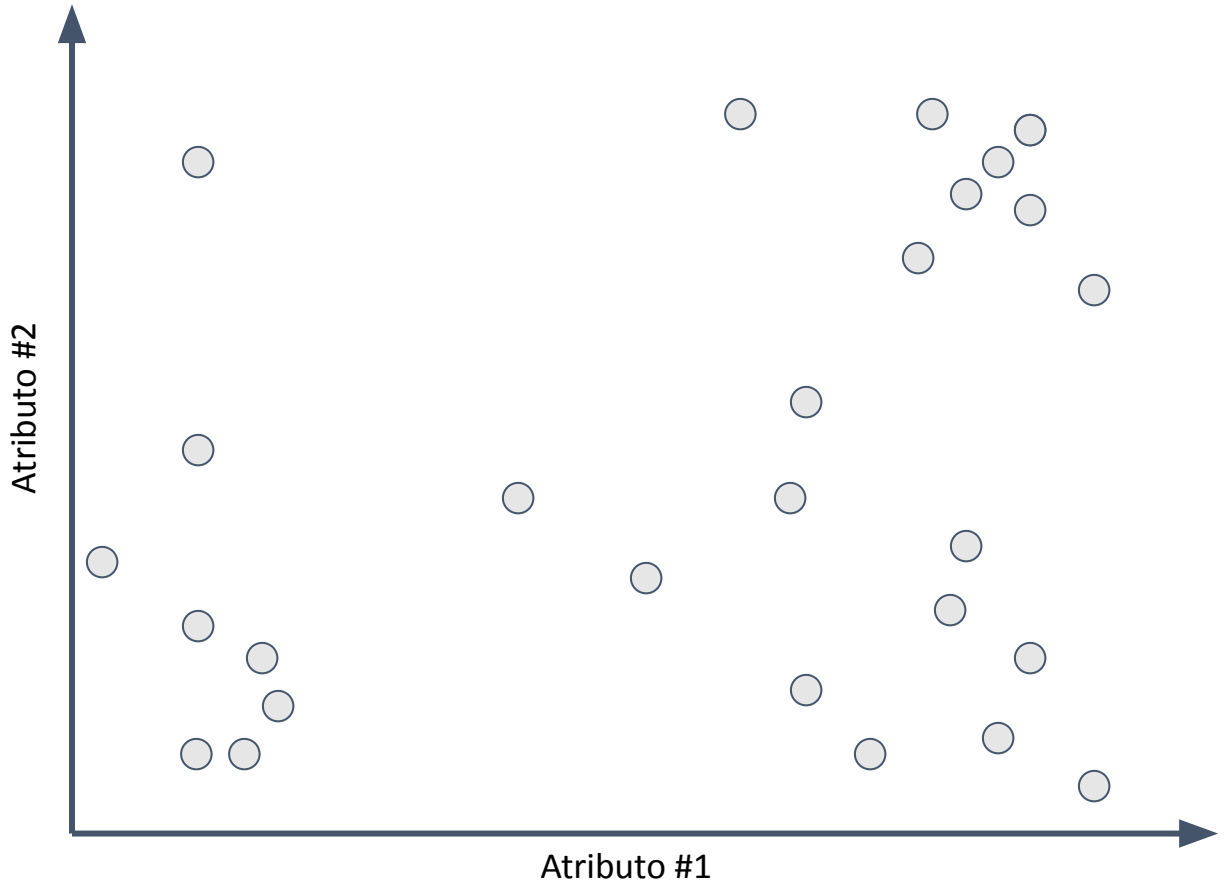
- Amplamente usado na indústria e academia
- Características desejáveis para Mineração de Dados
  - Simplicidade
  - Interpretabilidade
  - Eficiência Computacional

Algoritmo *k*-Médias ou *k-Means*

- Amplamente usado na indústria e academia
- Características desejáveis para Mineração de Dados
  - Simplicidade
  - Interpretabilidade
  - Eficiência Computacional

Vamos começar a estudar o *k-Means* a partir de um exemplo didático...

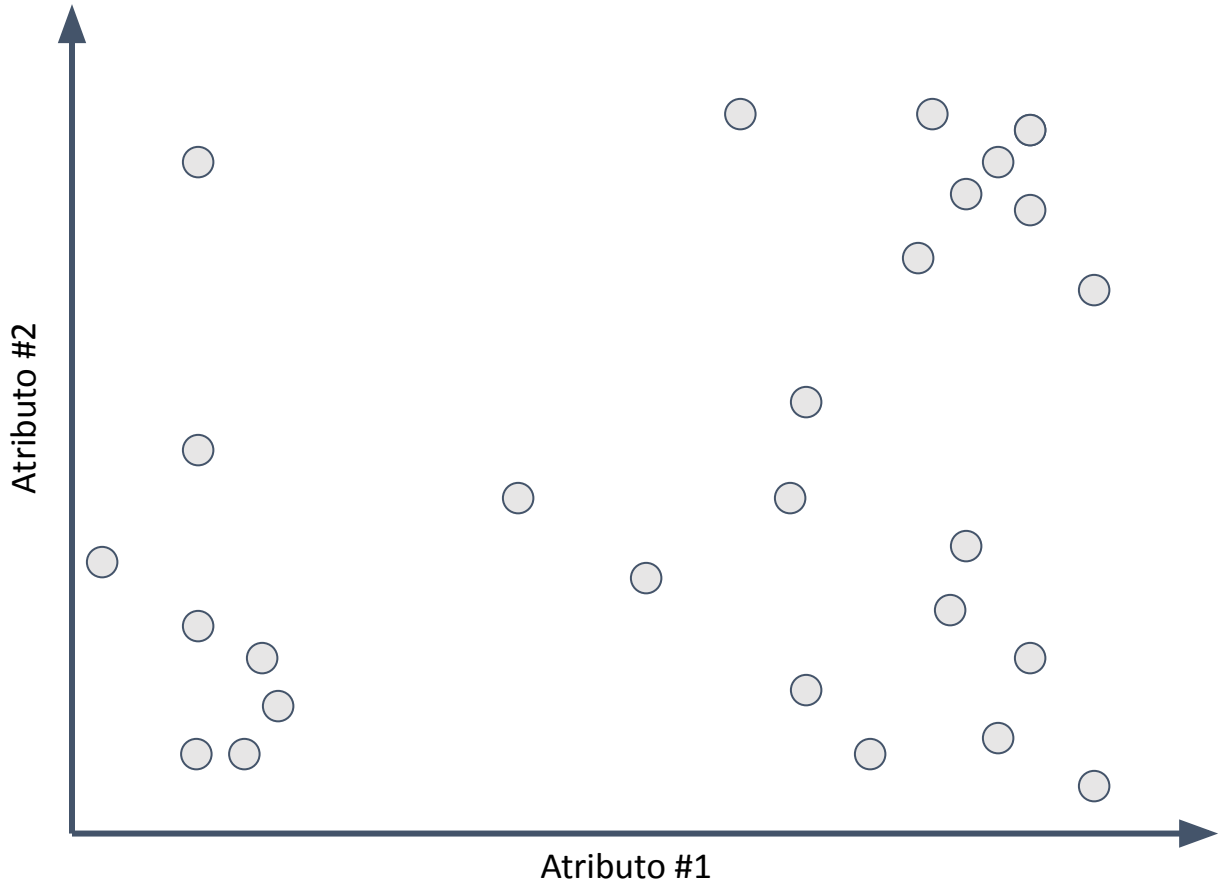
Algoritmo *k-Means*



Exemplo adaptado de Gregory Piatetsky-Shapiro & Gary Parker (www.kdnuggets.com)

Algoritmo *k-Means*

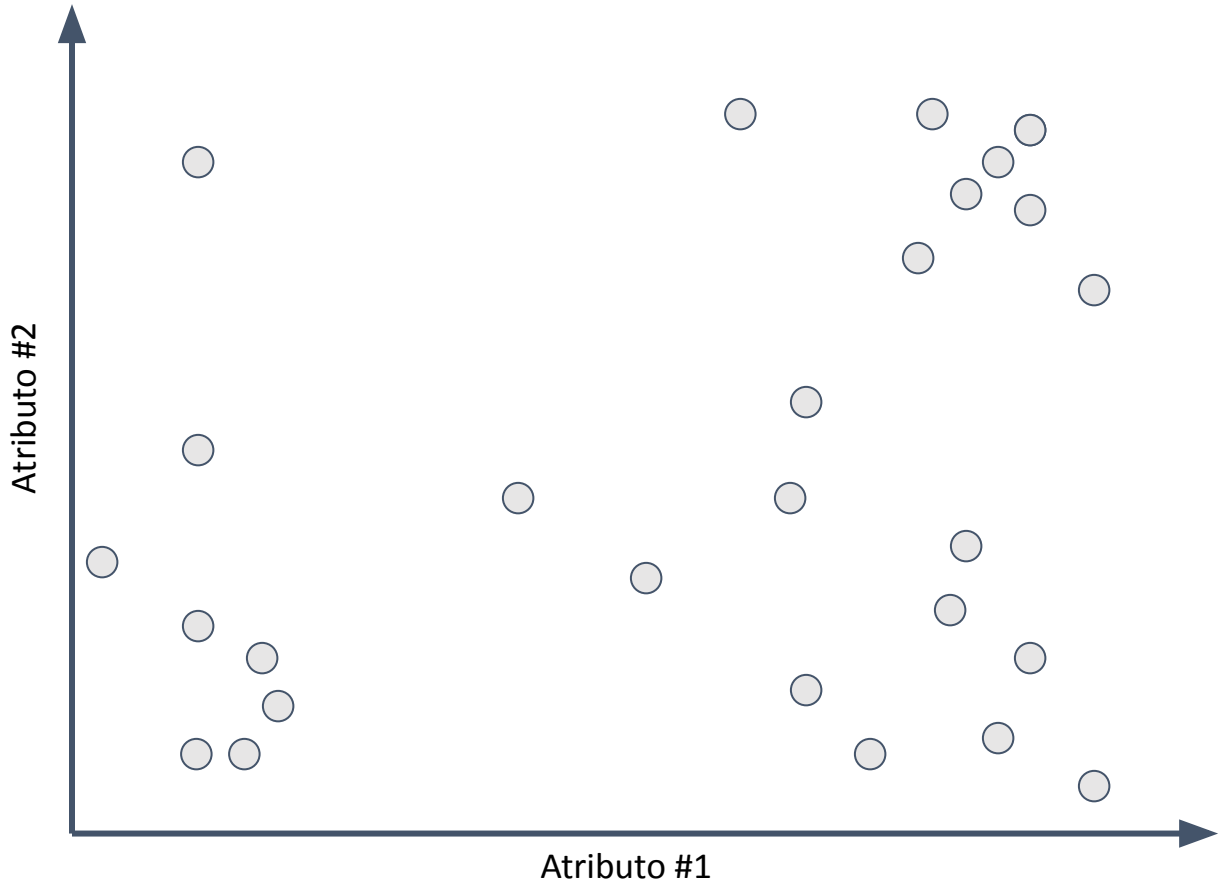
Primeiro passo é definir o número *k* de *clusters* que se deseja encontrar



Algoritmo *k-Means*

Primeiro passo é definir o número *k* de *clusters* que se deseja encontrar

Vamos definir **k=3**

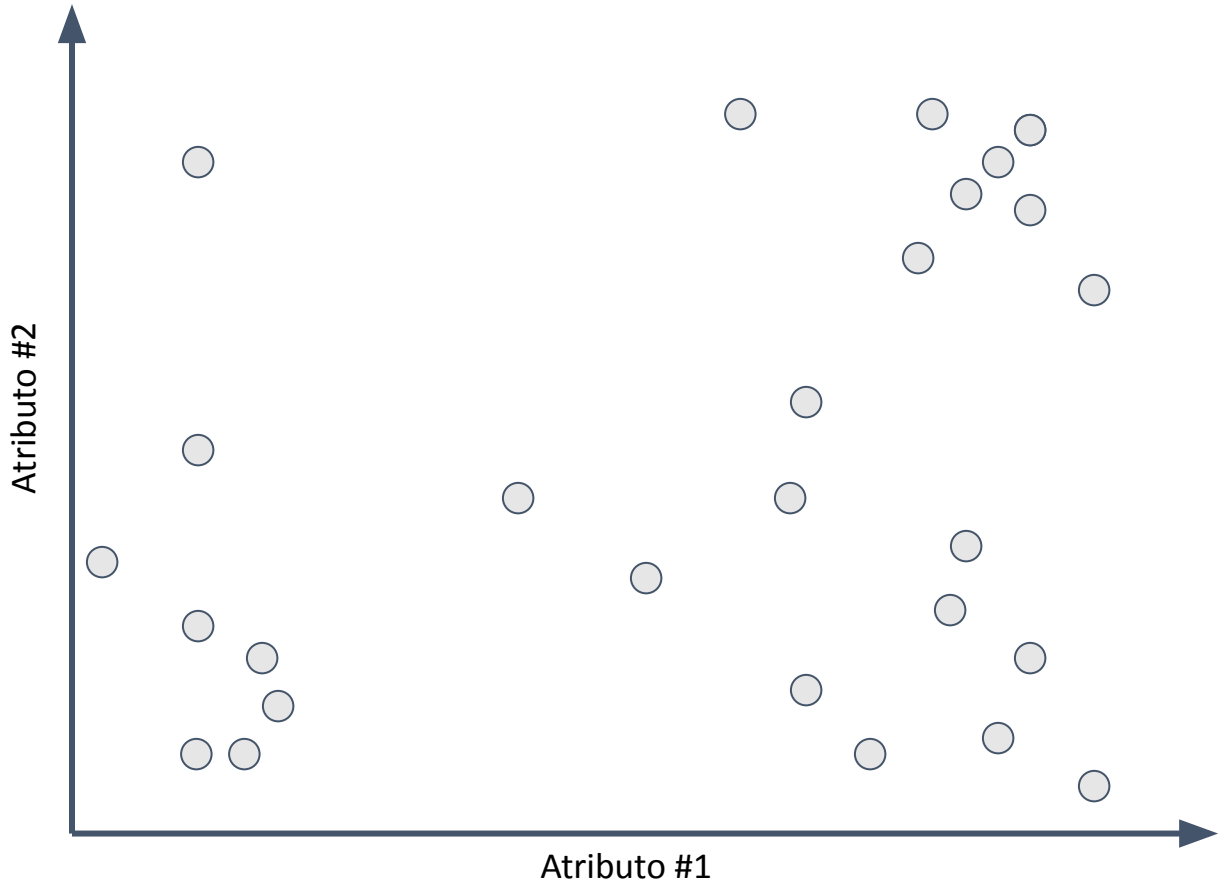


Algoritmo *k-Means*

Com k=3, nós vamos inicializar *k* centroides.

Cada centroide é um ponto e representa um cluster.

Inicialização aleatória de centroides.

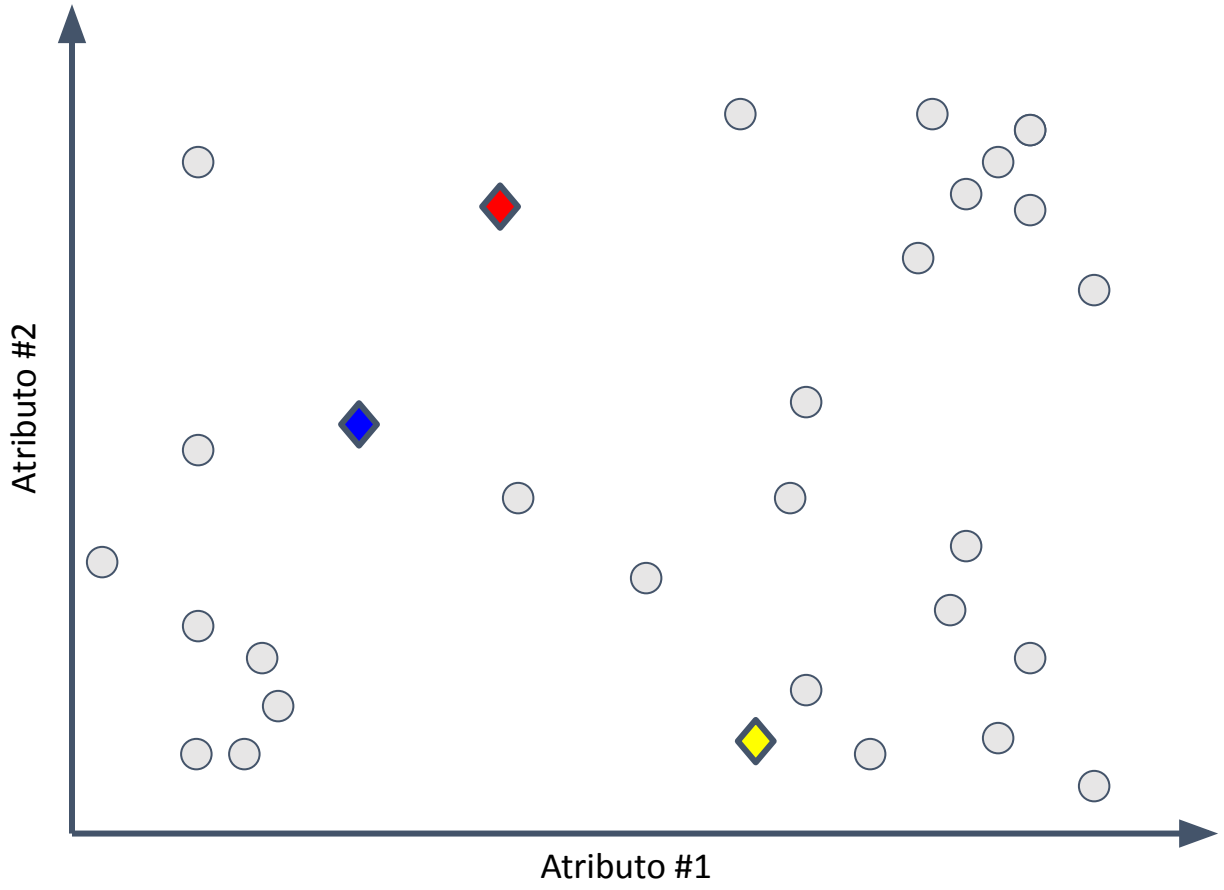


Algoritmo *k-Means*

Com k=3, nós vamos inicializar *k* centroides.

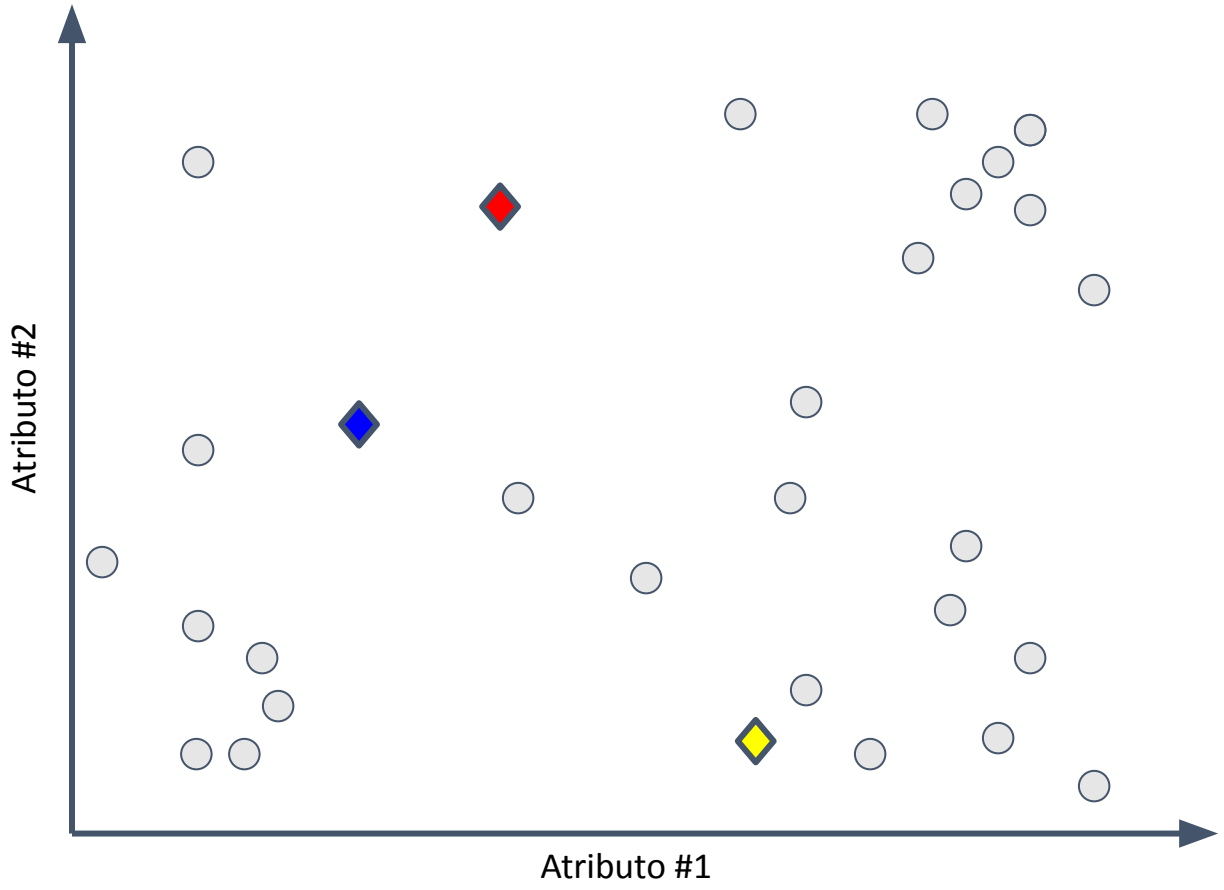
Cada centroide é um ponto e representa um cluster.

Inicialização aleatória de centroides.



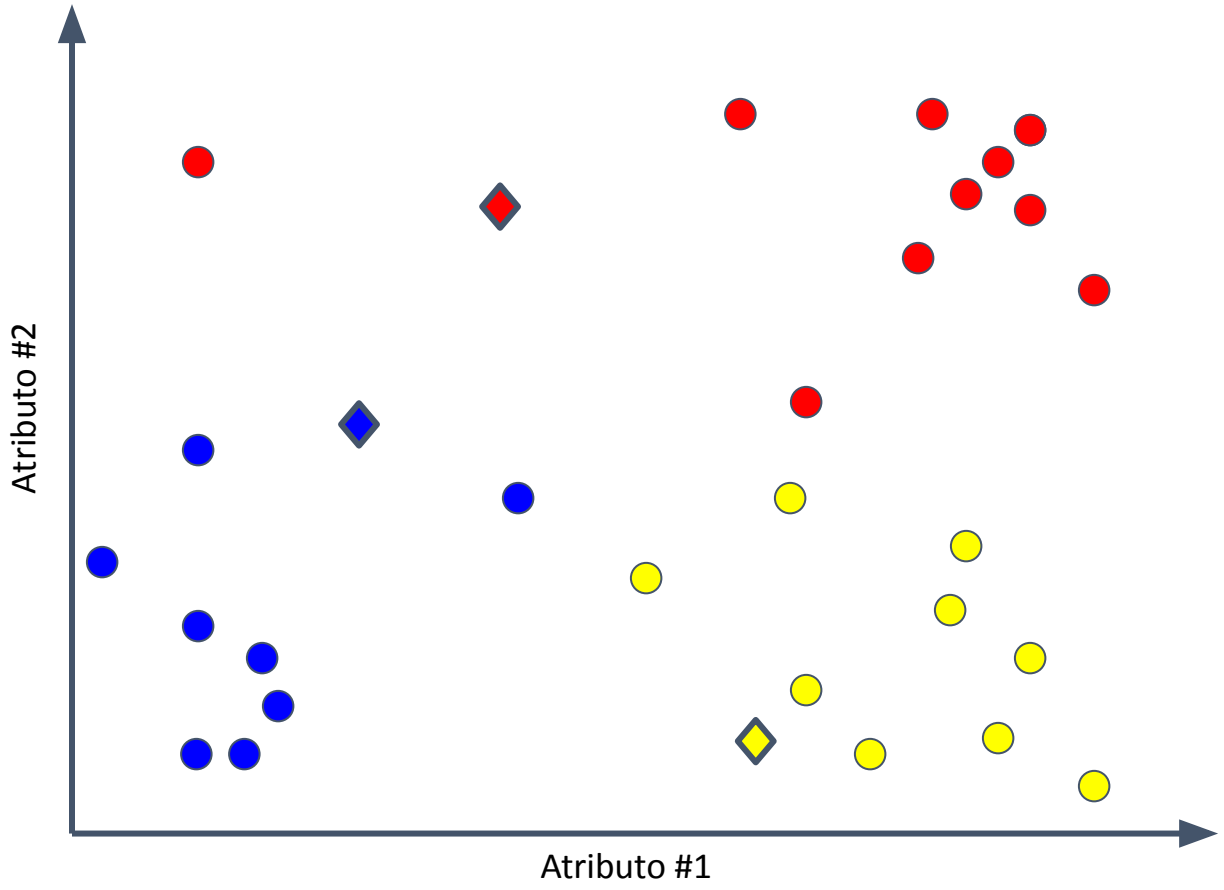
Algoritmo *k-Means*

Agora, nós associamos cada objeto ao centróide mais próximo.



Algoritmo *k-Means*

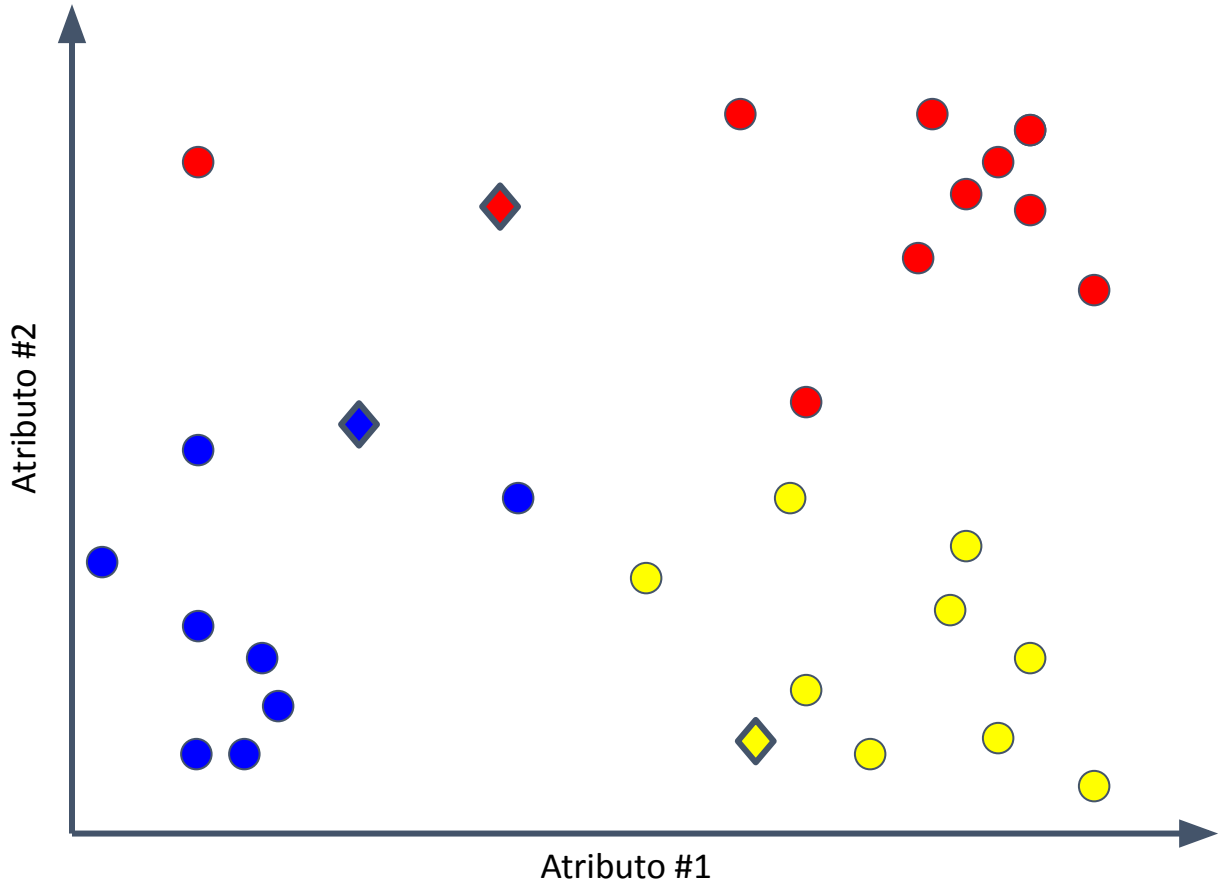
Agora, nós associamos cada objeto ao centróide mais próximo.



Algoritmo *k-Means*

Agora, nós atualizamos o centroide de cada cluster.

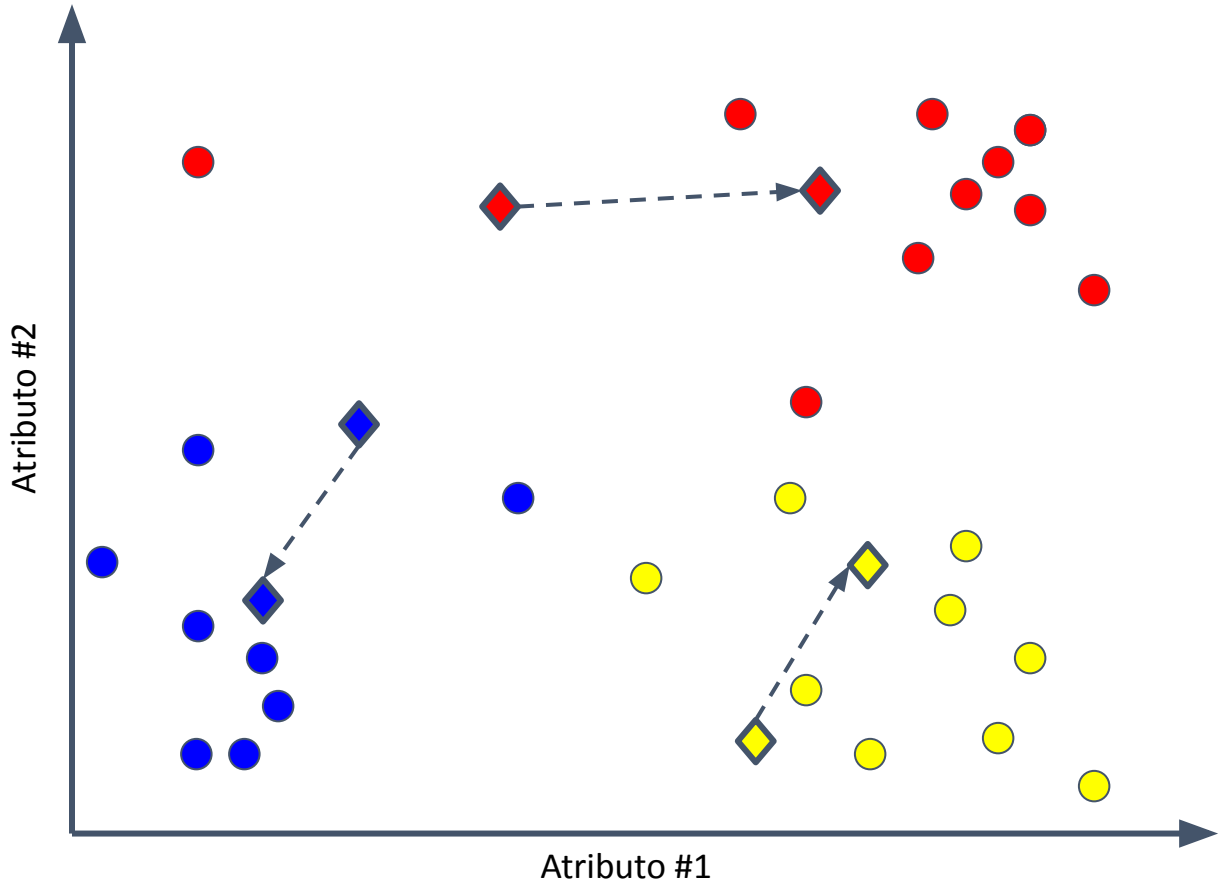
O centroide é calculado como o vetor médio do cluster.



Algoritmo *k-Means*

Agora, nós atualizamos o centroide de cada cluster.

O centroide é calculado como o vetor médio do cluster.

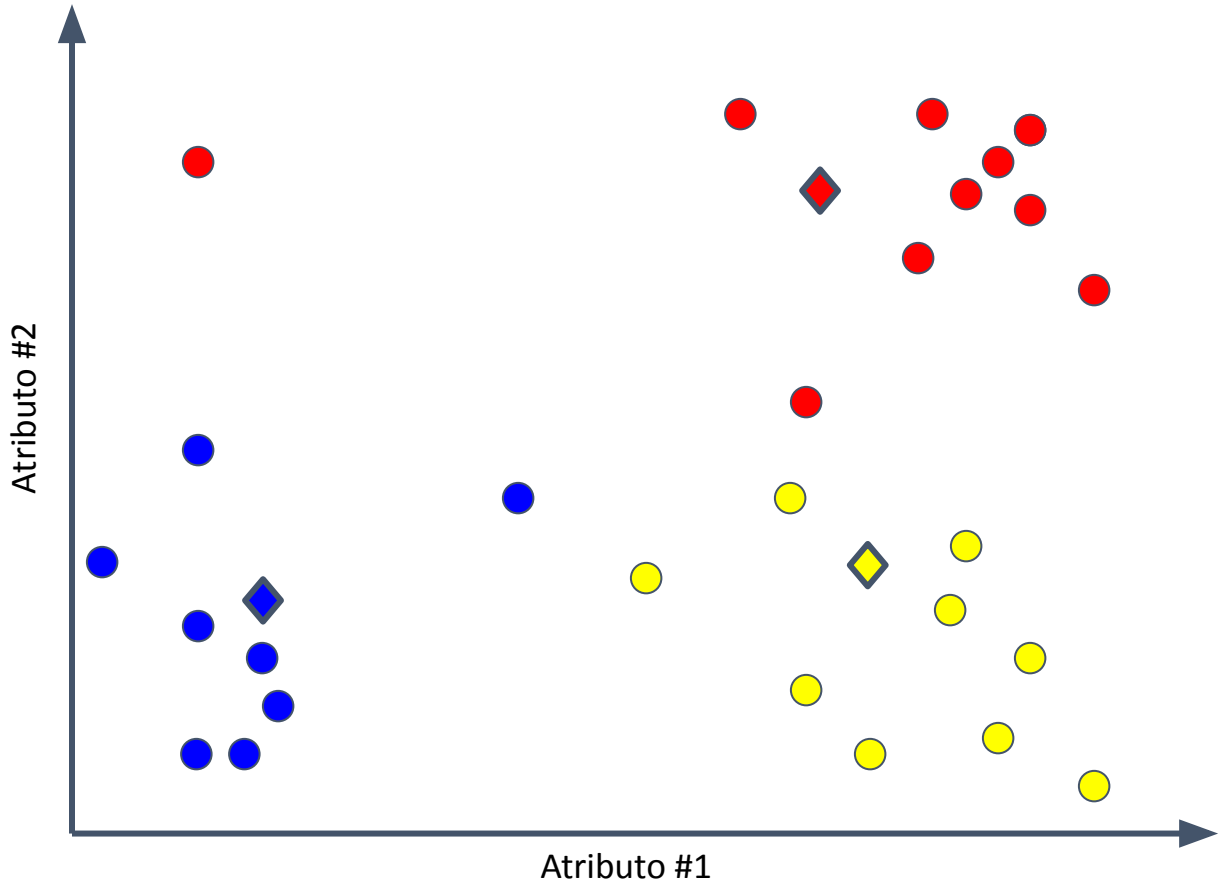


Algoritmo *k-Means*

Agora, nós atualizamos o centroide de cada cluster.

O centroide é calculado como o vetor médio do cluster.

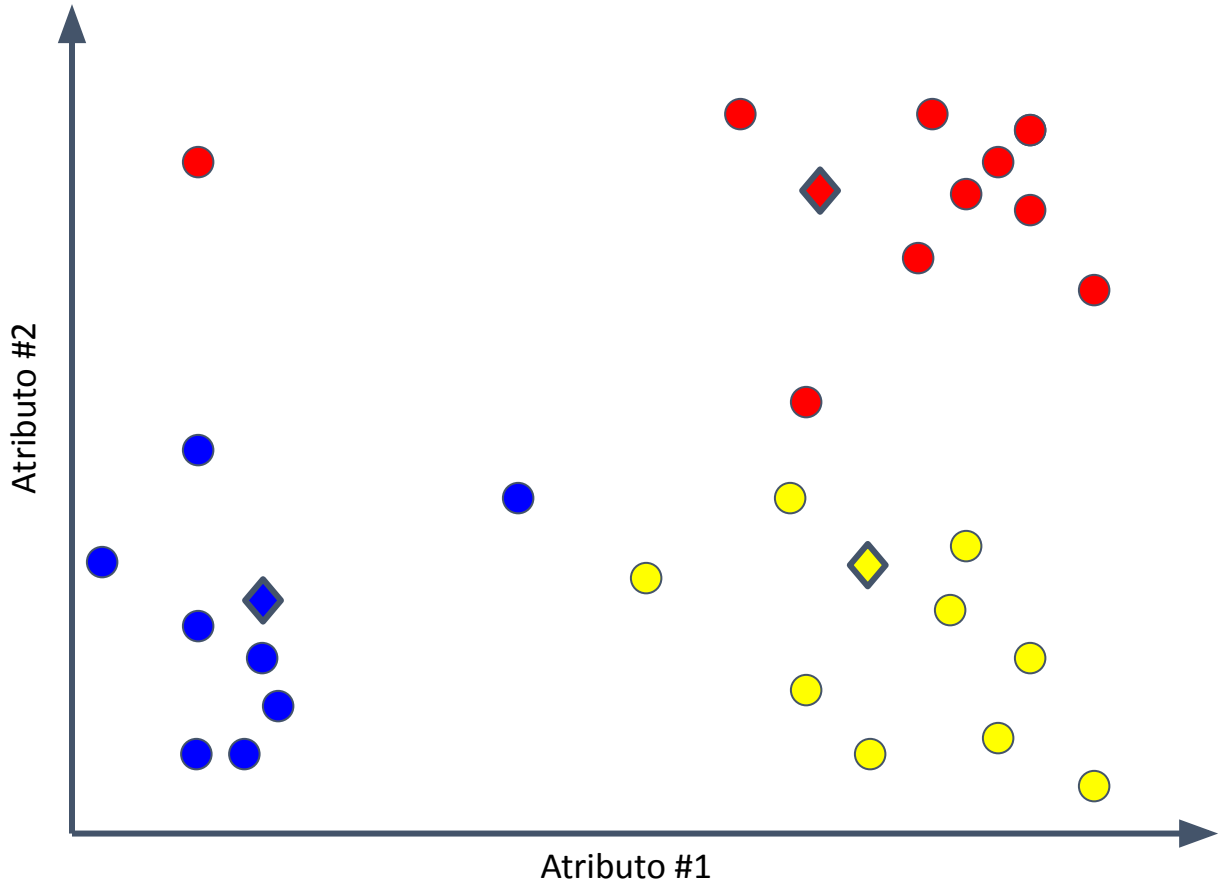
Confira os novos centroides obtidos.



Algoritmo *k-Means*

Repetir até convergir:

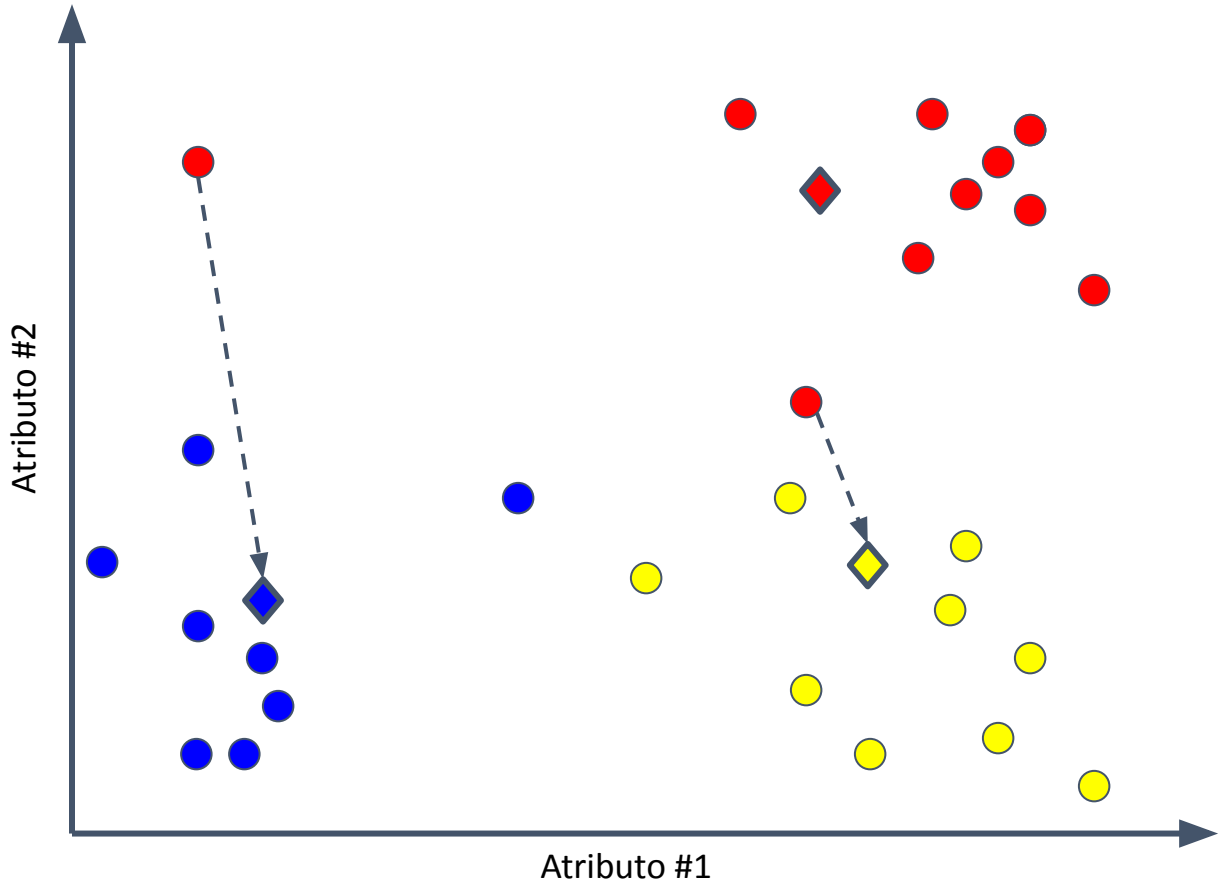
- Alocar objetos ao centróide mais próximo
- Atualizar centroides



Algoritmo *k-Means*

Repetir até convergir:

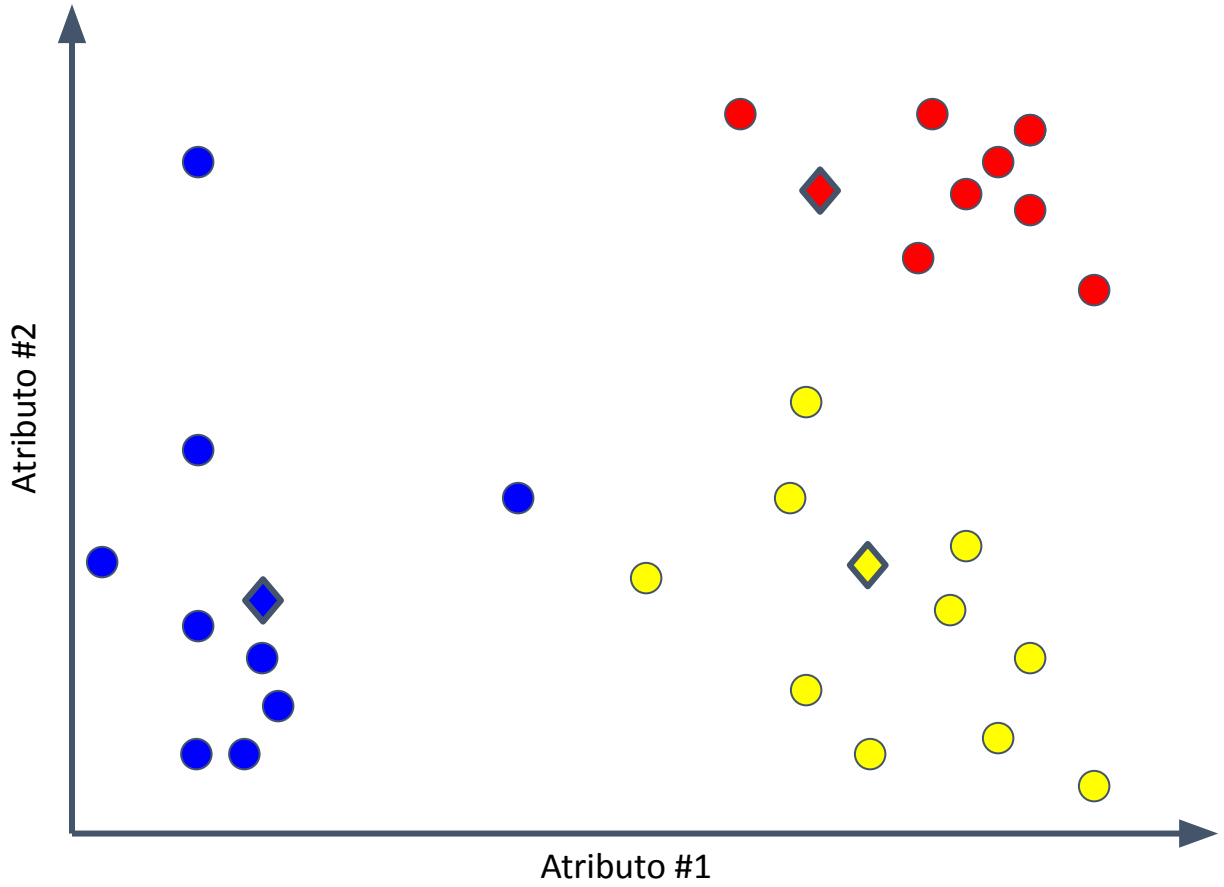
- Alocar objetos ao centróide mais próximo
- Atualizar centroides



Algoritmo *k-Means*

Algoritmo:

1. Selecionar k centroides iniciais
2. Repetir até convergir:
   2.1. Formar k clusters atribuindo cada objeto ao centroide mais próximo
   2.2. Atualizar o centroide de cada cluster



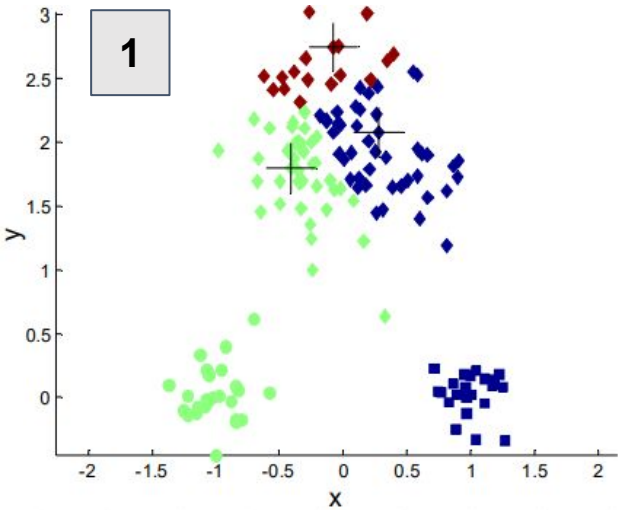
Critérios de convergência: (1) poucas mudanças nos clusters/centroides; (2) número máximo de iterações.

Algoritmo *k-Means*

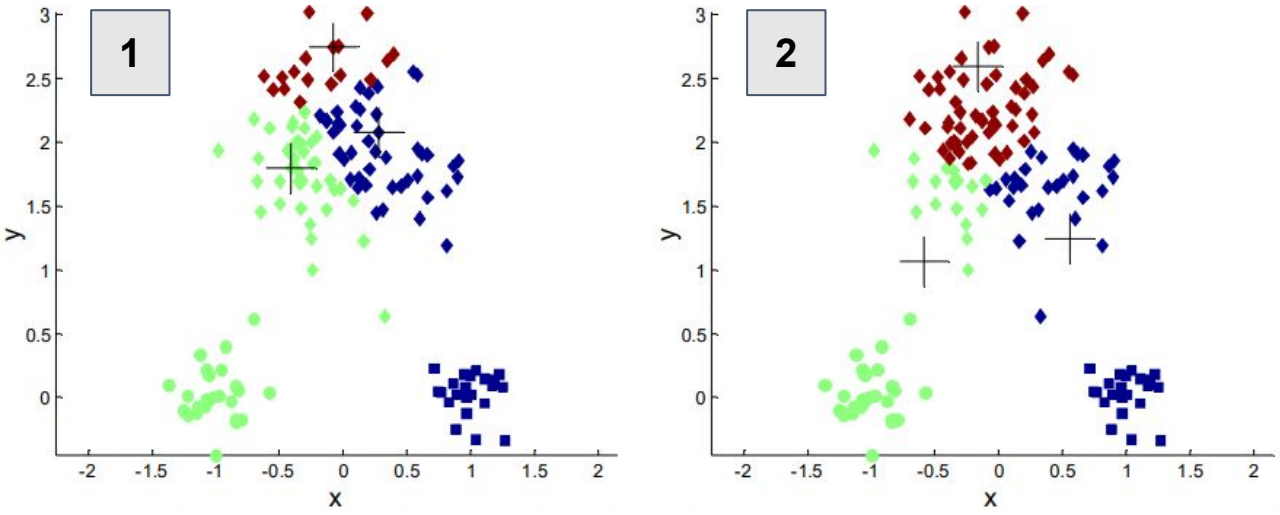
Em geral, os centroides iniciais são escolhidos aleatoriamente.

Clusters podem diferentes em cada execução do k-means.

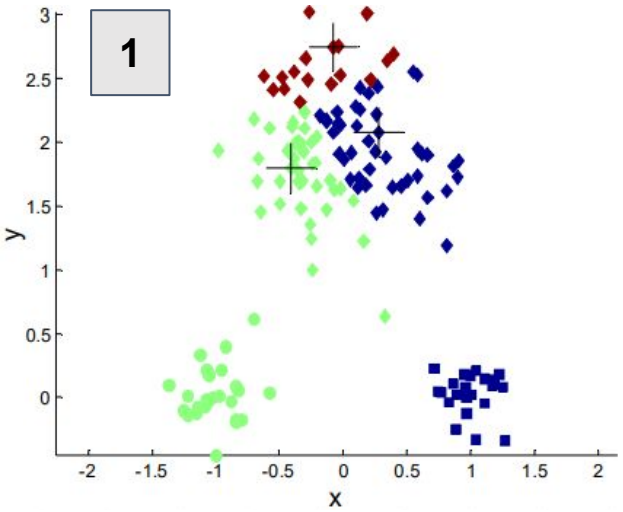
Algoritmo *k-Means*

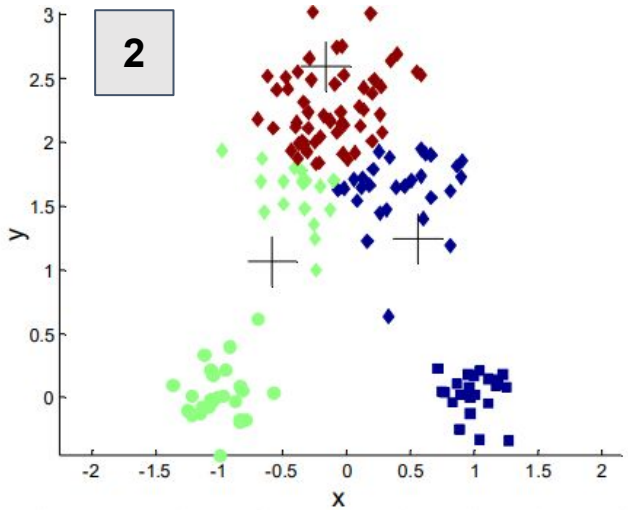


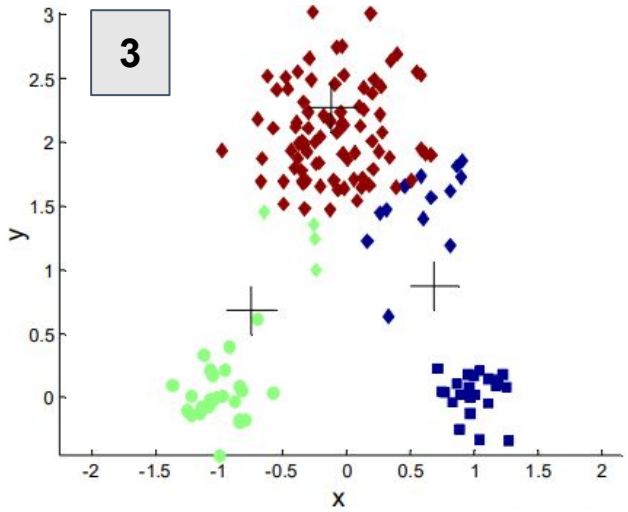
Algoritmo *k-Means*



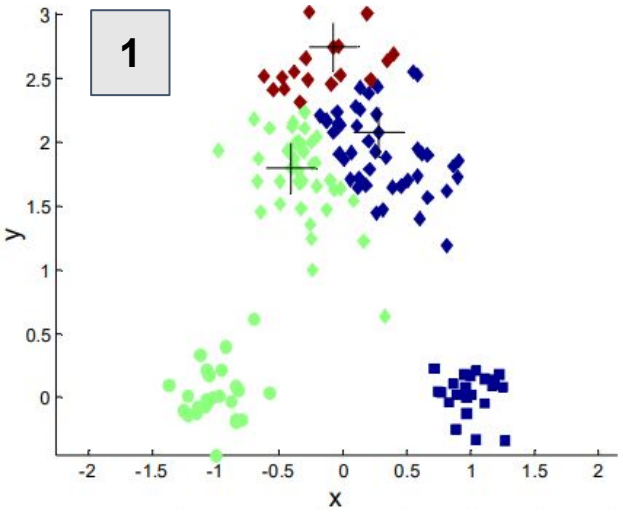
Algoritmo *k-Means*

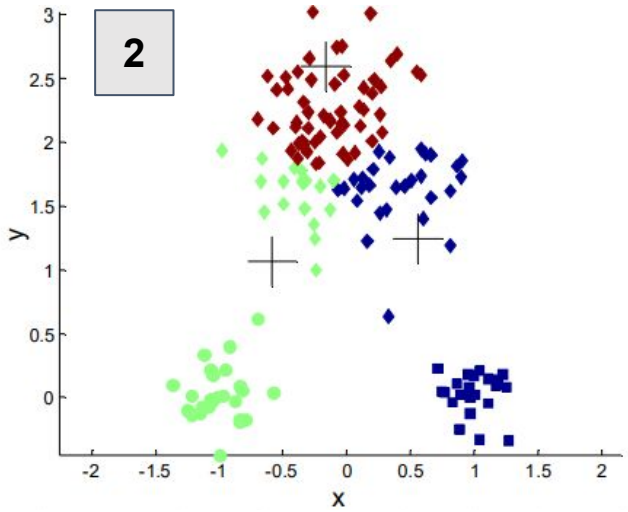


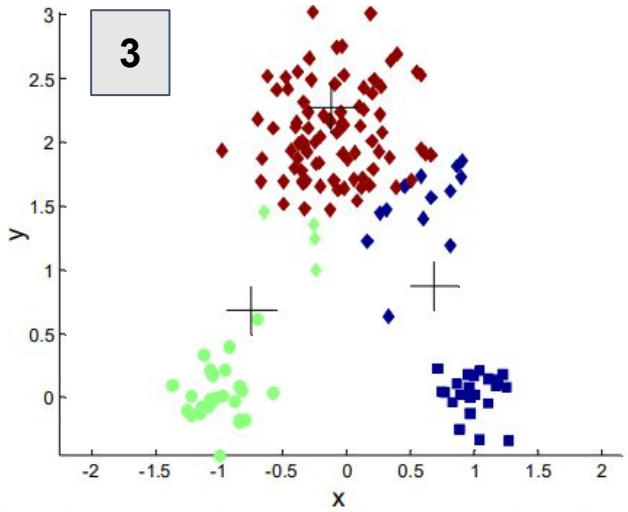


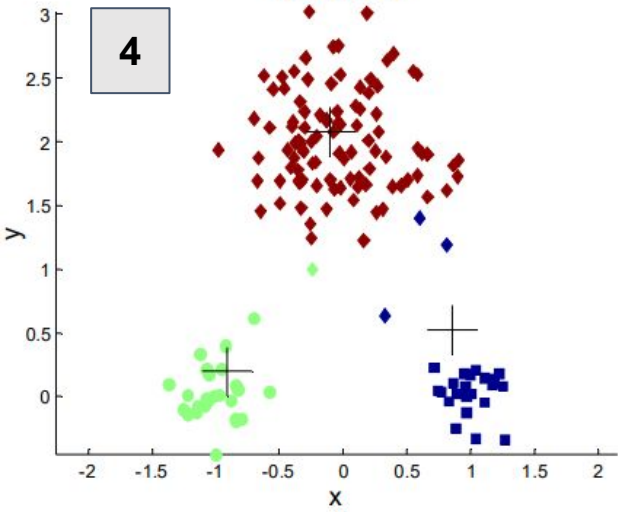


Algoritmo *k-Means*

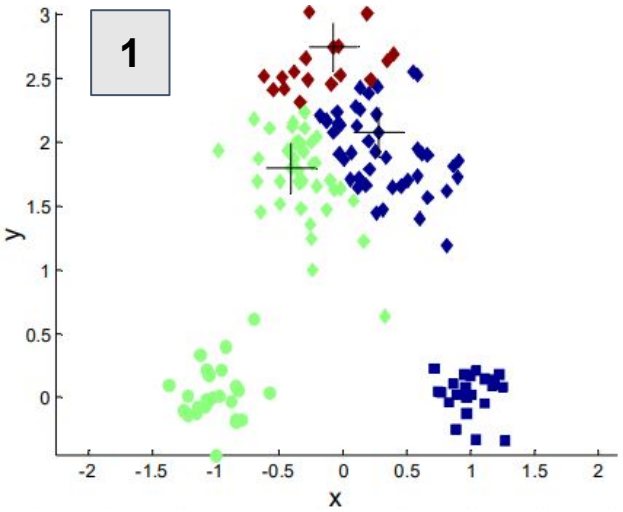


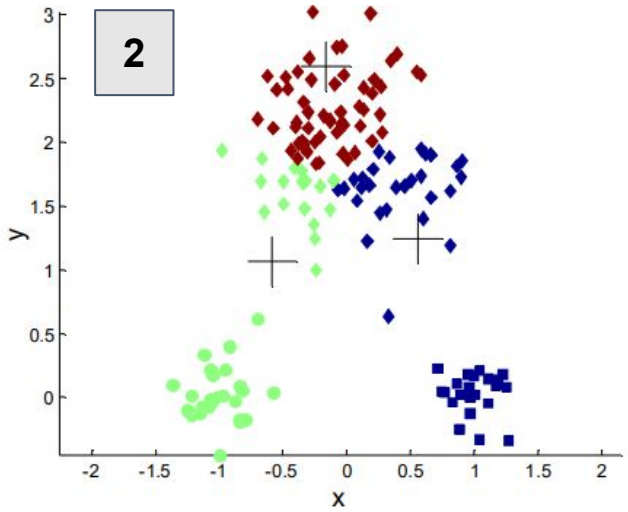


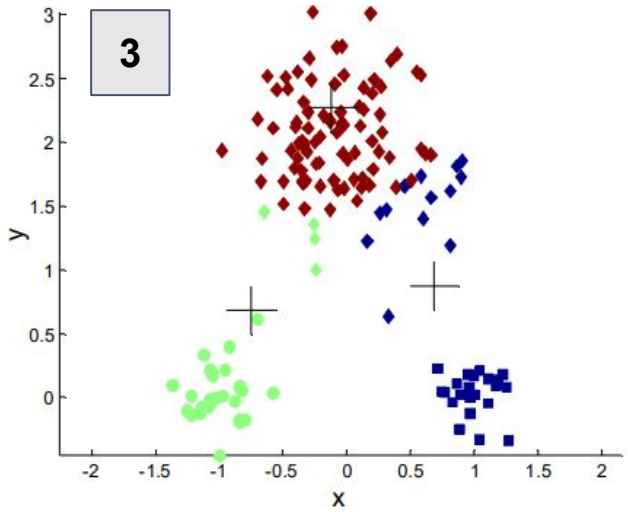


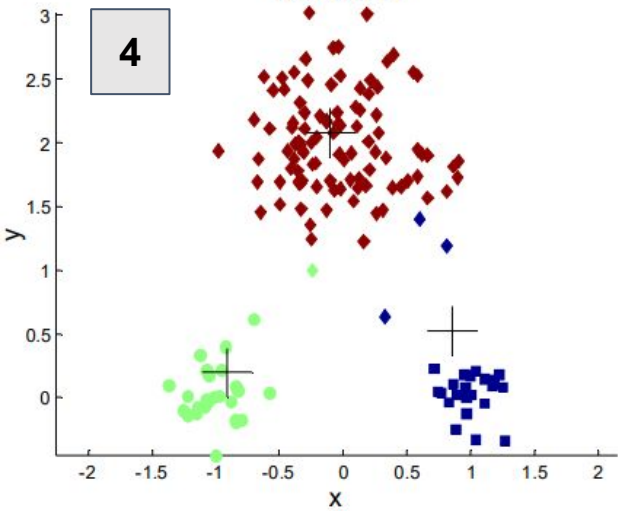


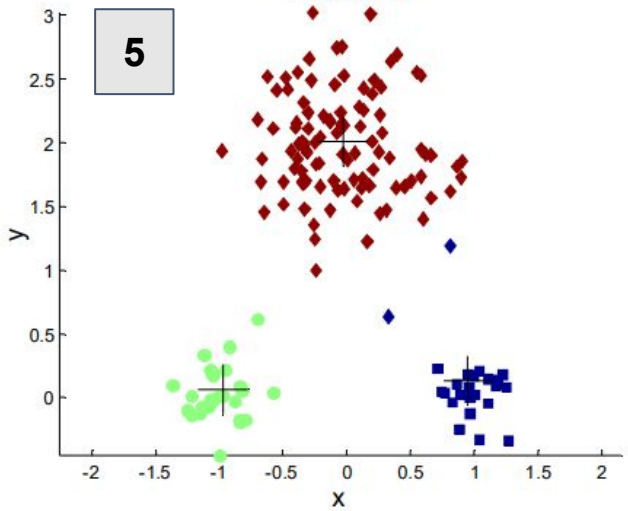
Algoritmo *k-Means*



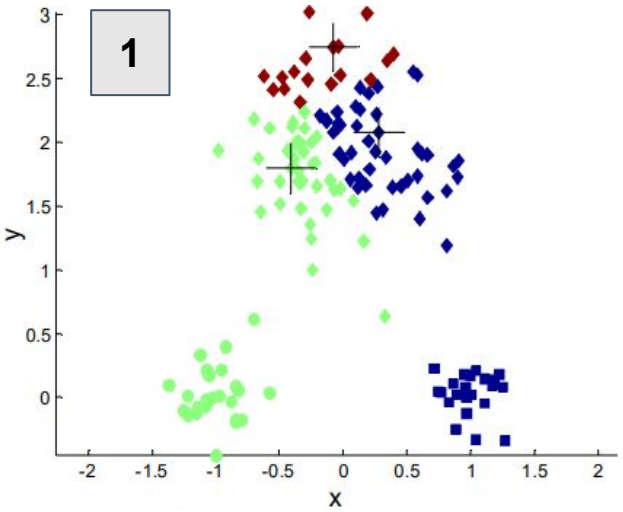


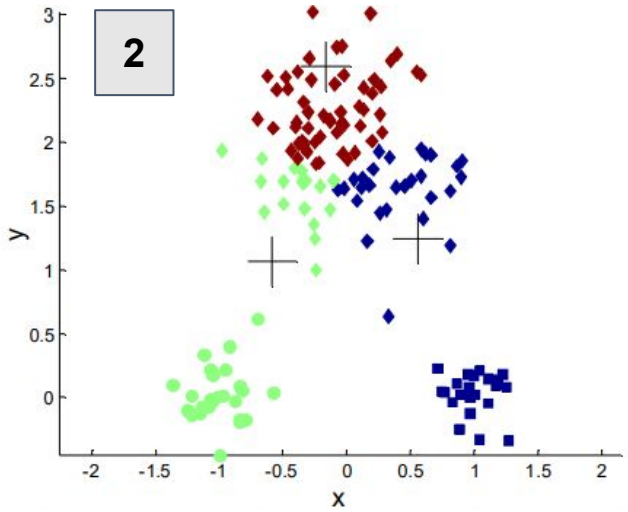


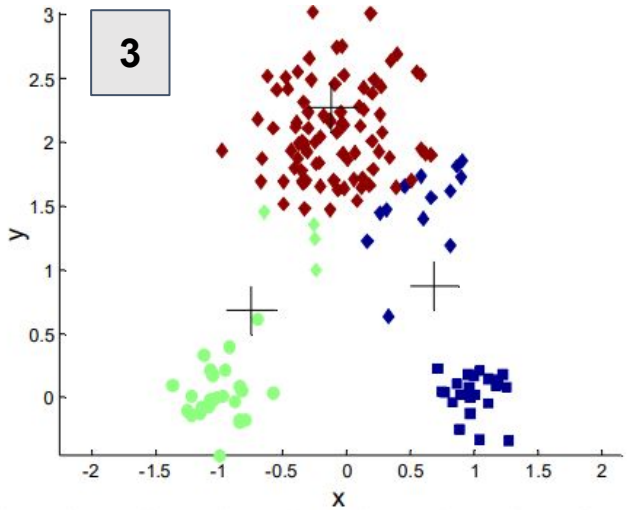


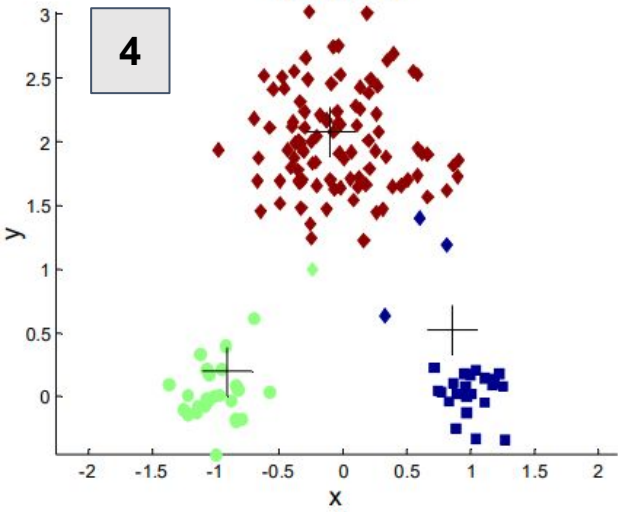


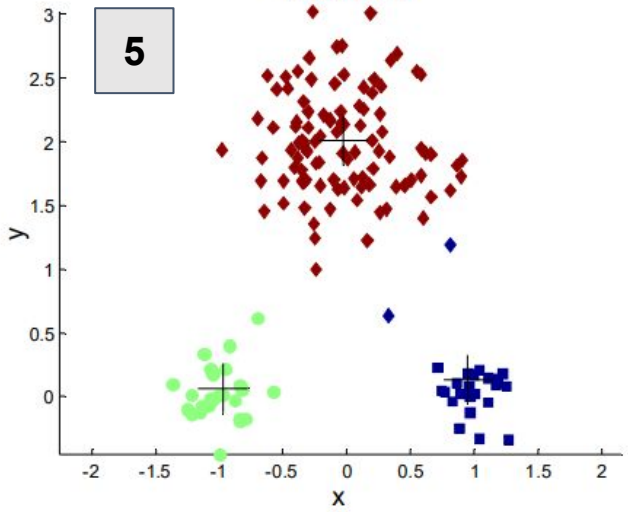
Algoritmo *k-Means*

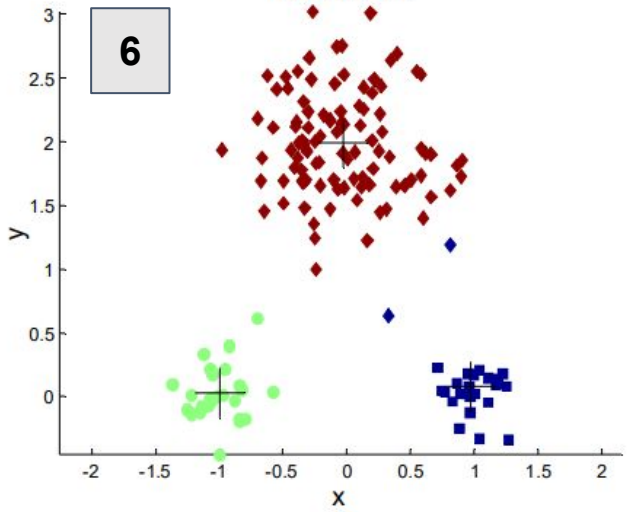




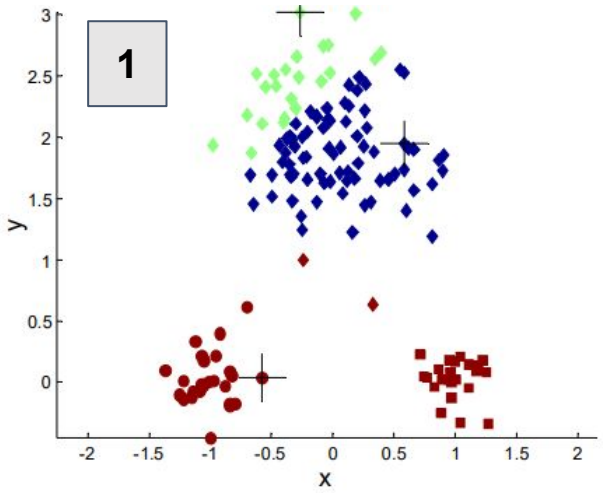




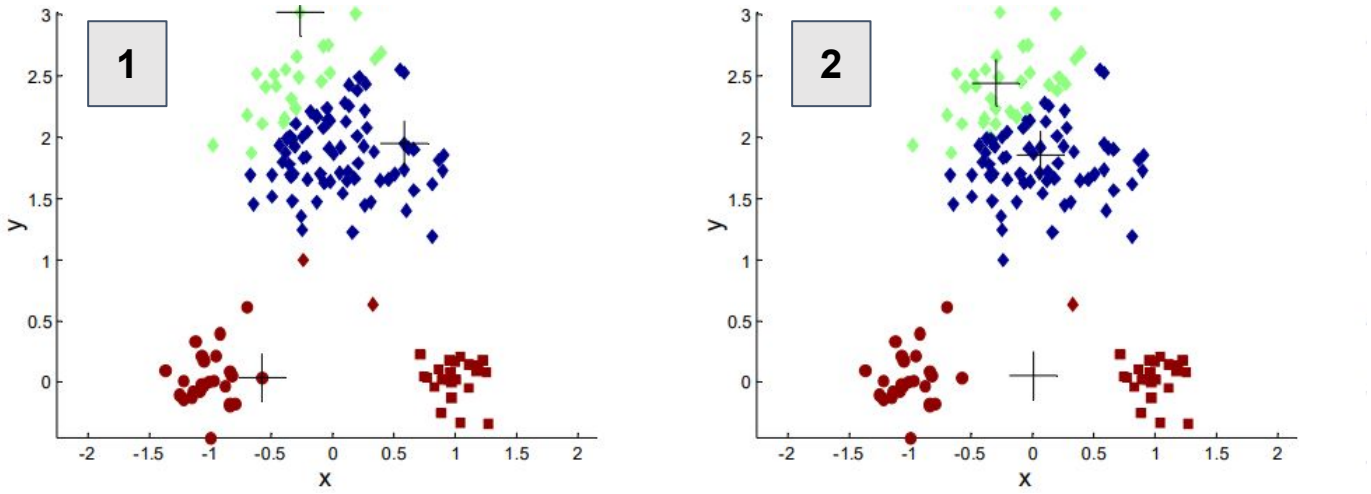




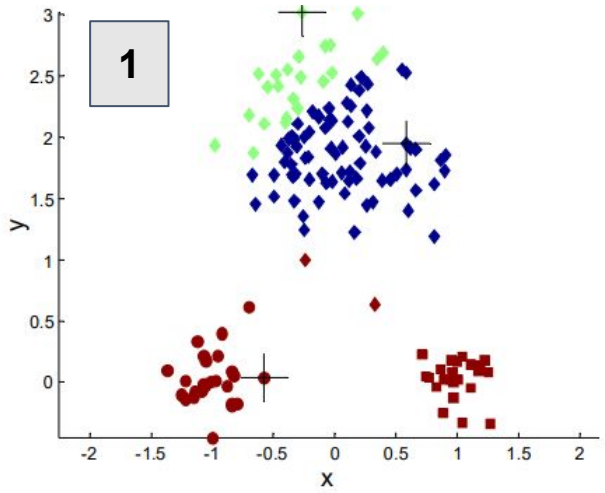
Algoritmo *k-Means*

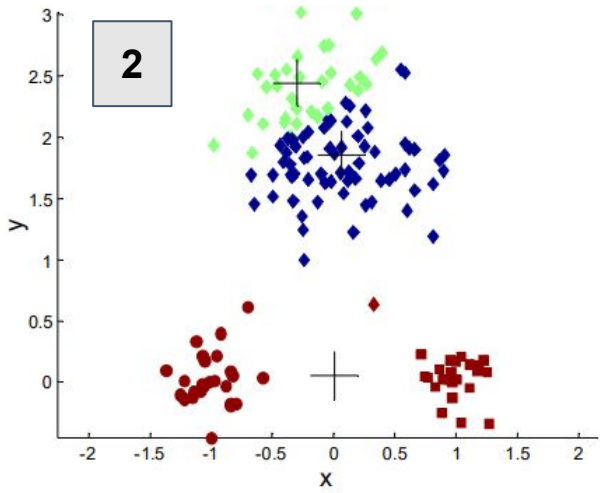


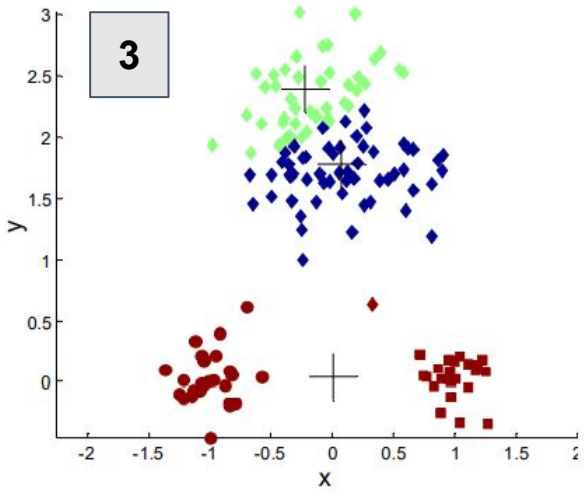
Algoritmo *k-Means*



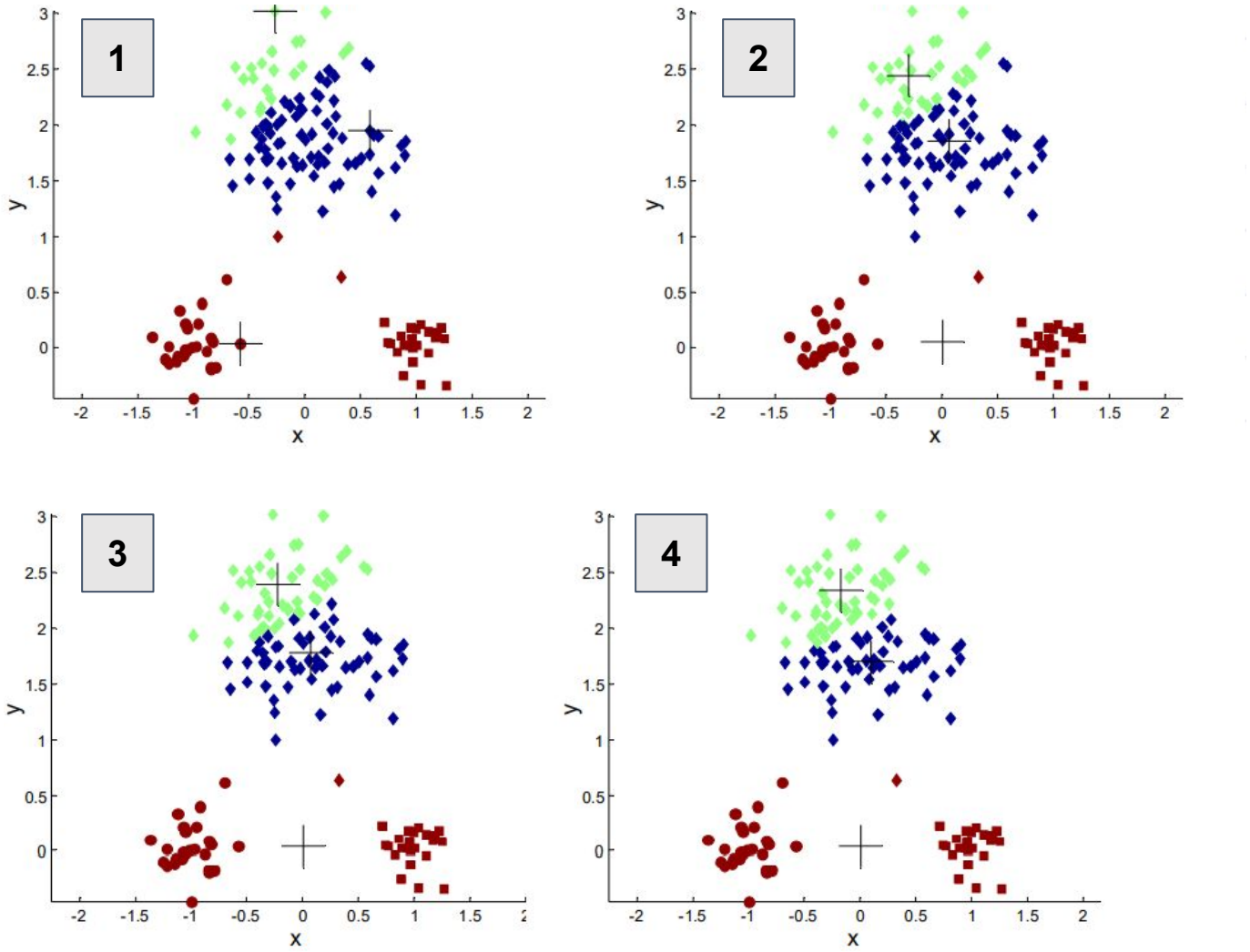
Algoritmo *k-Means*



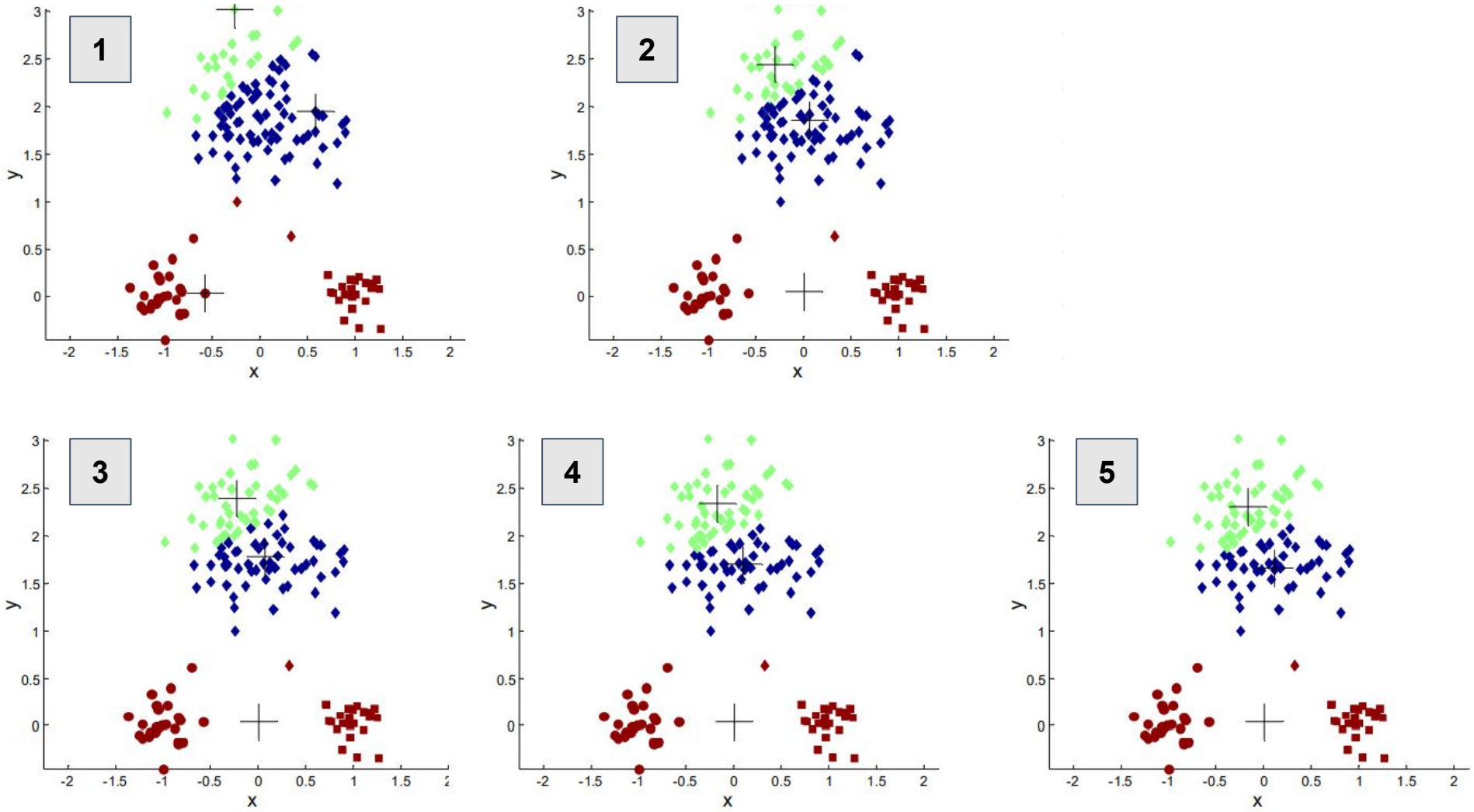




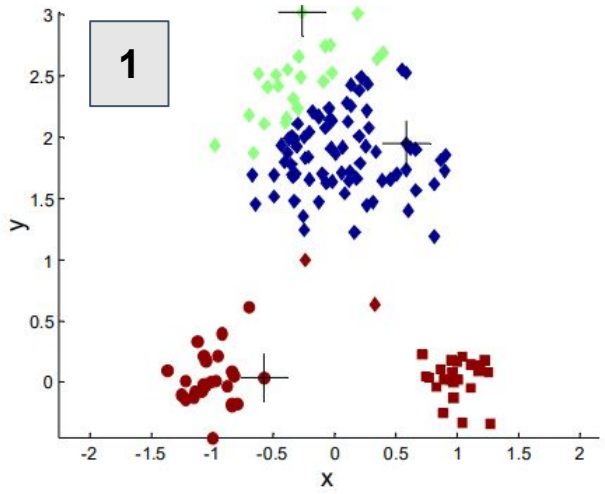
Algoritmo *k-Means*

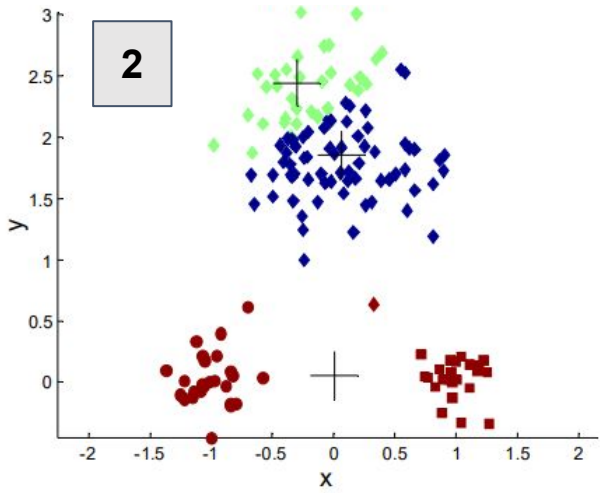


Algoritmo *k-Means*

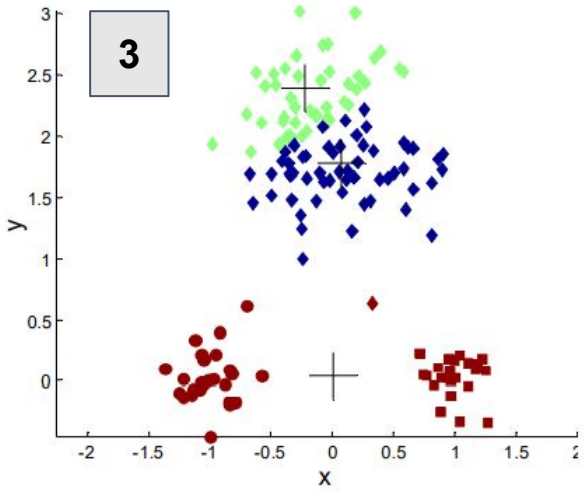


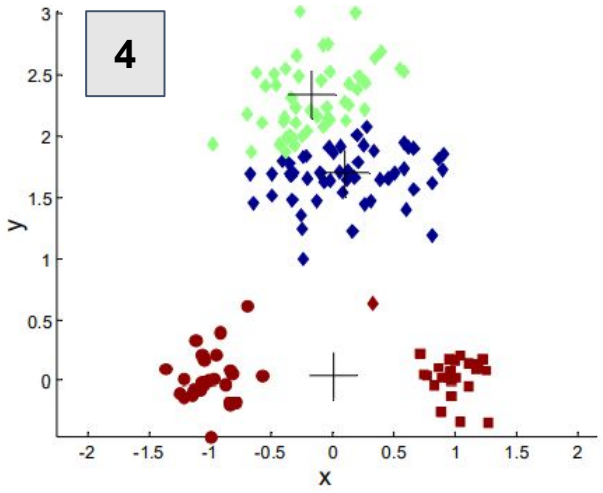
Algoritmo *k-Means*

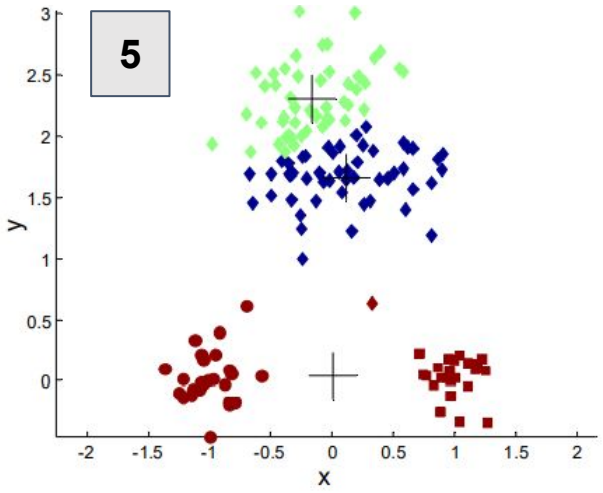




Nesta inicialização de centroides, o *k-means* obteve uma partição sub-ótima.







Algoritmo *k-Means*

- Importância da escolha dos centroides iniciais
- Soluções comuns:
  - Múltiplas execuções do *k-means* e escolher a “melhor” solução de agrupamento (minimizar erro quadrático *E*)

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

Algoritmo *k-Means*

- Importância da escolha dos centroides iniciais
- Soluções comuns:
  - Múltiplas execuções do *k-means* e escolher a “melhor” solução de agrupamento (minimizar erro quadrático *E*)

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

  - Seleção “informada” dos centroides:
    - Garantir que sejam distantes entre si
    - Analista pode indicar centroides considerando sua experiência sobre o domínio dos dados

Algoritmo *k-Means*

- **Limitações do *k-Means***
  - *Outliers*
  - Clusters de tamanhos muito diferentes
  - Clusters de densidades muito diferentes
  - Clusters de formatos não globulares

Algoritmos relacionados

- **Algoritmo *bisecting K-Means:***
  - Usa o *k-Means* para produzir um agrupamento hierárquico
  - Inicialmente, agrupar os dados com k=2
  - Escolher o maior cluster e repetir o agrupamento com k=2
  - Repetir até que obter a quantidade desejada de cluster

Agrupamento Hierárquico

- Dois métodos clássicos para agrupamento hierárquico

Aglomerativos:

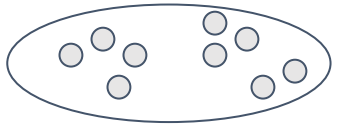
➔ Iniciar alocando cada objeto em um *cluster*
➔ Encontrar o melhor par de *clusters* para unir
➔ Repetir até formar um único *cluster*

Divisivos:

➔ Iniciar alocando todos os objetos em um único *cluster*
➔ Dividir um *cluster* em dois *subclusters*
➔ Repetir a divisão até que cada objeto seja um *cluster*

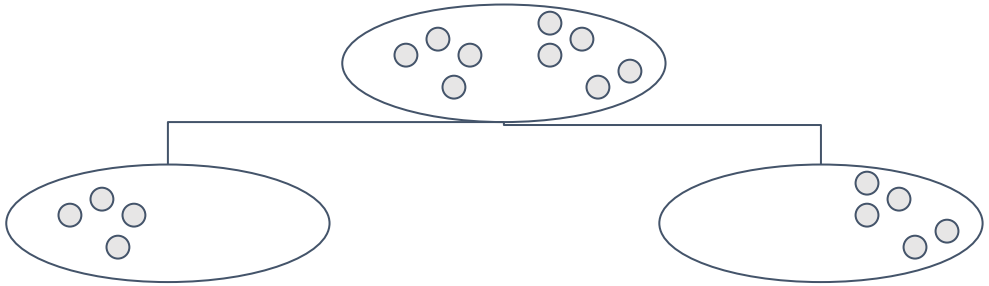
Algoritmos relacionados

- **Algoritmo *bisecting K-Means:***
  - Usa o *k-Means* para produzir um agrupamento hierárquico
  - Inicialmente, agrupar os dados com k=2
  - Escolher o maior cluster e repetir o agrupamento com k=2
  - Repetir até que obter a quantidade desejada de cluster



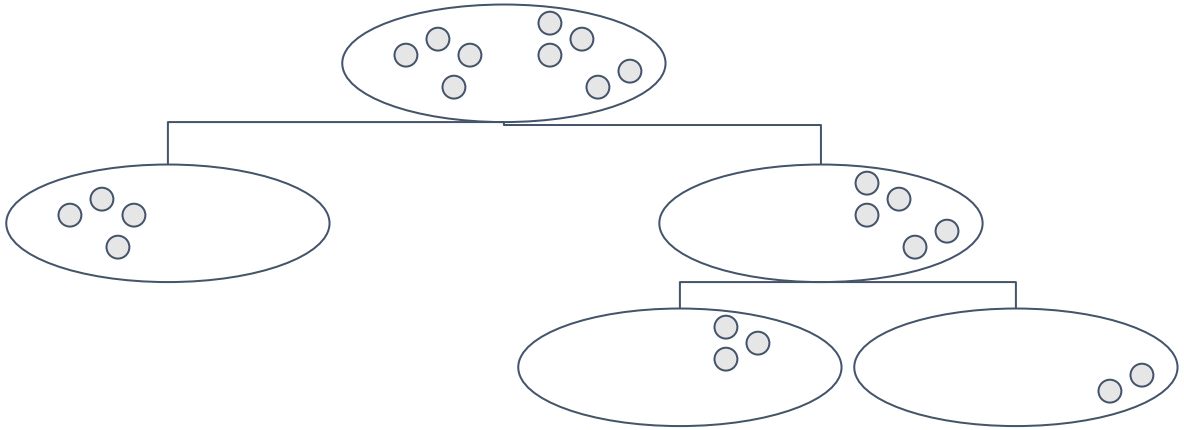
Algoritmos relacionados

- **Algoritmo *bisecting K-Means:***
  - Usa o *k-Means* para produzir um agrupamento hierárquico
  - Inicialmente, agrupar os dados com k=2
  - Escolher o maior cluster e repetir o agrupamento com k=2
  - Repetir até que obter a quantidade desejada de cluster



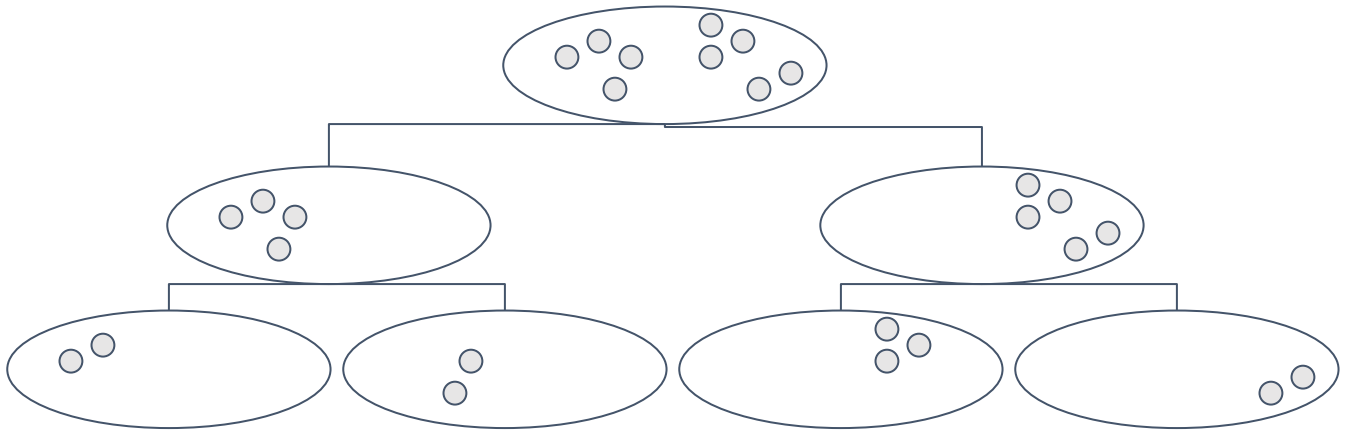
Algoritmos relacionados

- **Algoritmo *bisecting K-Means:***
  - Usa o *k-Means* para produzir um agrupamento hierárquico
  - Inicialmente, agrupar os dados com k=2
  - Escolher o maior cluster e repetir o agrupamento com k=2
  - Repetir até que obter a quantidade desejada de cluster



Algoritmos relacionados

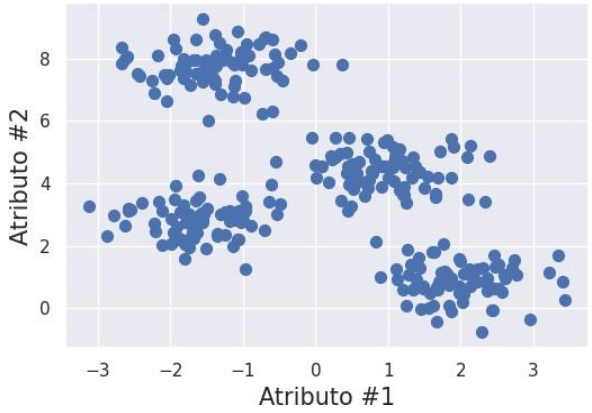
- **Algoritmo *bisecting K-Means:***
  - Usa o *k-Means* para produzir um agrupamento hierárquico
  - Inicialmente, agrupar os dados com k=2
  - Escolher o maior cluster e repetir o agrupamento com k=2
  - Repetir até que obter a quantidade desejada de cluster

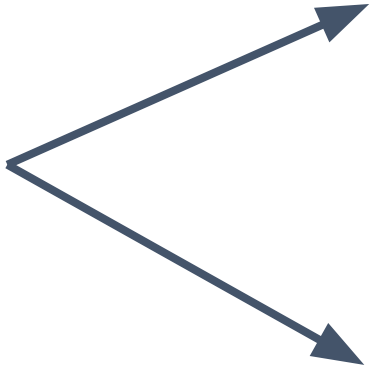


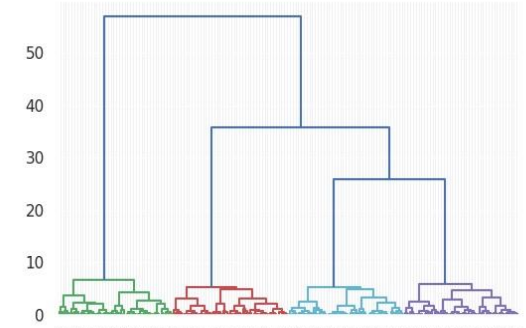
Métodos para Agrupamento de Dados

- Estudamos diferentes métodos e algoritmos
- Qual solução de agrupamento escolher?
- Qual o número apropriado de clusters para meus dados?

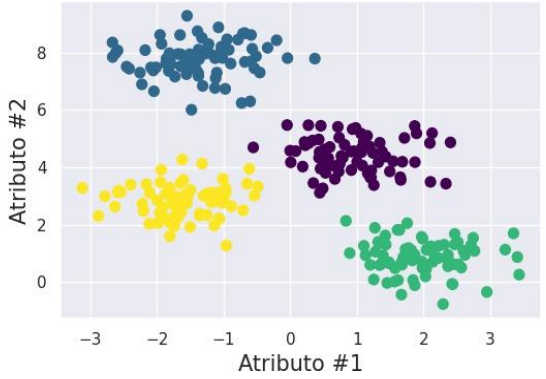
Conjunto de Dados







Agrupamento Hierárquico



Agrupamento Particional

Métodos para Agrupamento de Dados

- Estudamos diferentes métodos e algoritmos
- Qual solução de agrupamento escolher?
- Qual o número apropriado de clusters para meus dados?

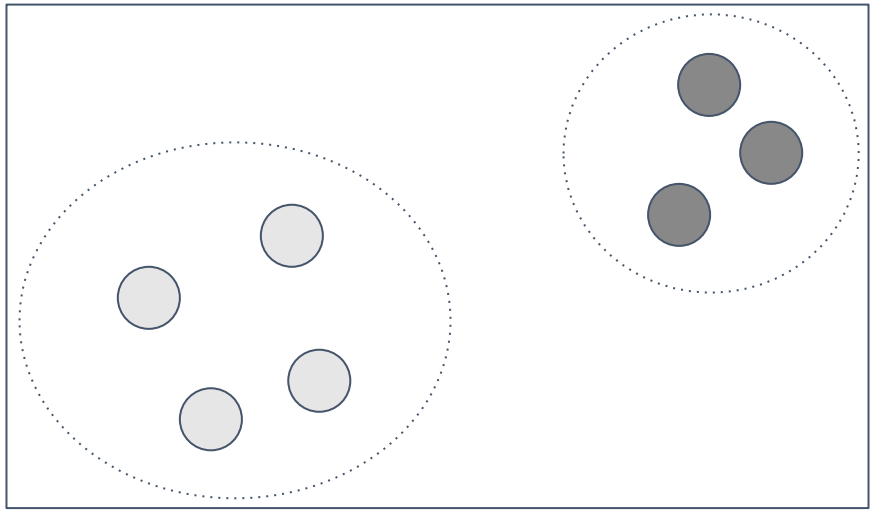
Validação de Agrupamentos

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*

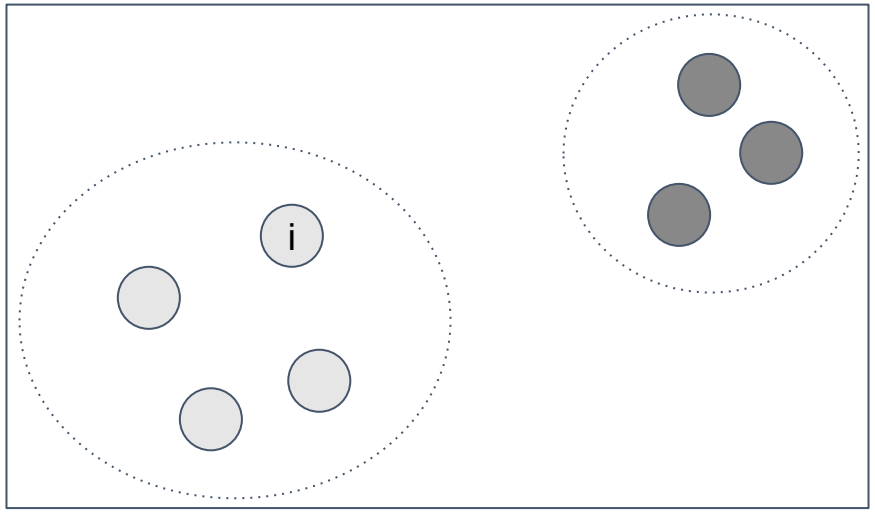
Validação de Agrupamentos

- Índices de validade relativa: Silhueta



Validação de Agrupamentos

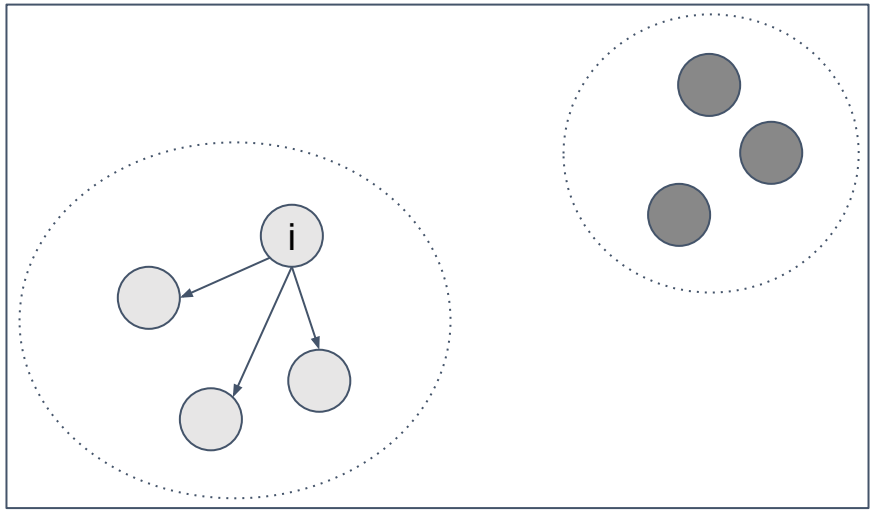
- Índices de validade relativa: Silhueta



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

Validação de Agrupamentos

- Índices de validade relativa: Silhueta

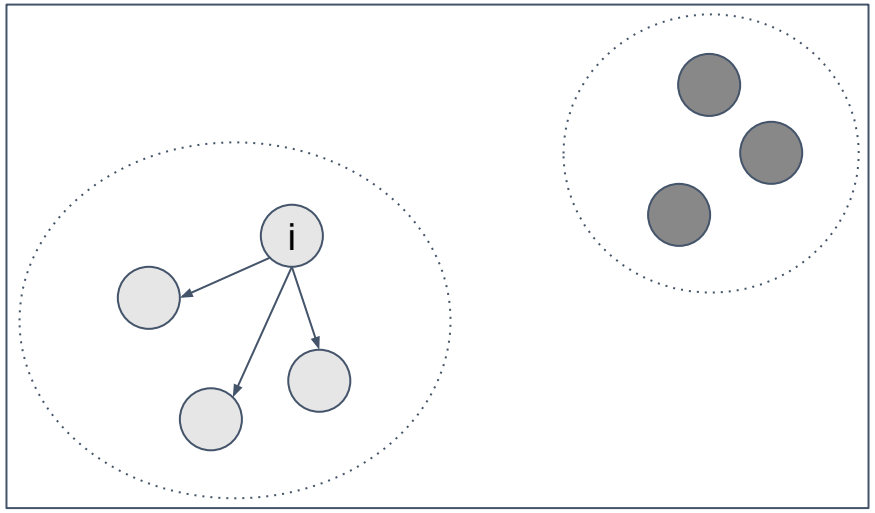


1. O quão bem o objeto *i* está alocado em seu próprio *cluster*?

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

Validação de Agrupamentos

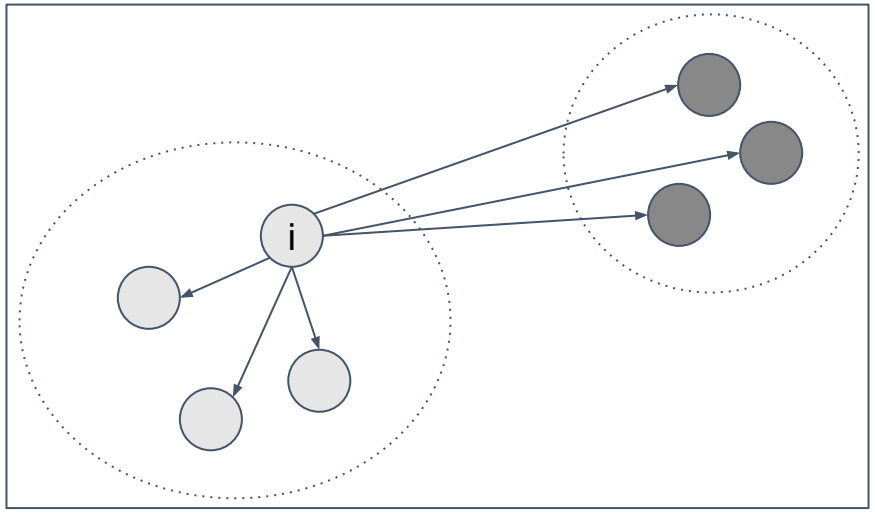
- Índices de validade relativa: Silhueta



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

Validação de Agrupamentos

- Índices de validade relativa: Silhueta

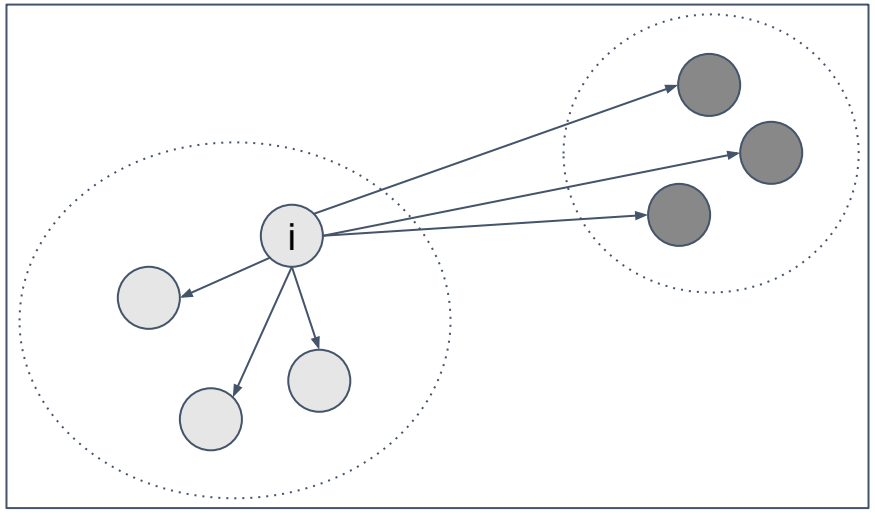


2. O quão próximo o objeto i está do seu *cluster* vizinho?

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

Validação de Agrupamentos

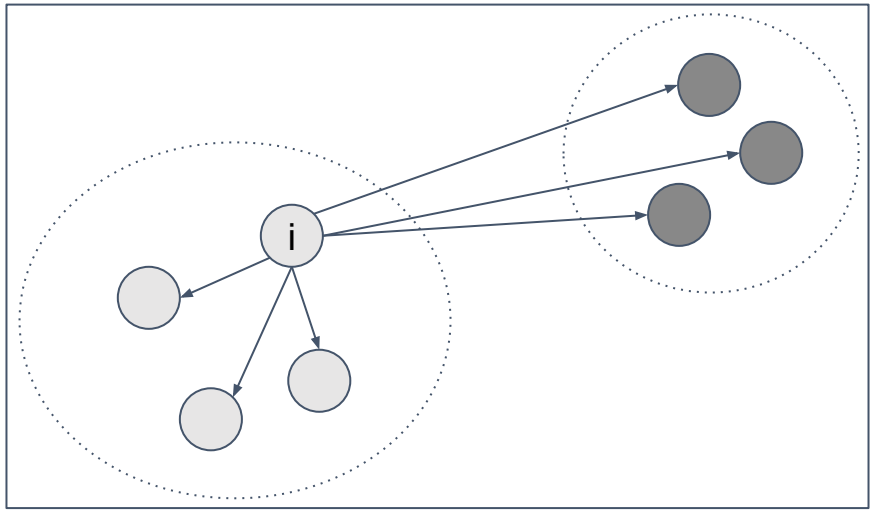
- Índices de validade relativa: Silhueta



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*?

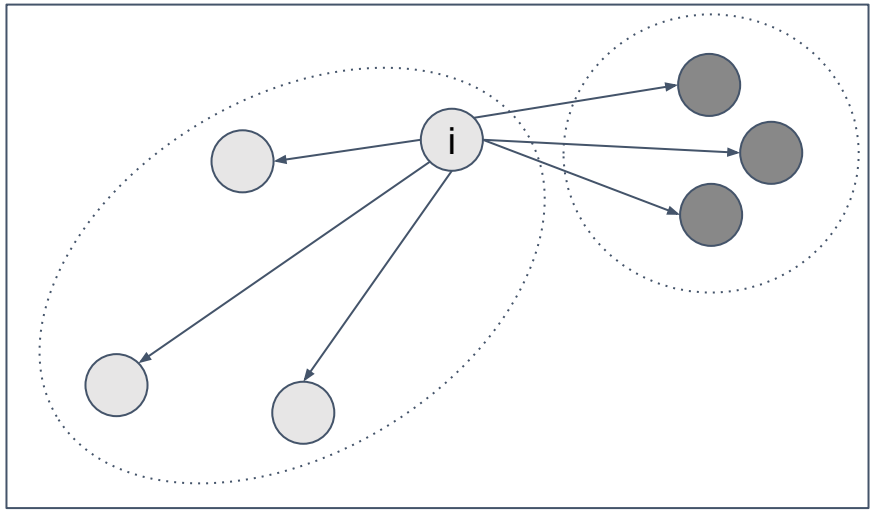
$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

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*?

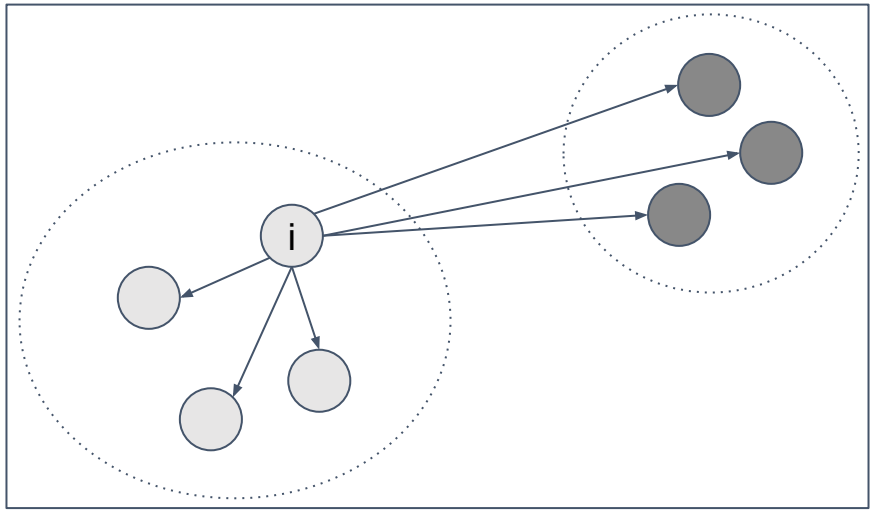
$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

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

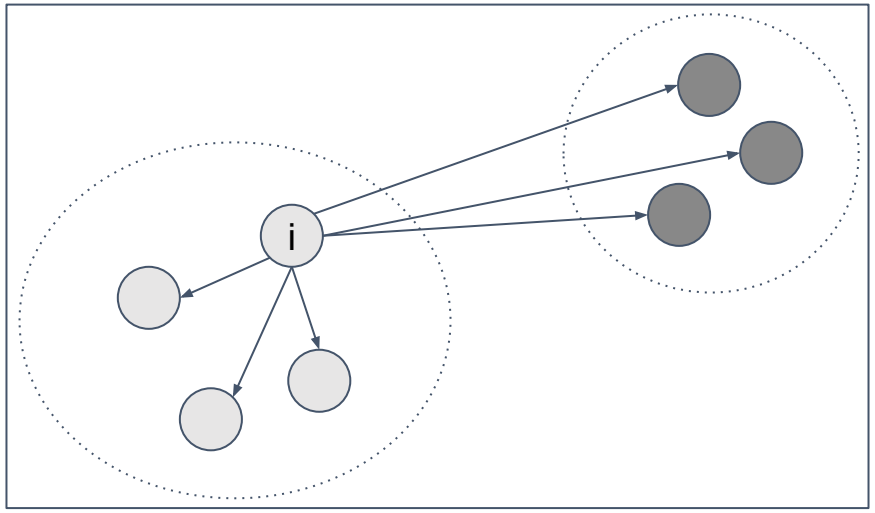


4. Calcular a silhueta de todos os objetos e computar a silhueta média

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

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.