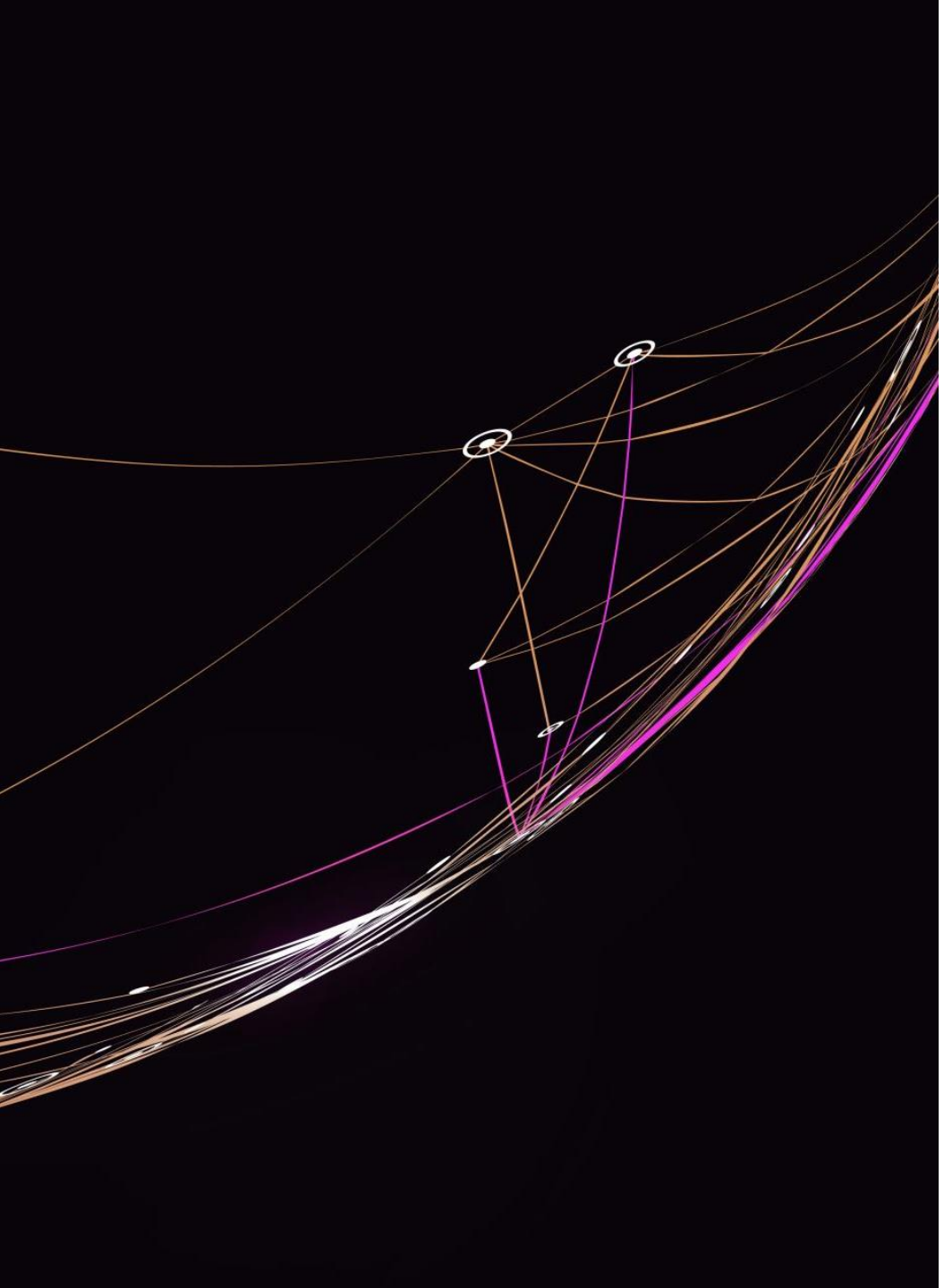


ADMINISTRAÇÃO

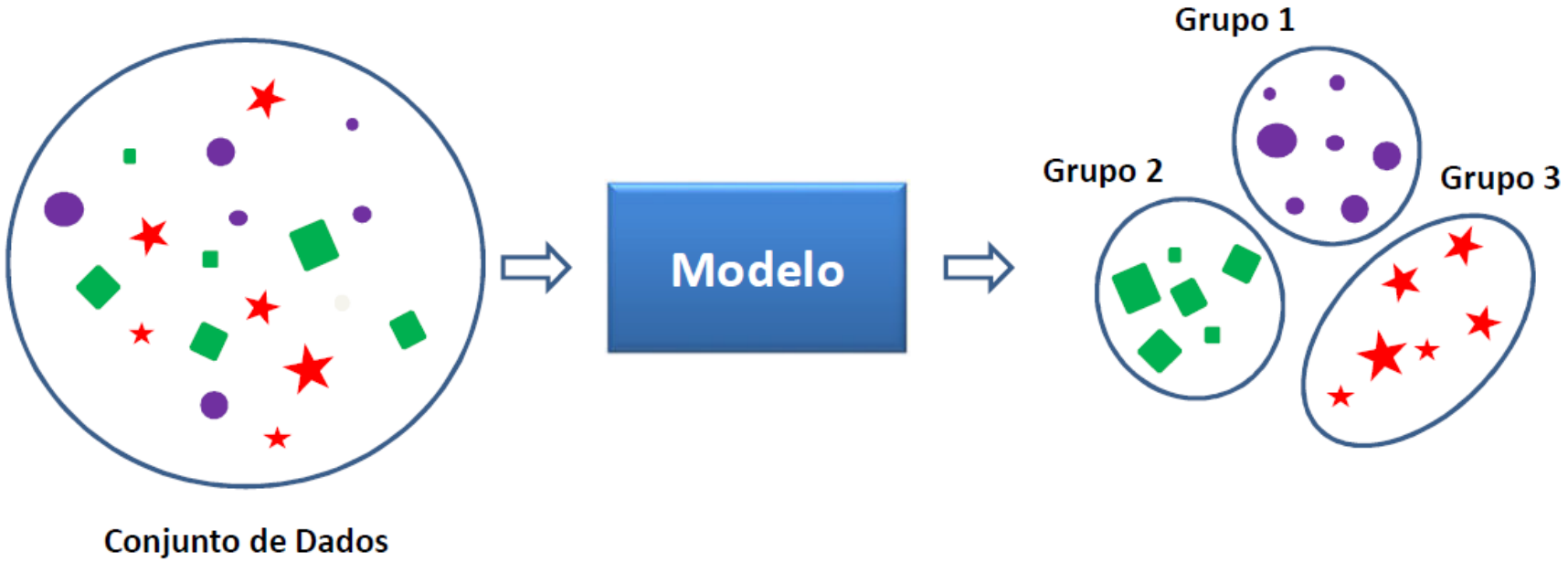
IBM0112 DATA MINING

Agrupamentos (Clustering)

Cassius Figueiredo

Definição

- O objetivo da análise de agrupamentos (ou segmentação) é encontrar grupos de objetos similares no conjunto de dados.



Análise de agrupamentos

- Um bom agrupamento tem grupos densos e separados entre si. Na prática, a noção de “grupo” é relativa e depende da aplicação.

Quantos grupos?

6 grupos?

2 grupos?

4 grupos?

Análise de agrupamentos

- O problema de análise de agrupamentos é diferente do problema de classificação:
- No problema de classificação, a informação sobre as classes é externa.
- Na análise de agrupamentos, a informação sobre os grupos é interna.

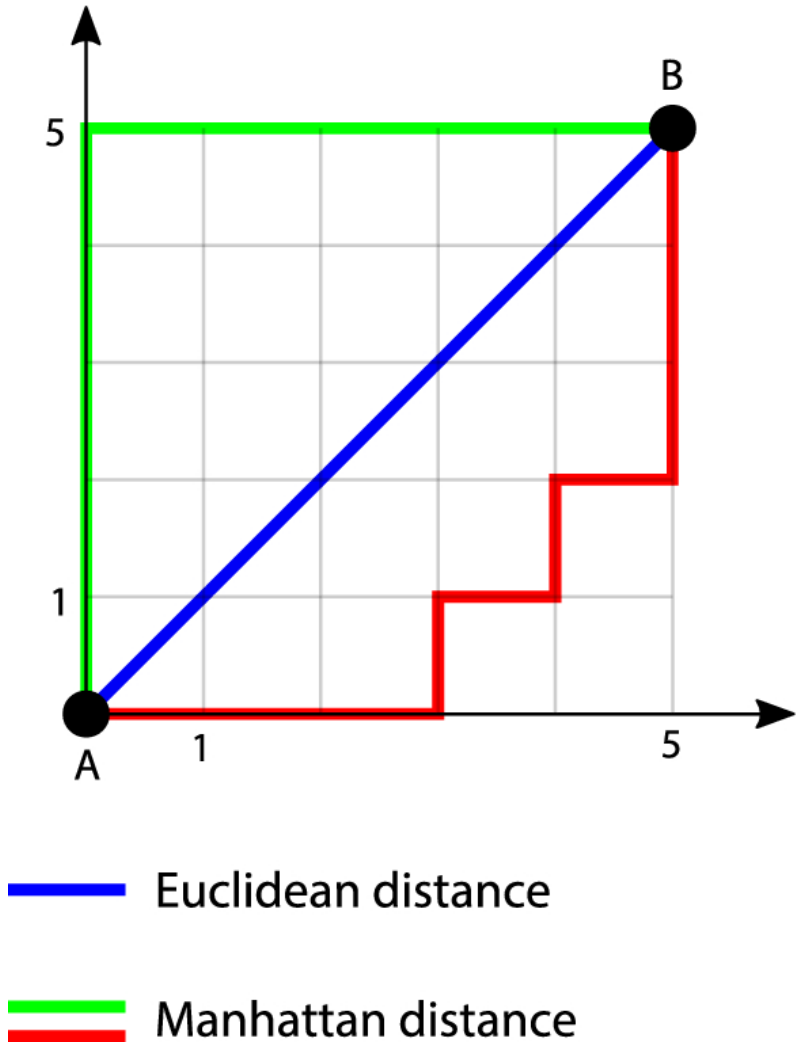
Métodos

- **Métodos de Particionamento**:
  - Os algoritmos constroem, a cada iteração, uma partição do conjunto de dados e a avaliam por um critério.
  - O critério mais comum é a soma das distâncias de todos os registros aos centros de grupos.
- **Métodos Hierárquicos**:
  - Os algoritmos geram uma decomposição hierárquica do conjunto de dados, através de estratégias “divisivas” (“top-down”) ou aglomerativas (“bottom-up”).
- **Métodos a base de densidades**:
  - Os grupos são gerados a partir da densidade dos registros (DBSCAN).

Principais funções de distância

- Distância Euclidiana

- Distância Manhattan



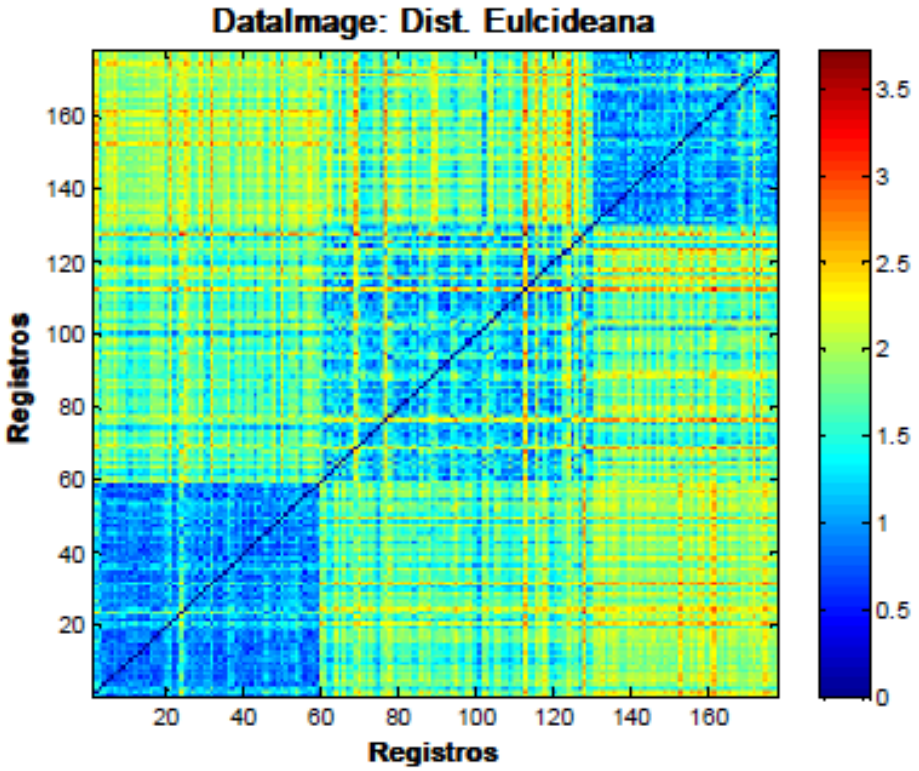
Principais funções de similaridade

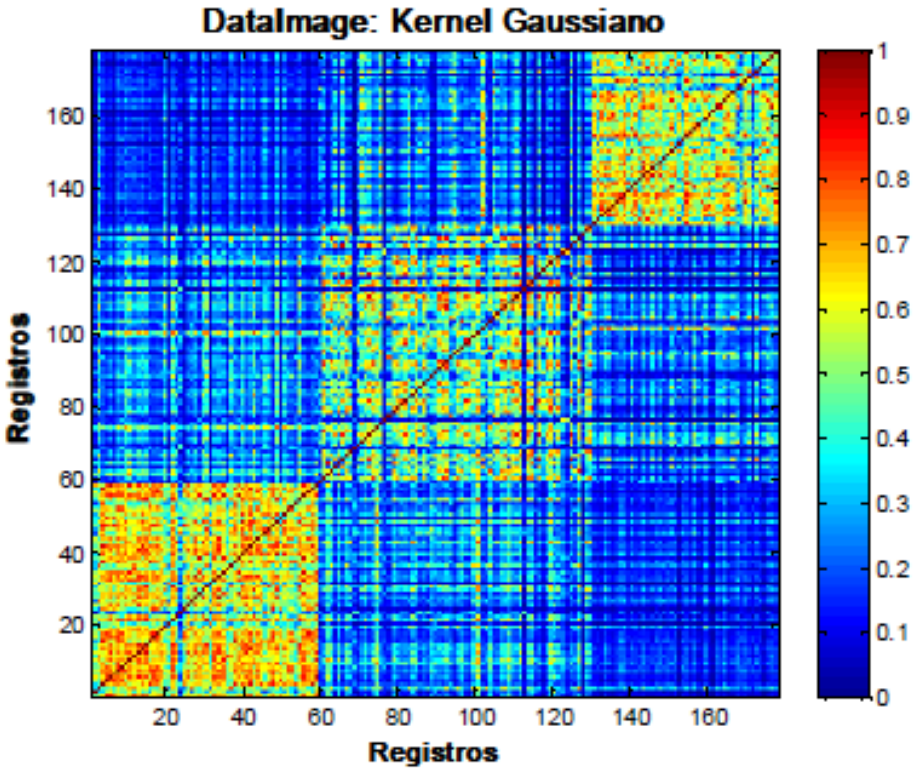
- Similaridade Gaussiana

- Similaridade do cosseno

Exemplo

Wine





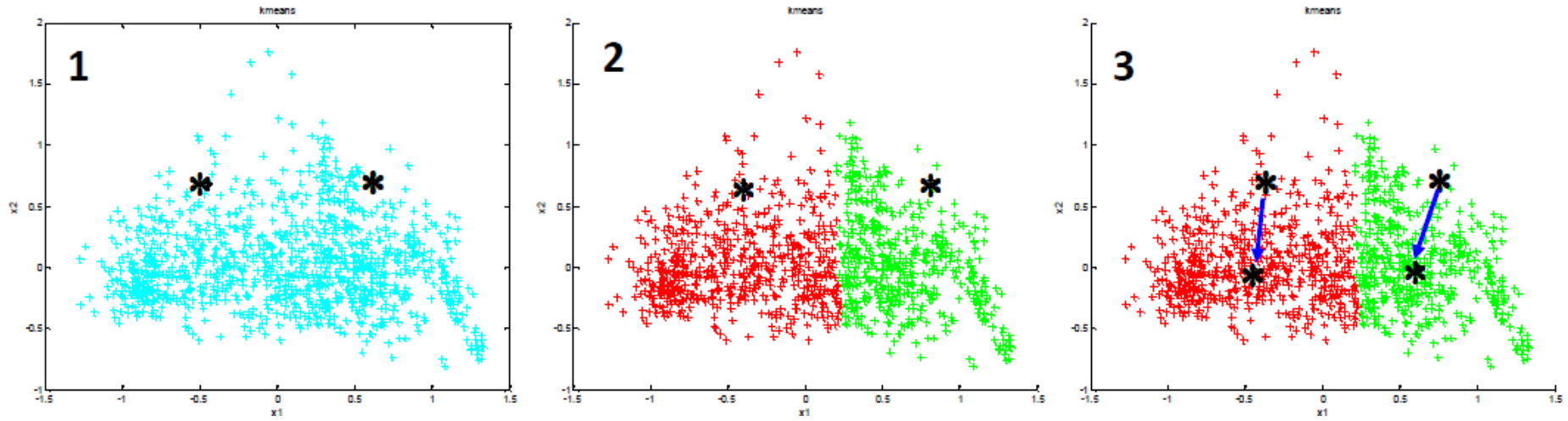
K-Médias
(K-Means)

K-Médias

- O problema de análise de agrupamentos é diferente do problema de classificação:
- No problema de classificação, a informação sobre as classes é externa.
- Na análise de agrupamentos, a informação sobre os grupos é interna.

K-Médias

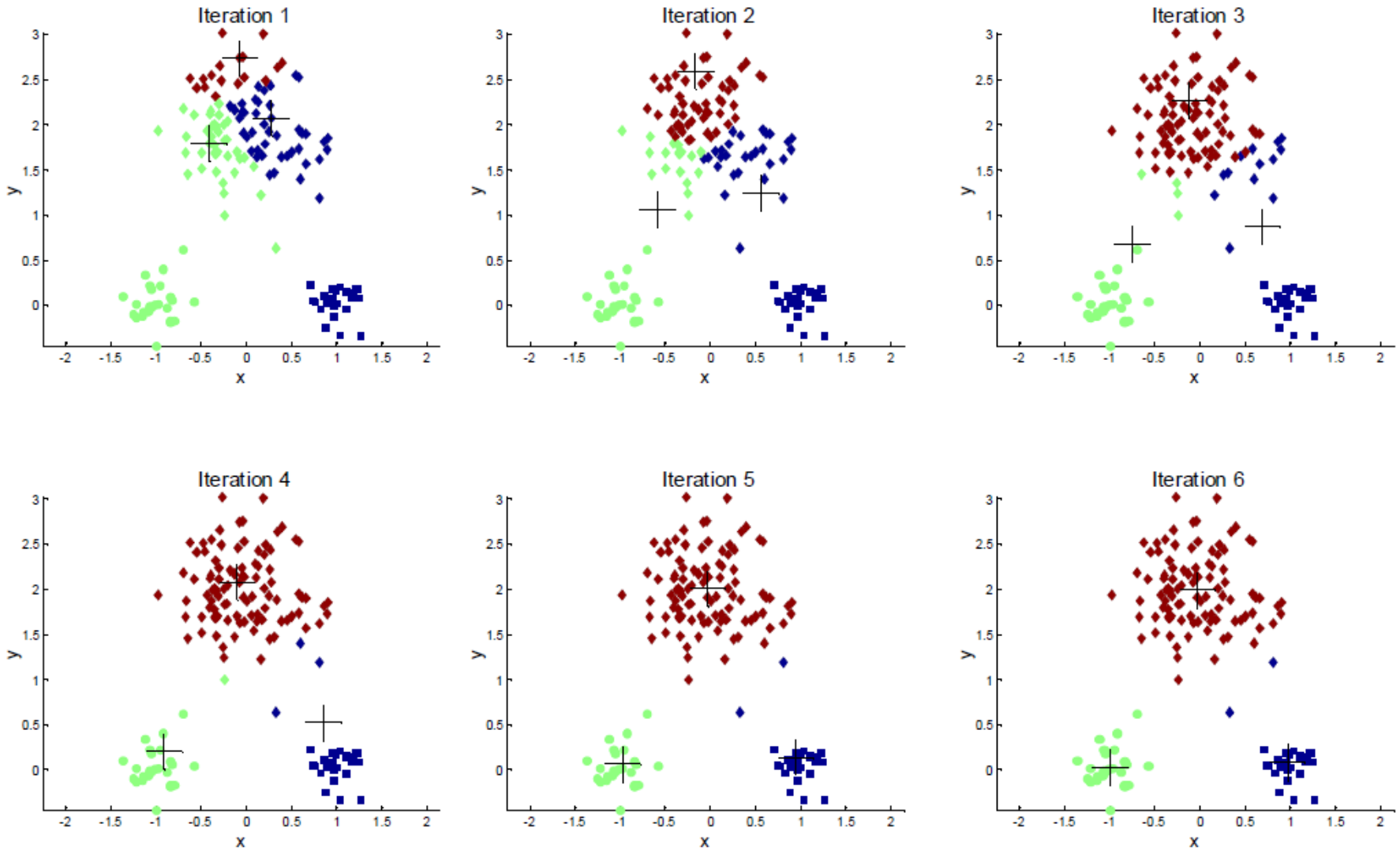
- Em cada iteração, o algoritmo k-médias:
- Calcula a distância de cada registro aos centros de grupo;
- Aloca cada registro ao grupo cujo centro é mais próximo;
- Atualiza as coordenadas do centro de cada grupo pela média dos registros alocados ao grupo.



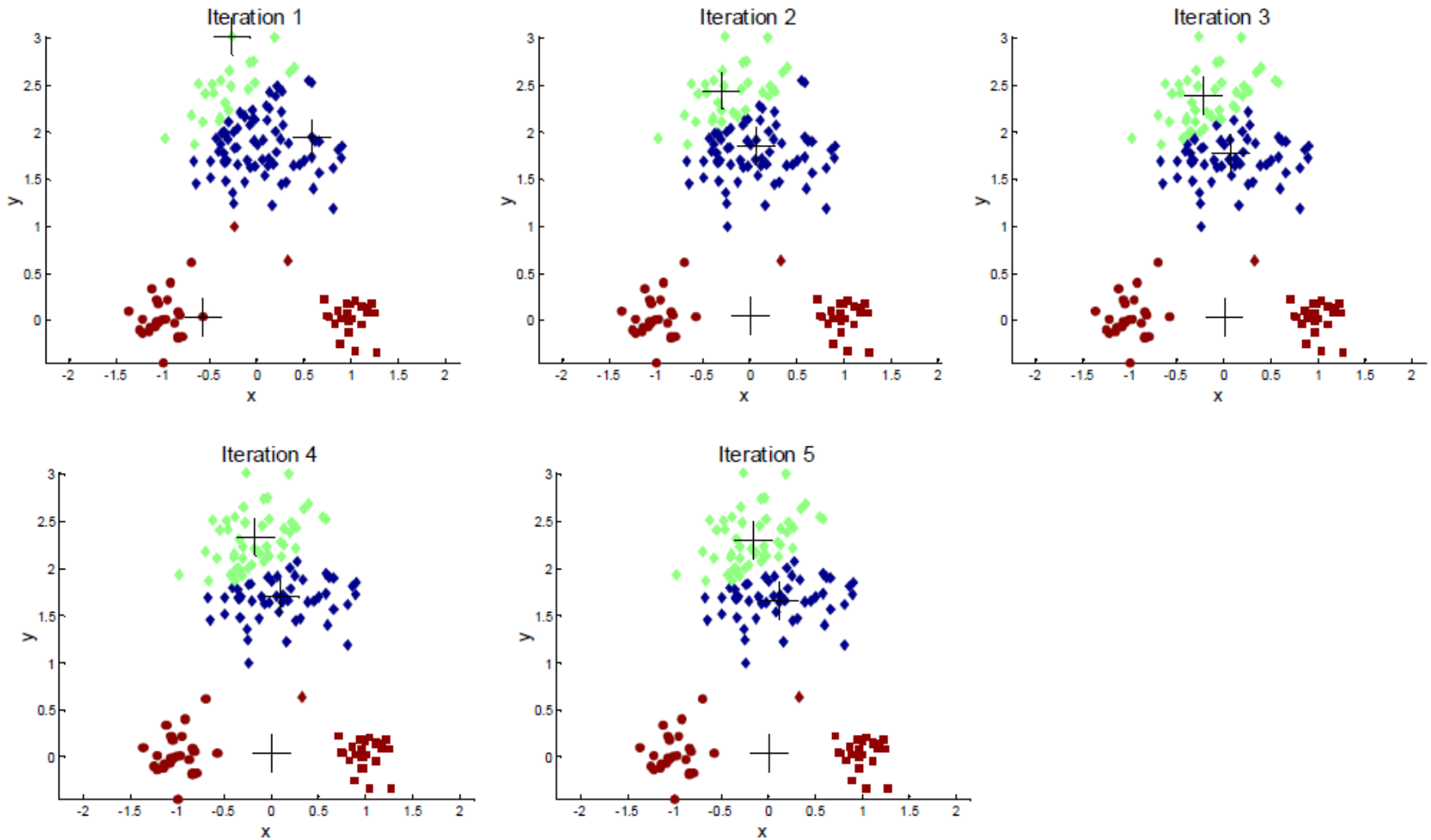
Observações

- As alternativas de inicialização:
  - Amostra do conjunto de dados escolhida aleatoriamente
  - Pontos sorteados aleatoriamente
  - Gerar um conjunto inicial a partir de um outro algoritmo determinístico.
- O algoritmo pode gerar um grupo vazio durante o processo. Neste caso, as alternativas mais comuns são:
  - Continuar o processo com K-1 grupos
  - Sortear um novo centro
  - Criar um novo grupo com o centro mais distante da iteração anterior.

Exemplo: inicialização funciona



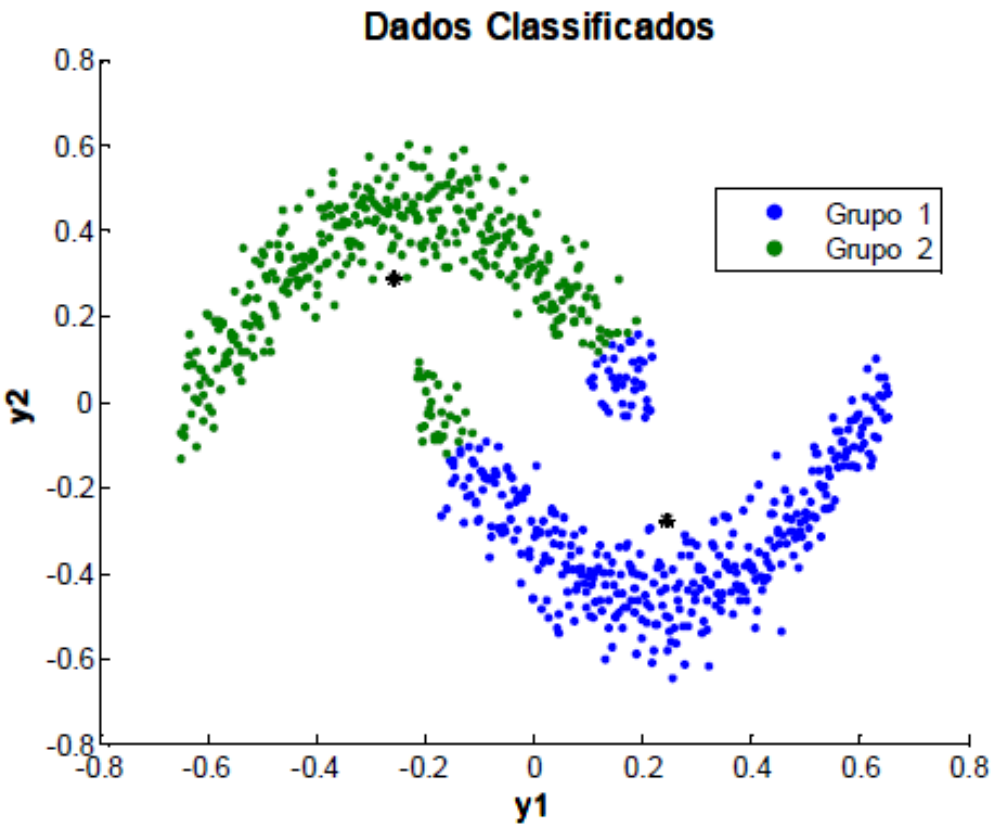
Exemplo: inicialização não funciona

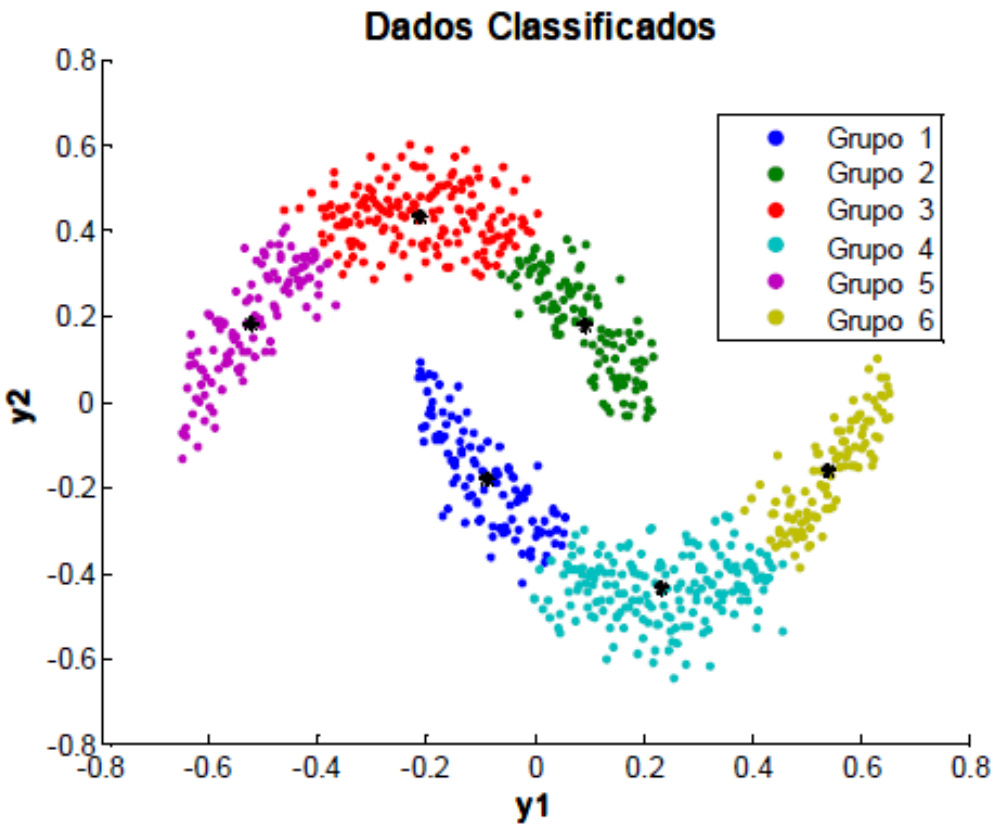


Limitações: formatos dos grupos

- Uma das maiores dificuldades do algoritmo é que a quantidade de grupos deve ser definida à priori.
- É possível usar o algoritmo para encontrar um número maior de grupos e posteriormente diminuir esta quantidade até um ponto que seja interpretável.
- O exemplo a seguir mostra que não foi possível identificar os dois grupos com K=2, porém ao usarmos K=6, isso torna-se possível pela agregação de grupos.

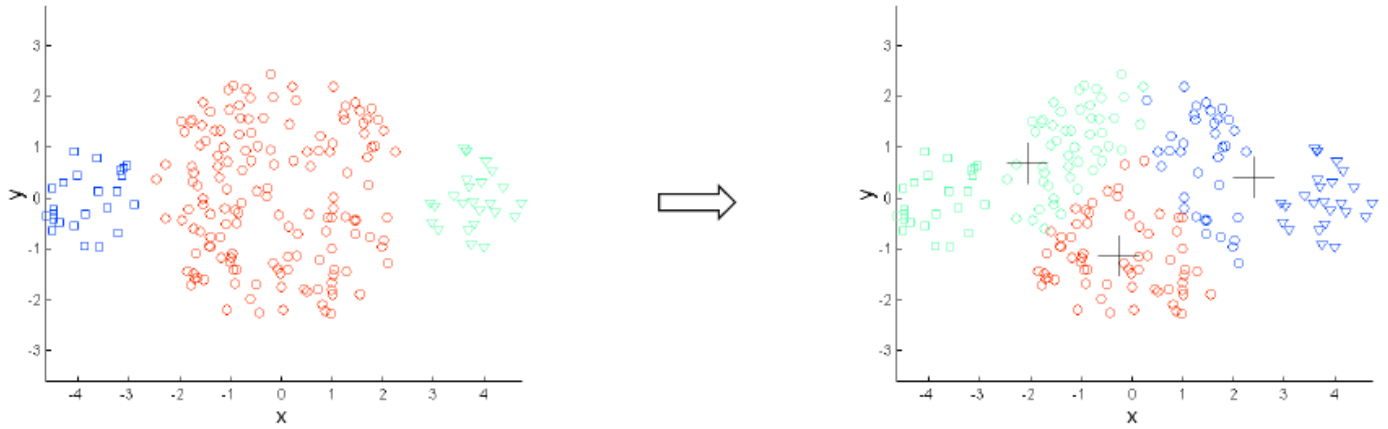
Limitações: formatos dos grupos





Limitações

- Tamanhos diferentes



- Densidades diferentes