Algoritmos de Agrupamento - K-means
Mineração de Dados

Ronaldo C. Prati

Universidade Federal do ABC (UFABC) - ronaldo.prati@ufabc.edu.br

Definição de Partição de Dados (Revisão)

- Consideremos um conjunto de N objetos a serem agrupados: $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$
- **Partição** (rígida): coleção de k grupos não sobrepostos $\mathbf{P} = \{\mathbf{C}_1, \mathbf{C}_2, ..., \mathbf{C}_k\}$ tal que:

$$\mathbf{C}_1 \cup \mathbf{C}_2 \cup ... \cup \mathbf{C}_k = \mathbf{X}$$

$$\mathbf{C}_i \neq \varnothing$$

$$\mathbf{C}_i \cap \mathbf{C}_j = \varnothing \text{ para } i \neq j$$

  - Exemplo: $\mathbf{P} = \{ (\mathbf{x}_1), (\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_6), (\mathbf{x}_2, \mathbf{x}_5) \}$

Matriz de Partição

- **Matriz de Partição** é uma matriz com k linhas (no. de grupos) e N colunas (no. de objetos) na qual cada elemento μ_{ij} indica o *grau de pertinência* do j-ésimo objeto ($\mathbf{x}_j$) ao i-ésimo grupo ($\mathbf{C}_i$)

$$U(X) = \begin{bmatrix} \mu_{11} & \mu_{12} & \cdots & \mu_{1N} \\ \mu_{21} & \mu_{22} & \cdots & \mu_{2N} \\ \vdots & & \ddots & \vdots \\ \mu_{k1} & \mu_{k2} & \cdots & \mu_{kN} \end{bmatrix}$$

- Se essa matriz for **binária**, ou seja, $\mu_{ij} \in \{0,1\}$, e ainda, se a restrição $\sum_i(\mu_{ij}) = 1$ $\forall j$ for respeitada, então denomina-se:
  - *matriz de partição* ***rígida***, ***exclusiva*** ou *sem* ***sobreposição***

Matriz de Partição

- **Exemplo:**
  - **P** = { $(\mathbf{x}_1)$, $(\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_6)$, $(\mathbf{x}_2, \mathbf{x}_5)$ }

$$U(X) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 \end{bmatrix}$$

Métodos Particionais (Sem Sobreposição)

- Métodos *particionais* sem sobreposição referem-se a algoritmos de agrupamento que buscam (explícita ou implicitamente) por uma matriz de partição rígida de um conjunto de objetos **X**

Encontrar uma Matriz de Partição U(X): Equivale a particionar o conjunto $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$ de N objetos em uma coleção $\mathbf{C} = \{\mathbf{C}_1, \mathbf{C}_2, ..., \mathbf{C}_k\}$ de k grupos disjuntos $\mathbf{C}_i$ tal que $\mathbf{C}_1 \cup \mathbf{C}_2 \cup ... \cup \mathbf{C}_k = \mathbf{X}$, $\mathbf{C}_i \neq \emptyset$, e $\mathbf{C}_i \cap \mathbf{C}_j = \emptyset$ para $i \neq j$

Particionamento como Problema Combinatório

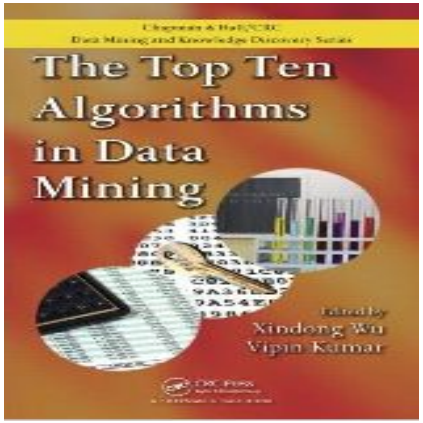
- **Problema**: Assumindo que *k* seja conhecido, o no. de possíveis formas de agrupar *N* objetos em *k clusters* é dado por (Liu, 1968):

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

- Por exemplo, NM(100, 5) ≈ 56.6×10^{67}.
  - Em um computador com capacidade de avaliar 10^9 partições/s, precisaríamos ≈ 1.8×10^{50} séculos para processar todas as avaliações
- Como k em geral é desconhecido, problema é ainda maior...
  - **NP-Hard**: Avaliação computacional exaustiva é impraticável...
- **Solução**: formulações alternativas...

Algoritmo k-Means

- Começaremos nosso estudo com um dos algoritmos mais clássicos da área de **mineração de dados** em geral
  - algoritmo das **k-médias** ou ***k-means***
  - listado entre os **Top 10 Most Influential Algorithms in DM**



- Wu, X. and Kumar, V. (Editors), *The Top Ten Algorithms in Data Mining*, CRC Press, 2009
- X. Wu et al., "Top 10 Algorithms in Data Mining", *Knowledge and Info. Systems*, vol. 14, pp. 1-37, 2008

Algoritmo k-Means

- **Referência Mais Aceita como Original**:

J. B. MacQueen, *Some methods of classification and analysis of multivariate observations*, In Proceedings 5th Berkeley Symposium on Mathematical Statistics and Probability, 1967.

- **Porém...**

"K-means has a rich and diverse history as it was independently discovered in different scientific fields by Steinhaus (1956), Lloyd (proposed in 1957, published in 1982), Ball & Hall (1965) and MacQueen (1967)" [**Jain, *Data Clustering: 50 Years Beyond K-Means*, Patt. Rec. Lett., 2010**]

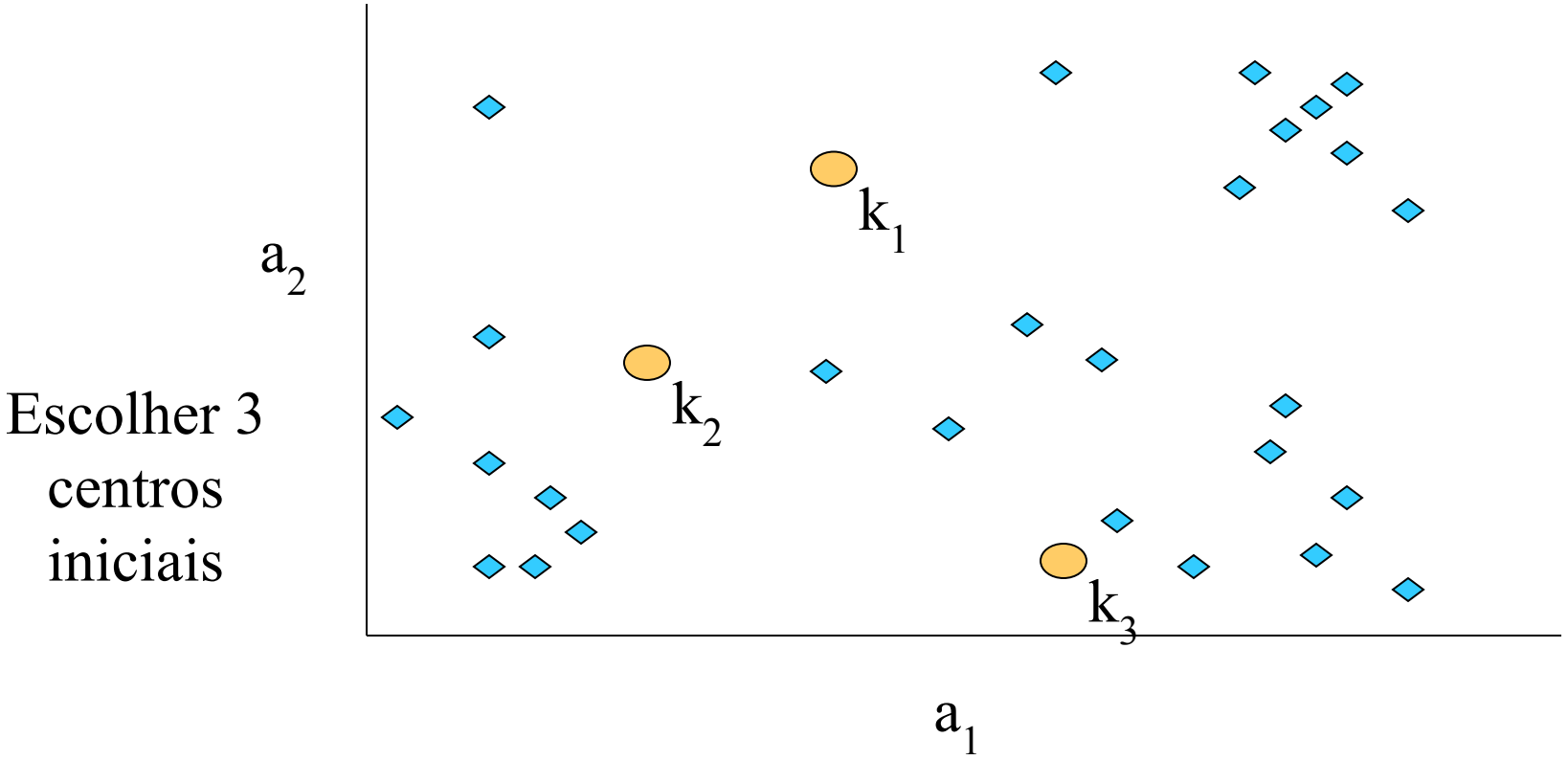
- **... e tem sido assunto por mais de meio século !**Douglas Steinley, *K-Means Clustering: A Half-Century Synthesis*, British Journal of Mathematical and Statistical Psychology, Vol. 59, 2006

k-Means

1) Escolher aleatoriamente *k* protótipos (centros) para os clusters
2) Atribuir cada objeto para o cluster de centro mais *próximo* (segundo alguma distância, e.g. Euclidiana)
3) Mover cada centro para a média (centróide) dos objetos do cluster correspondente
4) Repetir os passos 2 e 3 até que algum critério de convergência seja obtido:

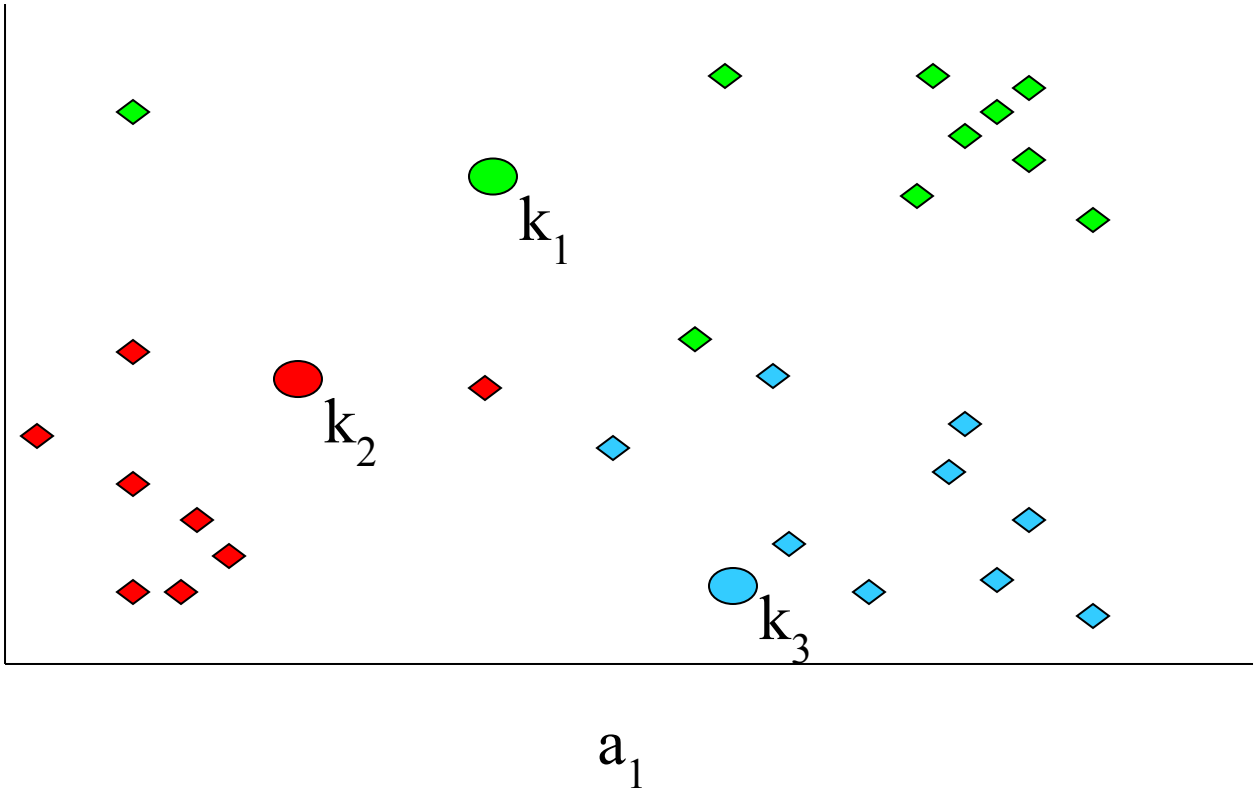
   - número máximo de iterações
   - limiar mínimo de mudanças nos centróides

k-Means - passo 1:



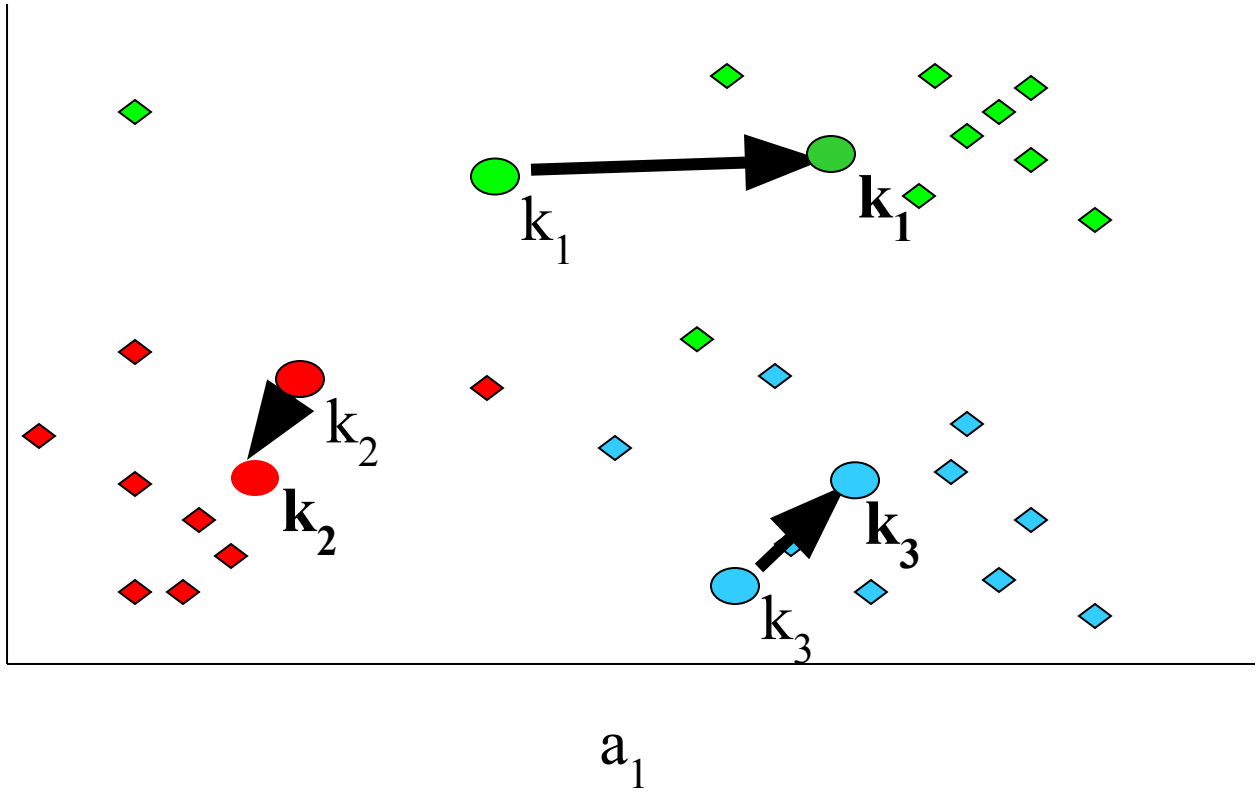
k-Means - passo 2:

Atribuir cada objeto ao cluster de centro + próximo



k-Means - passo 3:

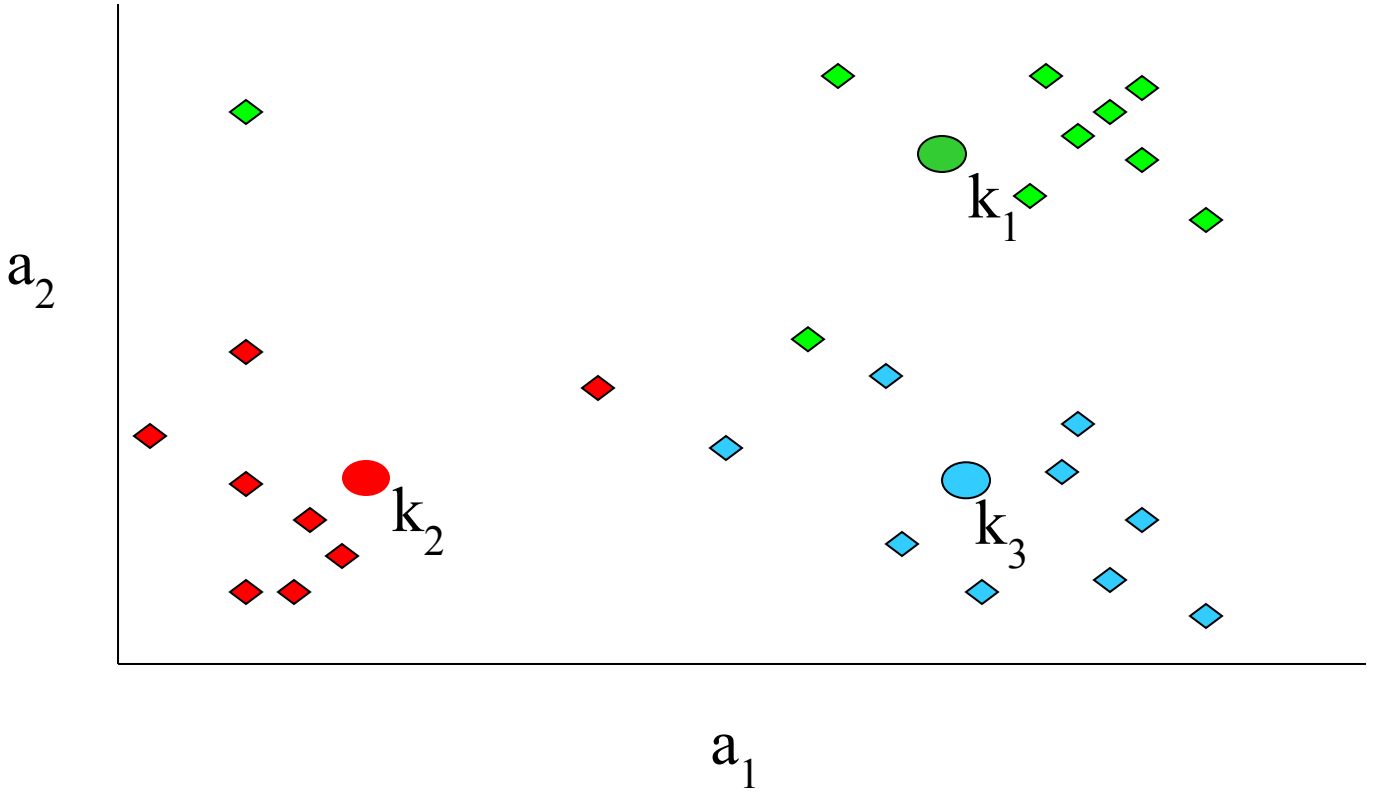
Mover cada centro para o vetor médio do cluster (centróide)



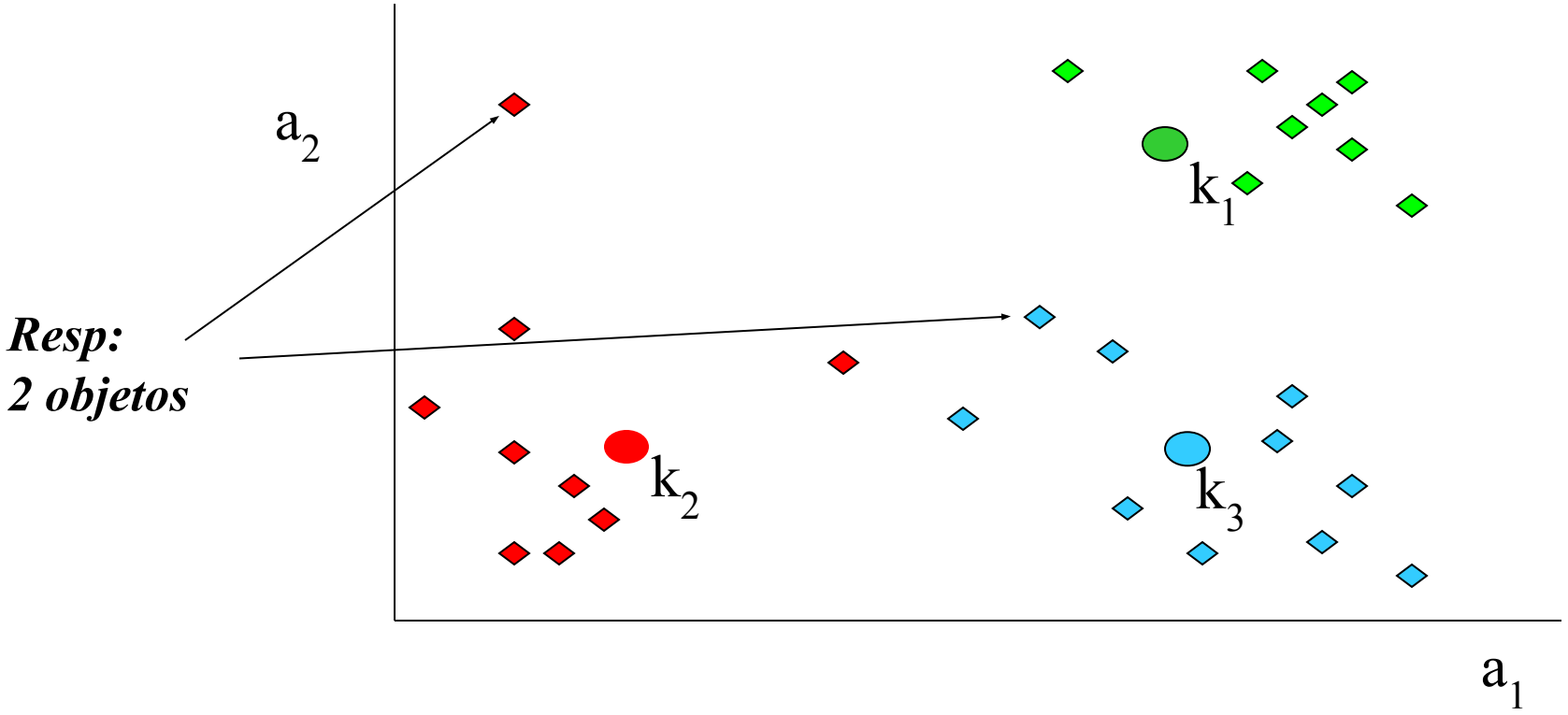
k-Means:

Re-atribuir objetos aos clusters de centróides mais próximos

Quais objetos mudarão de cluster?

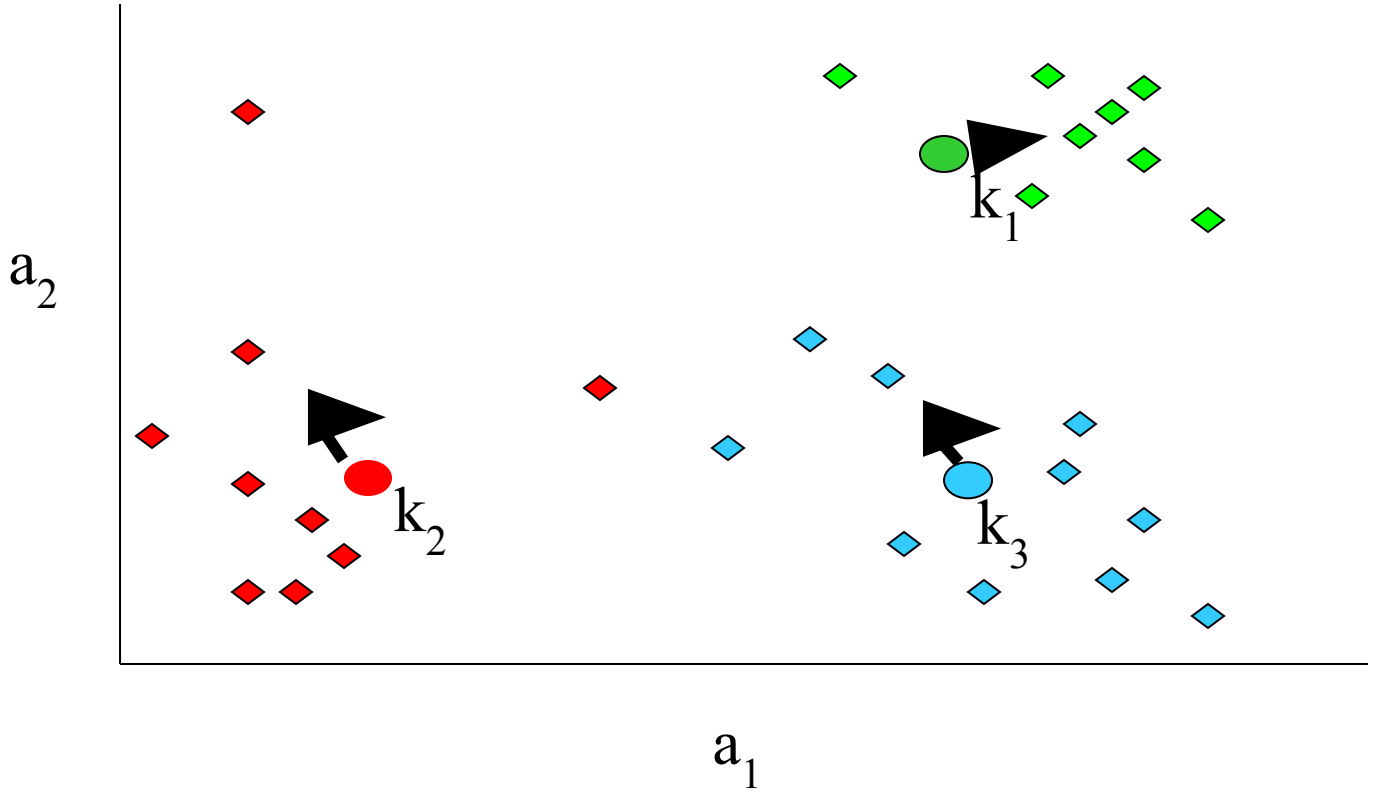


k-Means:



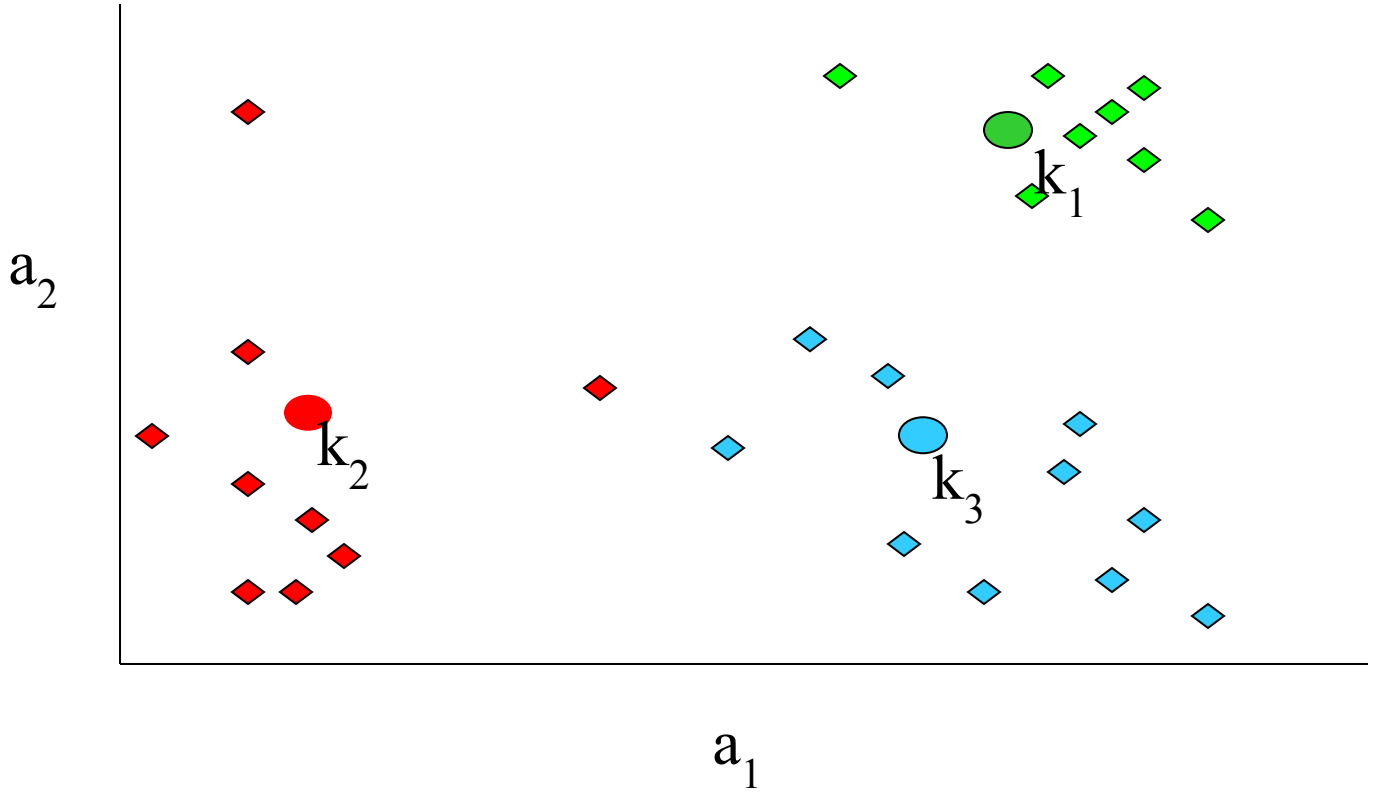
k-Means:

re-calcular
vetores
médios



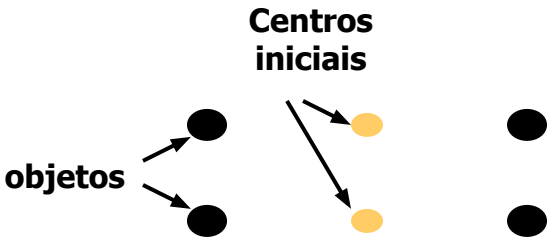
k-Means:

mover centros dos clusters…



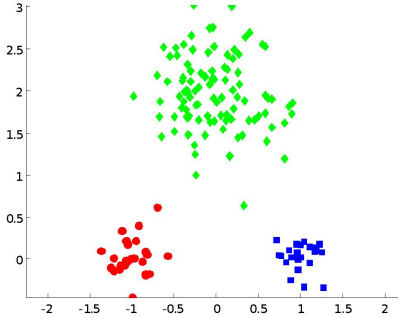
Discussão

- Resultado pode variar significativamente dependendo da escolha das sementes (protótipos) iniciais
- k-means pode "ficar preso" em ótimos locais
  - Exemplo:

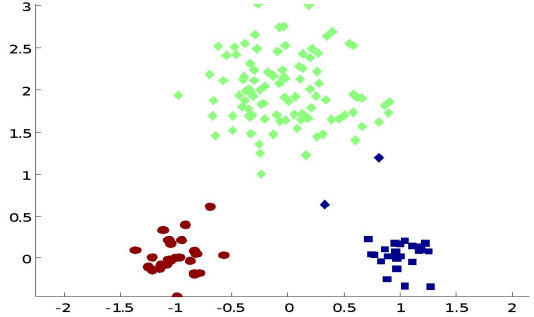


- Como evitar ... ?

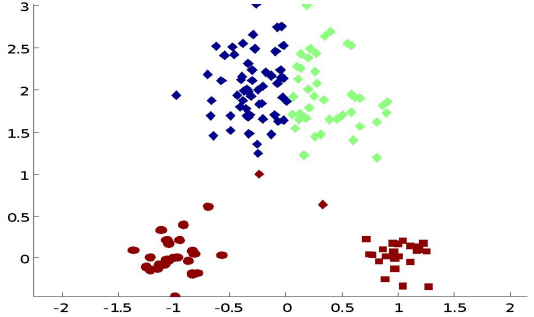
Dois agrupamentos diferentes com o k-means



Pontos originais

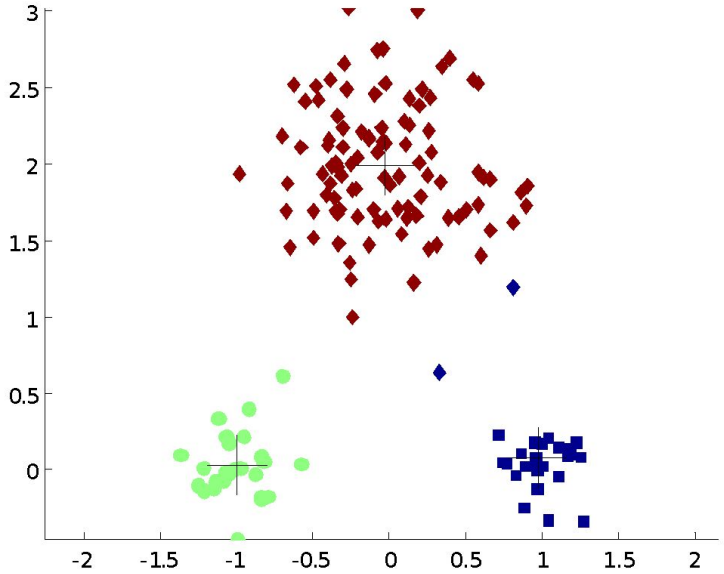


Clustering Ótimo

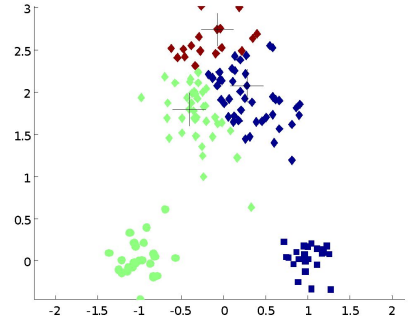
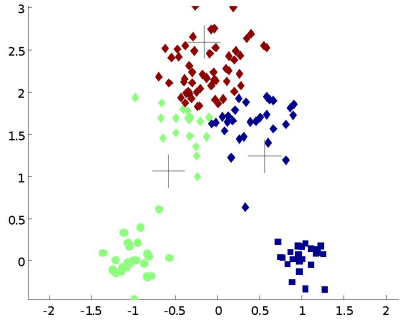
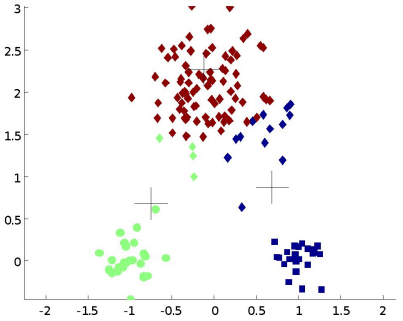
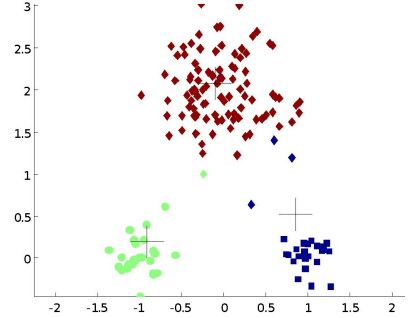
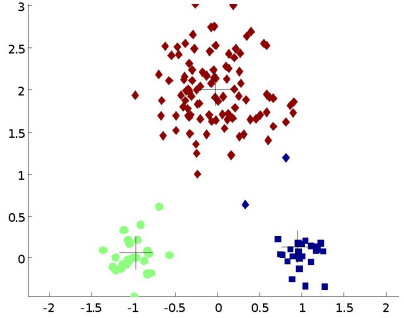
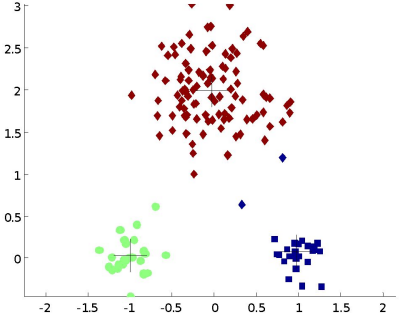


Clustering Sub-ótimo

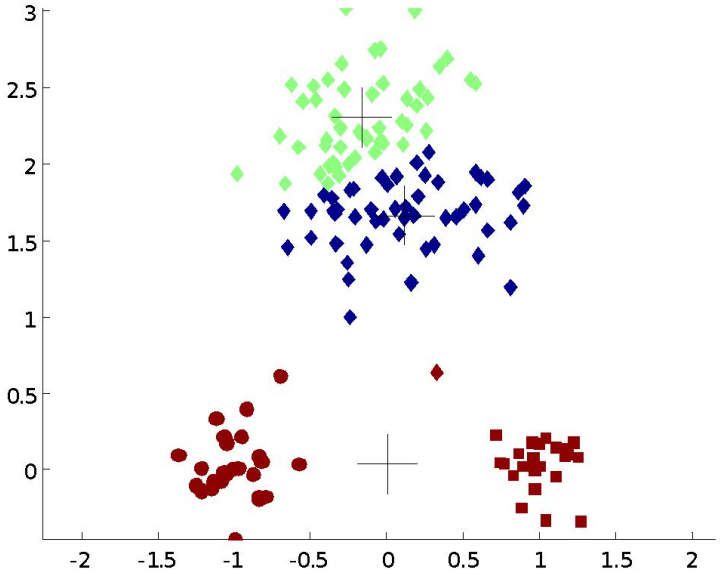
Importância da escolha dos centróides iniciais



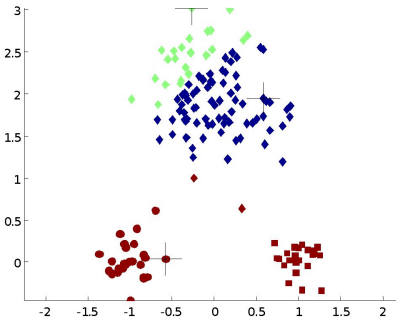
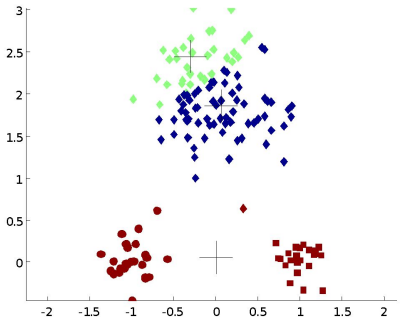
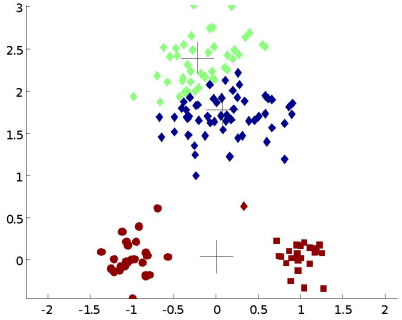
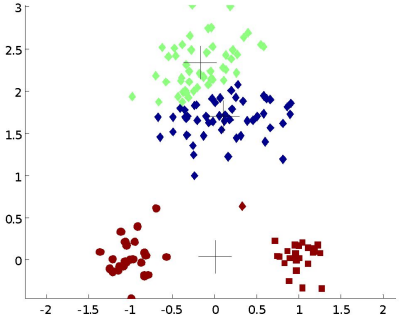
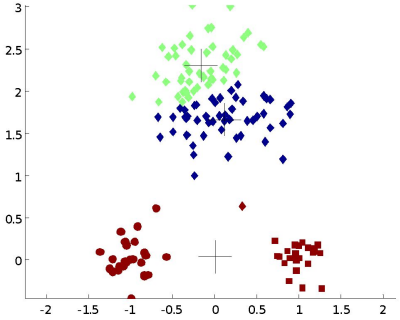
Importância da escolha dos centróides iniciais

Importância da escolha dos centróides iniciais



Importance of Choosing Initial Centroids …

Alternativas para Inicialização

- Múltiplas Execuções (inicializações aleatórias):
  - funciona bem em muitos problemas.
  - mas em bases de dados complexas, pode demandar um no. enorme de execuções.
  - em particular para no. de grupos grande.
    - especialmente porque k é, em geral, desconhecido
- Agrupamento Hierárquico:
  - agrupa-se uma amostra dos dados
  - tomam-se os centros da partição com k grupos

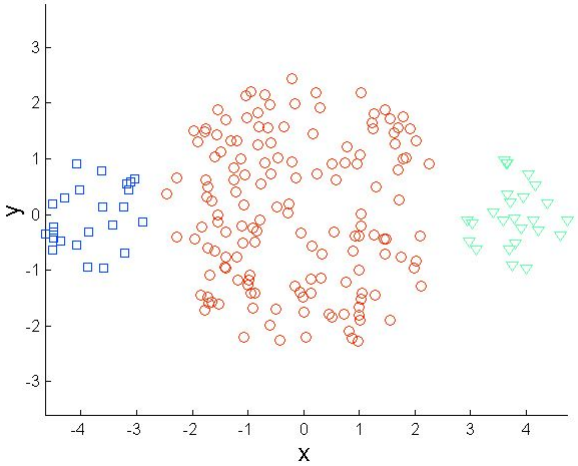
Alternativas para Inicialização

- Seleção "Informada" :
  - toma-se o 1º protótipo como um objeto aleatório
    - ou como o centro dos dados (*grand mean*)
  - sucessivamente escolhe-se o próximo protótipo
    - como o objeto mais distante dos protótipos correntes
  - **Nota**: para reduzir o esforço computacional e minimizar a probabilidade de seleção de outliers
    - processa-se apenas uma amostra dos dados
- Busca Guiada:
  - **X-means**, **k-means evolutivo**, …

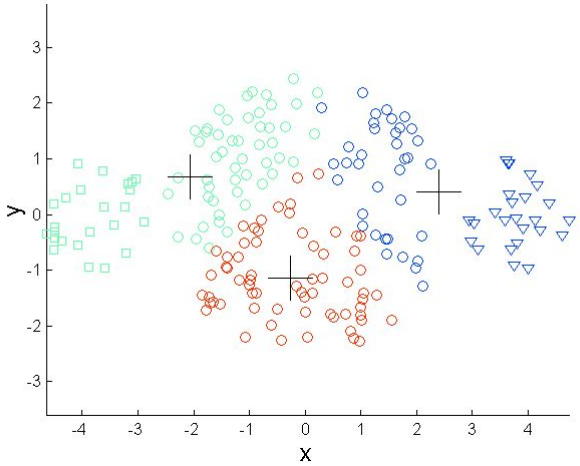
Discussão

- k-means é mais suscetível a problemas quando clusters são de diferentes
  - Tamanhos
  - Densidades
  - Formas não-globulares

Tamanhos diferentes

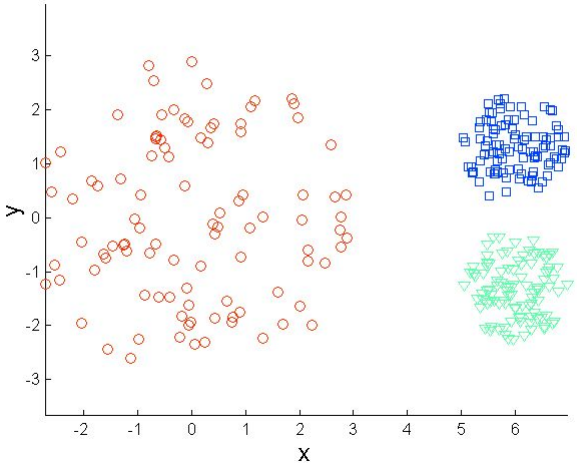


Pontos Originais

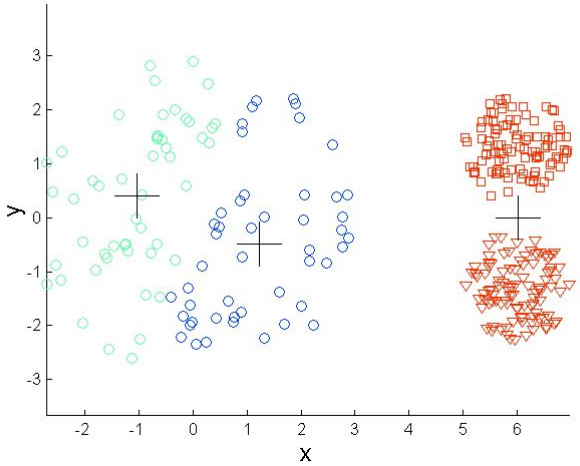


K-means (3 Clusters)

Densidade Diferente



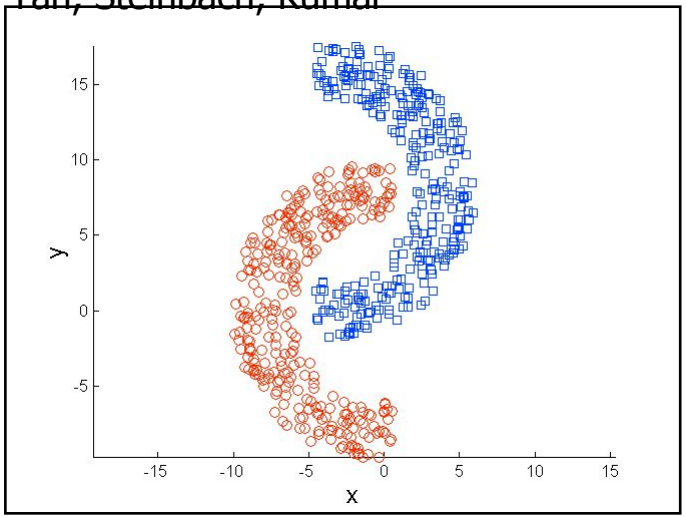
Pontos Originais

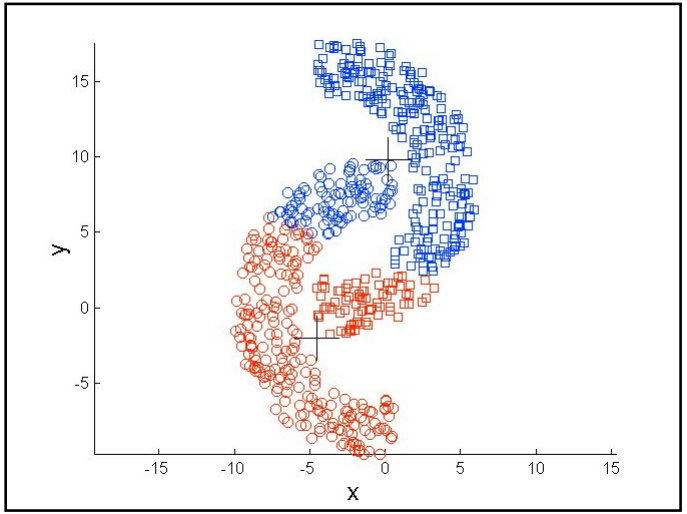


K-means (3 Clusters)

Formas Não-Globulares

Tan, Steinbach, Kumar



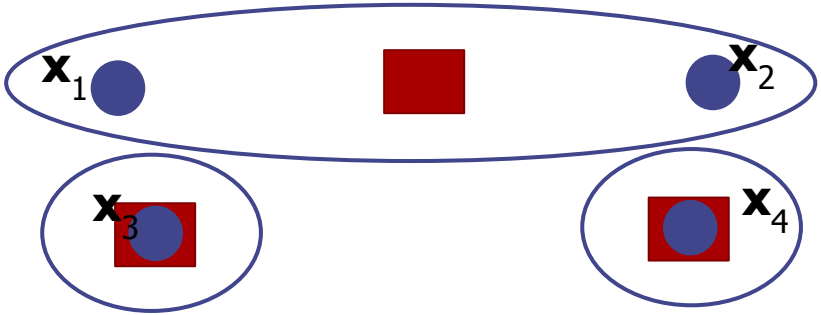


- **Nota**: na prática, esse problema em geral não é crítico, i.e., há pouco interesse na maioria das aplicações de mundo real

Como tratar esses casos?

- O k-means identifica bem grupos que possuem o mesmo tamanho/densidade ou que estão bem separados
- Quando isso não ocorre, existe solução?
  - Podemos dividir os grupos em subgrupos menores
  - O conjunto desses subgrupos permitem amenizar as dificuldades

- O que acontecerá na próxima iteração?

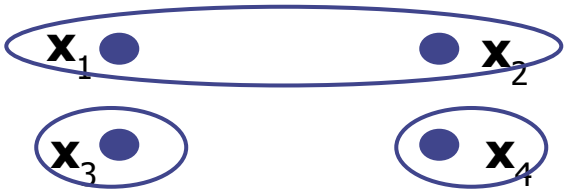


Grupos iniciais

***k* =3**

Manipulando Grupos Vazios

- k-means pode gerar **grupos vazios**
  - Por inicialização em pontos "dominados" do espaço
    - protótipos não representativos: nenhum objeto mais próximo
    - inicialização como objetos ao invés de pontos aleatórios resolve
  - Pela inicialização de grupos
    - cujos protótipos são não representativos; por exemplo:



Grupos iniciais

***k* =3**

  - Ao longo das iterações

Manipulando Grupos Vazios

- Estratégias para contornar o problema:
  - Eliminar os protótipos não representativos (reduz k)
    - viável se o número inicial de grupos, k, puder ser reduzido
    - pode ser útil para ajustar valores superestimados de k
  - Substituir cada protótipo não representativo (mantém k)
    - pelo objeto que mais contribui para o SSE da partição
    - por um dos objetos do grupo com maior MSE
      - visa dividir o grupo com maior erro quadrático médio
    - **Nota**: a execução do algoritmo prossegue após a substituição

Implementações Eficientes

- Desempenho computacional pode ser melhorado...
  - **Estruturas de Dados**, e.g.
    - **kd-trees**
  - **Algoritmos**, e.g.
    - **Atualização recursiva dos centróides**
      - Cálculo dos centróides só depende dos valores anteriores, dos nos. de objetos dos grupos e dos objetos que mudaram de grupo
        - Não demanda recalcular tudo novamente
      - **Uso da desigualdade triangular**
    - **Paralelização**

K-Means Paralelo / Distribuído

- Dados distribuídos em múltiplos *data sites* ou processadores
- **Algoritmo**:
  - Mesmos protótipos iniciais são distribuídos a cada sítio de dados
  - Cada sítio executa (em paralelo) uma iteração de k-means
  - Protótipos locais e nos. de objetos dos grupos são comunicados
  - Protótipos globais são calculados e retransmitidos aos sítios
  - Repete-se o processo

Resumo do k-means

Vantagens

- Simples e intuitivo
- Complexidade computacional **linear** em todas as variáveis críticas: O(N D k)
  - quadrático se D ≈ N ...
- Eficaz em muitos cenários de aplicação e produz resultados de interpretação simples
- Considerado um dos 10 mais influentes algoritmos em Data Mining (Wu & Kumar, 2009).

Desvantagens

- k = ?
- Sensível à inicialização dos protótipos (mínimos locais de J)
- Limita-se a encontrar clusters globulares
- Cada item deve pertencer a um único cluster (**partição rígida**, ou seja, sem sobreposição)
- Limitado a atributos numéricos
- Sensível a *outliers*

Algumas Variantes do k-means

- **K-medianas:** Substituir as médias pelas medianas
  - Média de 1, 3, 5, 7, 9 é **5**
  - Média de 1, 3, 5, 7, 1009 é **205**
  - Mediana de 1, 3, 5, 7, 1009 é **5**
  - **Vantagem:** menos sensível a outliers
  - **Desvantagem:** implementação mais complexa
    - cálculo da mediana em cada atributo...
- Pode-se mostrar que minimiza a soma das **distâncias de Manhattan** dos objetos aos centros (medianas) dos grupos

Algumas Variantes do k-means

- **K-medóides:** Substituir cada centróide por um objeto representativo do cluster, denominado **medóide**
  - Medóide = objeto mais próximo aos demais objetos do cluster
    - mais próximo em média (empates resolvidos aleatoriamente)
  - **Vantagens:**
    - menos sensível a outliers
    - permite cálculo relacional (apenas matriz de distâncias)
      - logo, pode ser aplicado a bases com atributos categóricos
    - convergência assegurada com qualquer medida de (dis)similaridade
  - **Desvantagem:** Complexidade quadrática com nº. de objetos (N)

Algumas Variantes do k-means

- **Métodos de Múltiplas Execuções de k-means:**
  - Executam k-means repetidas vezes a partir de diferentes valores de k e de posições iniciais dos protótipos
    - Ordenado: n_p inicializações de protótipos para cada $k \in [k_{min}, k_{max}]$
    - Aleatório: n_T inicializações de protótipos com k sorteado em $[k_{min}, k_{max}]$
  - Tomam a melhor partição resultante de acordo com algum critério de qualidade (**critério de validade de agrupamento**)
  - **Vantagens:** Estimam k e são menos sensíveis a mínimos locais
  - **Desvantagem:** Custo computacional pode ser elevado

Questão...

- A função de custo pode ser utilizada para escolher a melhor partição dentre um conjunto de candidatas ?
  - Resposta é sim se todas têm o mesmo no. k de clusters (fixo)
  - Mas e se k for desconhecido e, portanto, variável ?
- Para responder, considere, por exemplo, que as partições são geradas a partir de múltiplas execuções do algoritmo:
  - com protótipos iniciais aleatórios
  - com no. variável de grupos $k \in [k_{min}, k_{max}]$

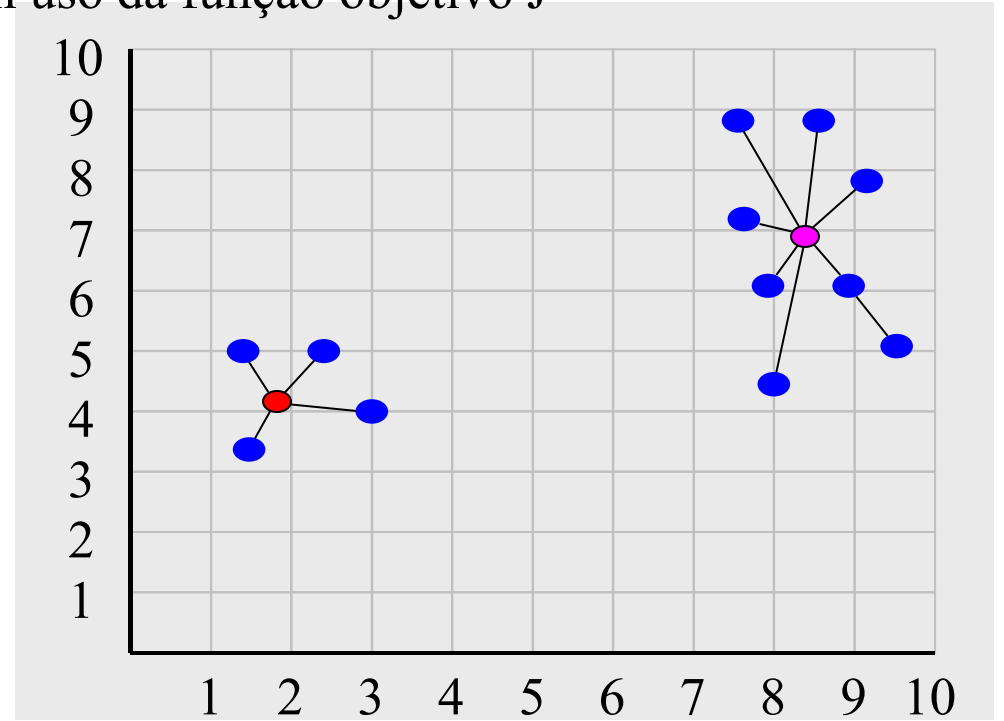
Questão...

- Para responder a questão anterior, vamos considerar o método de múltiplas execuções ordenadas de k-means, com uso da função objetivo J

Erro Quadrático:

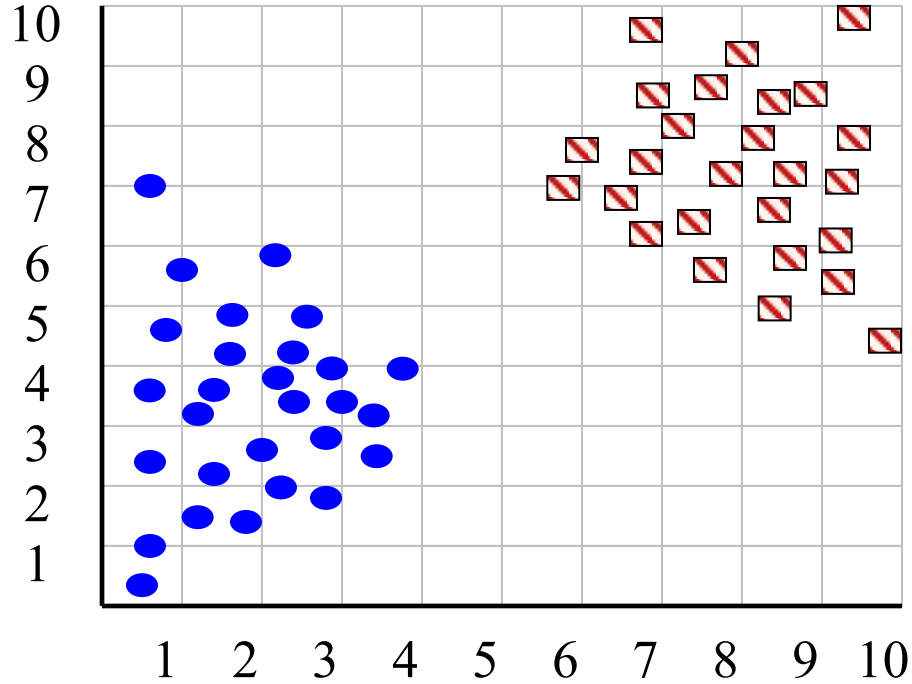
$$J = \sum_{i=1}^{k} \sum_{x_j \in C_i} d\left(x_j, \bar{x}_i\right)^2$$

Função Objetivo

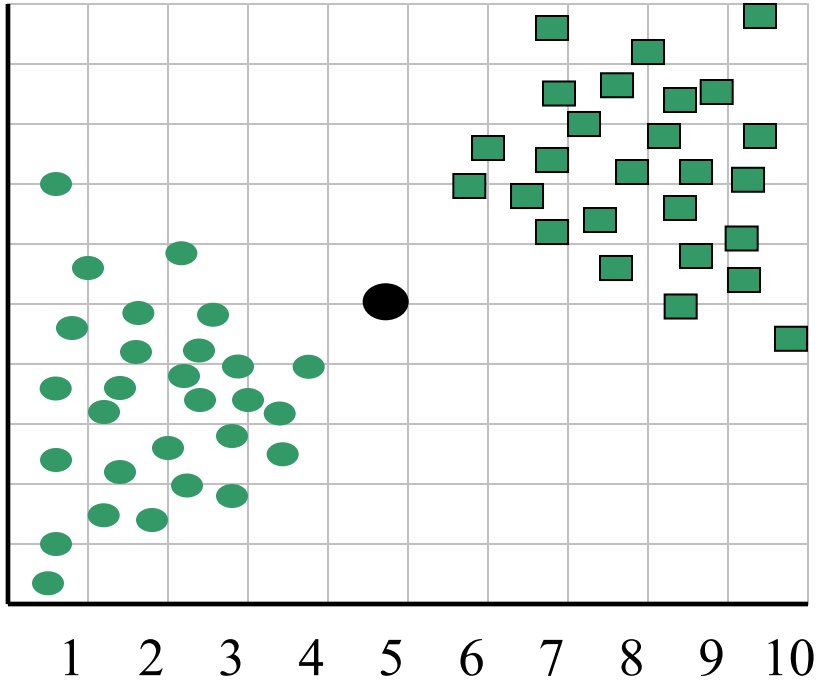


Questão...

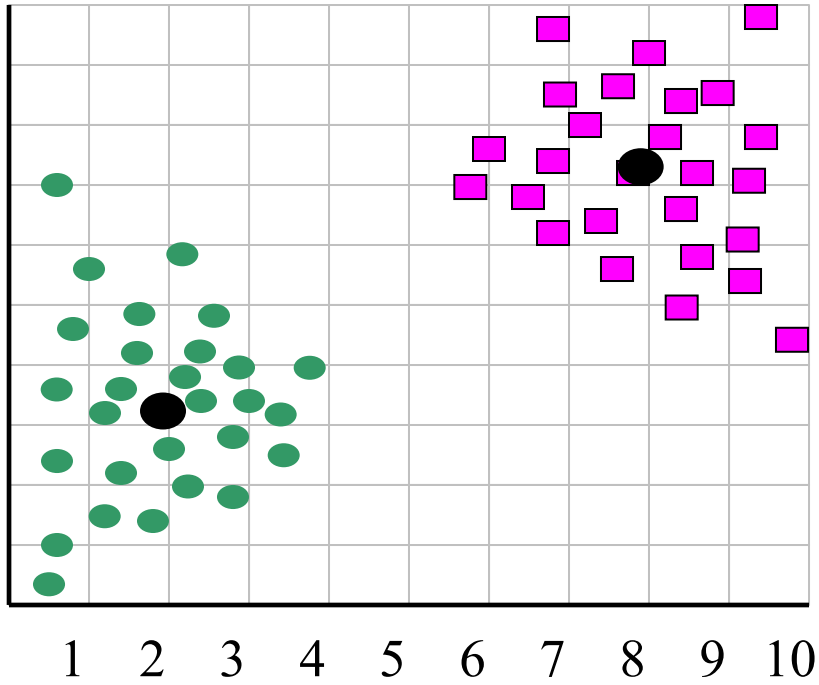
- Considere o seguinte exemplo:



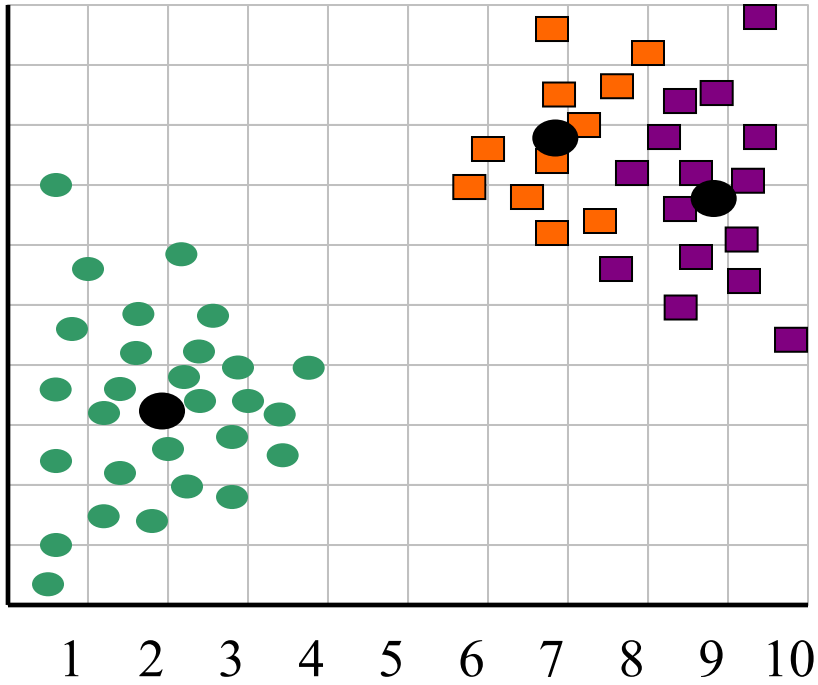
Para $k = 1$, o valor da função objetivo é 873,0



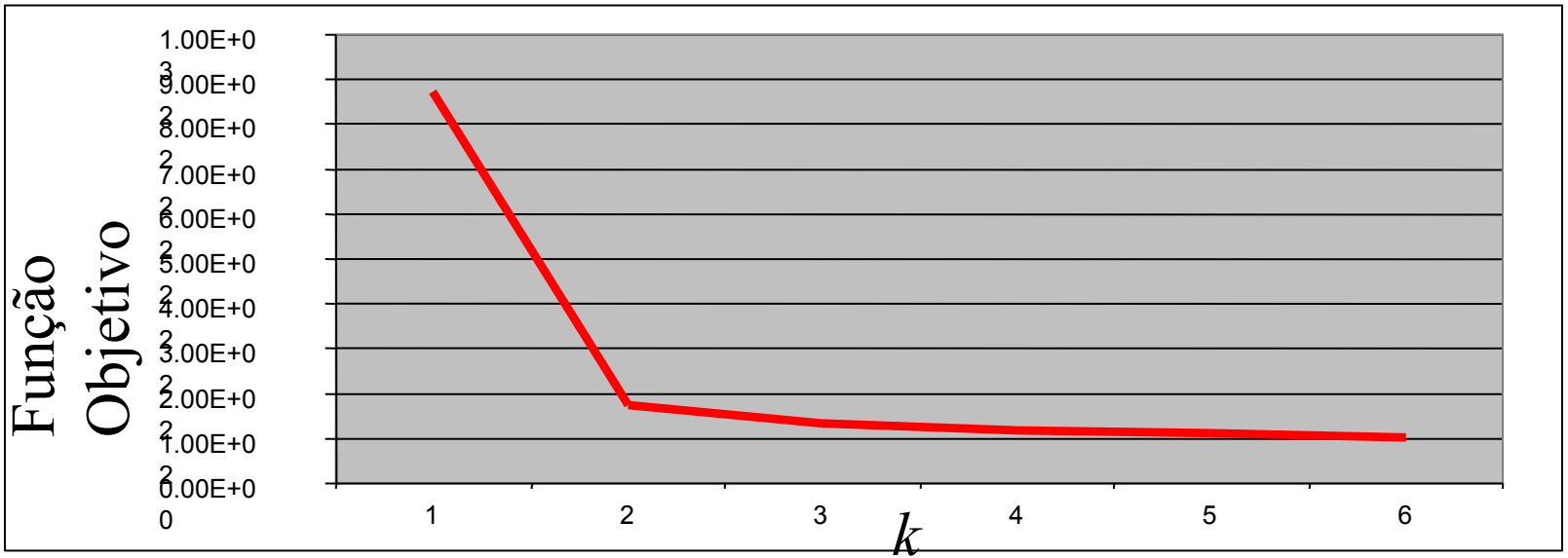
Para $k = 2$, o valor da função objetivo é 173,1



Para $k = 3$, o valor da função objetivo é 133,6

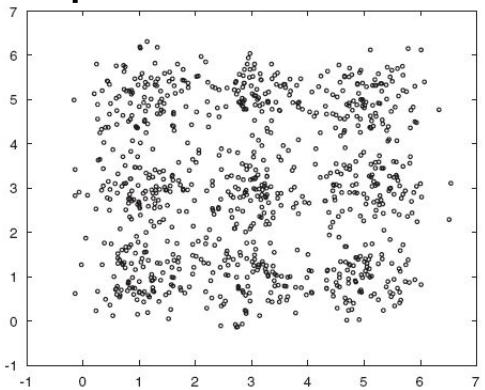


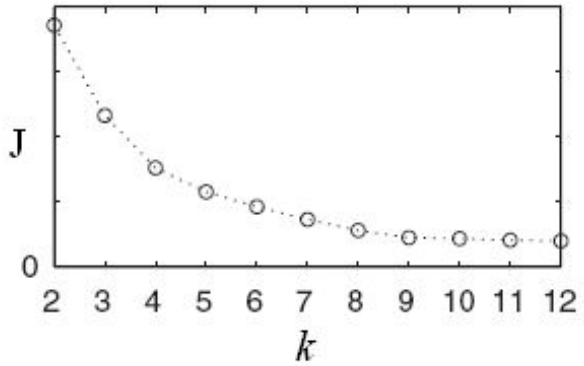
Podemos então repetir este procedimento e plotar os valores da função objetivo J para k =1,…,6, ... e tentar identificar um “joelho” :



Questão...

- Infelizmente os resultados não são sempre tão claros quanto no exemplo anterior... Vide exemplo abaixo...





- Além disso, como utilizar essa metodologia em variantes baseadas em busca guiada, que otimizam k ?
  - X-means, k-means evolutivo, ...
- Solução: **critérios de validação de agrupamento**.

Referências

- Jain, A. K. and Dubes, R. C., Algorithms for Clustering Data, Prentice Hall, 1988
- Kaufman, L., Rousseeuw, P. J., Finding Groups in Data – An Introduction to Cluster Analysis, Wiley, 2005.
- Tan, P.-N., Steinbach, M., and Kumar, V., *Introduction to Data Mining*, Addison-Wesley, 2006
- Wu, X. and Kumar, V., *The Top Ten Algorithms in Data Mining*, Chapman & Hall/CRC, 2009
- D. Steinley, *K-Means Clustering: A Half-Century Synthesis*, British J. of Mathematical and Stat. Psychology, V. 59, 2006
- Ester, M., H. P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR, AAAI Press, pp. 226-231. 1996