FIAP

MBA+ | Artificial Intelligence & Machine Learning


AI

Aprendizado não supervisionado

K-means / EM / Regras de Associação

Prof. Dr. Vinicius F. Caridá

Algoritmos Particionais

Métodos Particionais

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 $X = \{x_1, x_2, \dots, x_N\}$ de N objetos em uma coleção $C = \{C_1, C_2, \dots, C_k\}$ de k grupos disjuntos C_i tal que $C_1 \cup C_2 \cup \cdots \cup C_k = X$, $C_i \neq \emptyset$ e $C_1 \cap C_2 \neq \emptyset$ para $i \neq j$

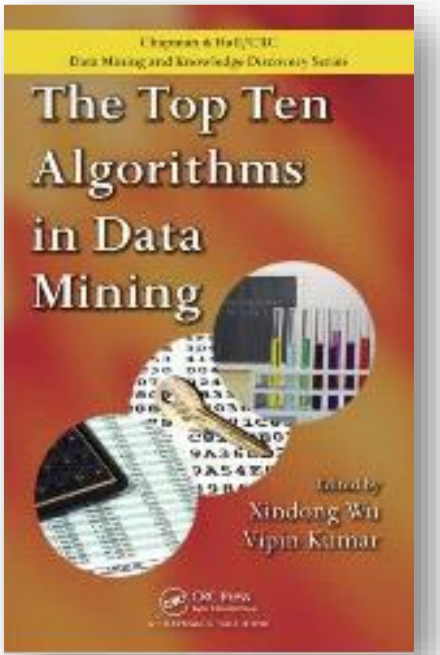
K-Means

Aqui veremos 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



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, Berkeley, California, USA, 1967, 281–297

Porém...

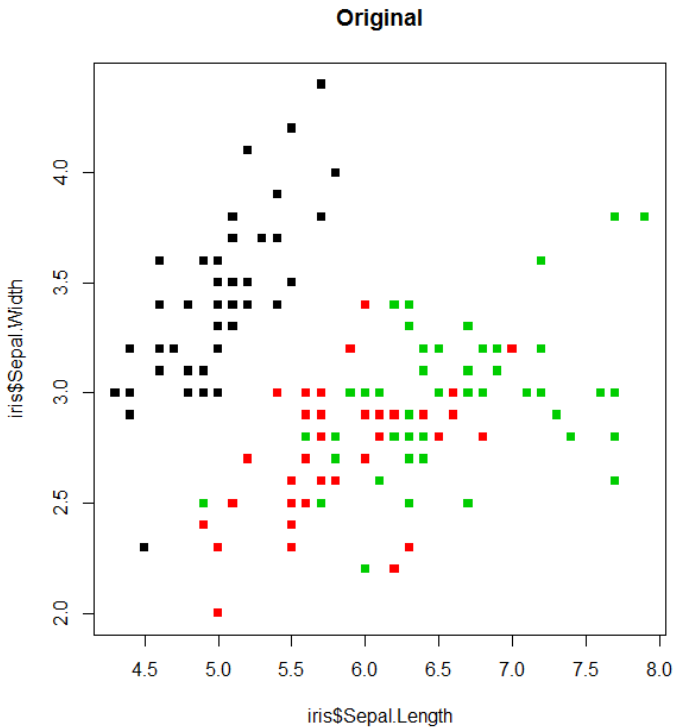
> “K-means has a rich and diverse history as it was independently discovered in different scientific fields by Steinhaus (1956), Lloyd (proposed in 1957, published in 1982), Ball & Hall (1965) and MacQueen (1967)” [Jain, Data Clustering: 50 Years Beyond K-Means, Patt. Rec. Lett., 2010]

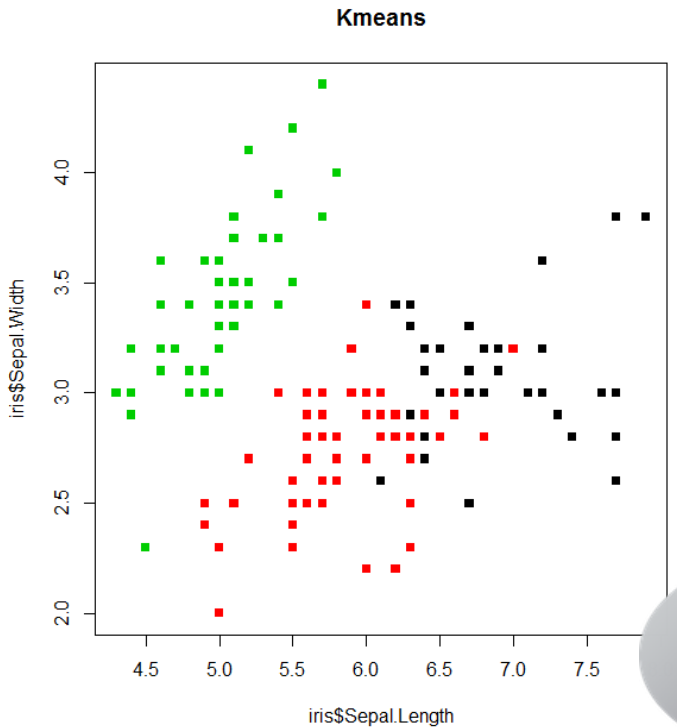
... e tem sido assunto por mais de meio século !

> Douglas Steinley, K-Means Clustering: A Half-Century Synthesis, British Journal of Mathematical and Statistical Psychology, Vol. 59, 2006

K-Means

```
1 data(iris) #Carrega os dados
2 groups = kmeans(iris[1:4], center=3, iter.max=10)
```

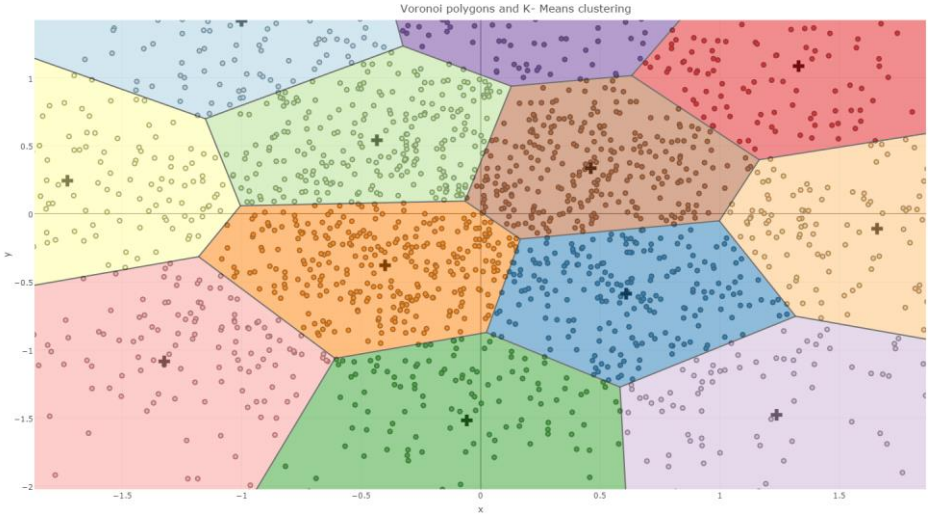


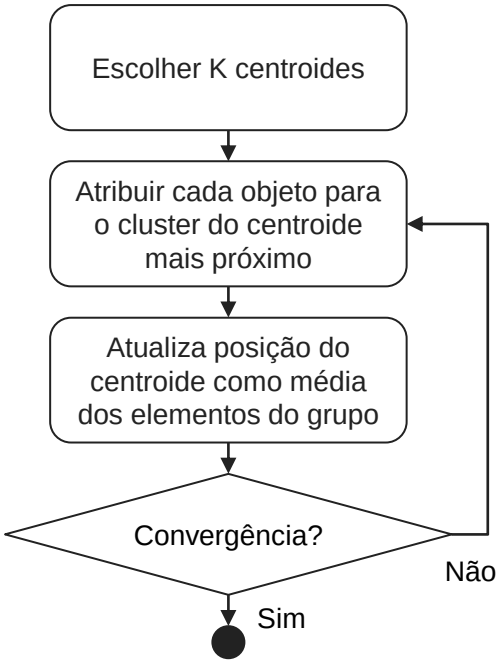


K-Means

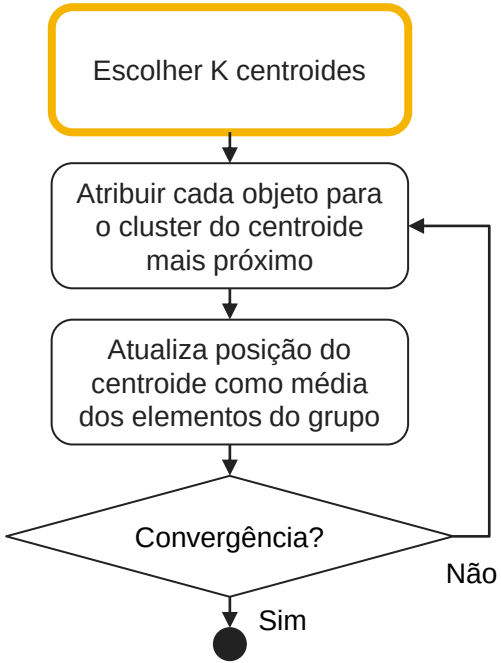
Objetiva particionar N observações dentre k grupos em que cada observação pertence ao grupo mais próximo da média. Isso resulta em uma divisão do espaço de dados em um Diagrama de Voronoi.

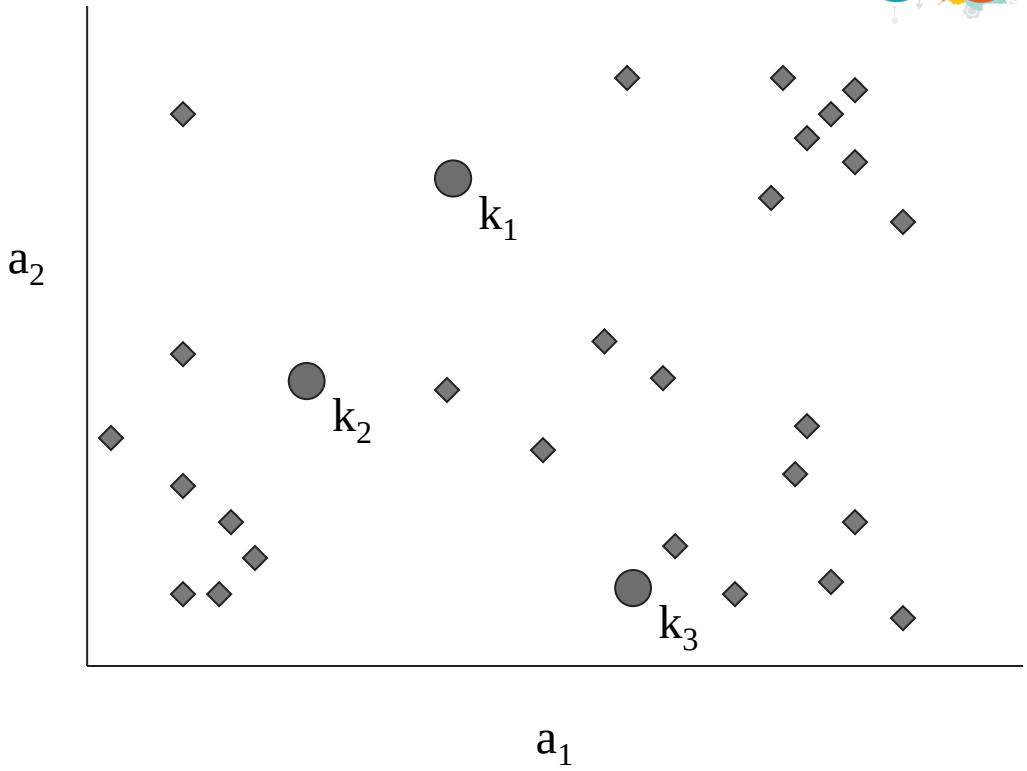
Calculado por meio da triangulação de Delaunay





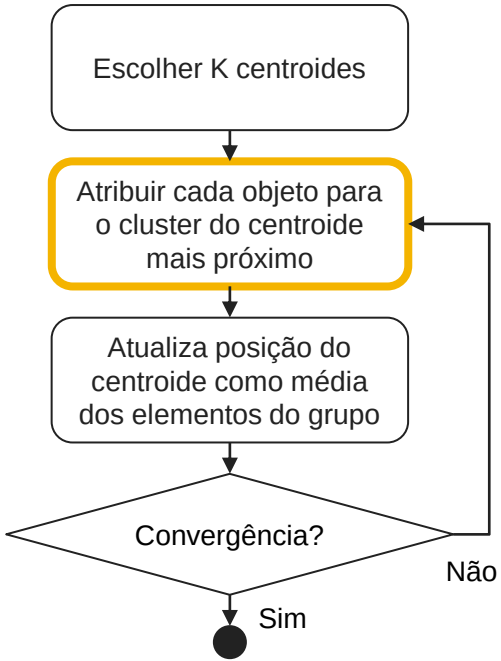
K-Means - Simulação

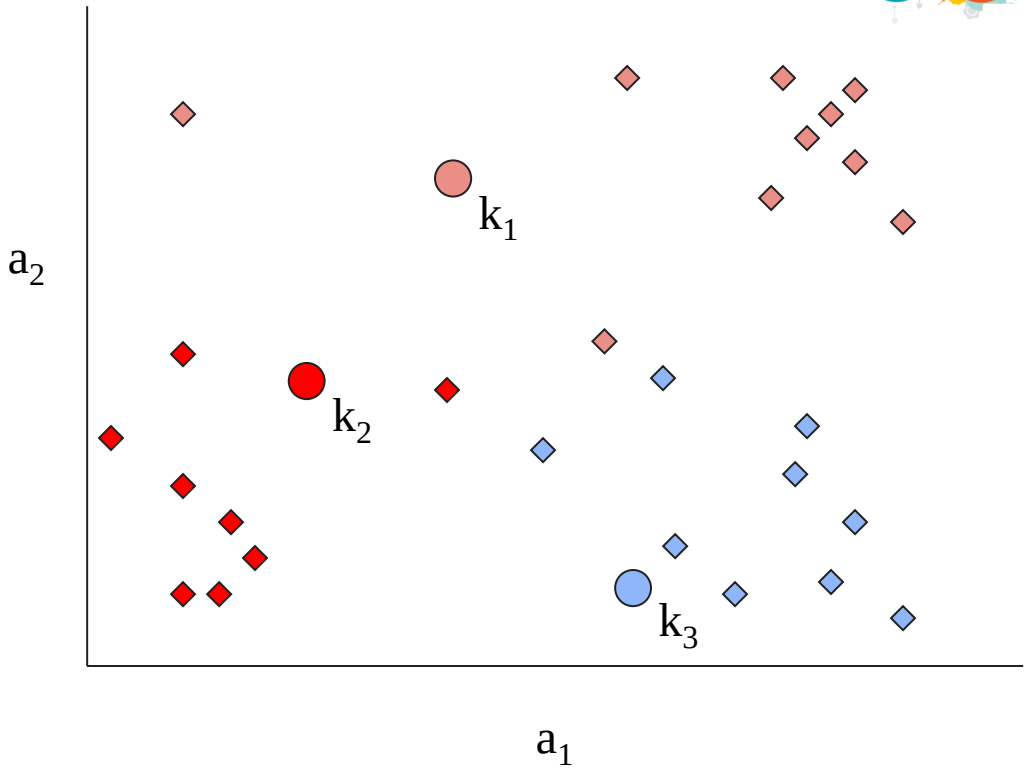




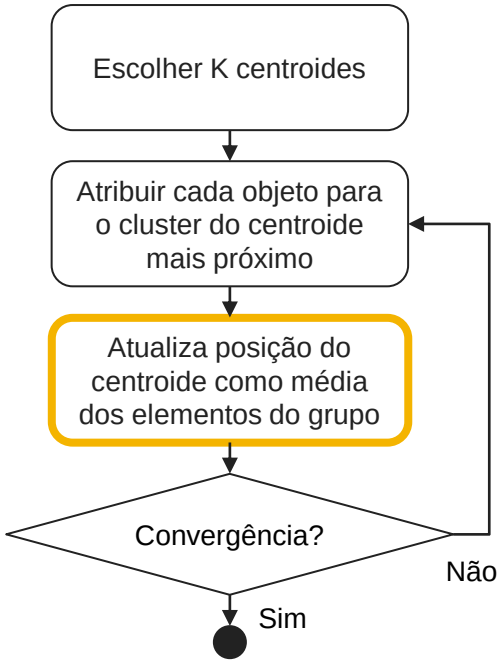
Slides desse exemplo são baseados no curso de Gregory Piatetsky-Shapiro, disponível em http://www.kdnuggets.com

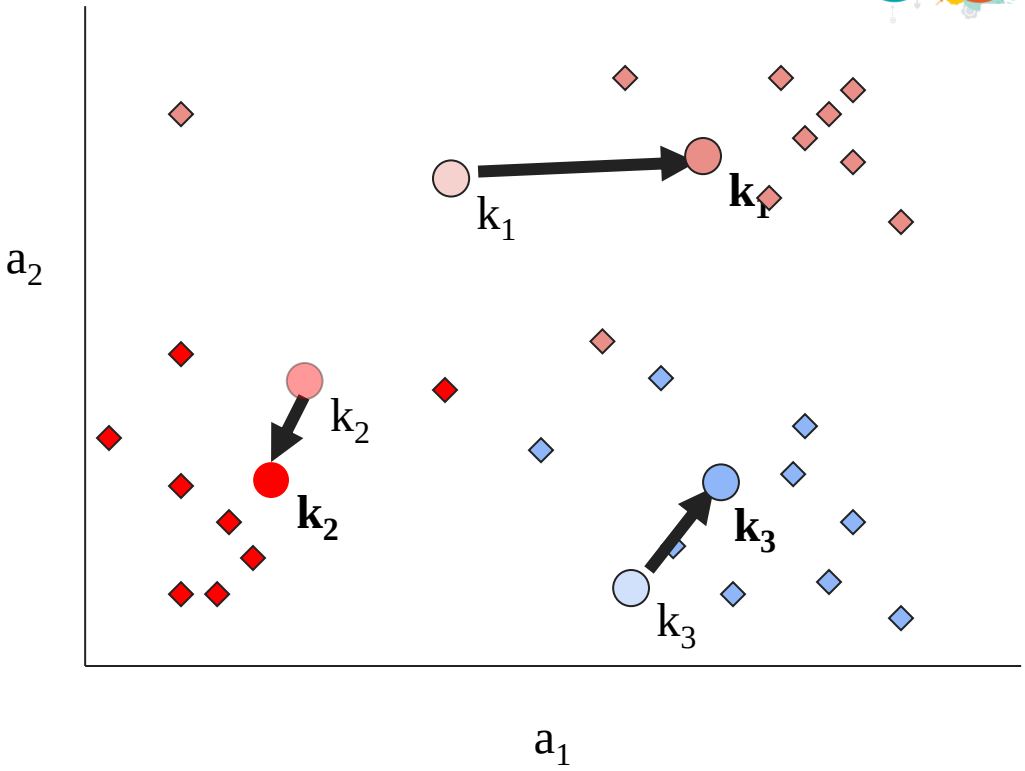
K-Means - Simulação



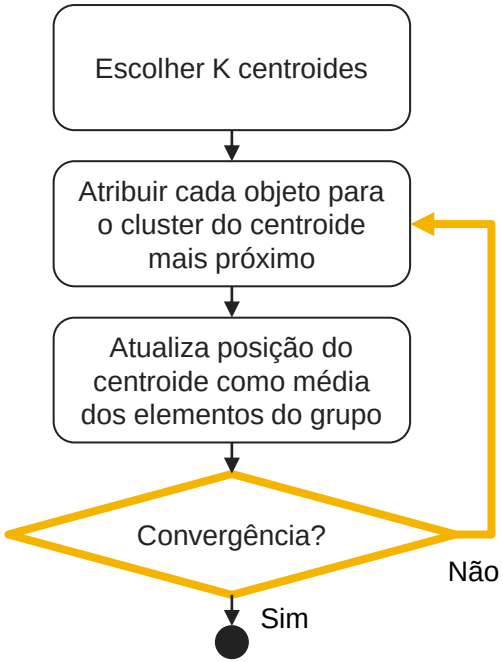


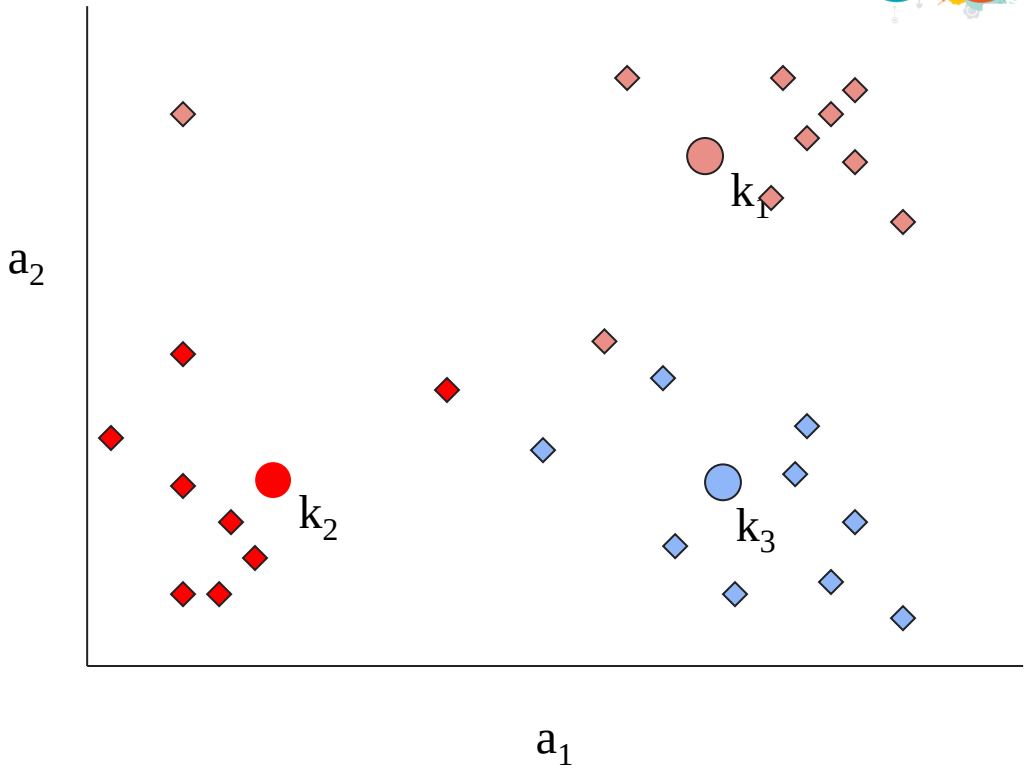
K-Means - Simulação



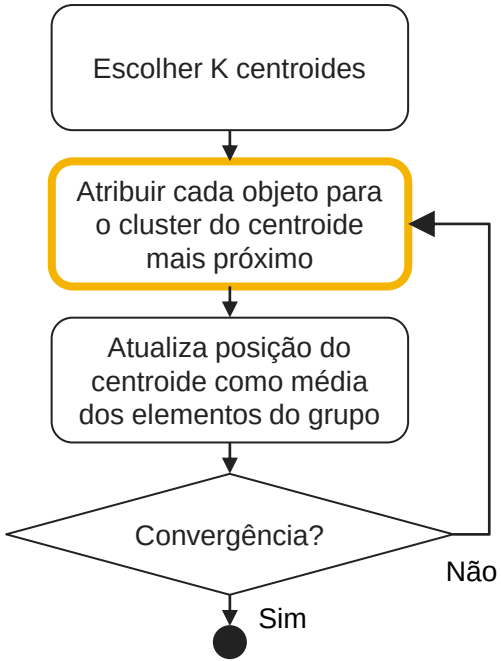


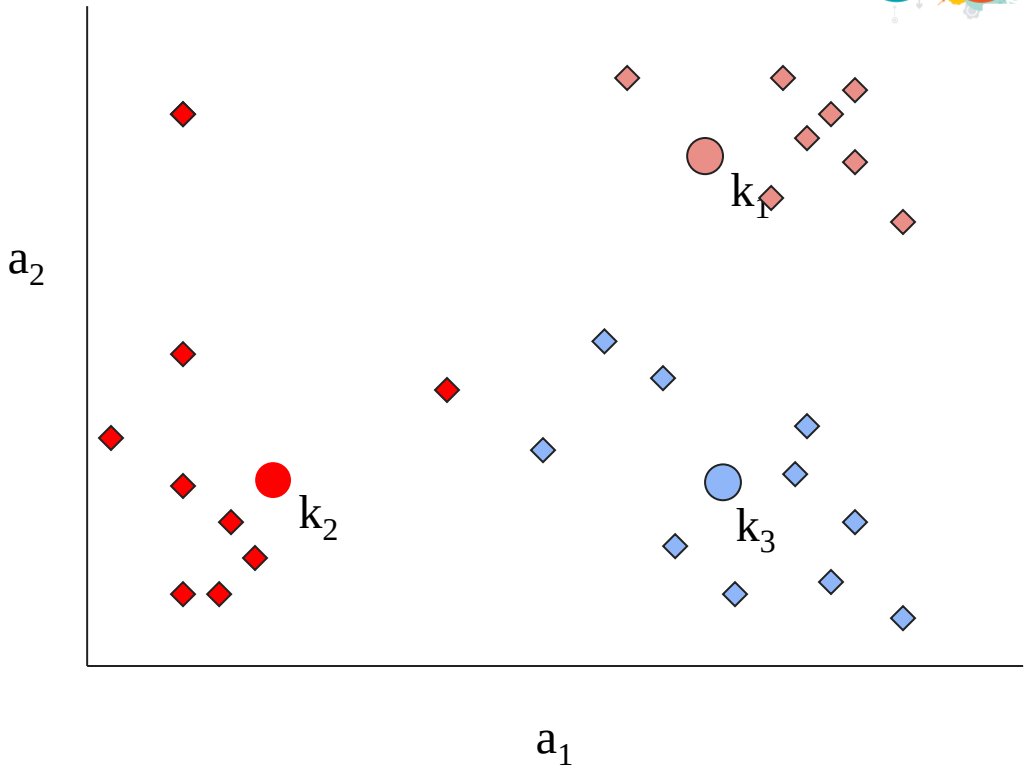
K-Means - Simulação



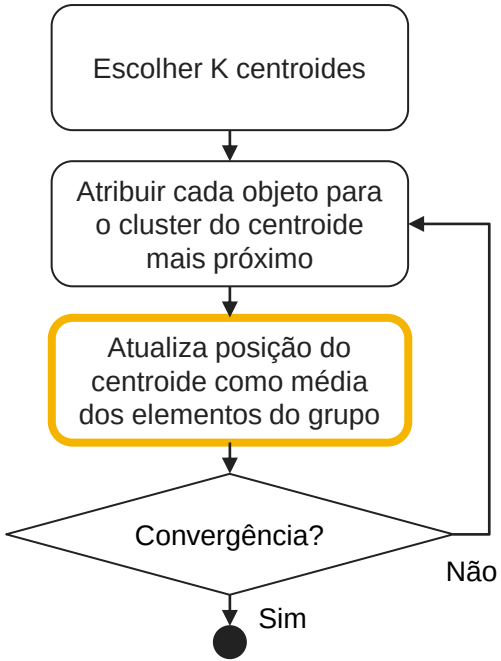


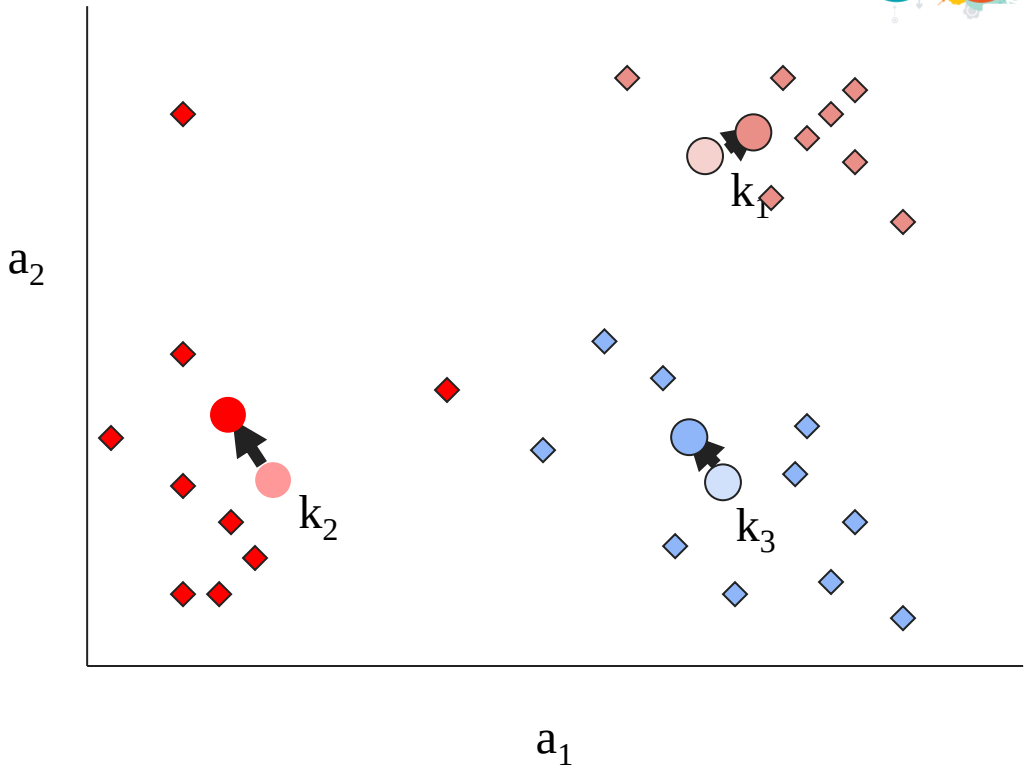
K-Means - Simulação



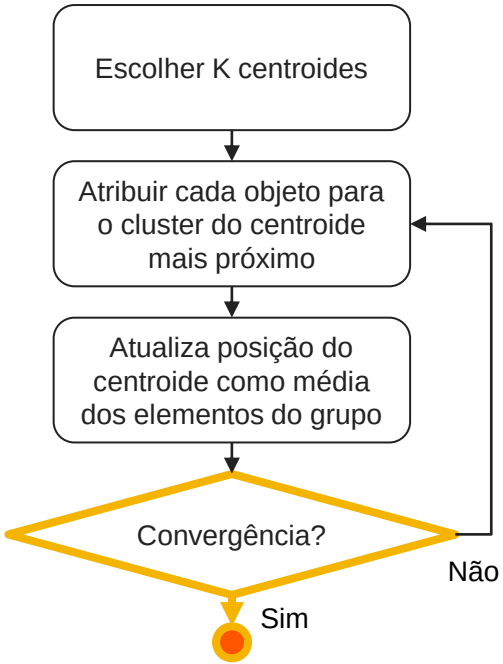


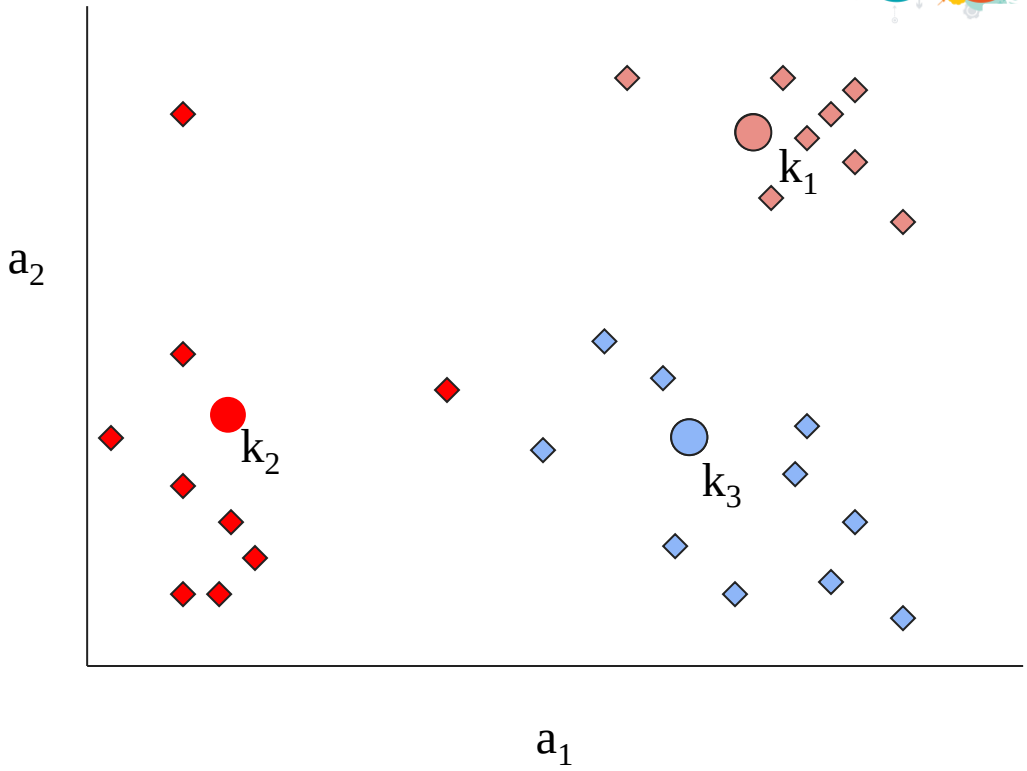
K-Means - Simulação





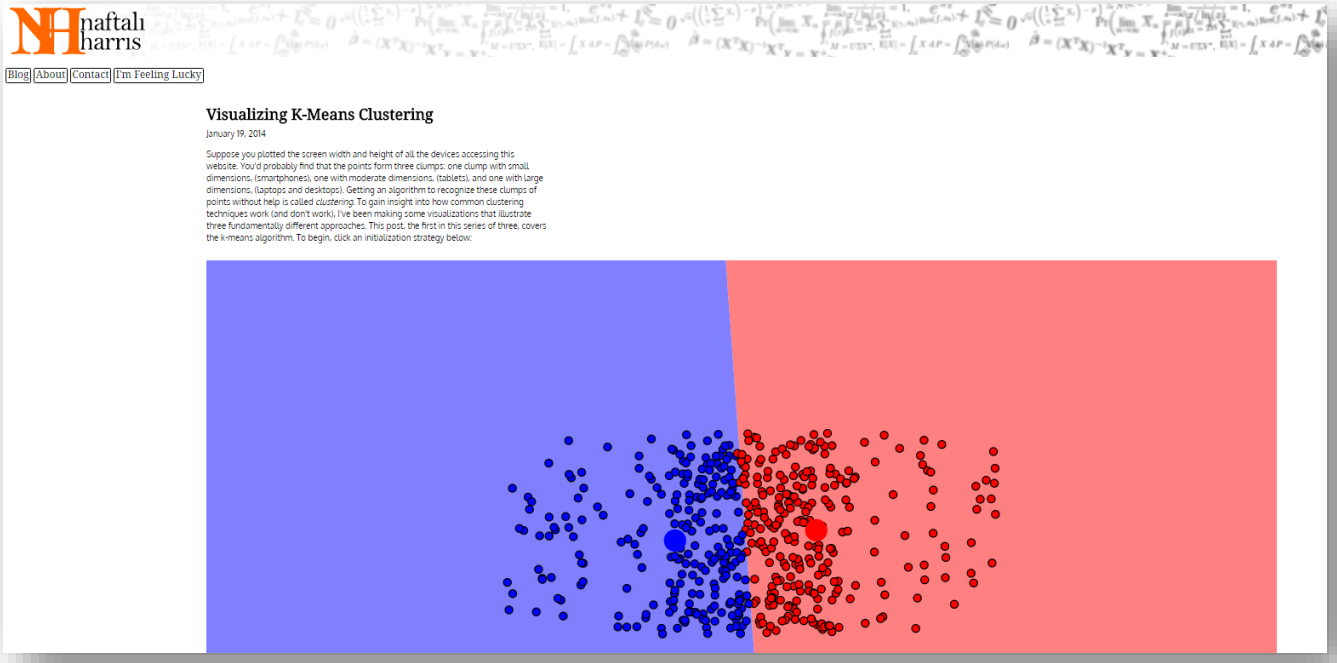
K-Means - Simulação





K-Means - Animação

https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

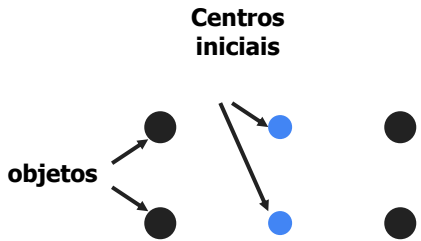


K-Means

K-Means: Sensibilidade em relação à inicialização

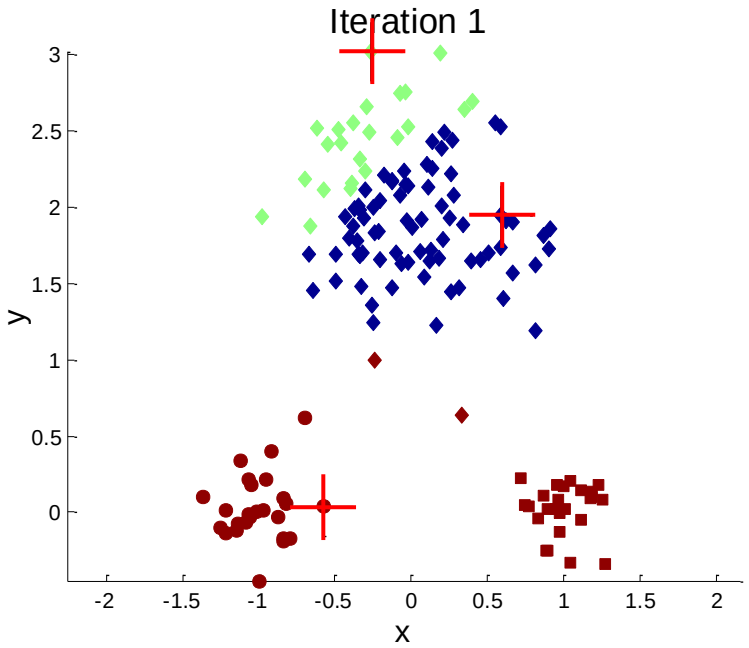
Resultado pode variar significativamente dependendo da escolha das sementes (protótipos) iniciais

k-means pode “ficar preso” em ótimos locais:



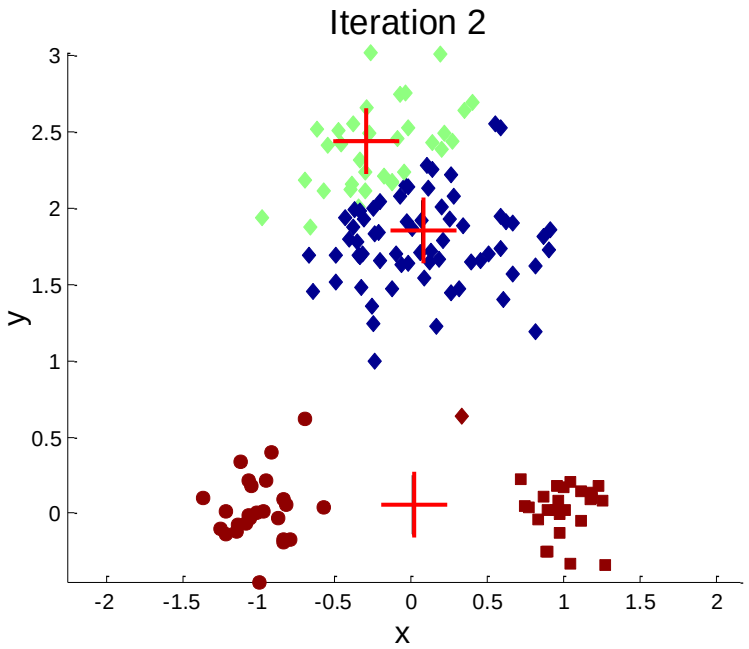
Slide baseado no curso de Gregory Piatetsky-Shapiro, disponível em http://www.kdnuggets.com

K-Means: Sensibilidade em relação à inicialização



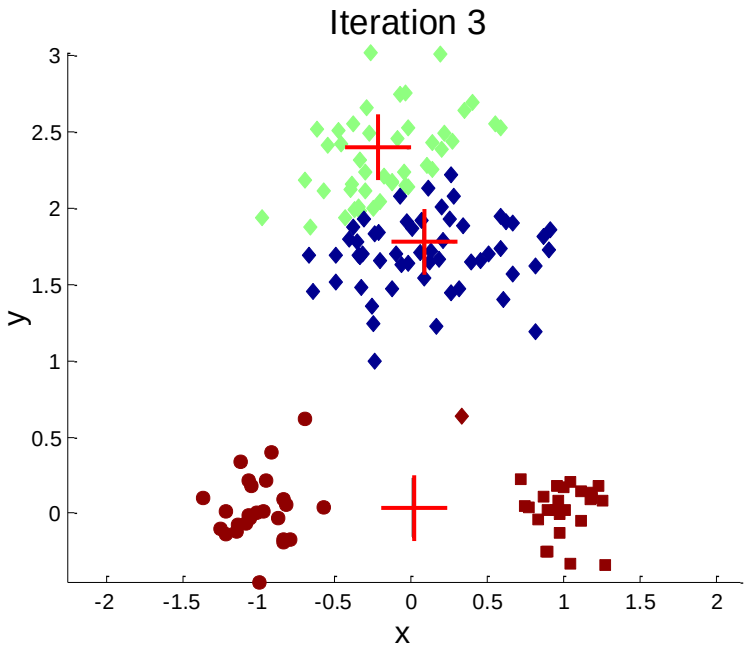
Tan, Steinbach & Kumar, Introduction to Data Mining, 2006.

K-Means: Sensibilidade em relação à inicialização



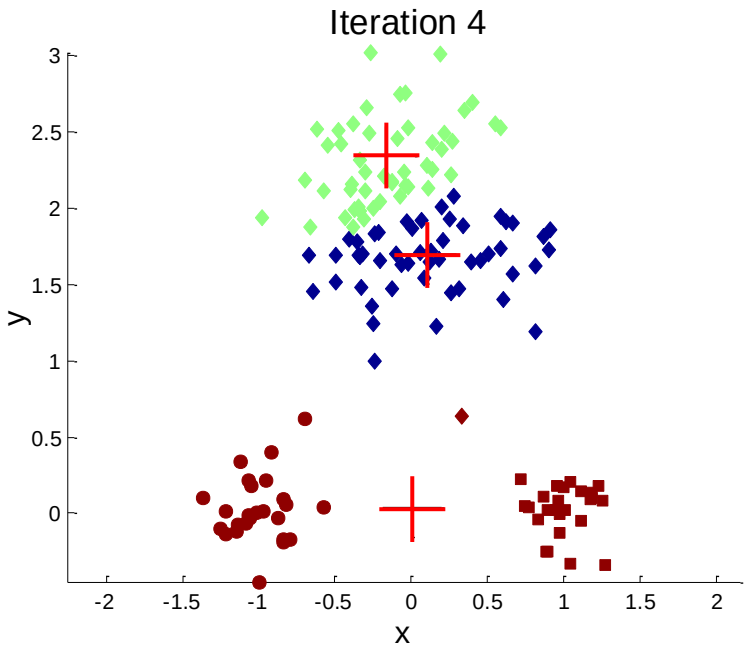
Tan, Steinbach & Kumar, Introduction to Data Mining, 2006.

K-Means: Sensibilidade em relação à inicialização



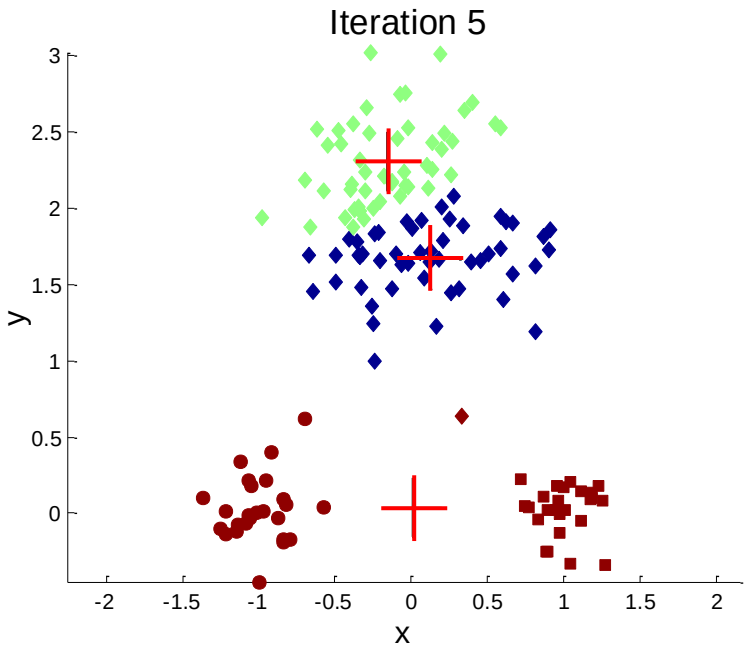
Tan, Steinbach & Kumar, Introduction to Data Mining, 2006.

K-Means: Sensibilidade em relação à inicialização



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

K-Means: Sensibilidade em relação à inicialização



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

K-Means: Sensibilidade em relação à inicialização

Premissa:

Uma boa seleção de *k* protótipos iniciais em uma base de dados com *k* grupos naturais é tal que cada protótipo é um objeto de um grupo diferente.

- No entanto, a chance de se selecionar um protótipo de cada grupo é pequena, especialmente para *k* grande.
- Consideremos grupos balanceados, com uma mesma quantidade $g = N / k$ de objetos cada. A probabilidade de selecionar um protótipo de cada grupo diferente é:

$$P = \frac{\text{no. de maneiras de selecionar 1 objeto de cada grupo (N / k objetos)}}{\text{no. de maneiras de selecionar k dentre N objetos}} = \frac{k!}{k^k}$$

Para $k = 10$ temos $P = 0.00036 \rightarrow 2.778$ inicializações.

K-Means: Sensibilidade em relação à inicialização

Múltiplas Execuções (inicializações aleatórias):

- Funciona bem em muitos problemas;
- Pode demandar muitas execuções (especialmente com *k* alto).

Agrupamento Hierárquico:

- agrupa-se uma amostra dos dados para tomar os centros da partição com *k* grupos.

Seleção “informada” em uma amostra dos dados:

- Tomar 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.

Busca Guiada:

- X-means, *k*-means evolutivo, ...

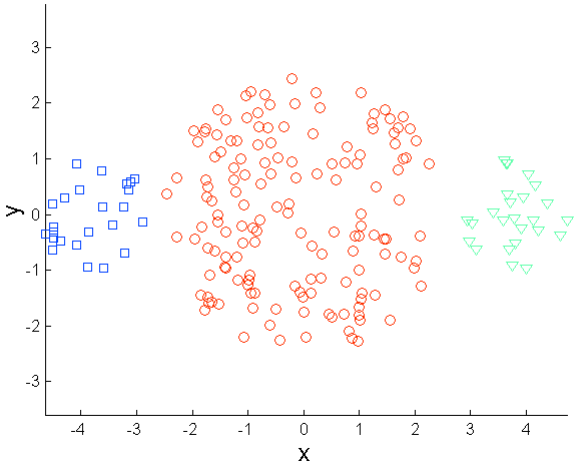
K-Means

K-Means: Problemas estruturais

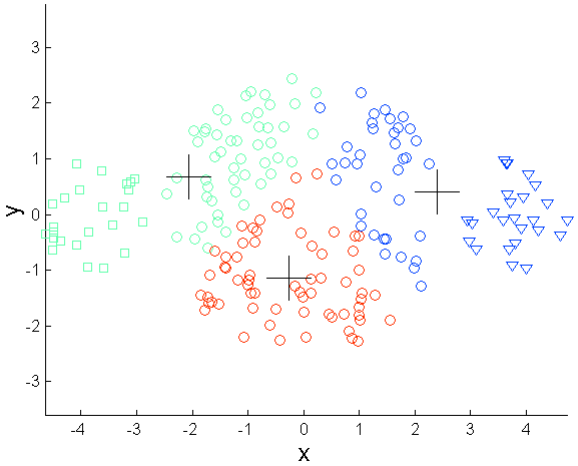
Algoritmo *k*-means funciona bem se:

- Clusters são (hiper)esféricos e bem separados
- Clusters de volumes aproximadamente iguais
- Cluster com quantidades de pontos semelhantes
- Formas Globulares

K-Means: Problemas estruturais

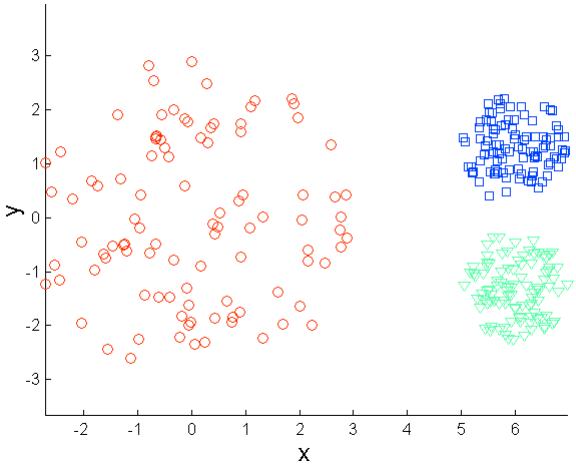


Estrutura correta

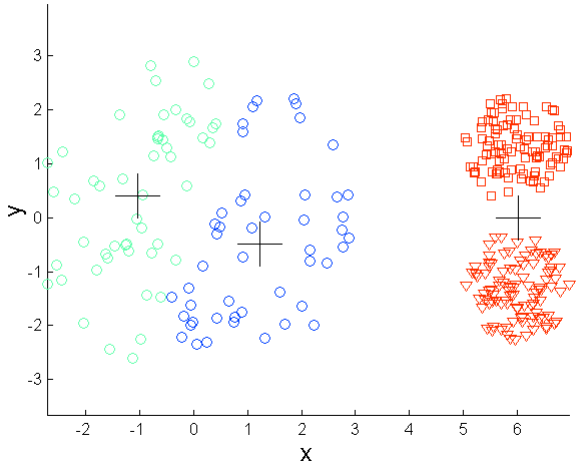


k-means (3 Clusters)

K-Means: Problemas estruturais

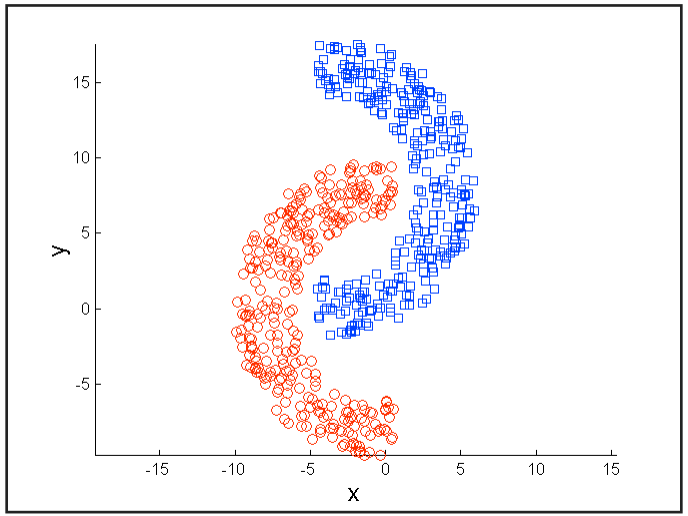


Estrutura correta

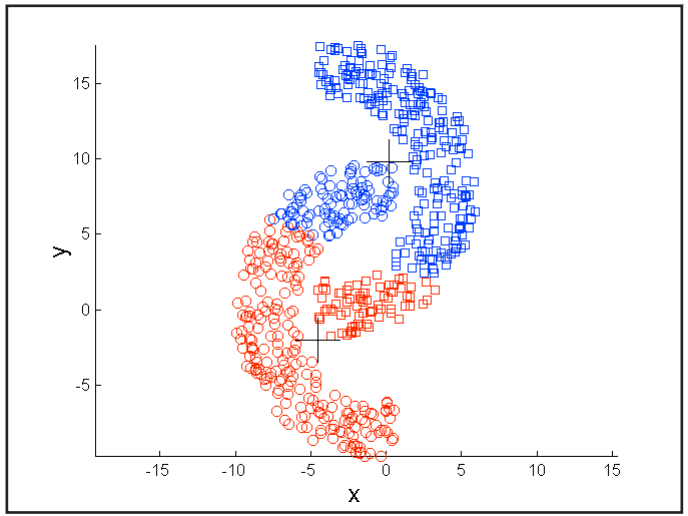


K-means (3 Clusters)

K-Means: Problemas estruturais



Estrutura correta



K-means (3 Clusters)

(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.

K-Means: Custo Computacional

Complexidade (assintótica) de tempo:

$$O(i \cdot K \cdot N \cdot n)$$

- O que isso significa?

O que dizer sobre a constante de tempo?
→ Computar Distância Euclidiana via aproximações sucessivas (Newton-Raphson) custa caro.

Se também tenho problema de espaço em memória...
→ Solução aproximada (*sampling*).
→ Paralelizar (mesmo computador) ou distribuir (e.g., map-reduce) o processamento.

Resumo das (des)vantagens do k-means

Vantagens

- Simples e intuitivo
- Complexidade **linear** em todas as variáveis críticas
- Eficaz em muitos cenários de aplicação
- Resultados de interpretação simples

Desvantagens

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

K-Medianas

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...

K-Medóides

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)

Questions and Feedback

Thank you!

Obrigado !

Vinicius Fernandes Caridá
vfcarida@gmail.com

@Vinicius Caridá

@Vinicius Caridá
@machine learning Brasil

@vfcarida