Aprendizagem de Máquina

O Algoritmo K-Means

Prof. Leandro Balby Marinho
http://leandro.lsd.ufcg.edu.br

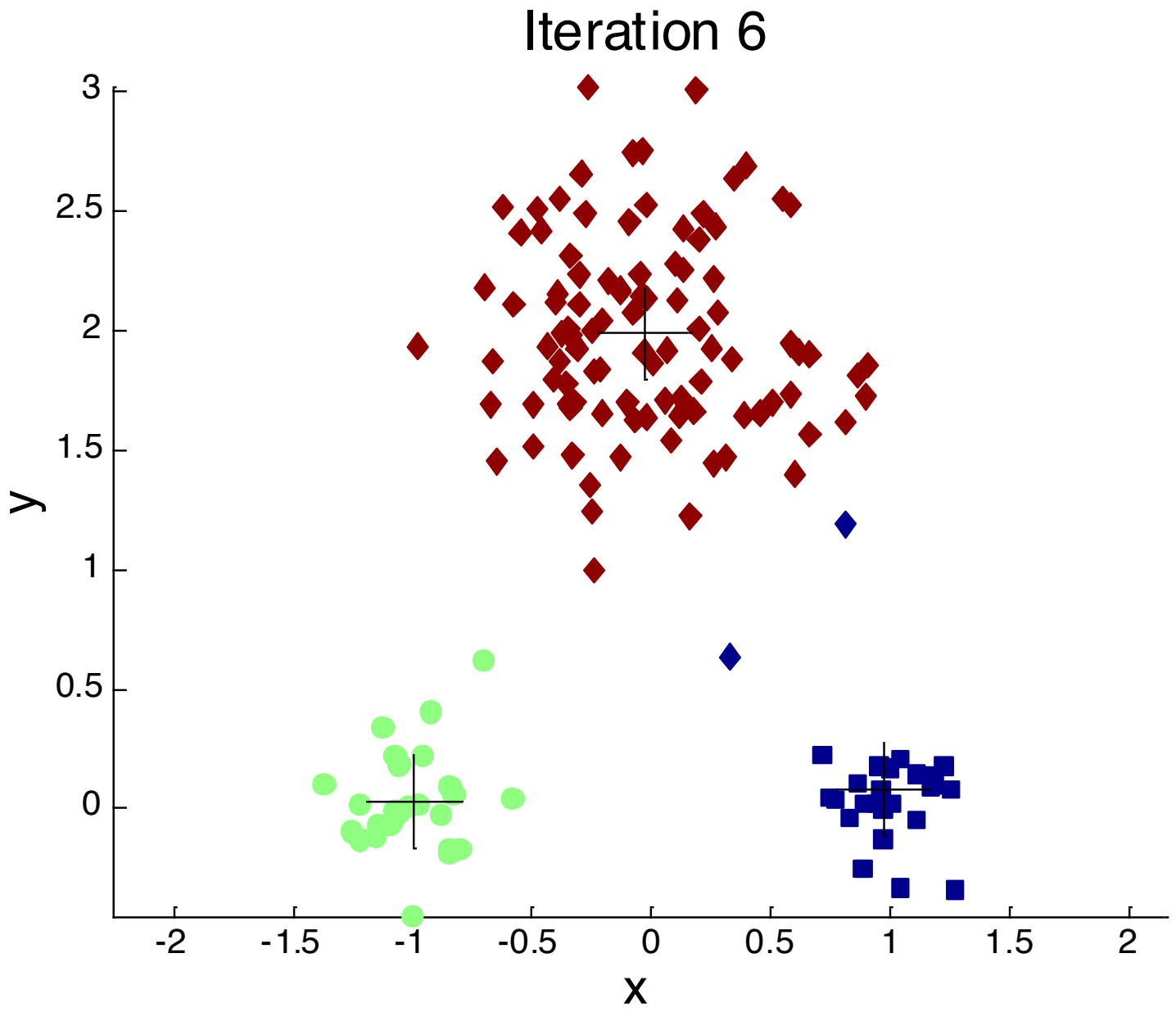
Conceitos e Tipos

Clutering com K-means

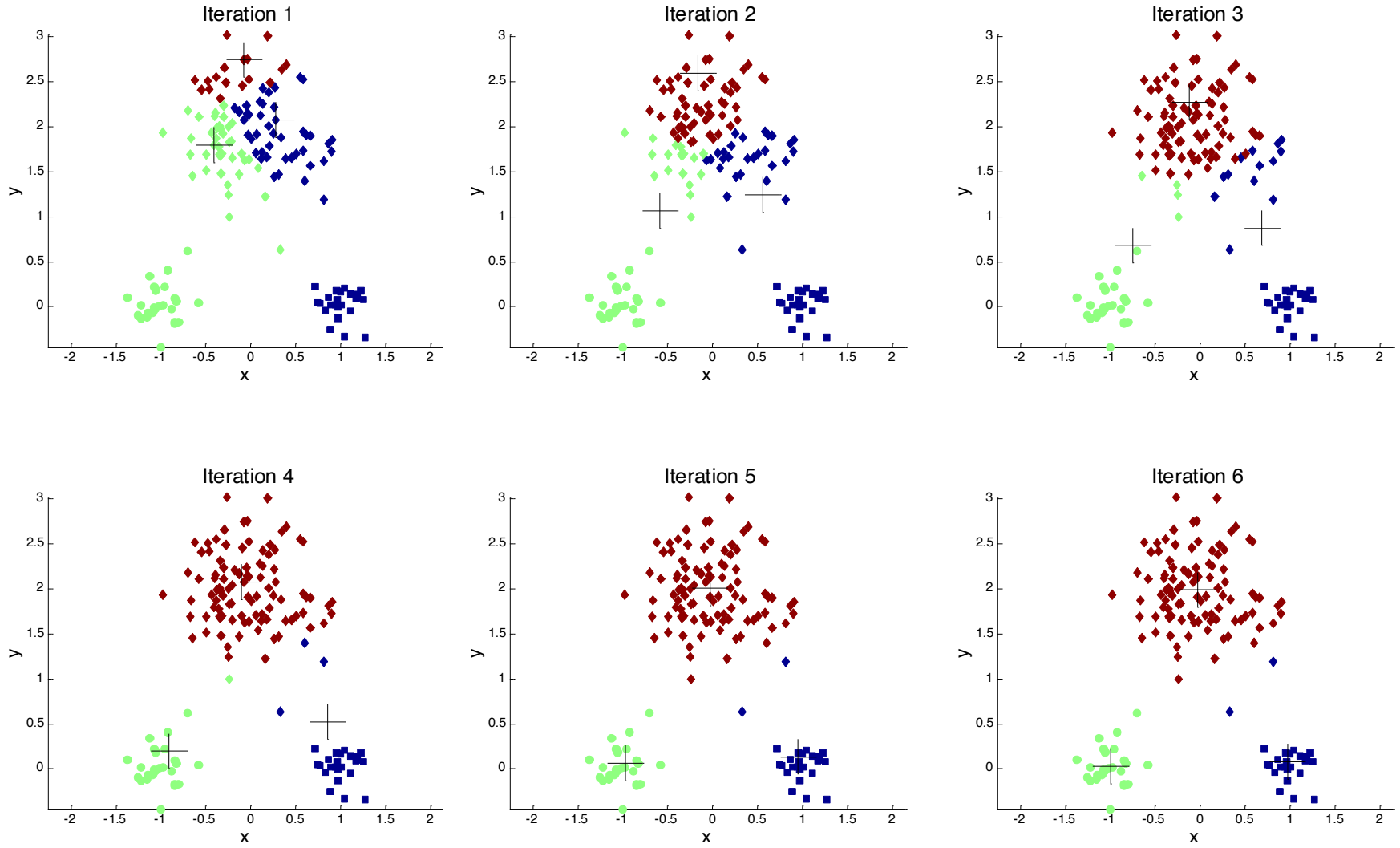
- Abordagem de agrupamento particional
- Número de clusters, K, deve ser especificado
- Cada cluster é associado à um centróide (ponto central)
- Cada ponto é atribuído ao cluster com o centróide mais próximo
- O algoritmo básico é muito simples

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

Examplo do Agrupamento K-means



Examplo do Agrupamento K-means



K-means Clustering – Detalhes

- Centróides iniciais são escolhidos randomicamente.
  - Clusters gerados variam entre execuções diferentes.
- O centróide é (tipicamente) a média dos pontos no cluster.
- ‘Proximidade’ é medida pela distância Euclidiana, similaridade do cosseno, etc.
- K-means converge para medidas de similaridade/distância apropriadas.
- Convergência normalmente aconece nas primeiras iterações.
  - Geralmente condições de parada são usadas para ‘Até poucos pontos mudarem de grupo’
- Complexidade é O(n * K * I * d)
  - n = número de instâncias, K = número de clusters,
    I = número de iterações, d = número de atributos

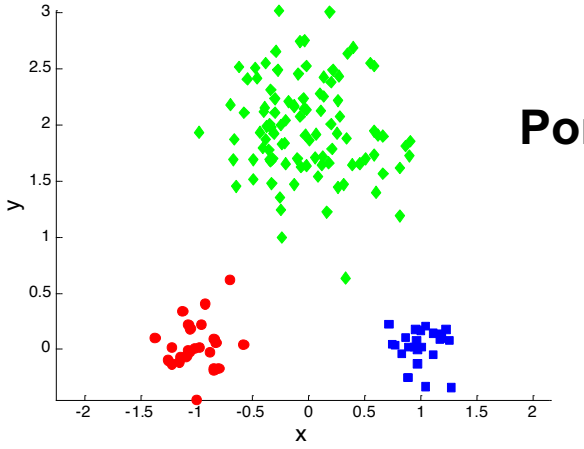
Avaliação de Clusters do K-means

- A medida mais comum é a Soma dos Erros Quadrados (SEQ)
  - Para cada ponto, o erro é a distância ao cluster mais próximo
  - Para calcular SEQ, elevamos esses erros ao quadrado e os somamos:

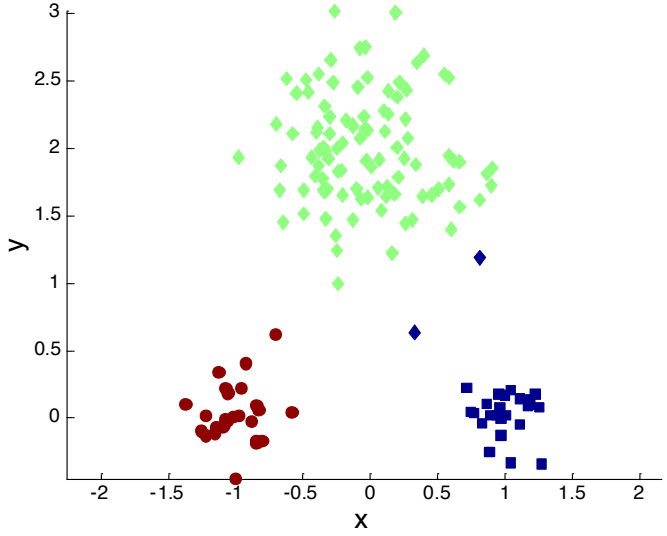
$$SEQ = \sum_{i=1}^{K} \sum_{x \in C_i} d(m_i, x)^2$$

  - x é um objeto no cluster C_i e m_i é o centróide do cluster C_i
  - Dados dois conjuntos de clusters, preferimos aquele com o menor erro
  - Uma forma fácil de reduzir SEQ é aumentar K
    - Cuidado pois pode levar a overfitting.

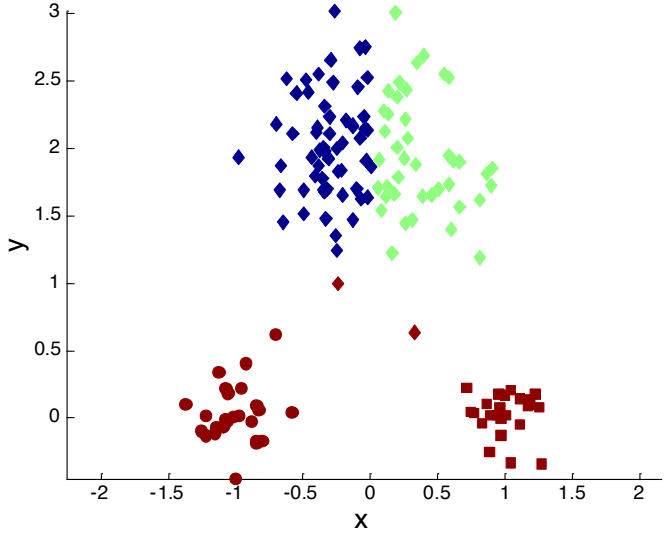
Dois Agrupamentos diferentes do K-means



Pontos Originais



Clustering Ótimo

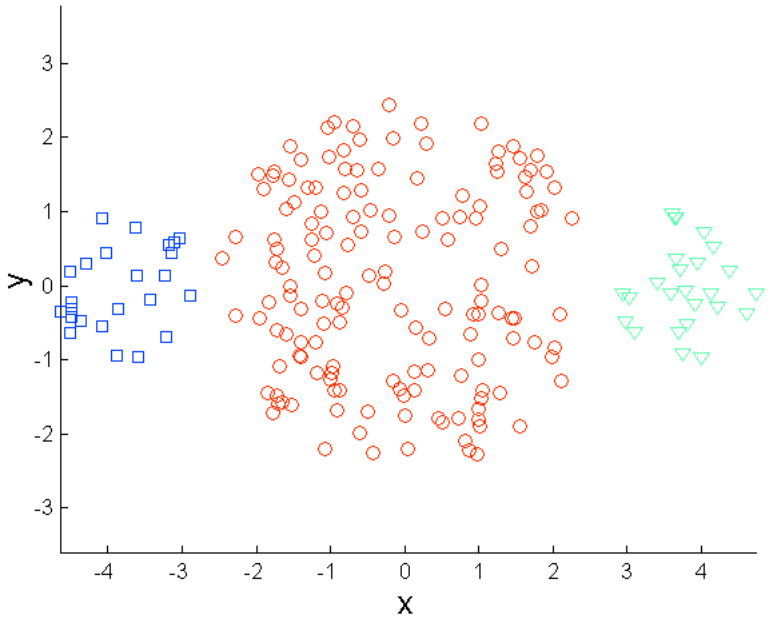


Clustering Sub-ótimo

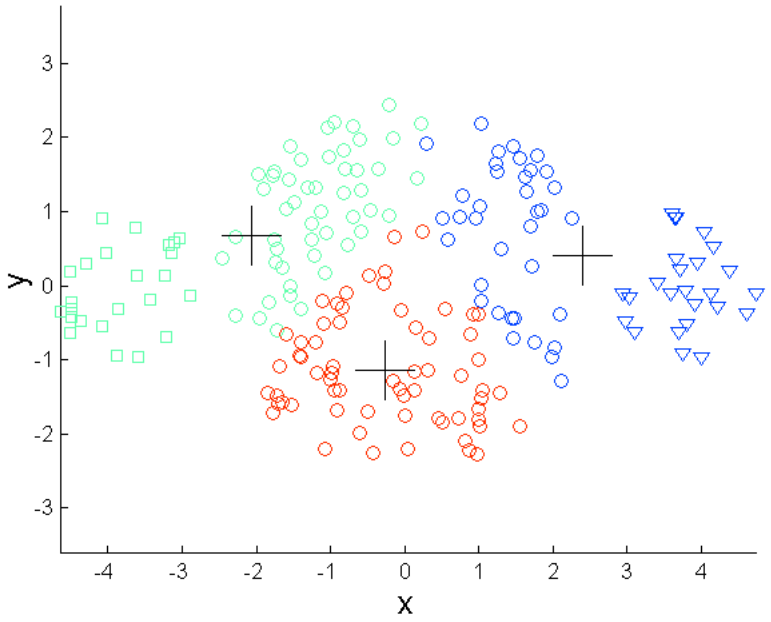
Limitações do K-means

- K-means tem problemas quando clusters tem diferentes
  - Tamanhos
  - Densidades
  - Formatos não globulares
- K-means tem problemas quando os dados contém outliers.

Limitações do K-means: Tamanhos Diferentes

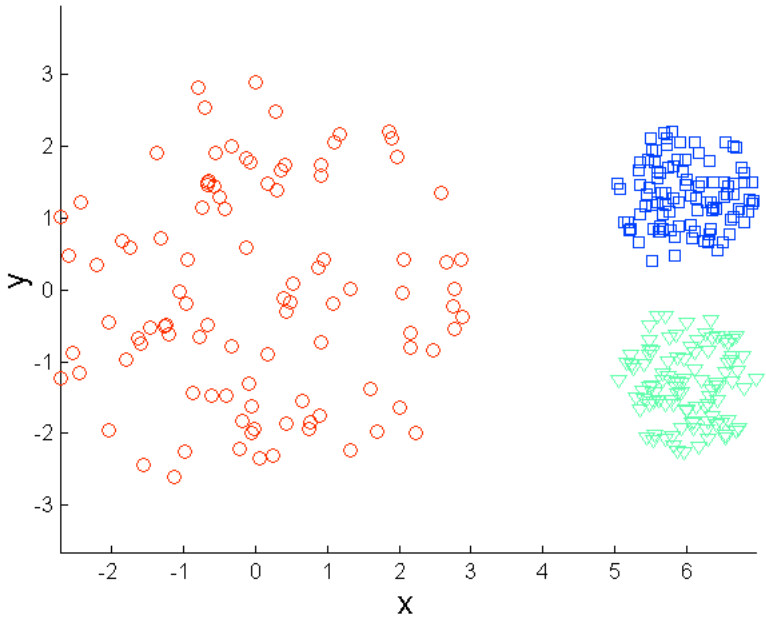


Pontos Originais

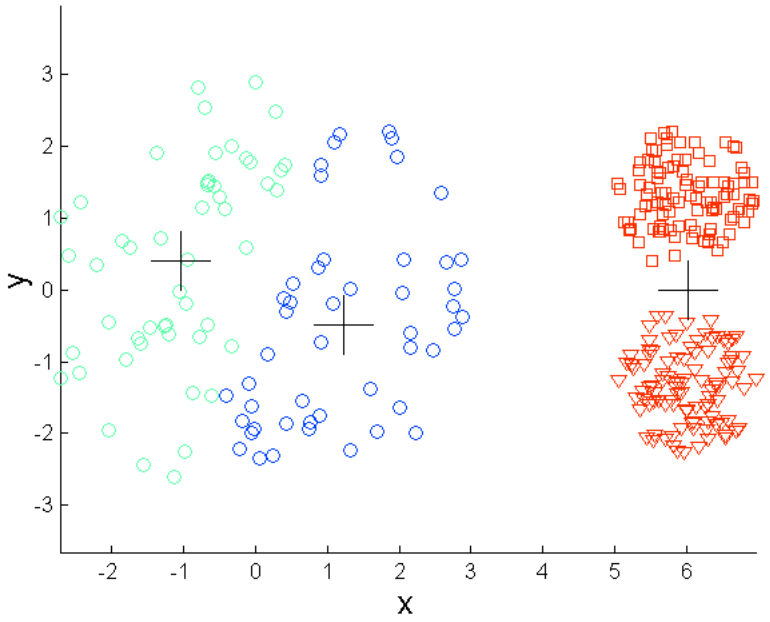


K-means (3 Clusters)

Limitações do K-means: Densidades Diferentes

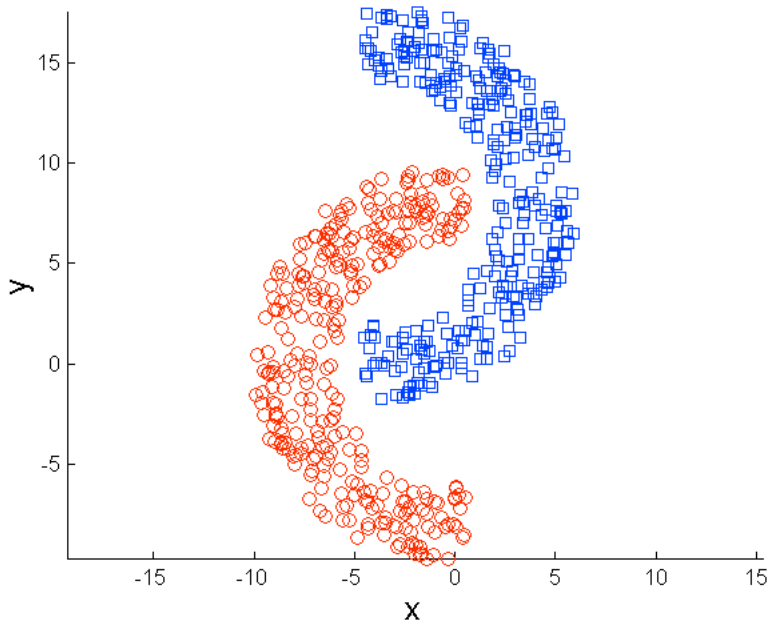


Pontos Originais

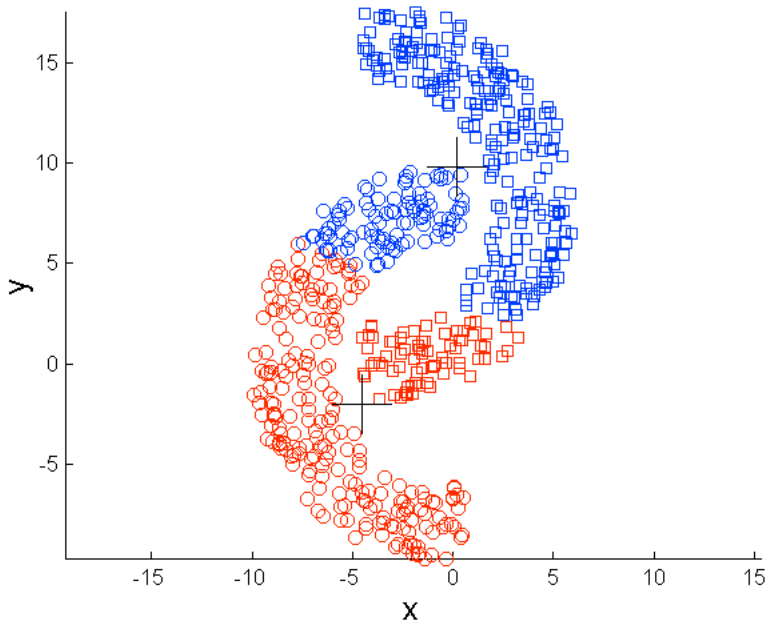


K-means (3 Clusters)

Limitações do K-means: Formatos Não Globulares

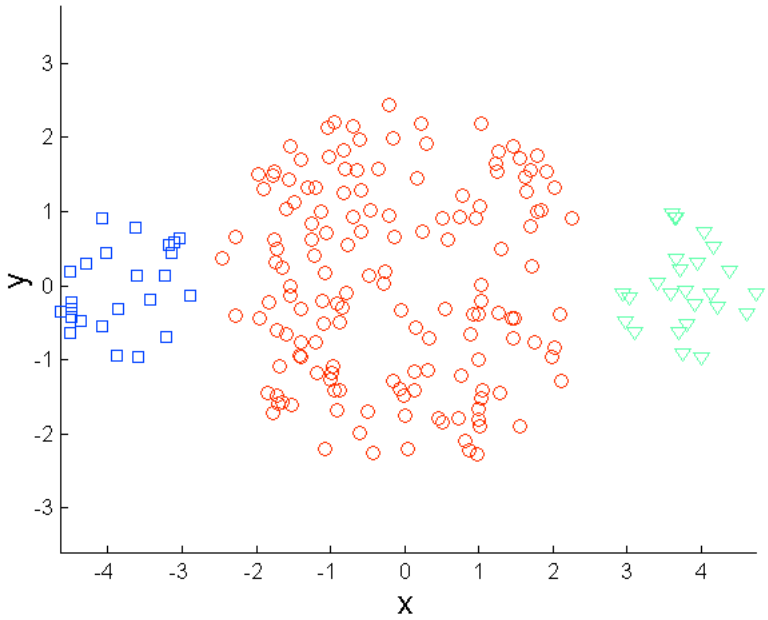


Pontos Original

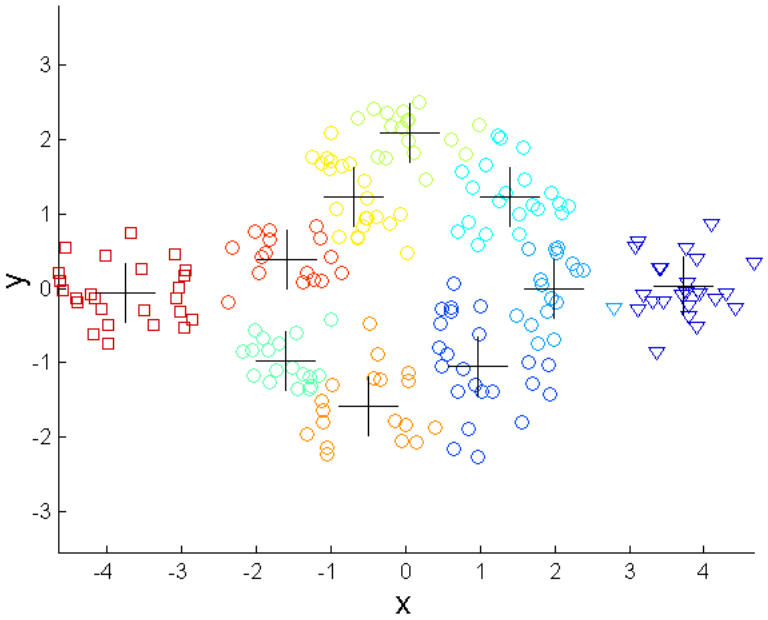


K-means (2 Clusters)

Superando Limitações do K-means



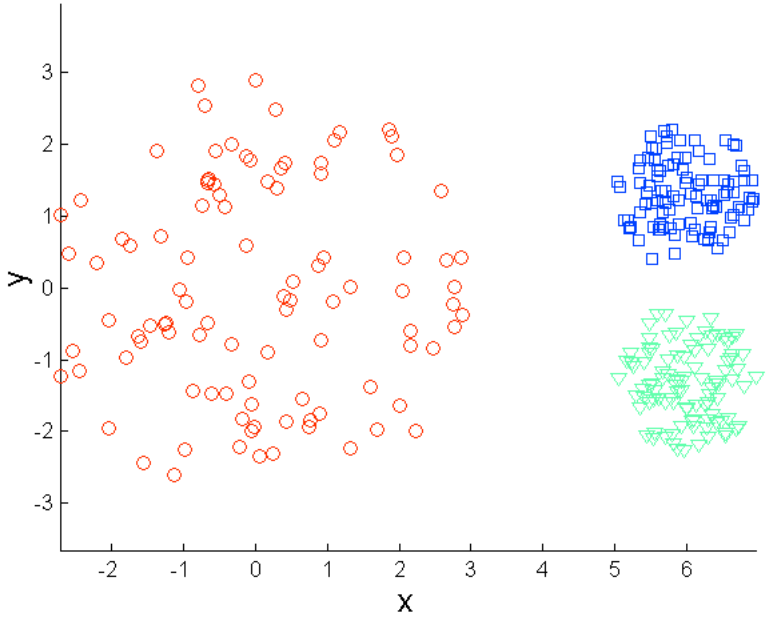
Pontos Originais



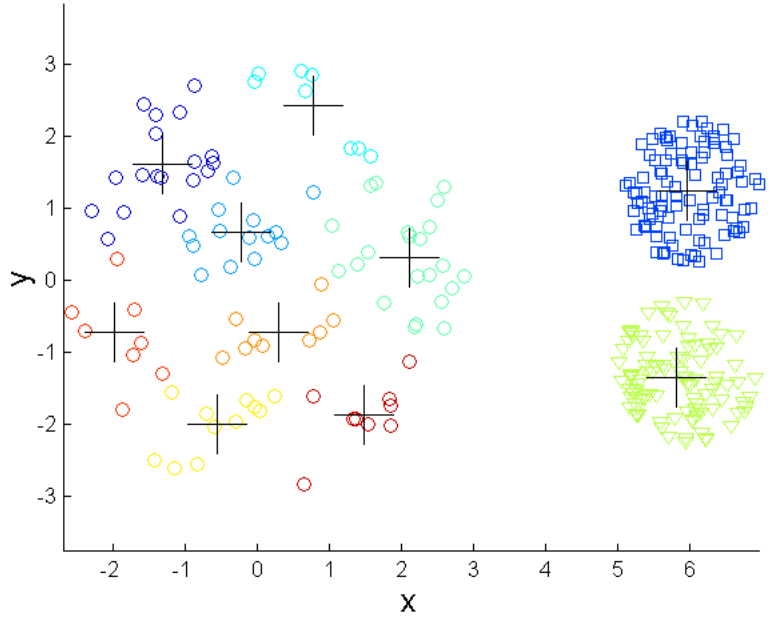
Clusters do K-means

Uma solução usar muitos clusters. Encontra partes de clusters, mas ainda precisa juntá-las.

Superando Limitações do K-means

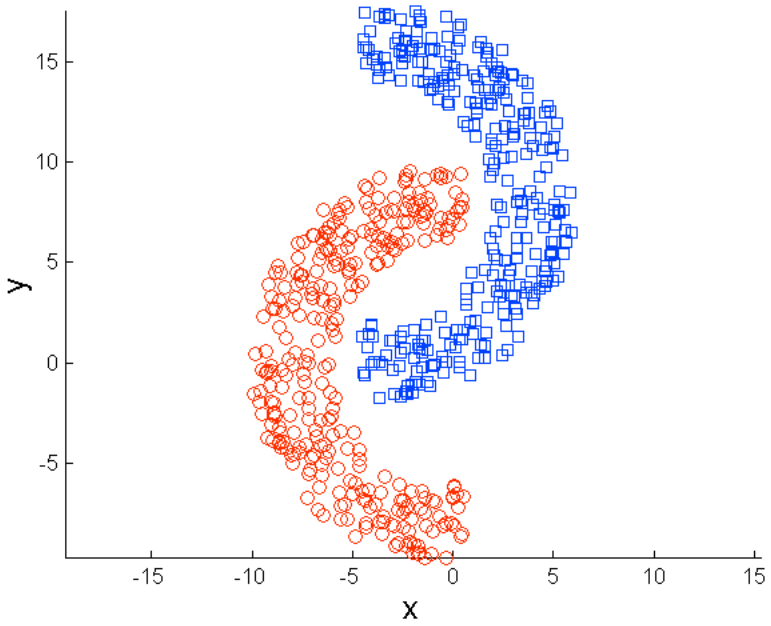


Pontos Originais

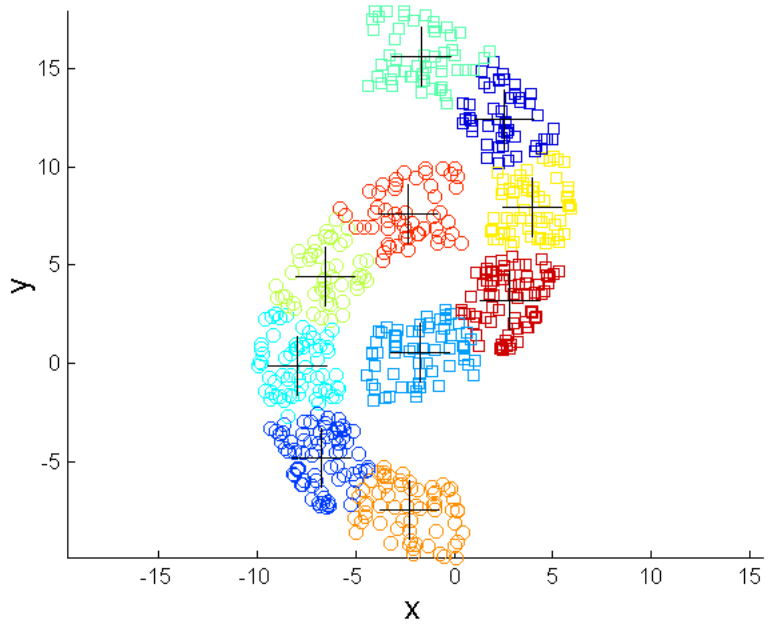


Clusters do K-means

Superando Limitações do K-means

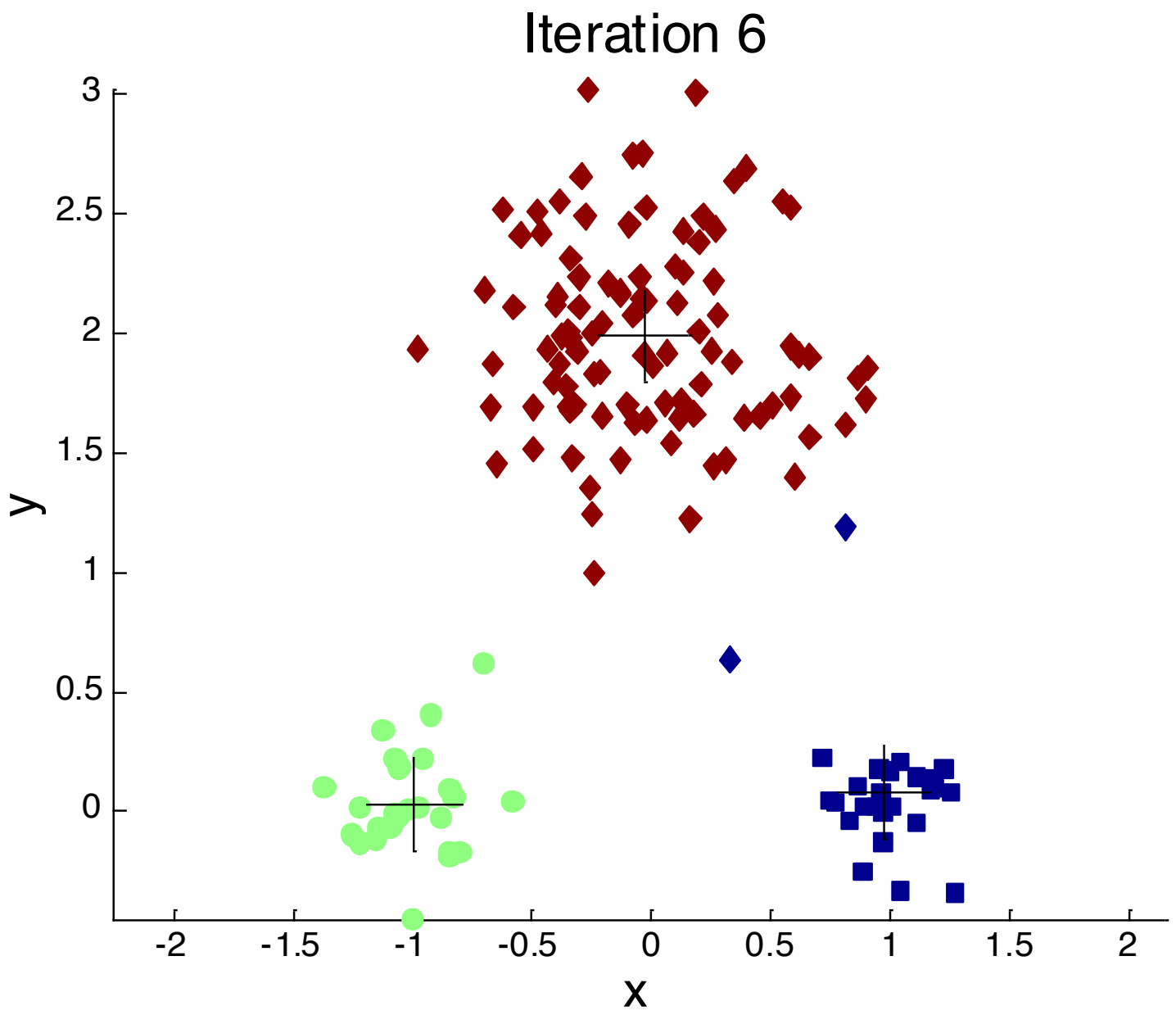


Pontos Originais

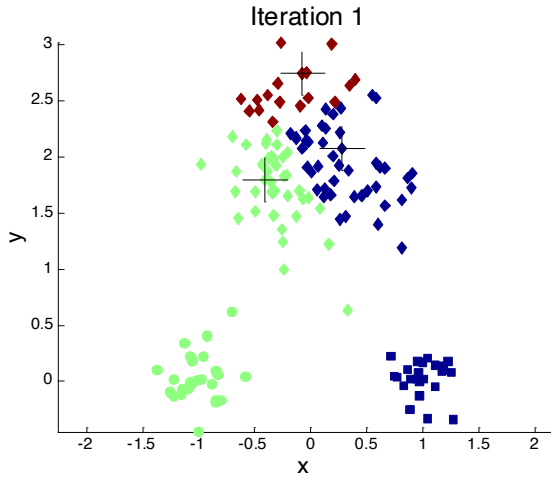


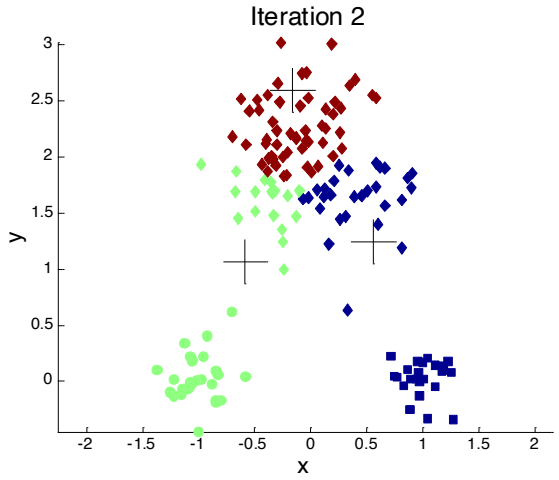
Clusters do K-means

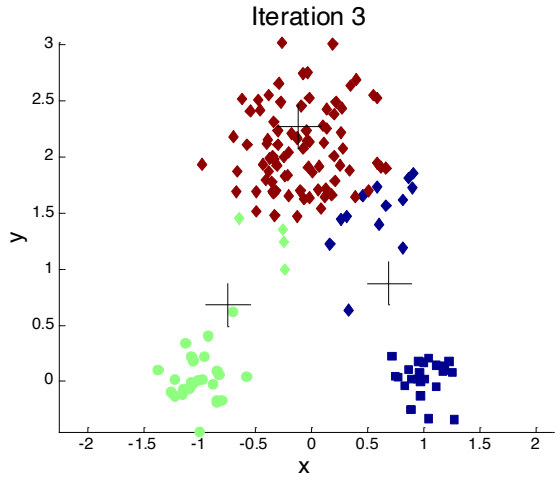
Importância de Escolher Centróides Iniciais

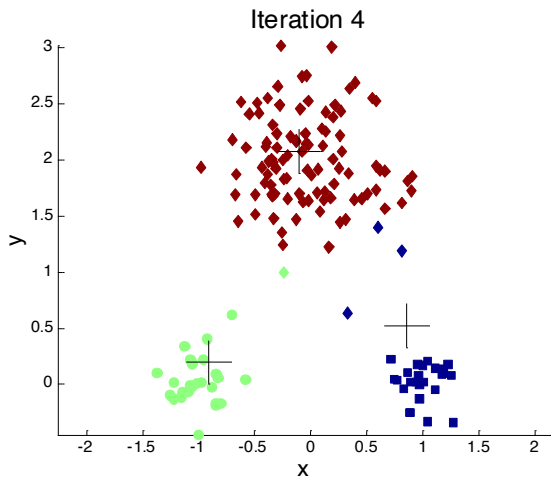


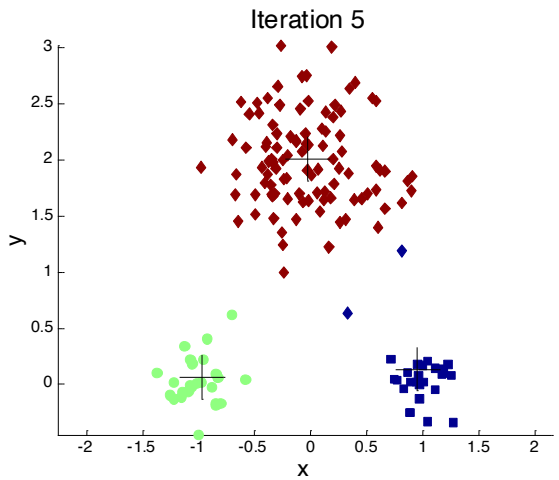
Importância de Escolher Centróides Iniciais

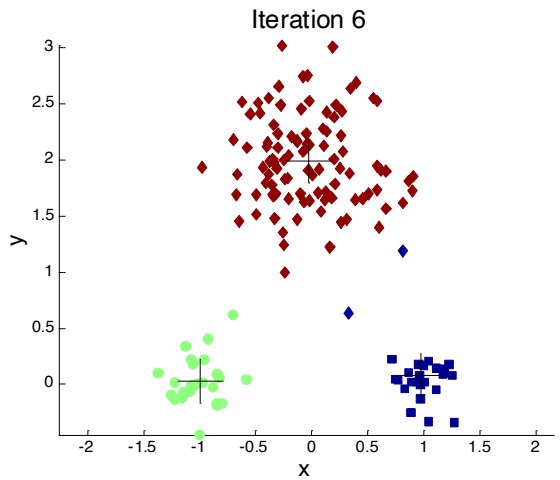




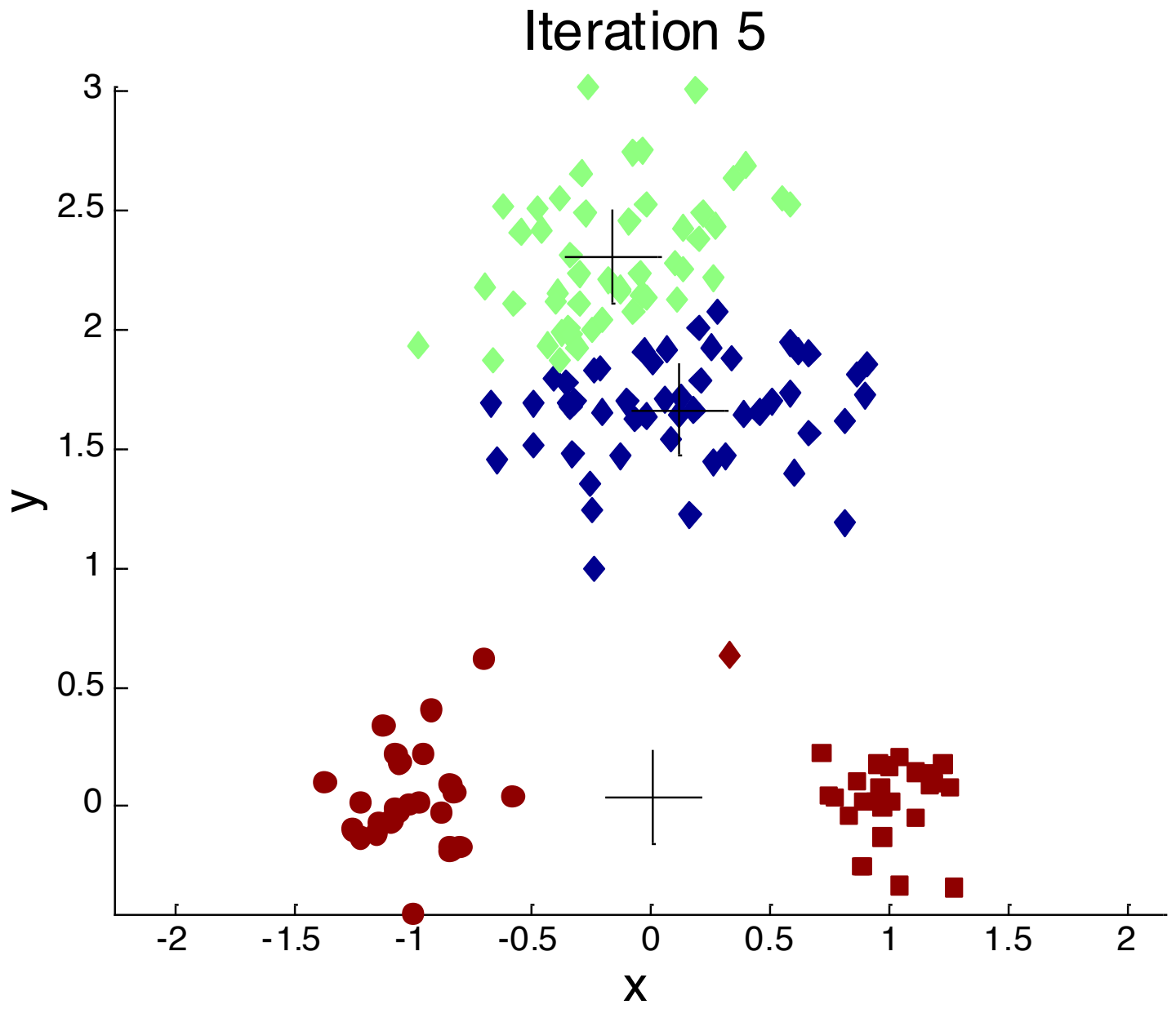




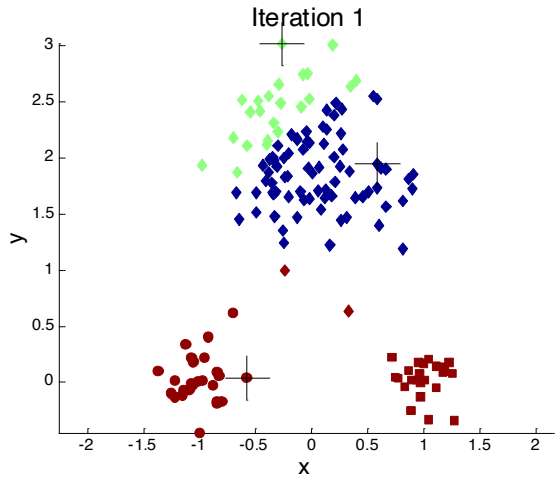


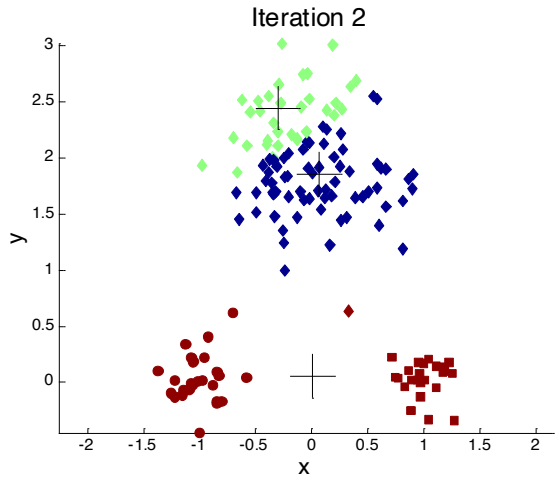


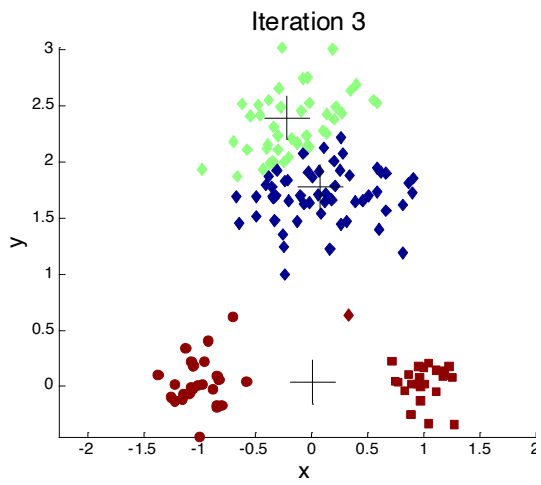
Importância de Escolher Centróides Iniciais

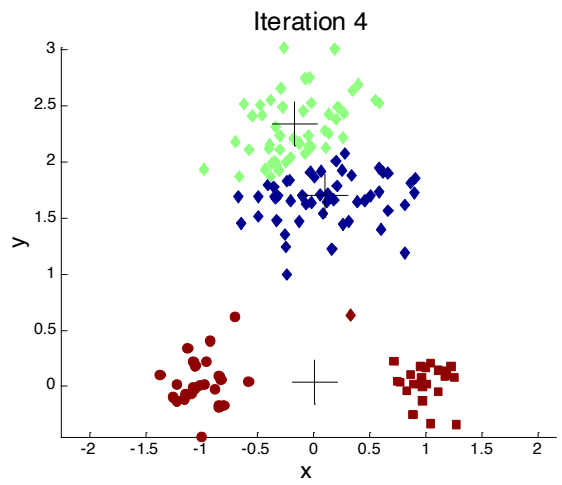


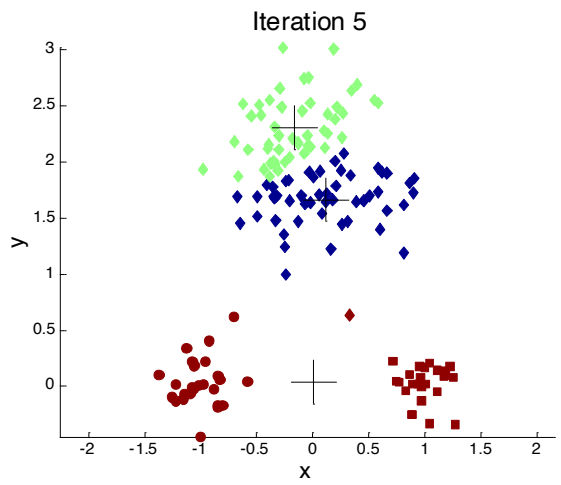
Importância de Escolher Centróides Iniciais











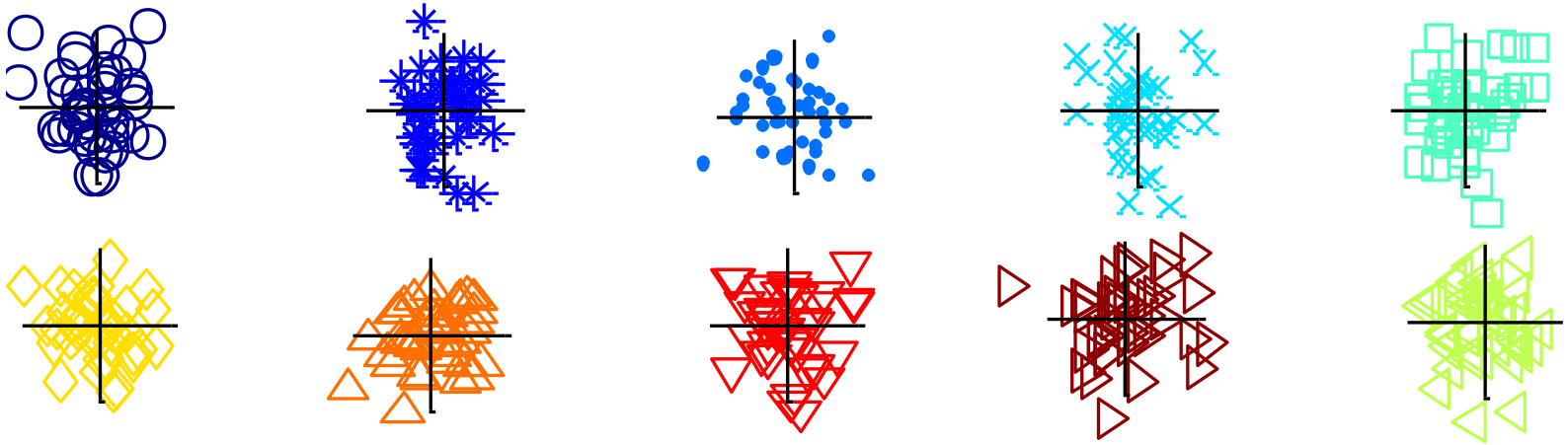
Problemas com a Seleção de Pontos Iniciais

- Se há K clusters ‘reais’ então a chance de selecionar um centróide de cada cluster é pequena.
  - A chance é relativamente pequena quando K é grande
  - Se os clusters tem o mesmo tamanho, n, então’1

$$P = \frac{\text{number of ways to select one centroid from each cluster}}{\text{number of ways to select } K \text{ centroids}} = \frac{K!n^K}{(Kn)^K} = \frac{K!}{K^K}$$

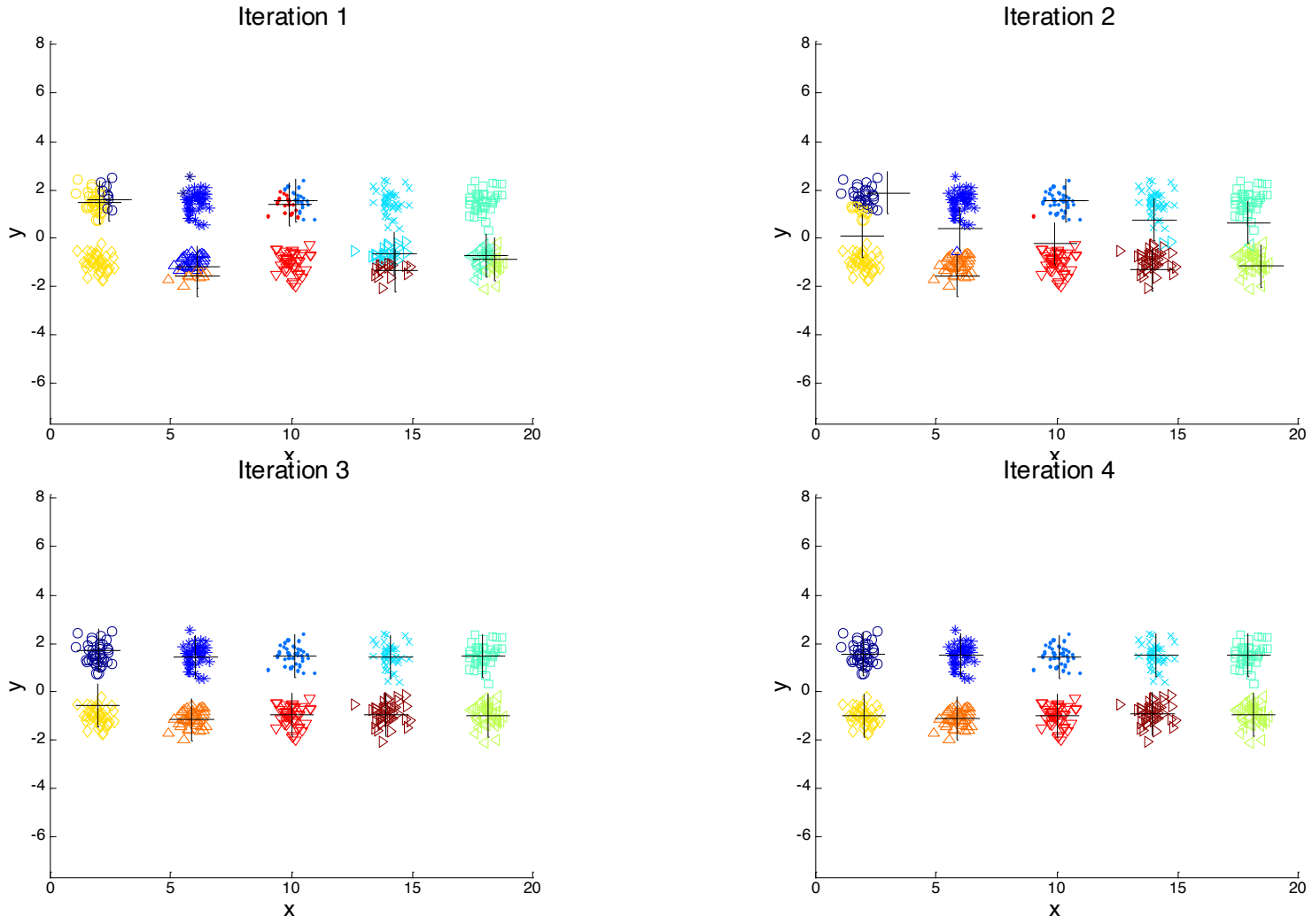
  - Por exemplo, se K = 10, então a probabilidade = $10!/10^{10}$ = 0.00036
  - Algumas vezes os centróides iniciais se reajustam da forma “correta”, outras vezes não
  - Considere um examplo de cinco pares de clusters

Examplo: 10 Clusters



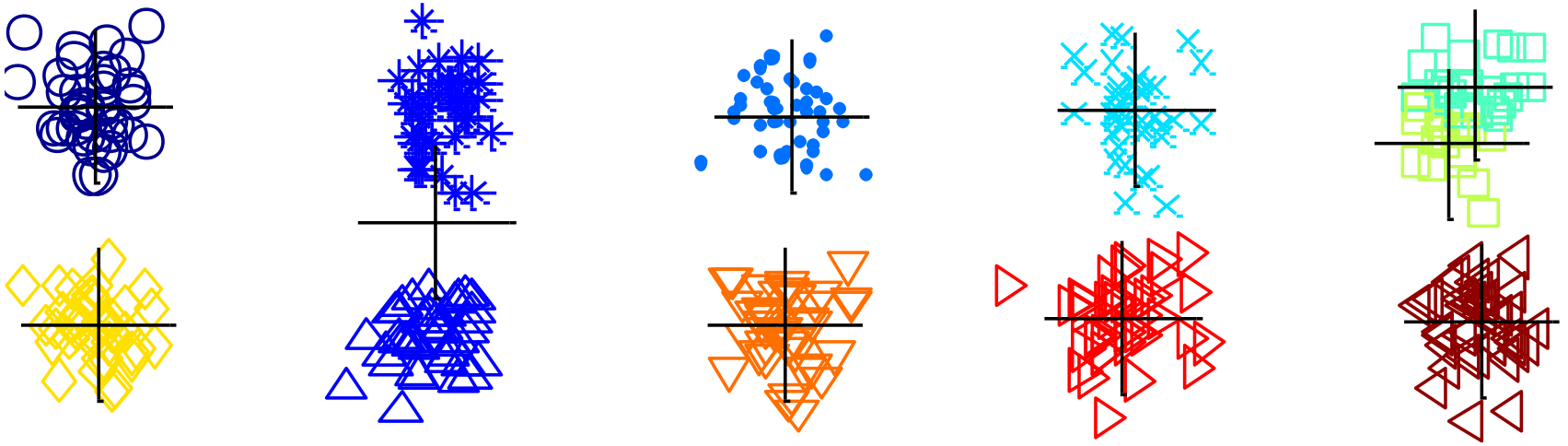
Começando com dois centróides iniciais em um cluster de cada par de clusters

Exemplo: 10 Clusters



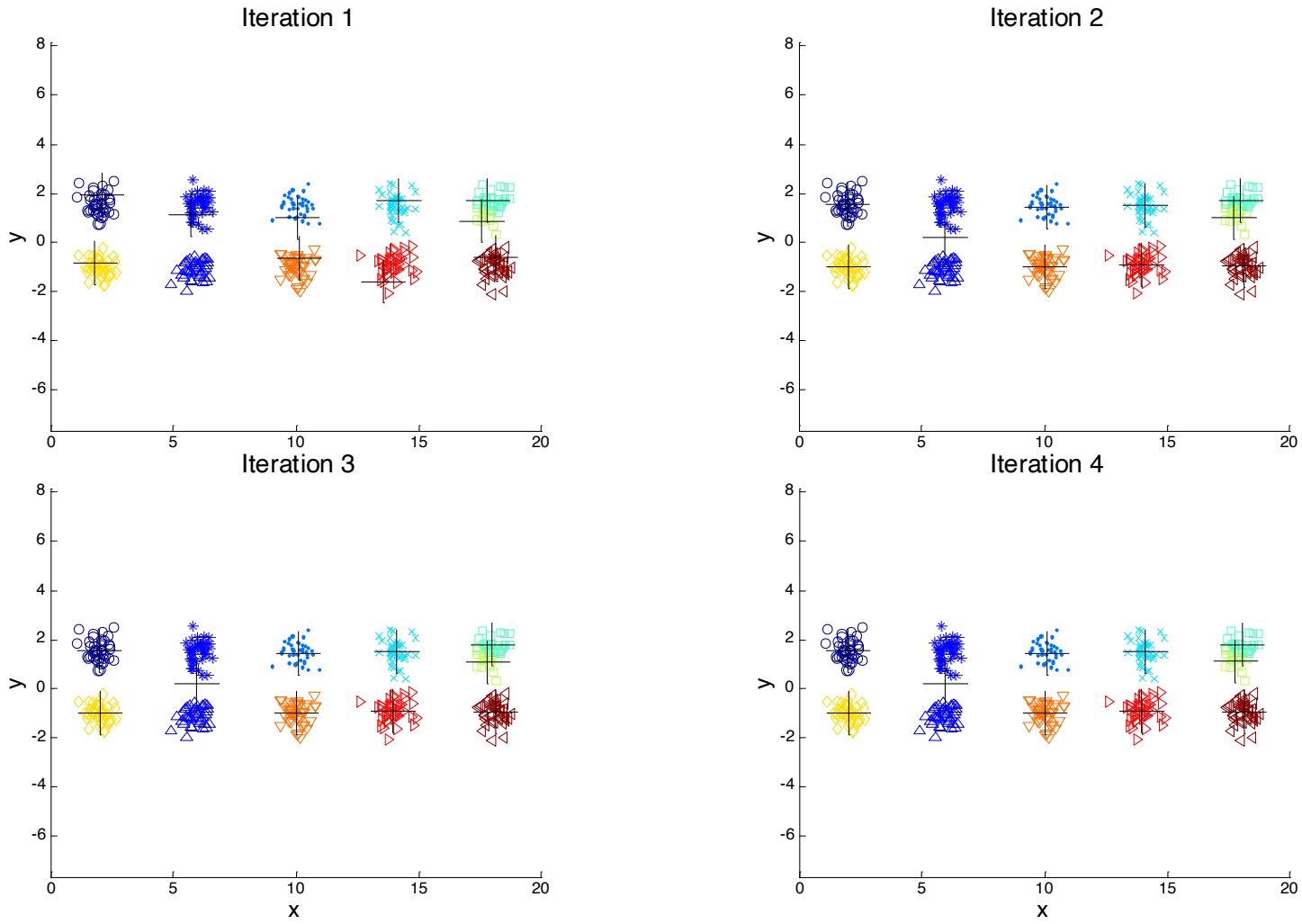
Começando com dois centróides iniciais em um cluster de cada par de clusters

Exemplo: 10 Clusters



Começando com alguns pares de clusters tendo três centroids iniciais, enquanto outros tendo apenas um.

Exemplo: 10 Clusters



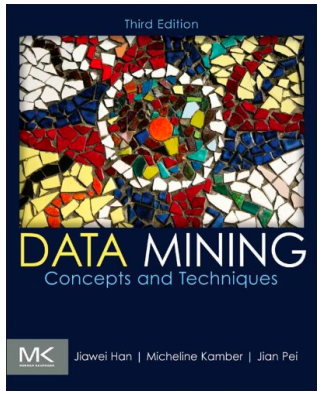
Começando com alguns pares de clusters tendo três centroids iniciais, enquanto outros tendo apenas um.

Solução para o Problema dos Centróides Iniciais

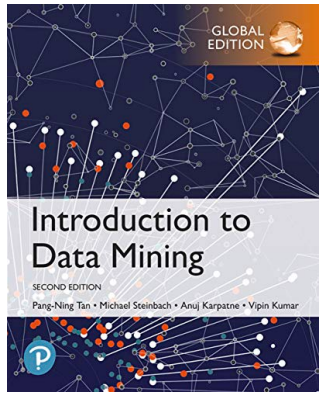
- Múltiplas execuções
  - Ajuda, mas a probabilidade não está do nosso lado
- Amostre e use clustering hierárquico para determiner centróides iniciais
- Selecione mais so que K centróides iniciais e então selecione entre esses centróides iniciais
  - Selecione os mais separados
- Pós-processamento
- Gere um grande número de clusters e então realize um agupamento hierárquico
- Bisecting K-means
  - Não tão suscetível à problemas de inicialização

Referências

- Data Mining: Concepts and Techniques: Jiawei Han, Jian Pei, Micheline Kamber. 2011



- Introduction to Data Mining: Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Anuj Karpatne. 2019



- Slides adaptados dos slides dos Autores.