Análise de Agrupamento de Dados
(Aula 3.2 – Métodos Particionais)

Prof. Eduardo Raul Hruschka

Departamento de Ciências de Computação
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo

Agenda

3. Métodos para Agrupamento de Dados.

   3.1 Métodos Hierárquicos

   3.2 Métodos Particionais:

3.2.1 Partições Rígidas

- Considere um conjunto de objetos $\mathbf{X}=\{\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_N\}$ a serem agrupados;

- Numa partição rígida, os objetos devem ser reunidos em k grupos não sobrepostos $\mathbf{C}=\{\mathbf{C}_1,\mathbf{C}_2,...,\mathbf{C}_k\}$ tal que:
$\mathbf{C}_1 \cup \mathbf{C}_2 \cup ... \cup \mathbf{C}_k = \mathbf{X}$; $\mathbf{C}_i \neq \varnothing$; $\mathbf{C}_i \cap \mathbf{C}_j = \varnothing$ para $i \neq j$.

- Definir um mapeamento $f:\mathbf{X}\rightarrow\{1,2,...,k\}$ para o qual cada $\mathbf{x}_i$ é atribuído a um *cluster* $\mathbf{C}_j$, $1 \leq j \leq k$. Um *grupo de dados* $\mathbf{C}_j$ contém precisamente os objetos mapeados para o mesmo, i.e.:$\mathbf{C}_j = \{ \mathbf{x}_i \mid f(\mathbf{x}_i)=j, 1 \leq i \leq N, 1 \leq j \leq k \text{ e } \mathbf{x}_i \in \mathbf{X} \}$.

- Assumindo-se que *k* seja conhecido, o número de maneiras (*NM*) de se agrupar *n* objetos em *k clusters* é dado por (Liu, 1968):

$$NM(N,k) = \frac{1}{k!}\sum_{i=0}^{k}(-1)^i \binom{k}{i}(k-i)^N$$

- Por exemplo, *NM* (100,5)≈56.6x10^{67}. Supondo que um computador tenha capacidade de avaliar 10^9 partições/s, precisaríamos de aproximadamente 1.8x10^{50} séculos para processar todas as avaliações.

- Claramente, em grande parte das aplicações reais, enumerar e avaliar todas as partições possíveis é inviável sob o ponto de vista computacional.

- Otimizar uma função objetivo, usando técnicas baseadas em subida de encosta, é uma abordagem comum.

Iniciaremos por estudar um algoritmo amplamente usado na prática:

- *k*-médias (*k-means*);

E uma extensão:

- *Bisecting k-means*.

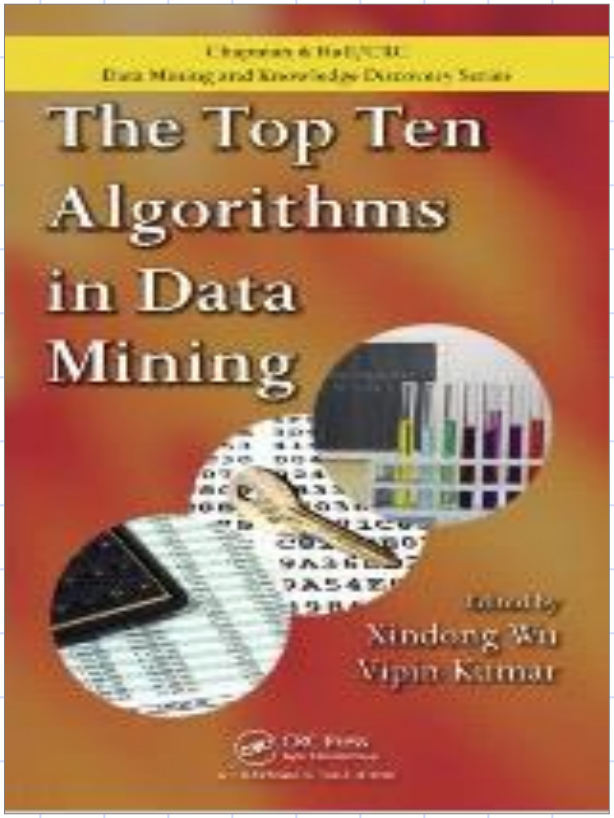
Algoritmo *k*-médias (MacQueen, 1967)

- Amplamente usado na prática:

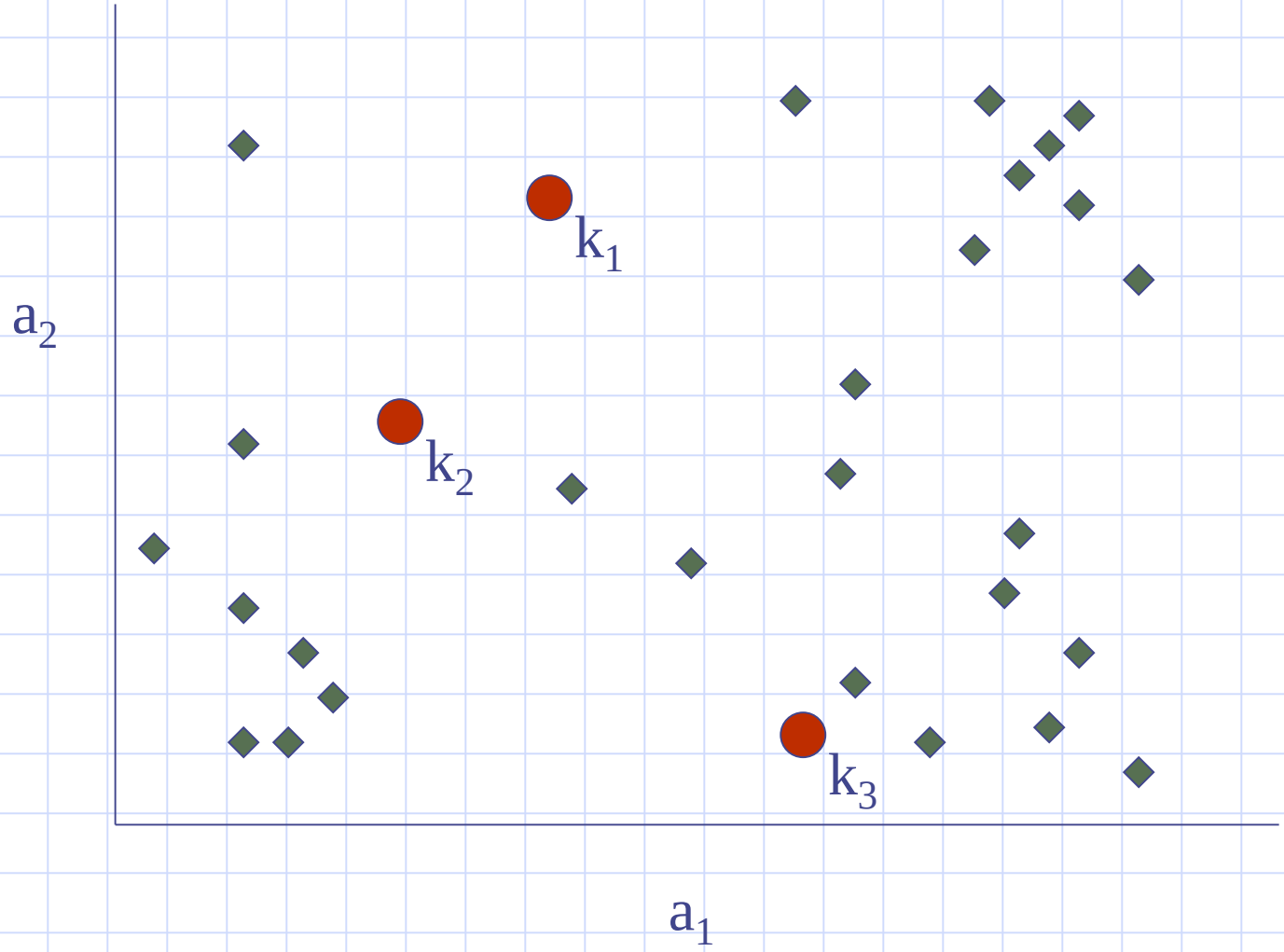
➢ Simplicidade;

➢ Interpretabilidade;

➢ Eficiência Computacional.

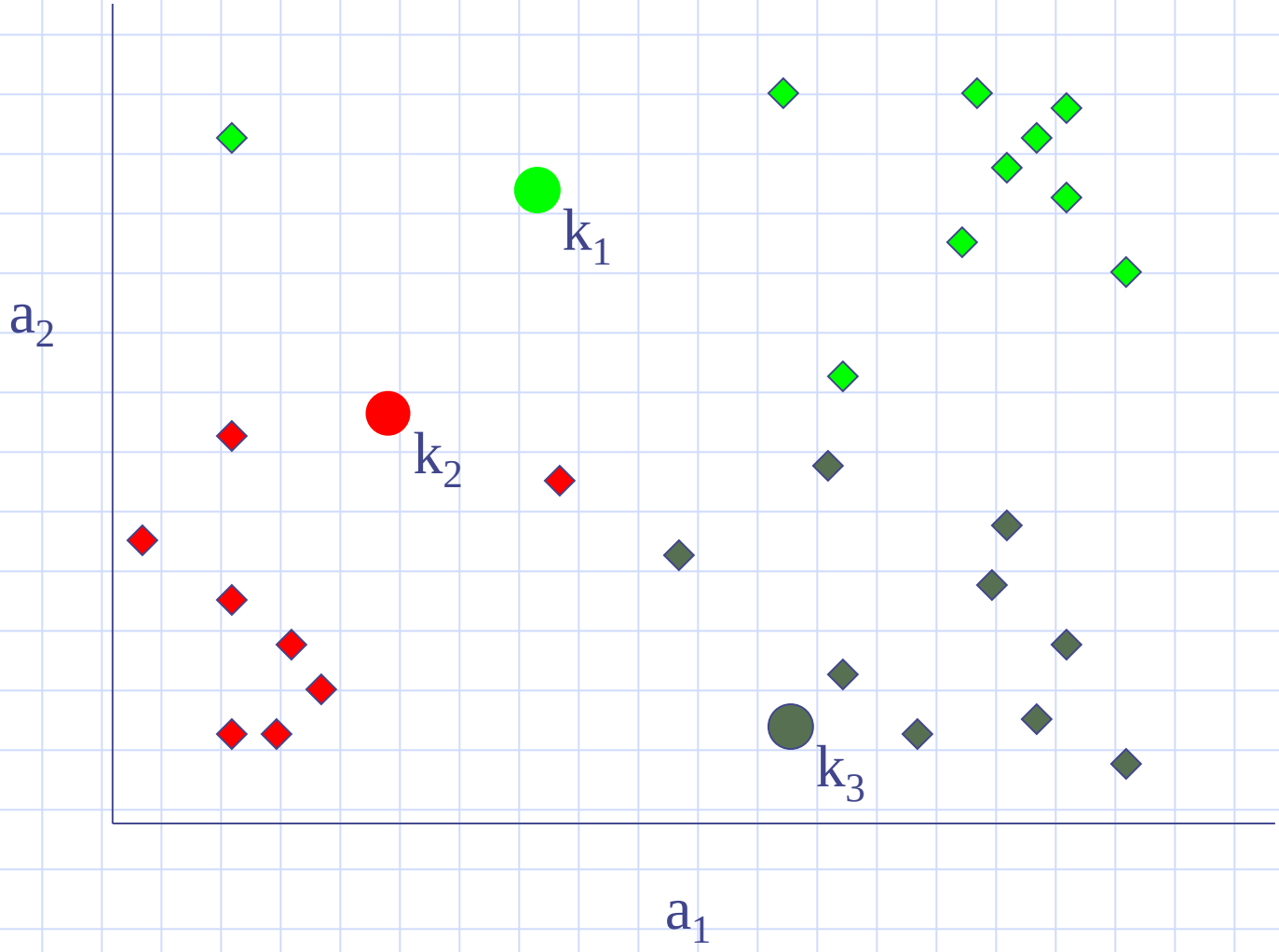


- Assumamos que queremos encontrar três *clusters* (k = 3) para uma base de dados bi-dimensional:

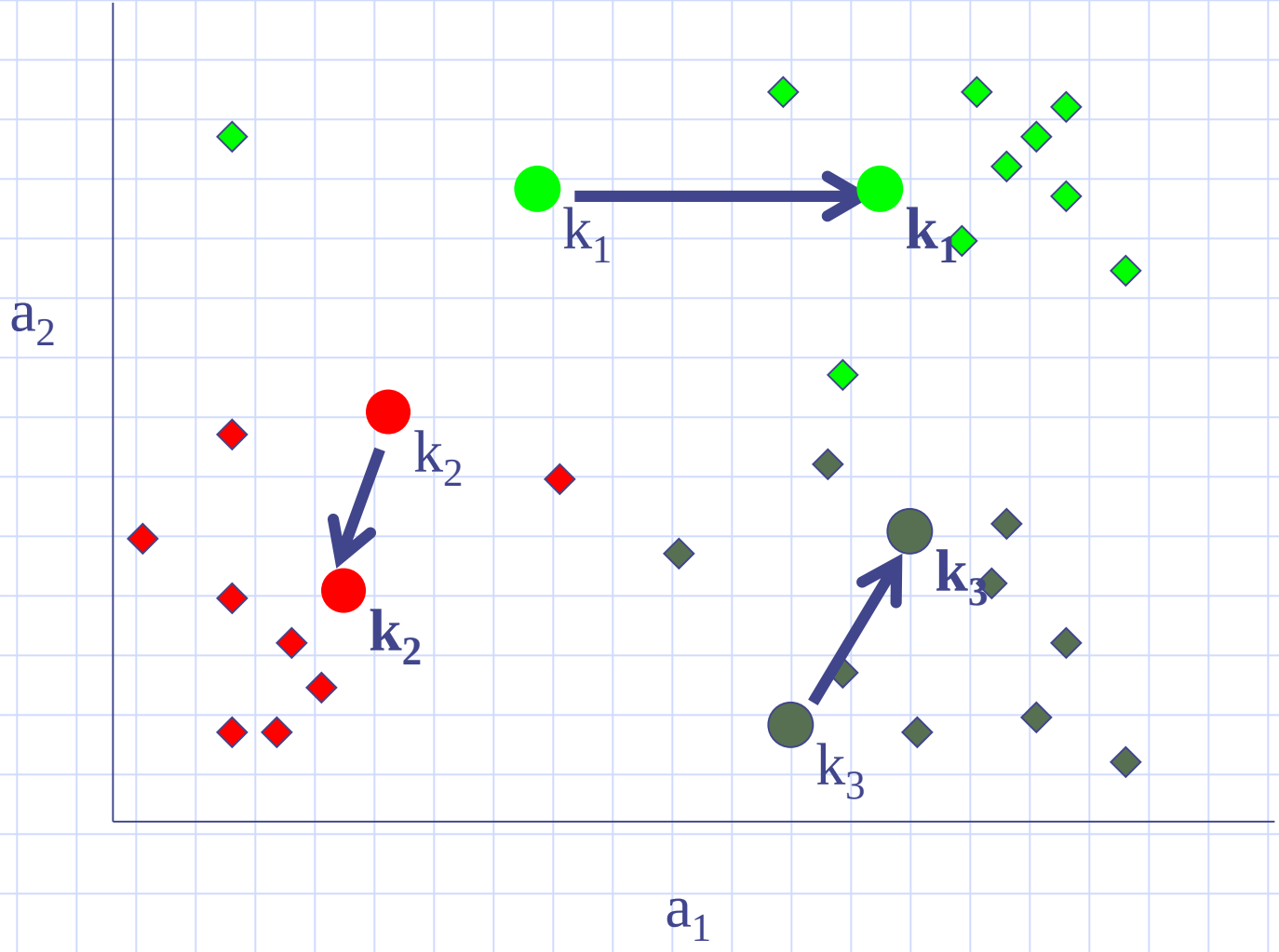


From the slides by Gregory Piatetsky-Shapiro & Gary Parker - available at www.kdnuggets.com

- Calcular dissimilaridades entre objetos e protótipos (k_1,k_2,k_3), encontrando grupos iniciais pela regra do vizinho mais próximo:

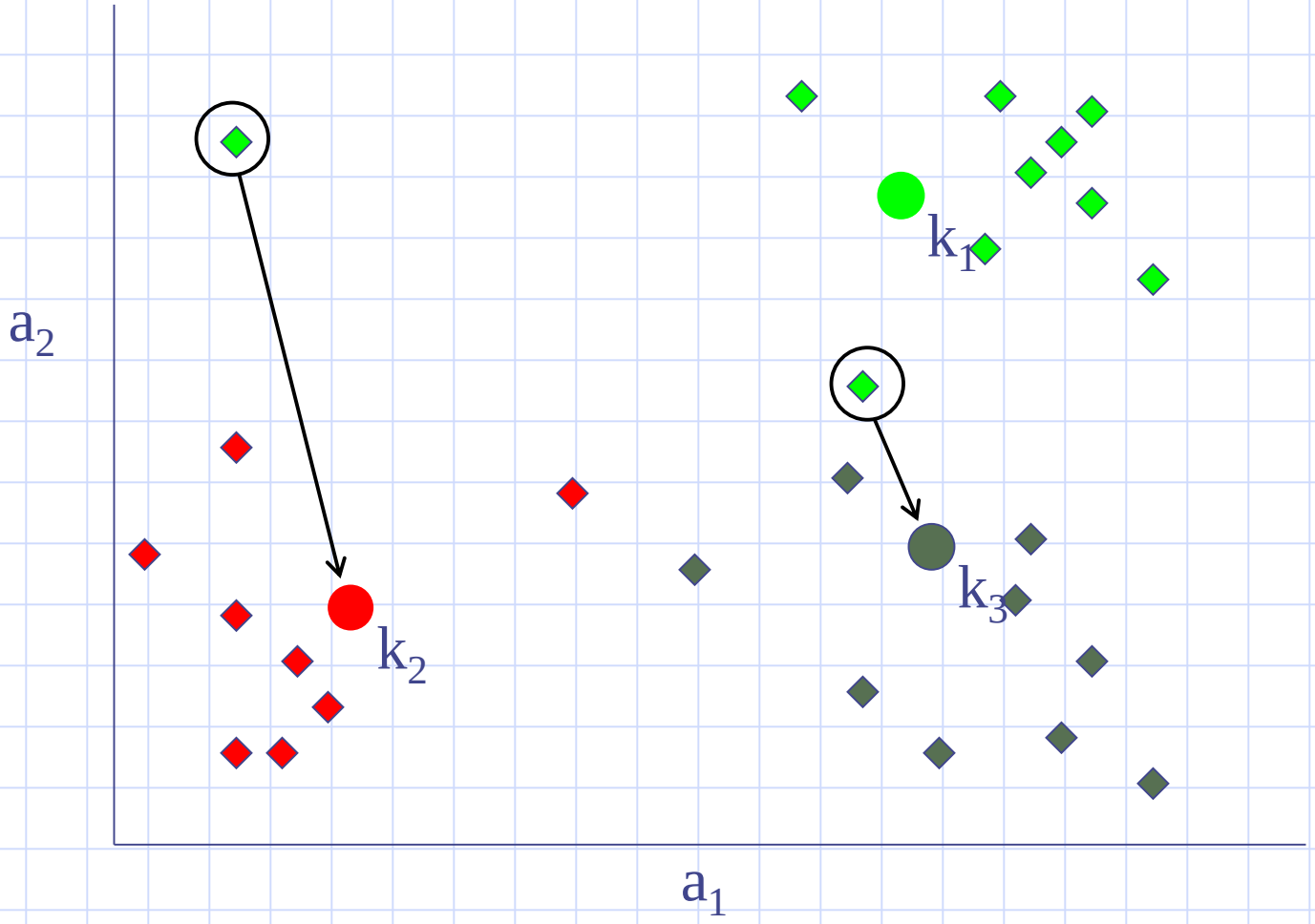


- Atualizar os protótipos (centróides) dos grupos:



From the slides by Gregory Piatetsky-Shapiro & Gary Parker - available at www.kdnuggets.com

- Calcular dissimilaridades entre objetos e centróides;
- Atualizar *clusters* (regra do vizinho mais próximo);



- Repetir até convergência/ número de iterações.

From the slides by Gregory Piatetsky-Shapiro & Gary Parker - available at www.kdnuggets.com

Algoritmo básico:

1. Selecionar *k* pontos (*centróides* iniciais);
2. Repetir até "convergir":
   2.1 Formar *k* grupos atribuindo cada ponto ao seu centróide mais próximo;
   2.2 Re-computar o centróide (média) de cada grupo;

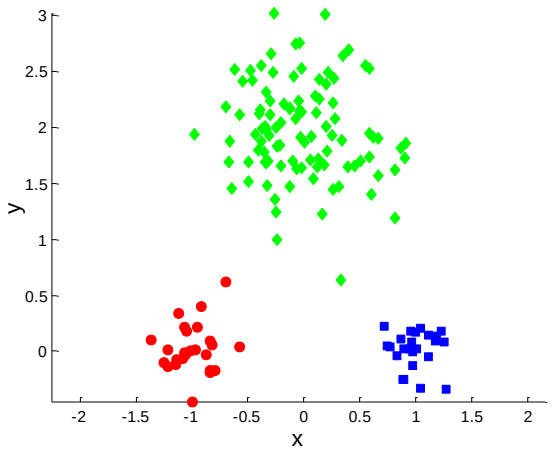
Obs. Convergência: estrita, aproximada, número de iterações.

Detalhes sobre o *k*-médias:

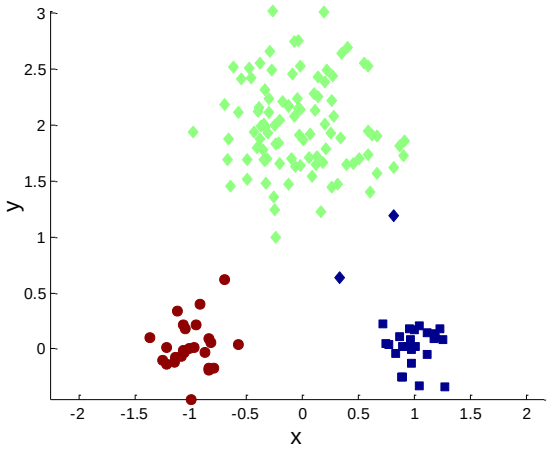
- Centróides iniciais são freqüentemente escolhidos aleatoriamente.
  - *Clusters* obtidos podem variar de uma rodada para outra.
- *Proximidade* medida por meio de Distância Euclidiana (ao quadrado).
- *k*-médias converge, geralmente em poucas iterações;
- Complexidade de tempo é O($n \cdot k \cdot I \cdot d$).

- Vejamos alguns exemplos interessantes...

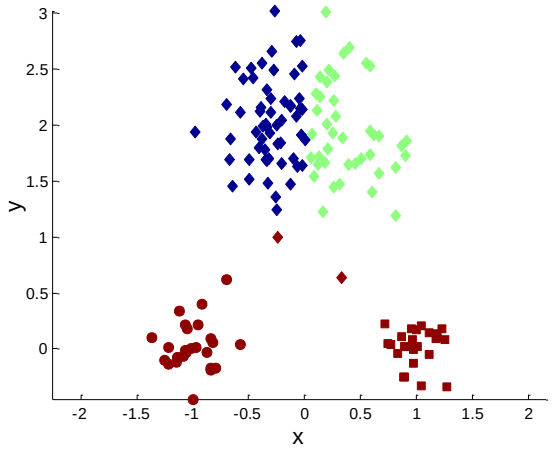
Consideremos duas partições diferentes obtidas para $k = 3$:



Pontos originais



Partição ótima



Partição Sub-ótima

Tan, Steinbach & Kumar, Introduction to Data Mining, Pearson, 2006.

Avaliando os grupos obtidos:

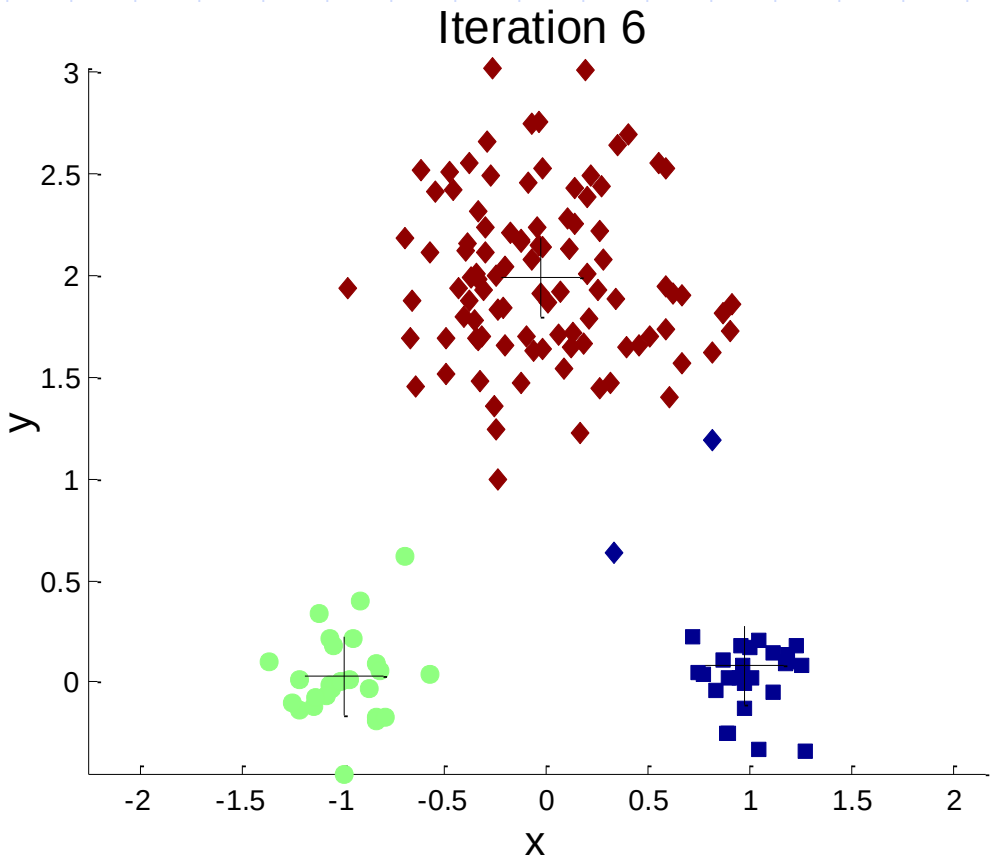
- Soma dos erros quadráticos:

$$SEQ = \sum_{i=1}^{k} \sum_{x \in C_i} dist^2(m_i, x) \qquad m_i = \frac{\sum_{x \in C_i} x}{|C_i|}$$

- Dadas duas partições, escolher aquela que apresenta SEQ menor;

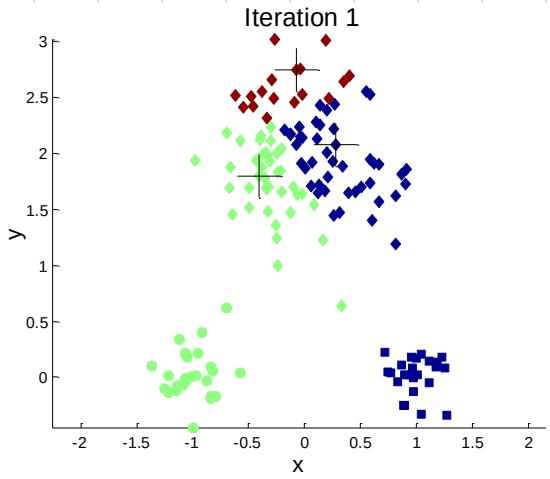
- E se o número de *clusters* for diferente?
  - Aumento de *k* : tende a diminuir, por si só, SEQ;

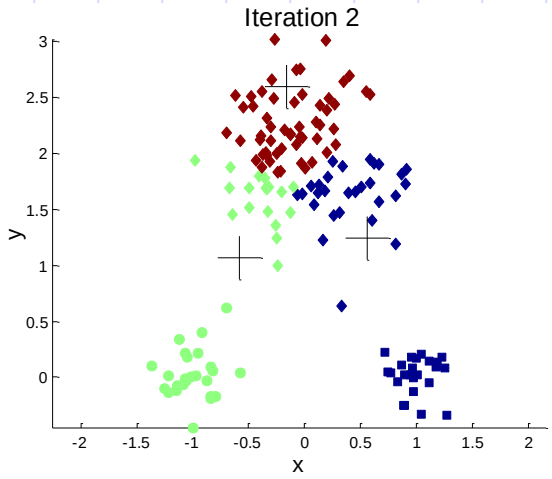
Importância da escolha dos centróides iniciais...

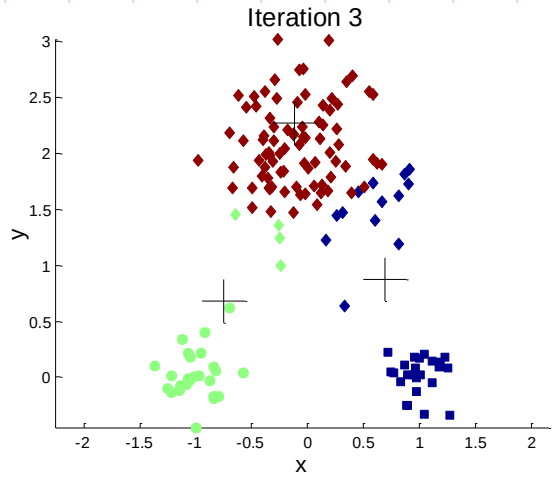


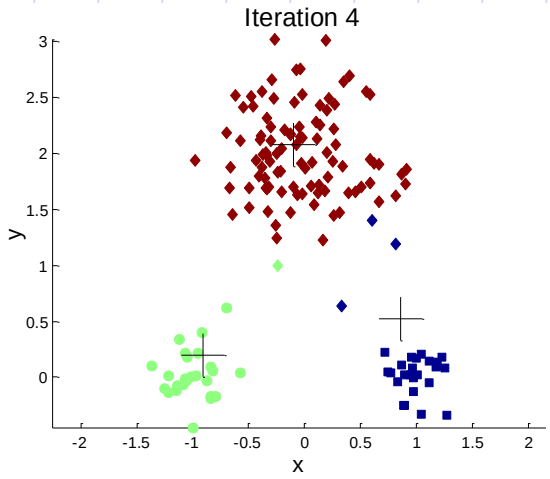
Tan, Steinbach & Kumar, Introduction to Data Mining, Pearson, 2006.

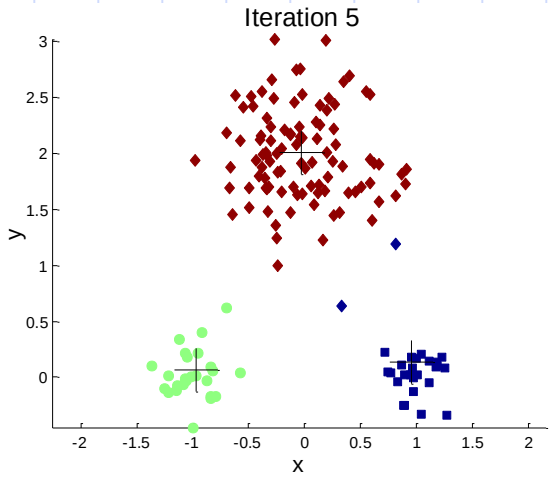
Importância da escolha dos centróides iniciais...

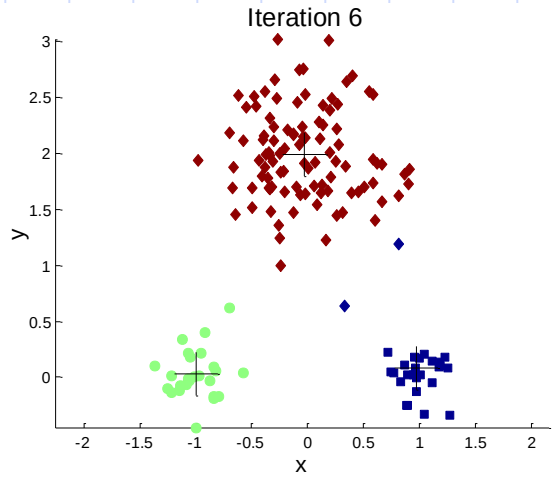






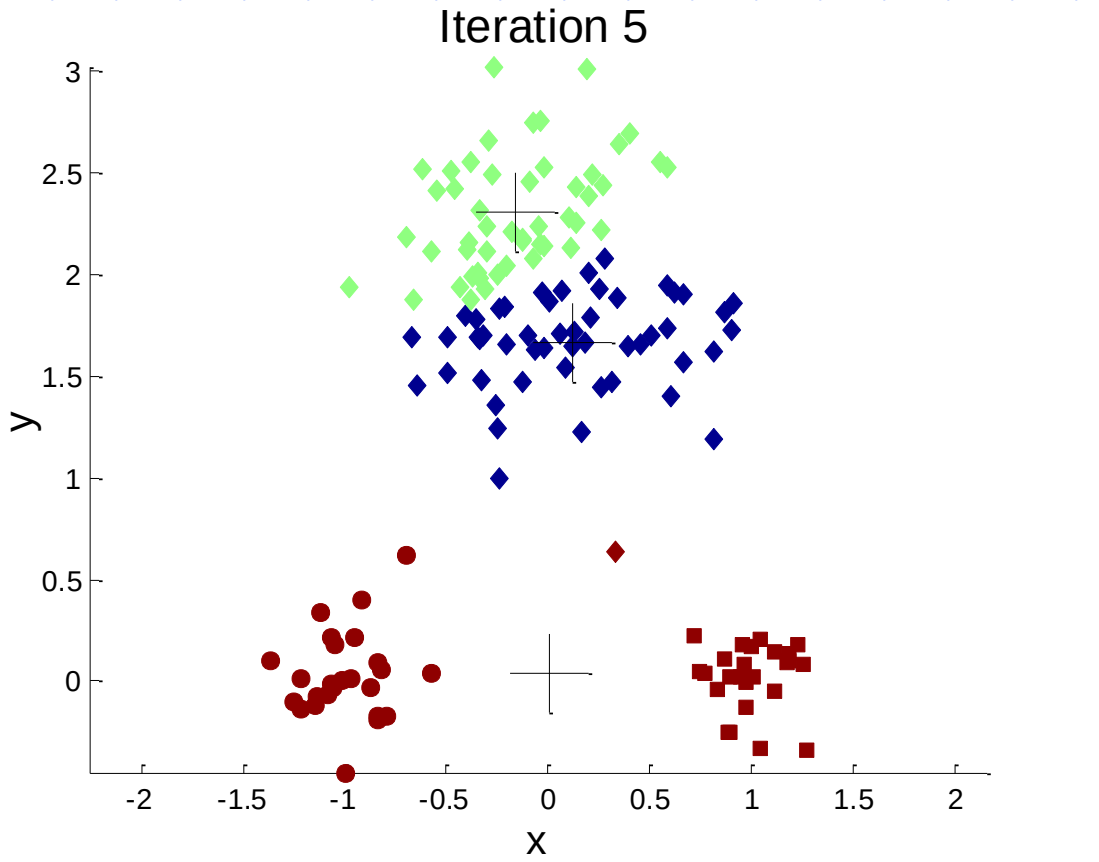






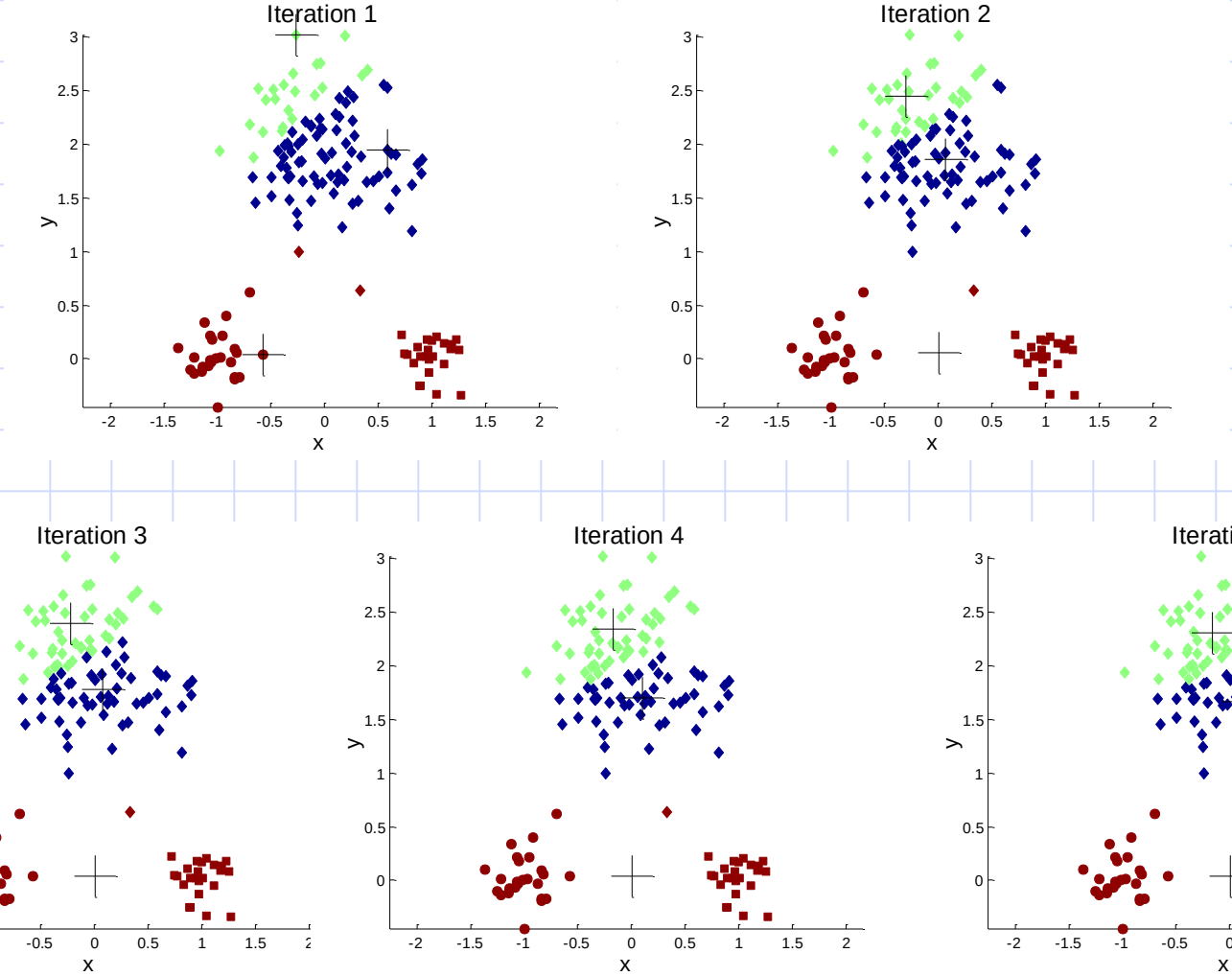
Tan, Steinbach & Kumar, Introduction to Data Mining, Pearson, 2006.

Importância da escolha dos centróides iniciais...



Tan, Steinbach & Kumar, Introduction to Data Mining, Pearson, 2006.

Importância da escolha dos centróides iniciais...



Tan, Steinbach & Kumar, Introduction to Data Mining, Pearson, 2006.

Soluções para inicialização?

- Múltiplas execuções:
  - Ajuda, mas pequena $P_{sucesso}$;
- Amostragem via métodos hierárquicos;
- Seleção "informada" de centróides distantes entre si;
- Algoritmos de busca (e.g., evolutivos);

Pré/Pós-processamento

- Pré-processamento:
  - Normalização;
  - Eliminação de *outliers.*
- Pós-processamento:
  - Eliminar pequenos *clusters* (*outliers*)?
  - Dividir grupos com EQ relativamente alto?
  - Unir grupos próximos e com EQ pequeno?
  - Usar tais passos durante iterações do algoritmo das *k*-médias?

Bisecting k-means :

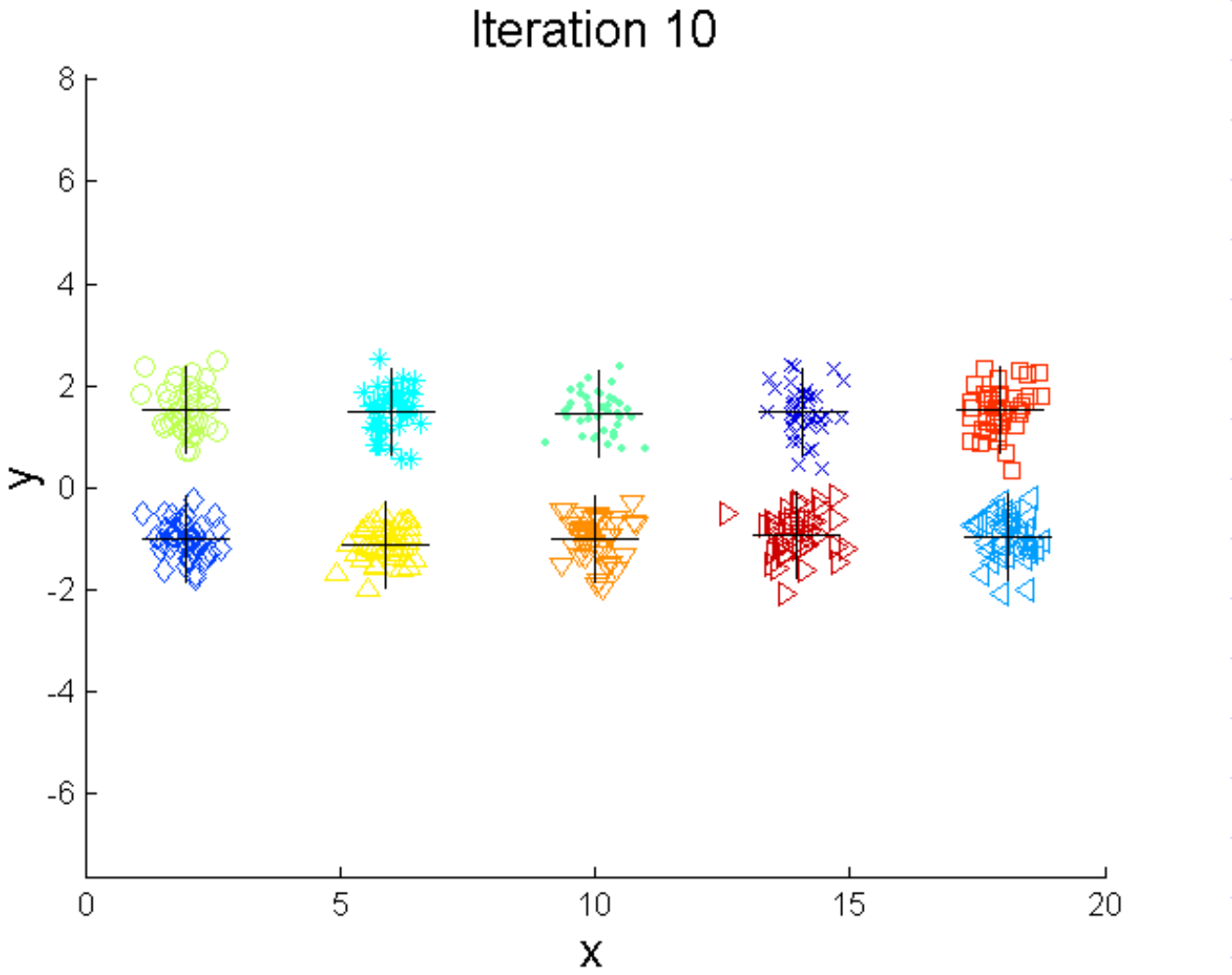
- Variante do *k*-médias que pode produzir uma partição ou uma hierarquia:

Algoritmo básico:

1. Inicializar uma lista de grupos (inicialmente um único grupo);
2. Repetir (até que a lista de grupos contenha *k* grupos):
    2.1. Selecionar um grupo da lista;
    2.2. Para i=1 até número de inicializações (N_i) fazer:
        - Dividir o grupo selecionado usando *2*-médias (k=2);
    2.3 Adicionar os dois grupos de menor SEQ à lista de grupos;

➢ Dado que a probabilidade de se selecionar um protótipo inicial para cada um dos dois grupos (balanceados) é $k!/k^k$, qual seria um valor razoável para N_i ?

Exemplo:

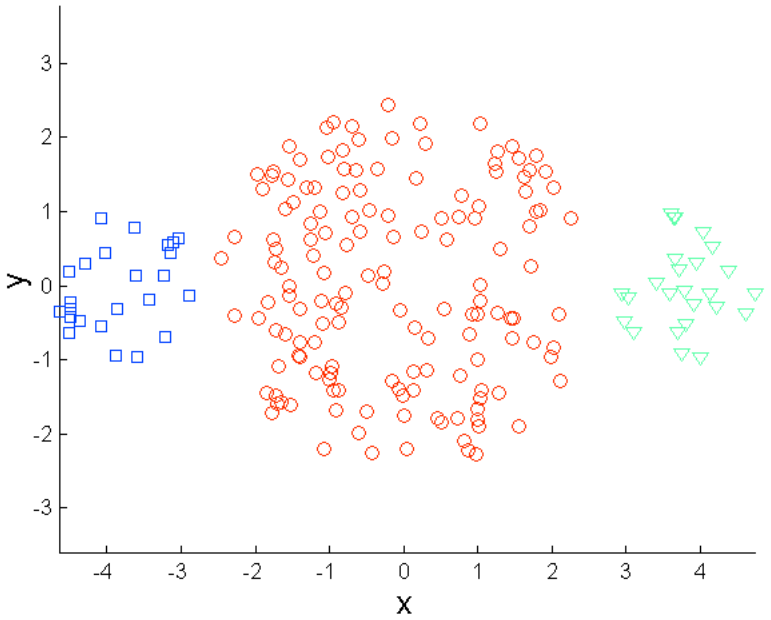


Limitações do *k*-médias:

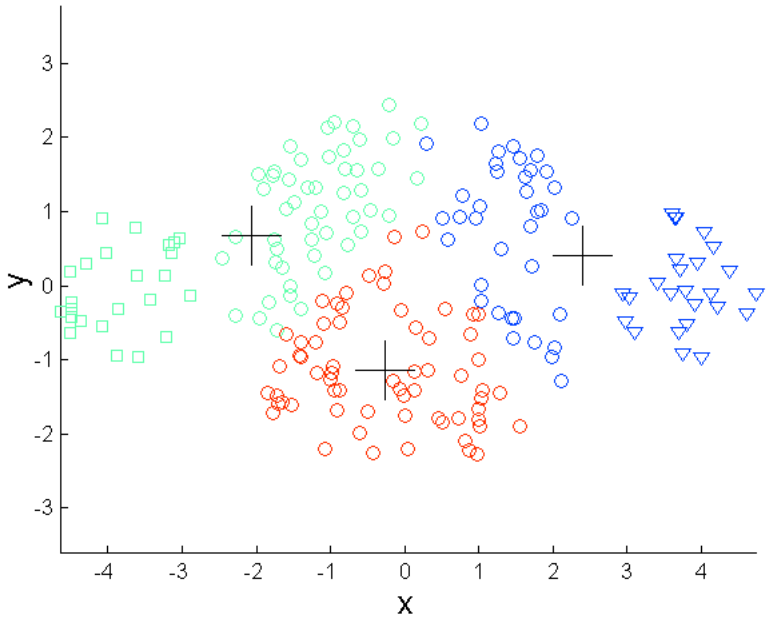
- Grupos de diferentes:
  - Tamanhos;
  - Densidades;
  - Formas não globulares.

- *Outliers*.

Limitações do *k*-médias: grupos de tamanhos diferentes



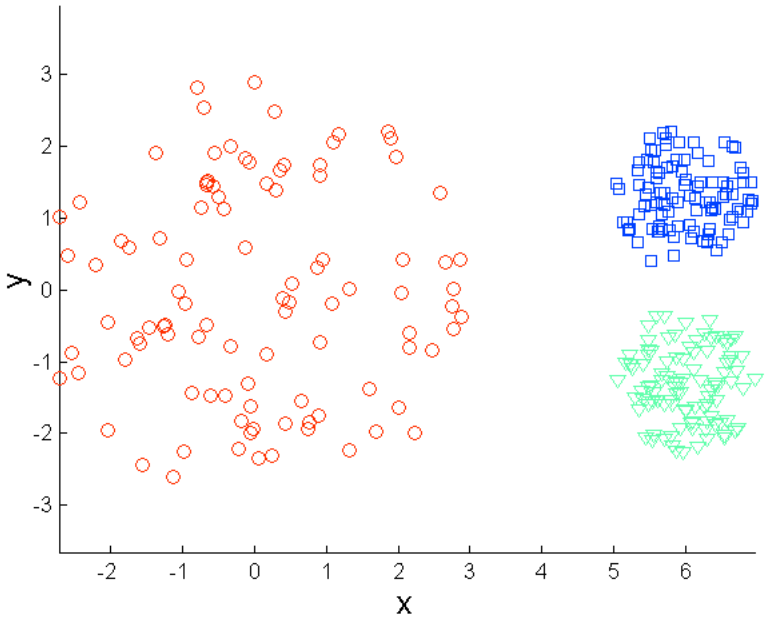
Pontos originais



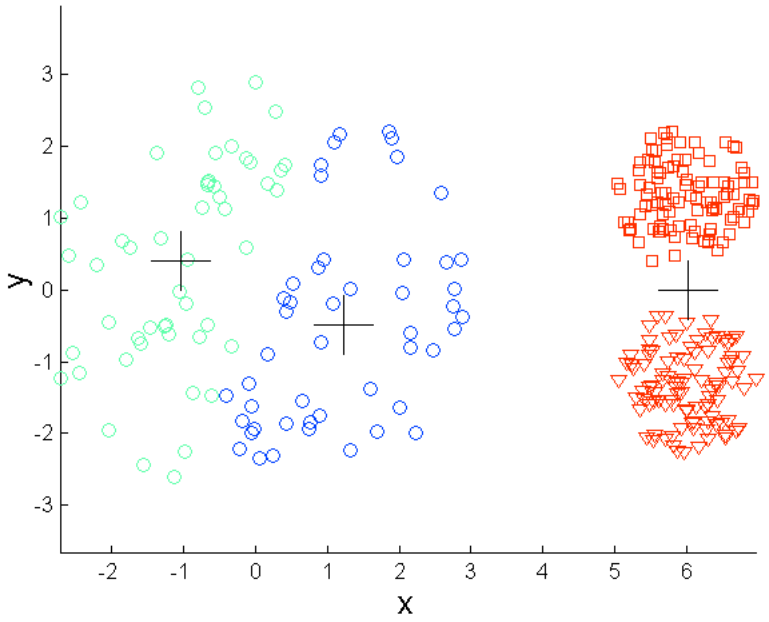
3-médias

Tan, Steinbach & Kumar, Introduction to Data Mining, Pearson, 2006.

Limitações do *k*-médias: densidades diferentes



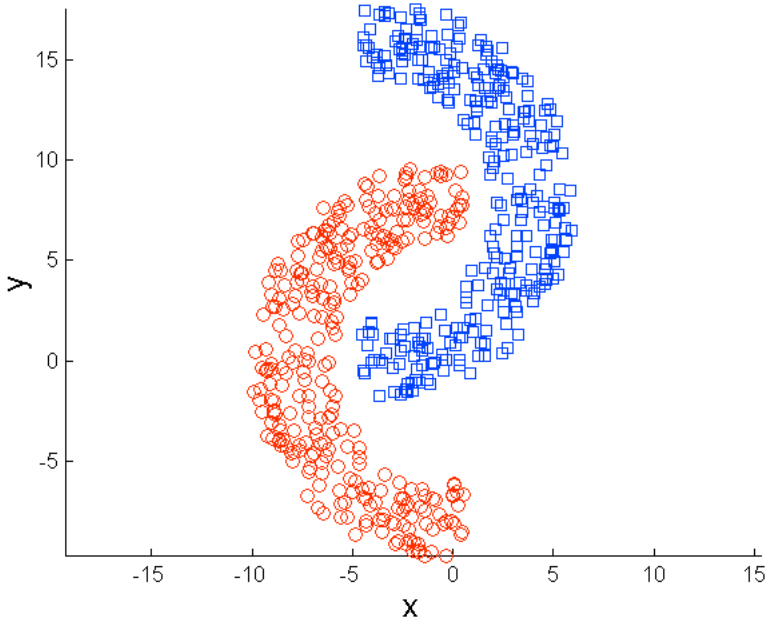
Pontos originais



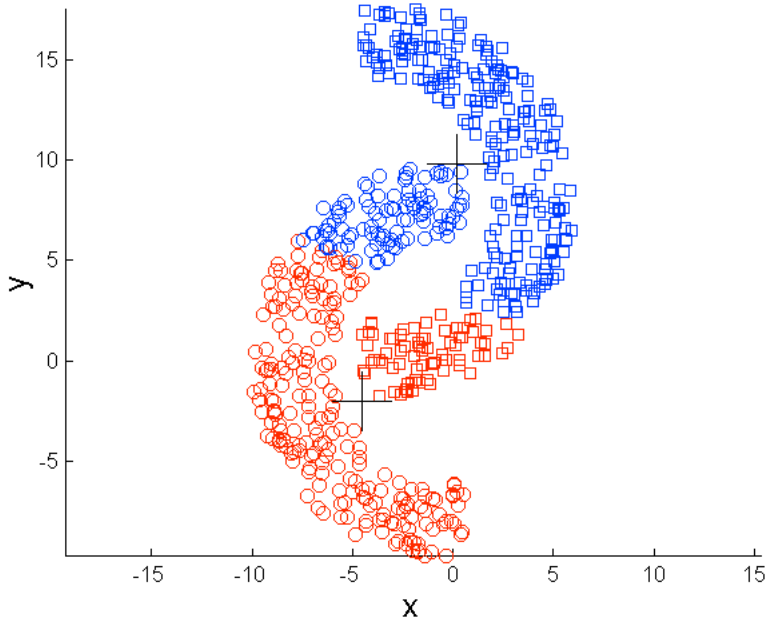
3-médias

Tan, Steinbach & Kumar, Introduction to Data Mining, Pearson, 2006.

Limitações do *k*-médias: formas não globulares



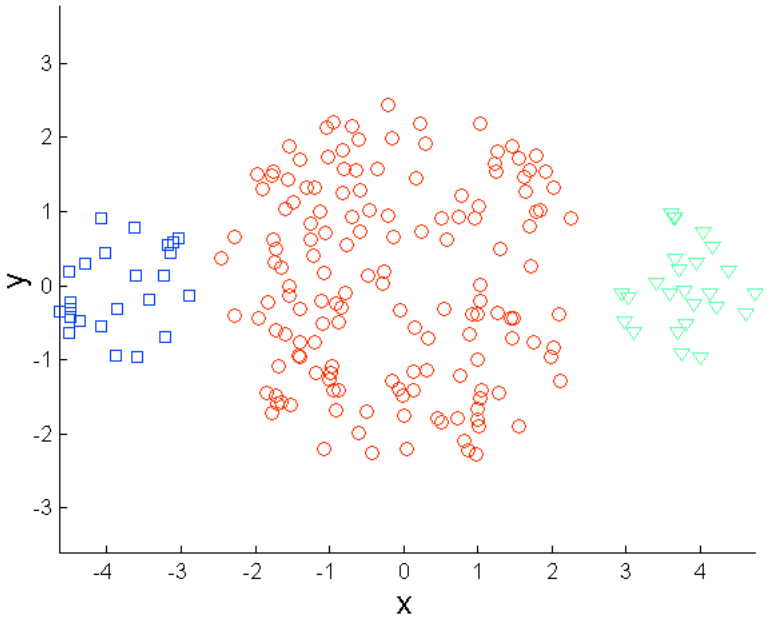
Pontos originais



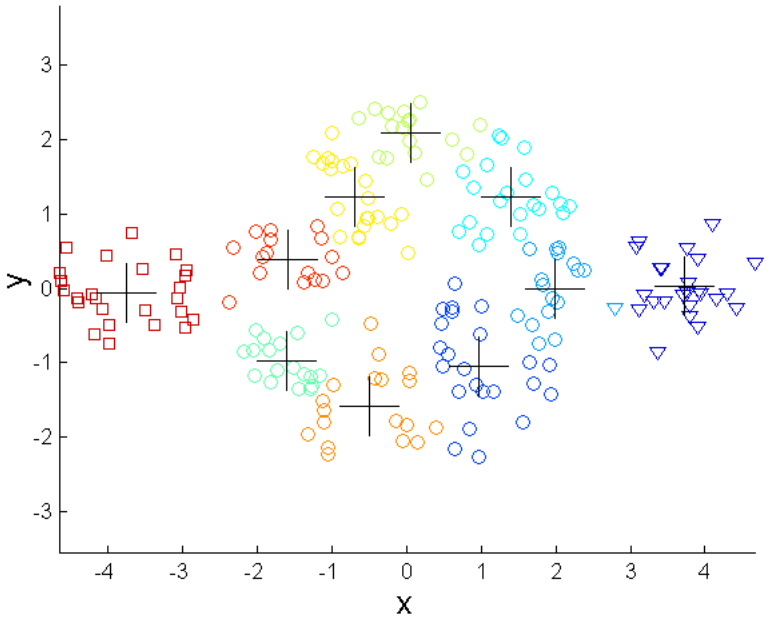
2-médias

Tan, Steinbach & Kumar, Introduction to Data Mining, Pearson, 2006.

Superando algumas limitações do *k-médias*



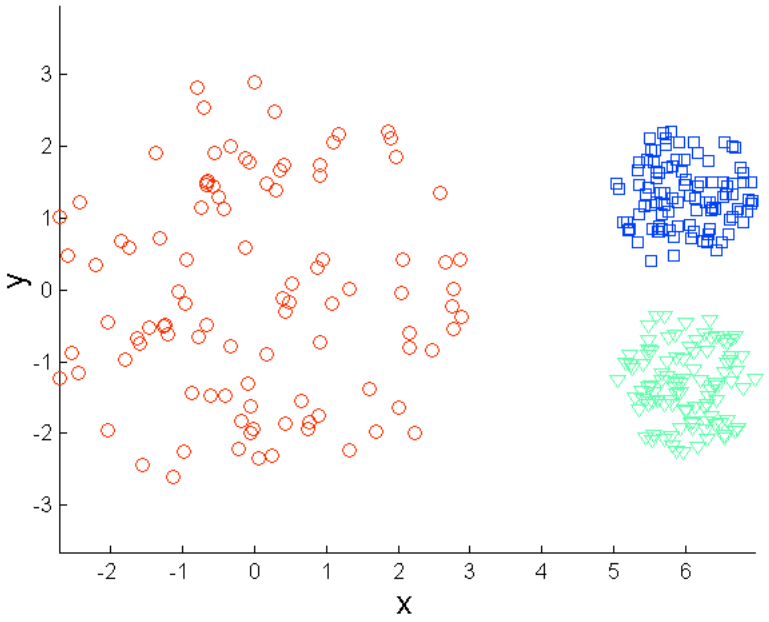
Pontos originais



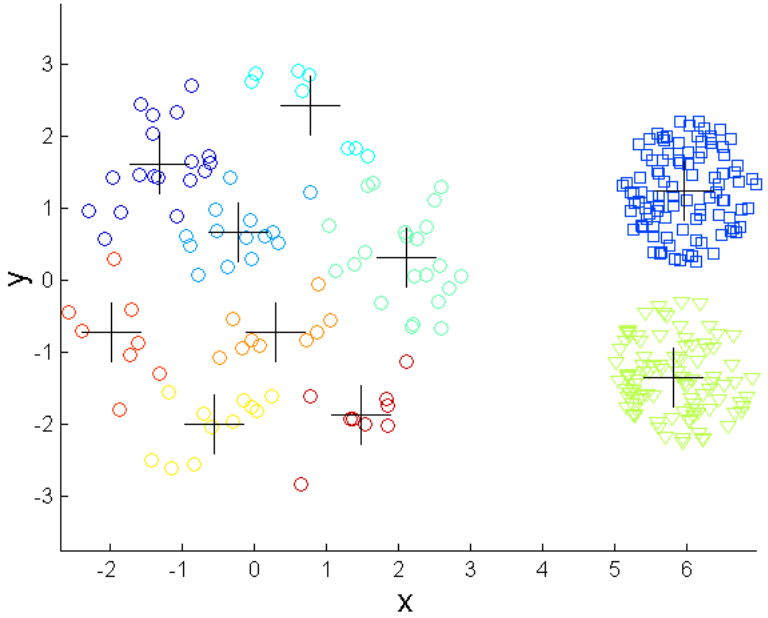
"Mais Grupos"

Tan, Steinbach & Kumar, Introduction to Data Mining, Pearson, 2006.

Superando algumas limitações do *k-médias*...



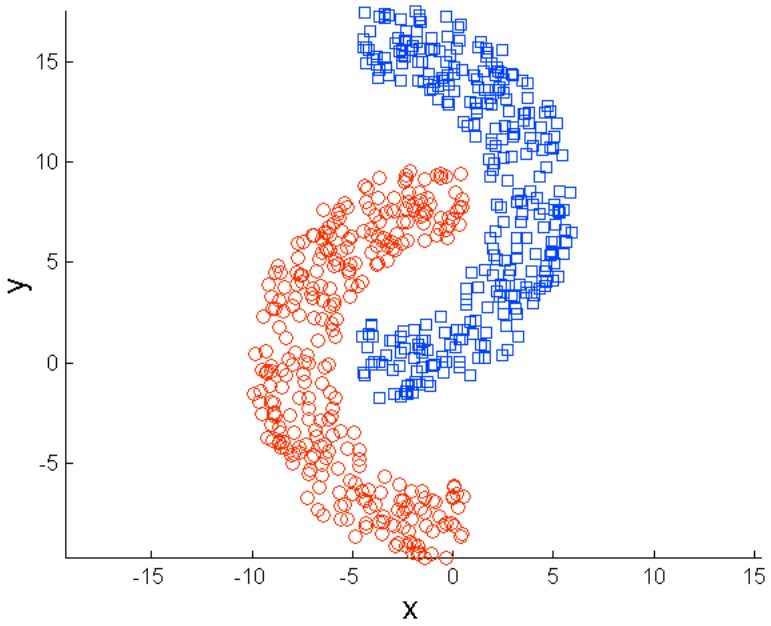
Pontos originais

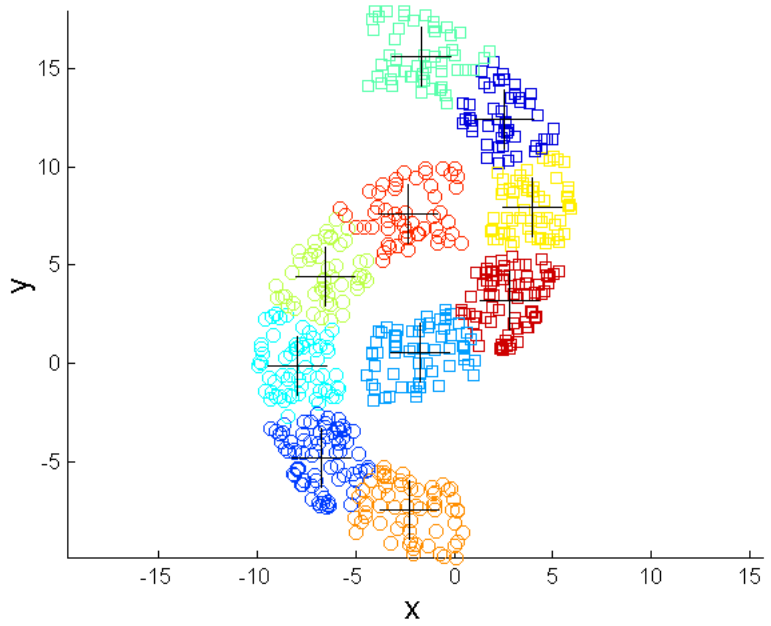


"Mais grupos"

Tan, Steinbach & Kumar, Introduction to Data Mining, Pearson, 2006.

Superando algumas limitações do *k-médias*...





Pontos originais

"Mais grupos"

Tan, Steinbach & Kumar, Introduction to Data Mining, Pearson, 2006.

Principais referências usadas para preparar essa aula:

- Xu, R., Wunsch, D., **Clustering**, IEEE Press, 2009.
  - Capítulo 4.

- Tan, Steinbach & Kumar, **Introduction to Data Mining**, Pearson, 2006.
  - Capítulo 8, pp. 496-515.

- Jain, A. K., Dubes, R. C., **Algorithms for Clustering Data**, Prentice Hall, 1988.
  - Capítulo 3, pp. 89-142.

- Bishop, C. M., **Pattern Recognition and Machine Learning**, 2006.
  - Capítulo 9, pp. 423-439.

Onde estamos?

3. Métodos para Agrupamento de Dados.
   - 3.1 Métodos Hierárquicos;
   - 3.2 Métodos Particionais.
4. Agrupamento de textos:
   - Medida de similaridade adequada:
   - Adaptação do k-médias