APRENDIZADO DE MÁQUINA

Clustering: Algoritmos Particionais & Validação

Aula de Hoje

- Algoritmos Particionais Sem Sobreposição
  - k-means
  - Variantes do k-means
  - Estimativa do número de grupos k
- Critérios relativos de validade de agrupamento
- Critérios externos de validade de agrupamento
- Algoritmos Particionais Com Sobreposição
  - Fuzzy c-Means (FCM)
  - Expectation Maximization (EM)

Métodos de Partição (Sem Sobreposição)

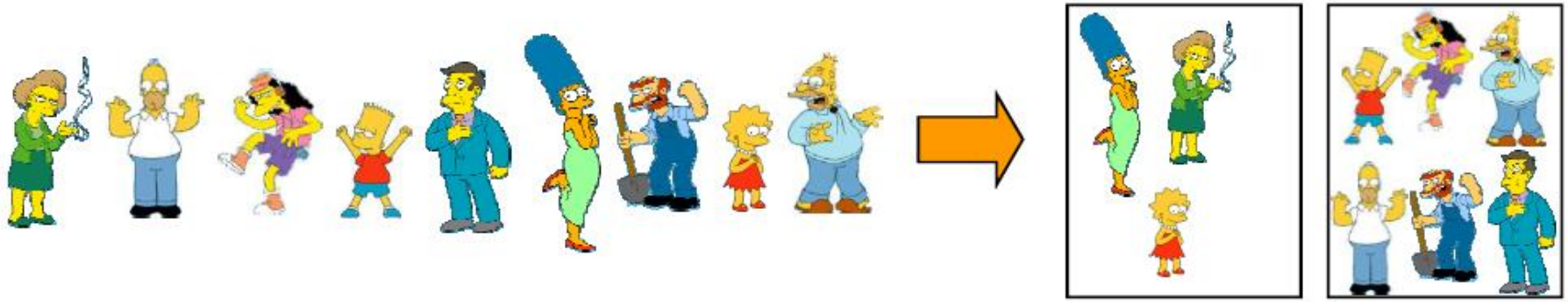
- Matriz de Dados:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Nn} \end{bmatrix}$$

Problema: Particionar o conjunto $\mathbf{X}=\{\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_N\}$ de objetos $\mathbf{x}_l \in \Re^n$ em uma coleção $\mathbf{C}=\{\mathbf{C}_1,\mathbf{C}_2,...,\mathbf{C}_k\}$ de k sub-conjuntos mutuamente disjuntos $\mathbf{C}_i$ de $\mathbf{X}$ tal que $\mathbf{C}_1 \cup \mathbf{C}_2 \cup ... \cup \mathbf{C}_k = \mathbf{X}$, $\mathbf{C}_i \neq \varnothing$, e $\mathbf{C}_i \cap \mathbf{C}_j = \varnothing$ para $i \neq j$

Métodos de Partição (Sem Sobreposição)

- Cada exemplo pertence a um cluster dentre k clusters possíveis;
- Usuário normalmente deve fornecer o número de clusters (k);
- Normalmente envolvem a otimização de algum índice (critério numérico) que reflete a qualidade de determinada partição;



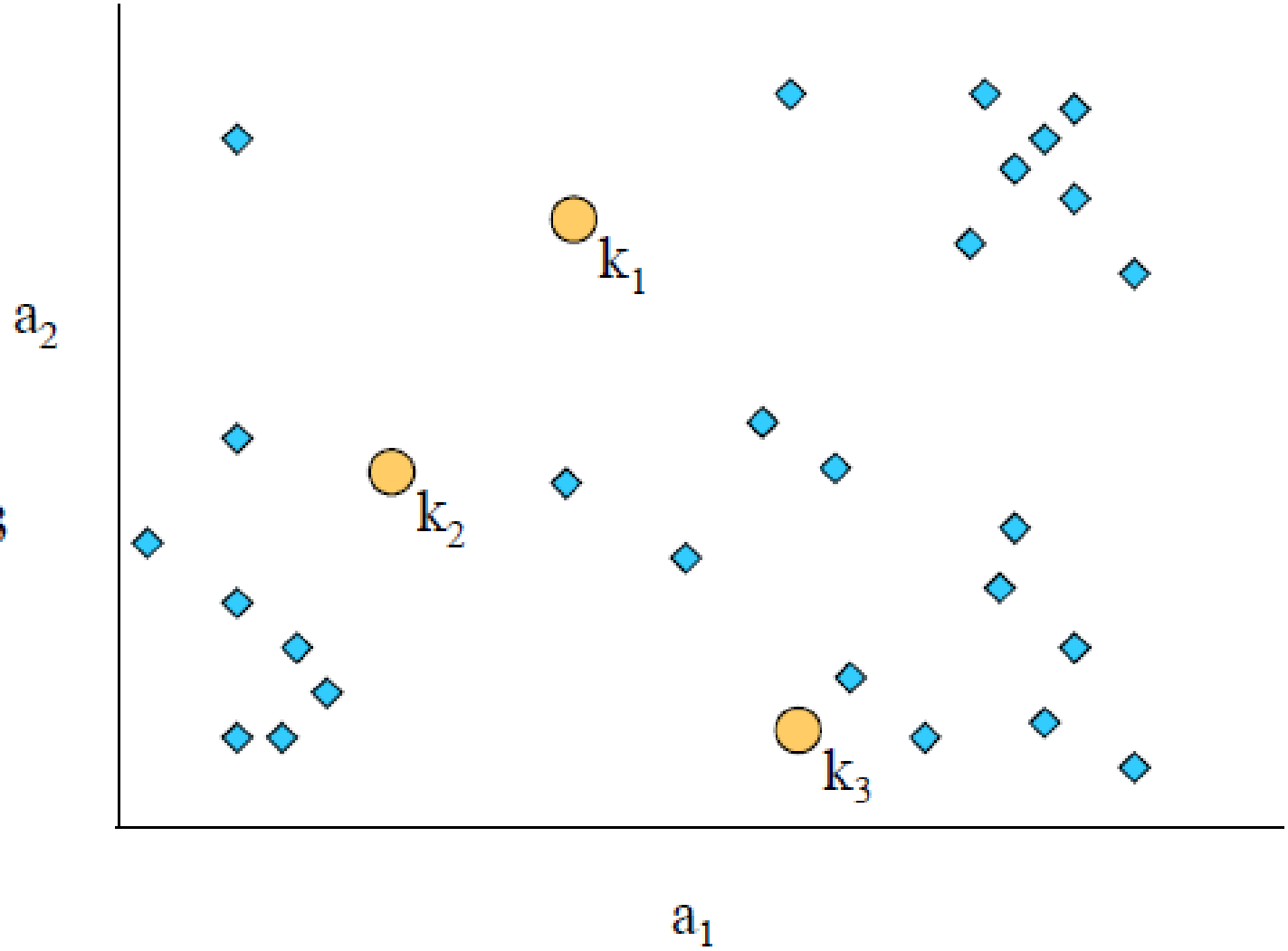
- Vamos iniciar por um algoritmo amplamente utilizado (k-means), o qual fornecerá uma noção mais intuitiva do problema a ser resolvido.

k-means

- Escolher aleatoriamente um número *k* de protótipos (centros) para os clusters
- Atribuir cada objeto para o cluster de centro mais *próximo* (segundo alguma distância, e.g. Euclidiana)
- Mover cada centro para a média (centroide) dos objetos do cluster correspondente
- Repetir os passos 2 e 3 até que algum critério de convergência seja obtido:
  - número máximo de iterações
  - limiar mínimo de mudanças nos centroides

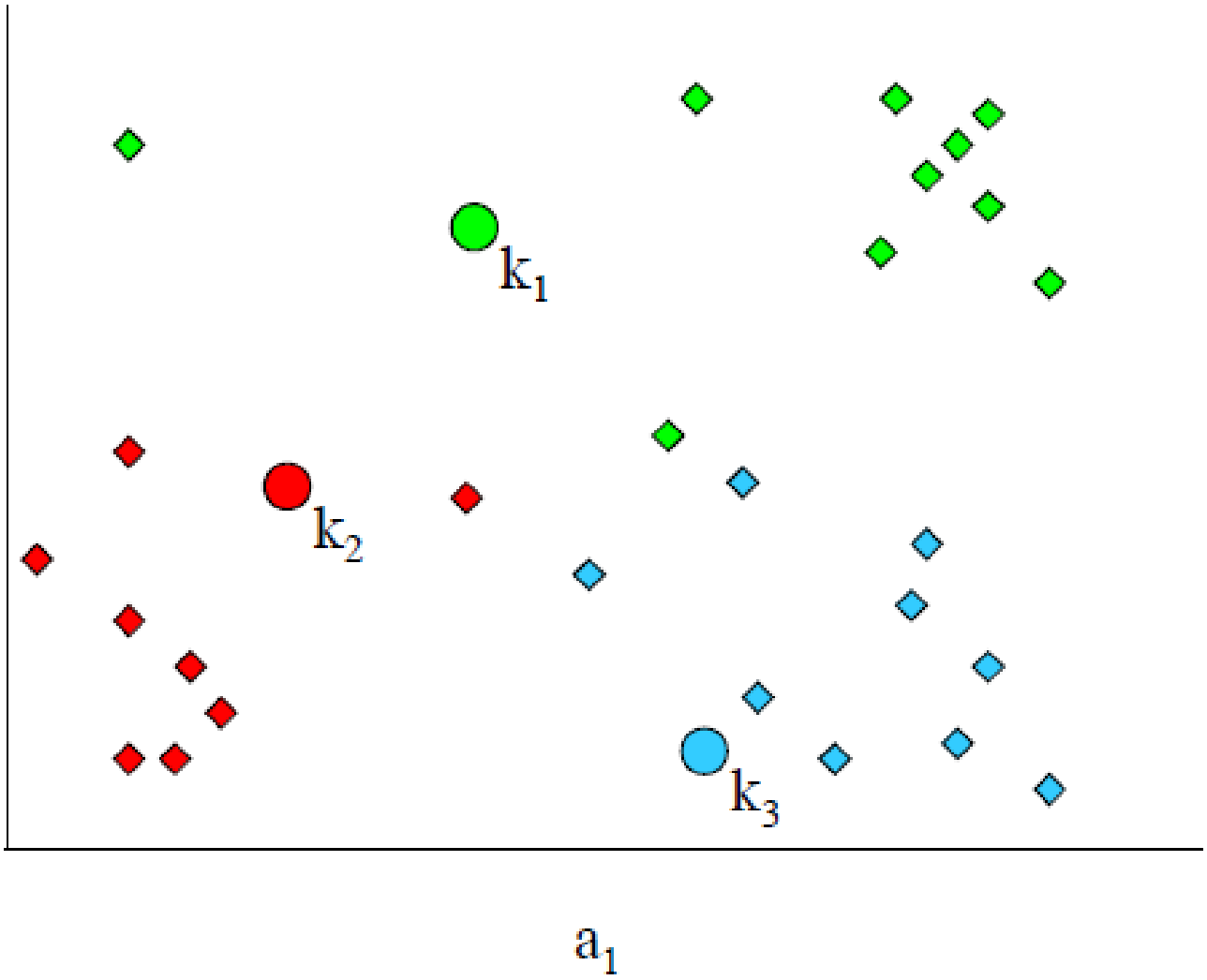
k-means - passo 1:

Escolher 3 centros iniciais



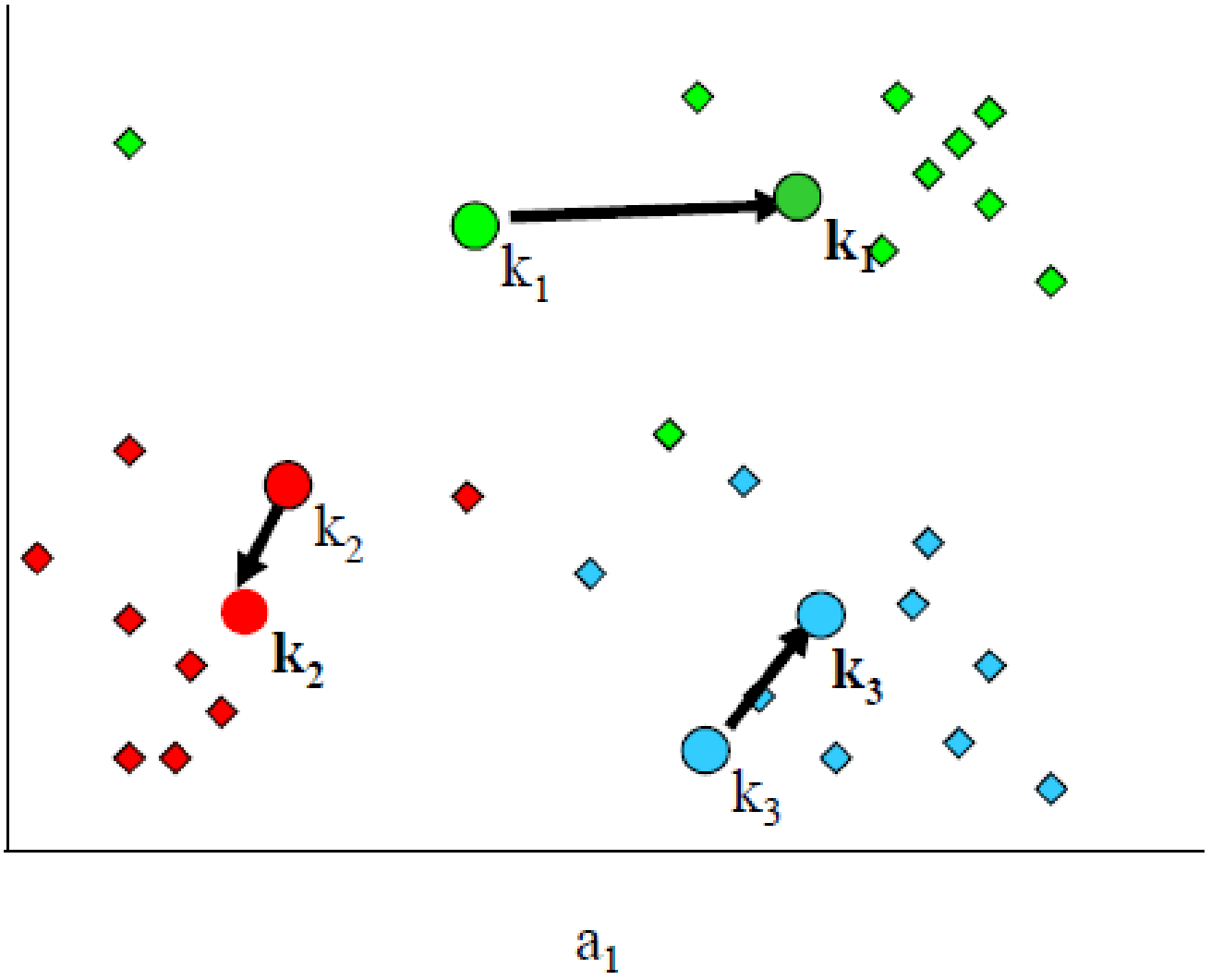
k-means - passo 2:

Atribuir cada objeto ao cluster de centro + próximo



k-means - passo 3:

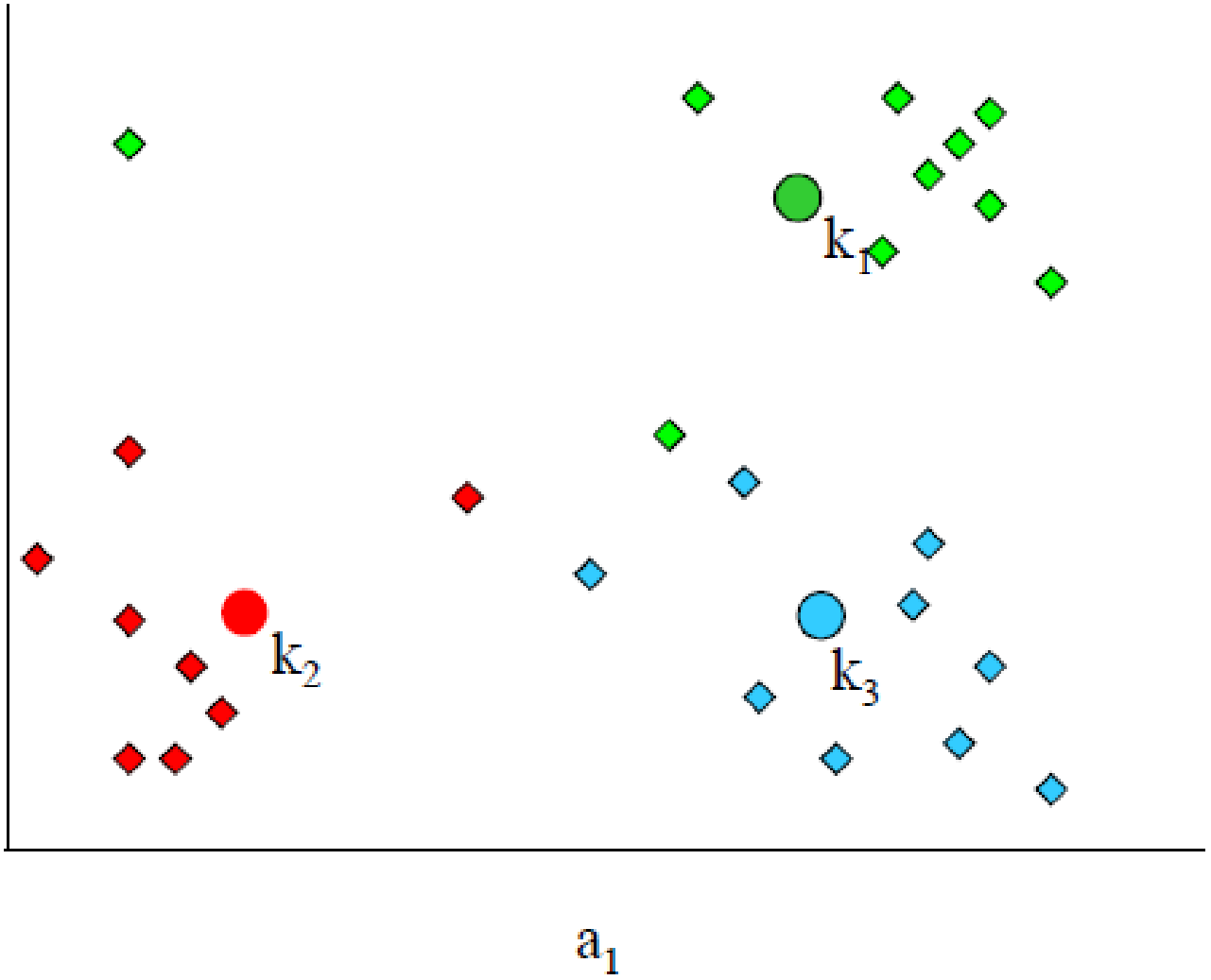
Mover cada centro para o vetor médio do cluster (centróide)



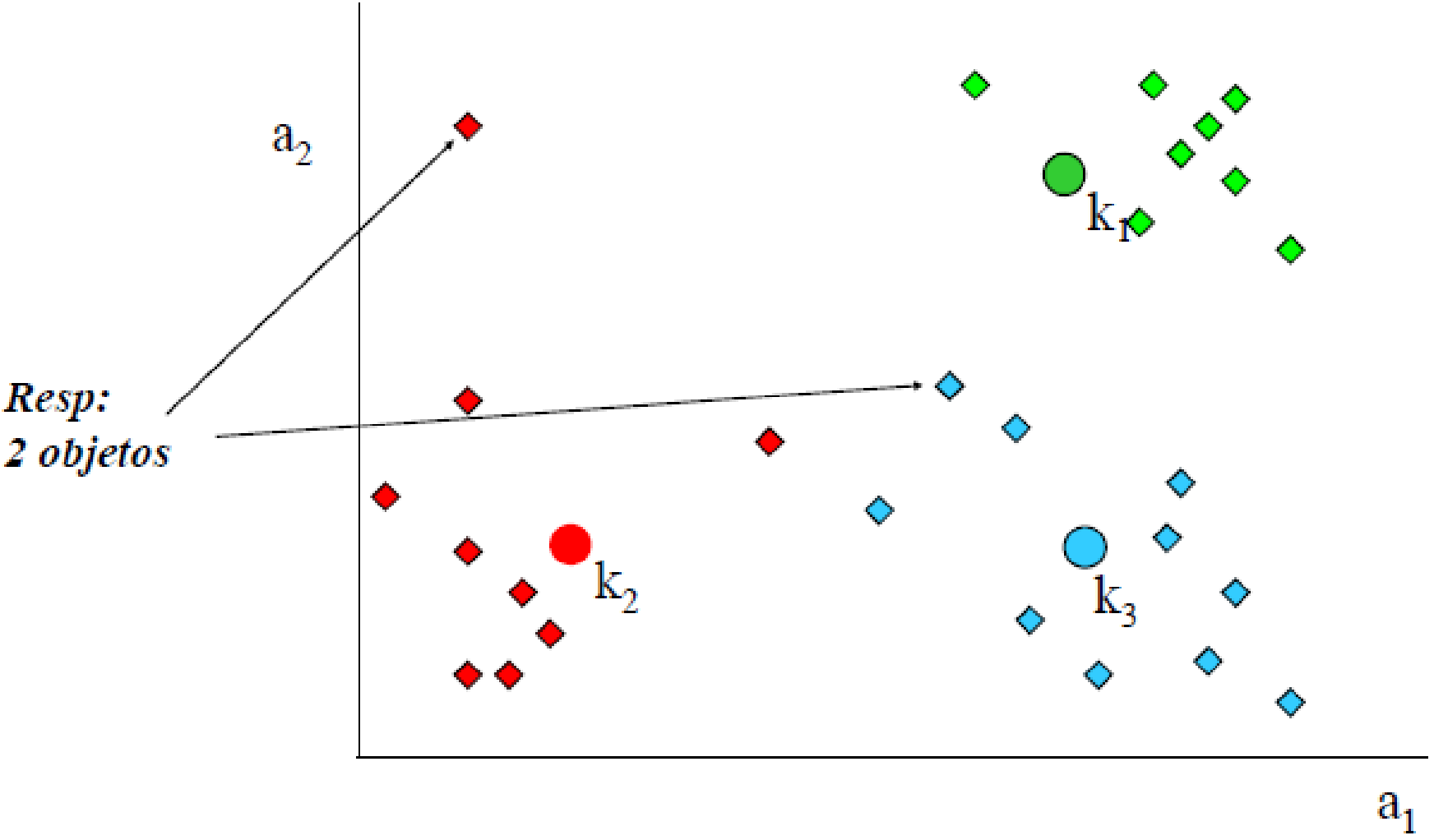
k-means:

Re-atribuir objetos aos clusters de centróides mais próximos

Quais objetos mudarão de cluster?

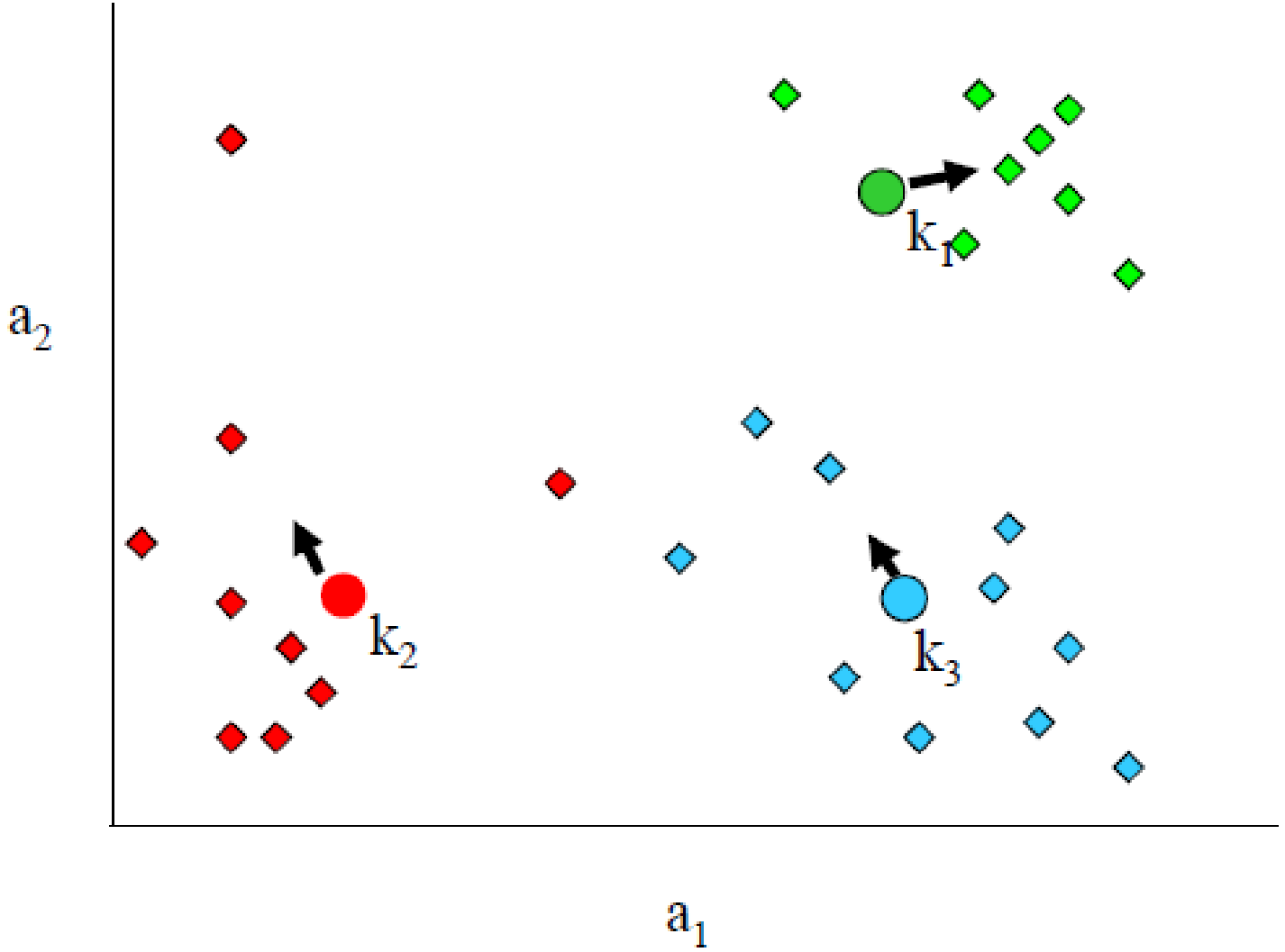


k-means:



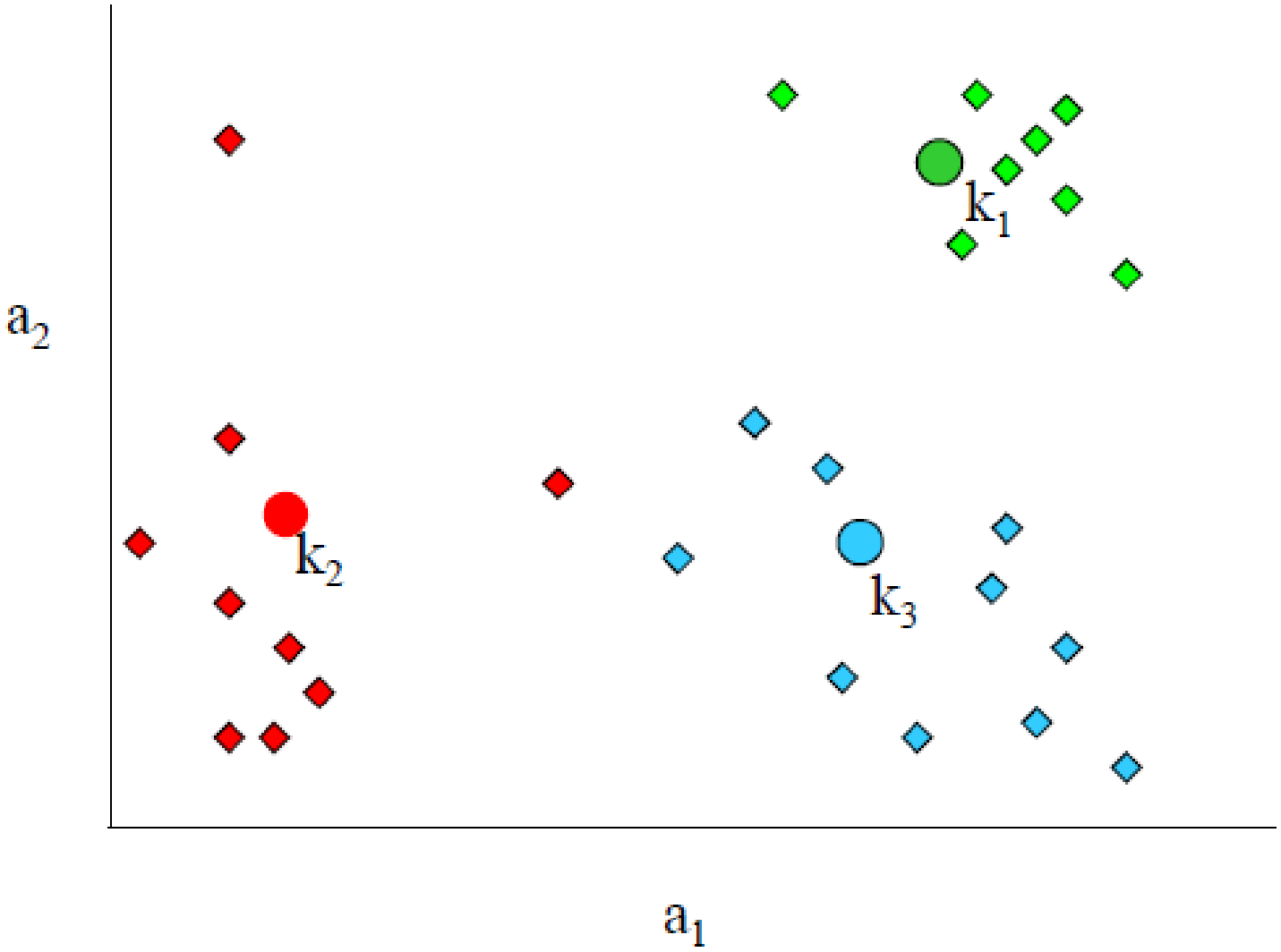
k-means:

re-calcular
vetores
médios



k-means:

mover centros dos clusters…



Discussão

- Pode-se demonstrar que o algoritmo minimiza a seguinte função objetivo (variâncias intra-cluster):

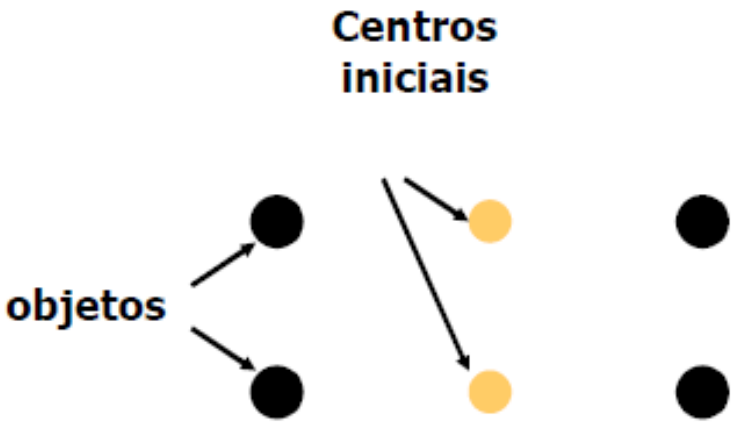
$$J = \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in \mathbf{C}_i} d\left(\mathbf{x}_j, \bar{\mathbf{x}}_i\right)^2$$

onde $\mathbf{x}_i$ é o centróide do i-ésimo cluster:

$$\bar{\mathbf{x}}_i = \frac{1}{\left|\mathbf{C}_i\right|} \sum_{\mathbf{x}_i \in \mathbf{C}_i} \mathbf{x}_i$$

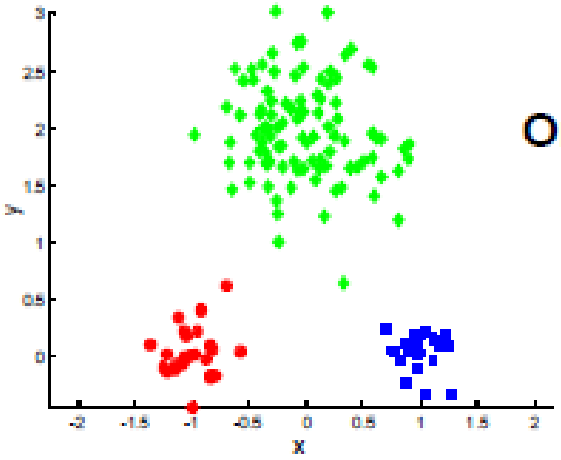
Discussão

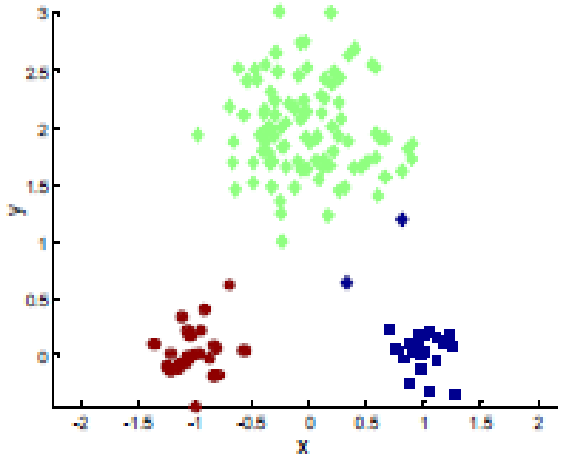
- Resultado pode variar significativamente dependendo da escolha das sementes (protótipos) iniciais;
- k-means pode “ficar preso” em ótimos locais;
  - Exemplo:

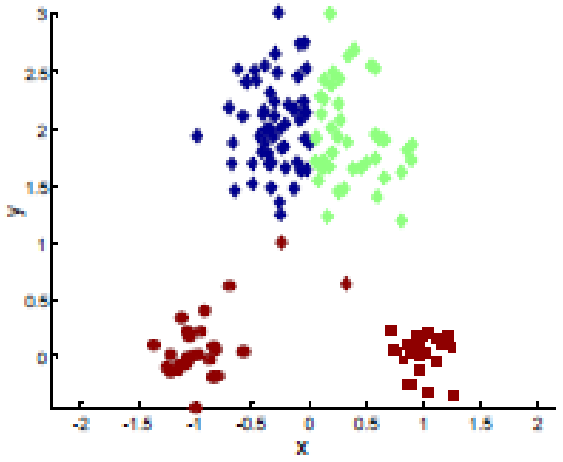


  - Como evitar... ?

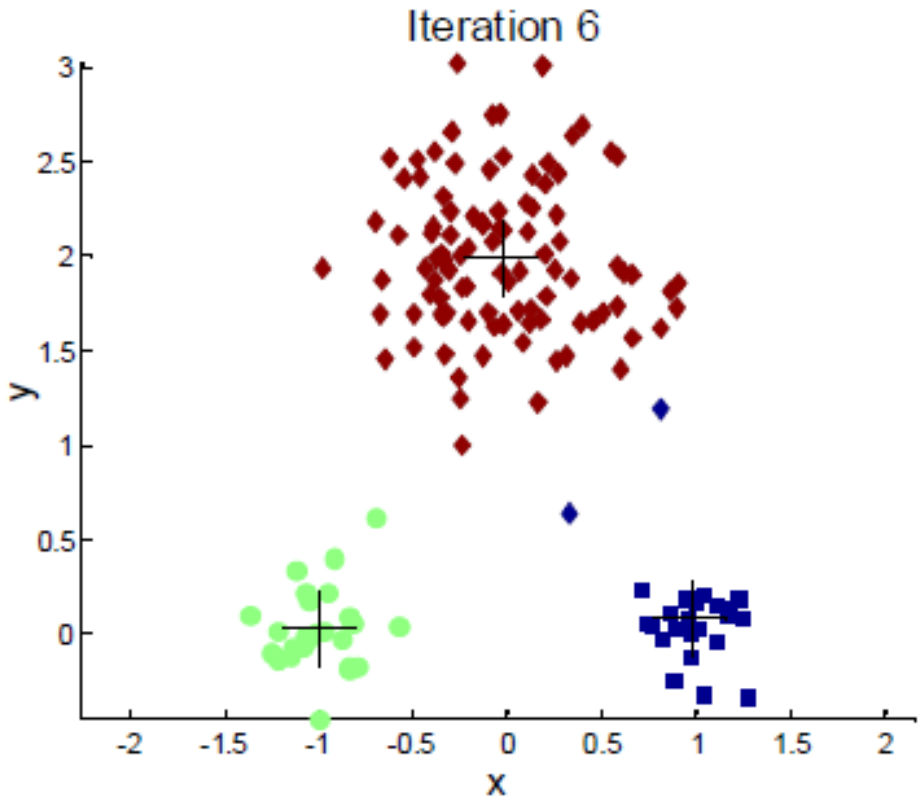
Two different K-means Clusterings



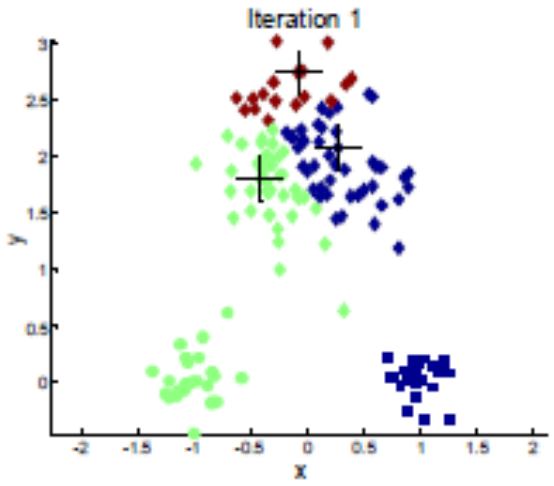


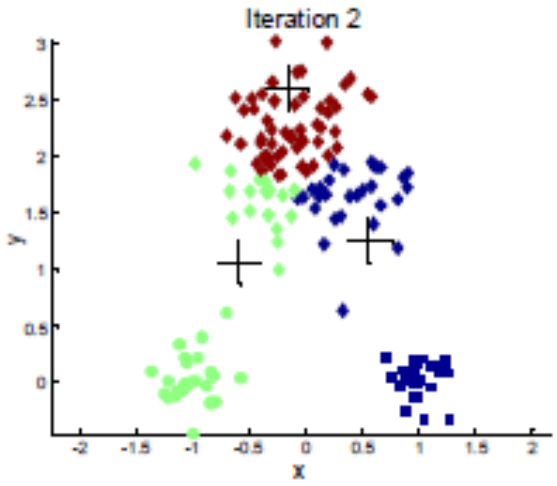


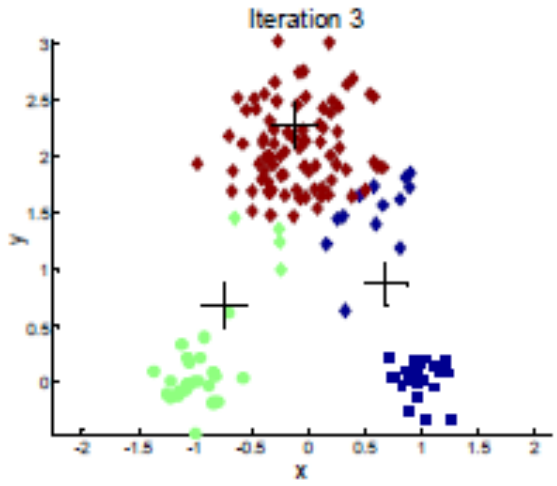
Importance of Choosing Initial Centroids

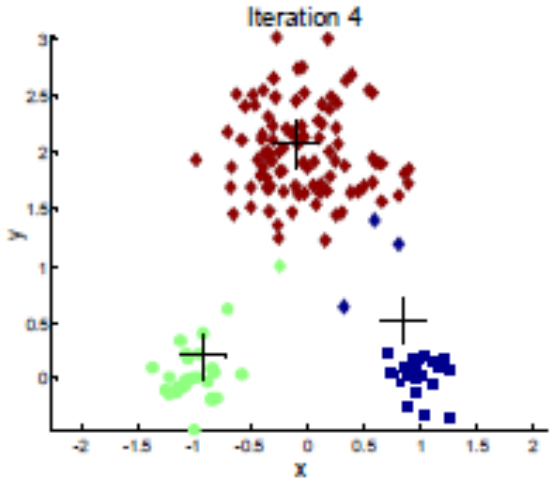


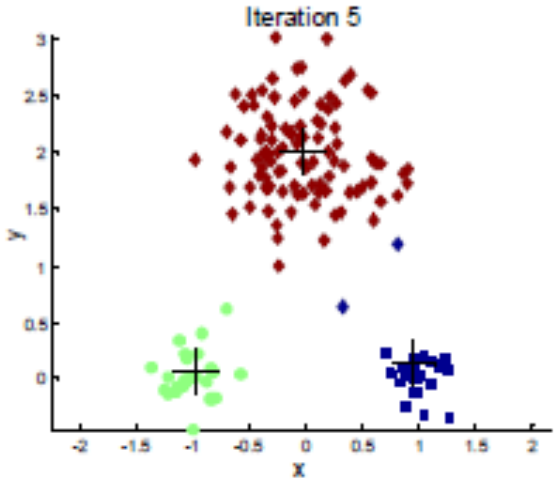
Importance of Choosing Initial Centroids

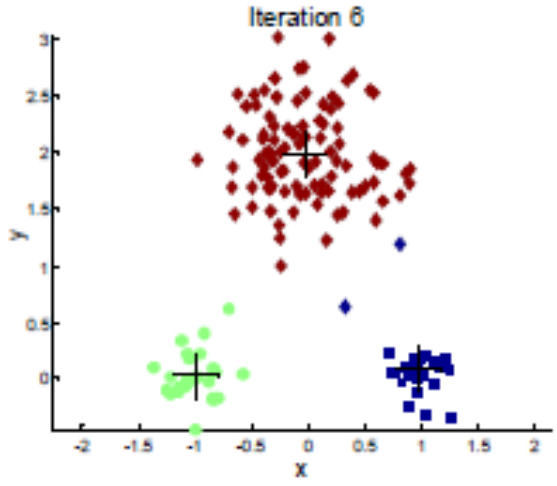




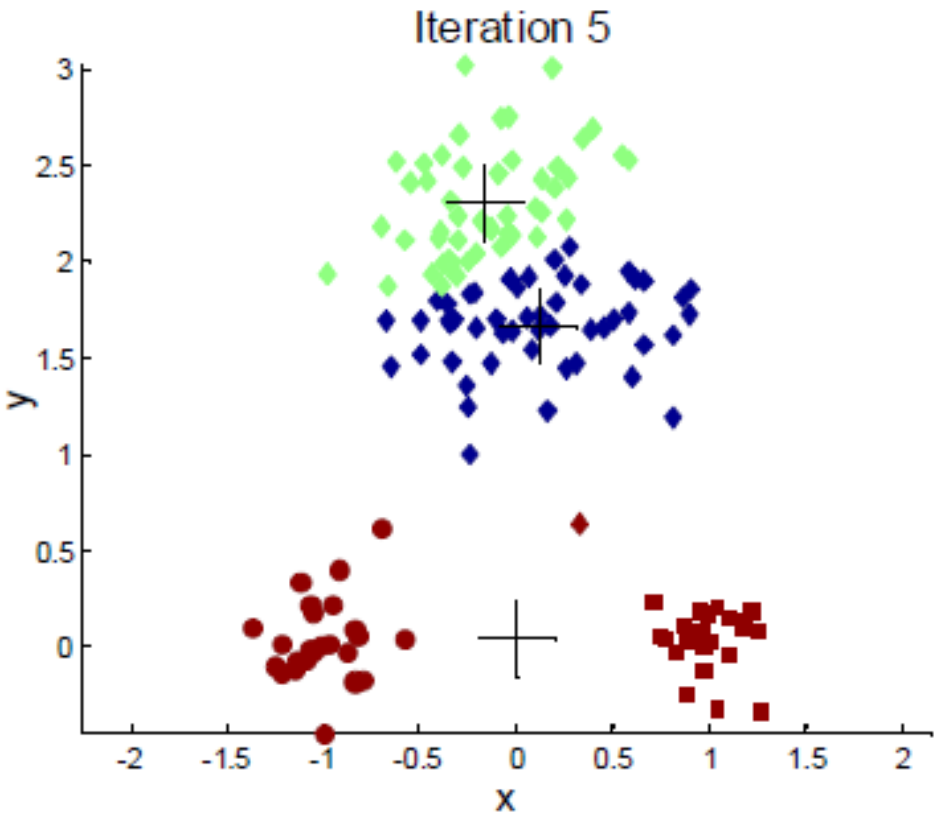




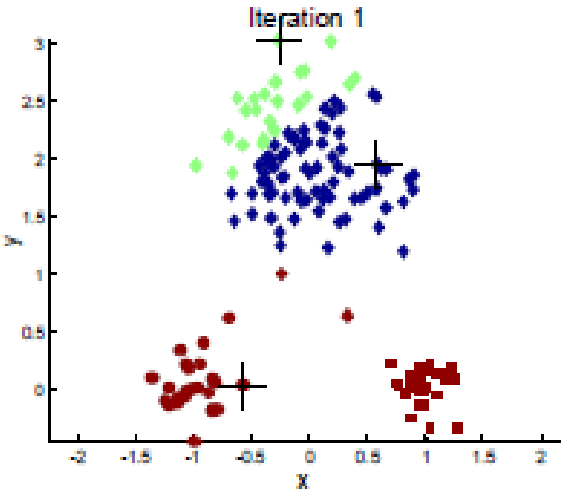


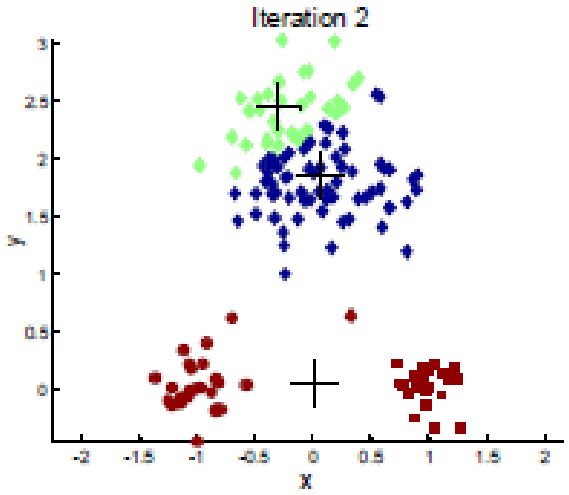


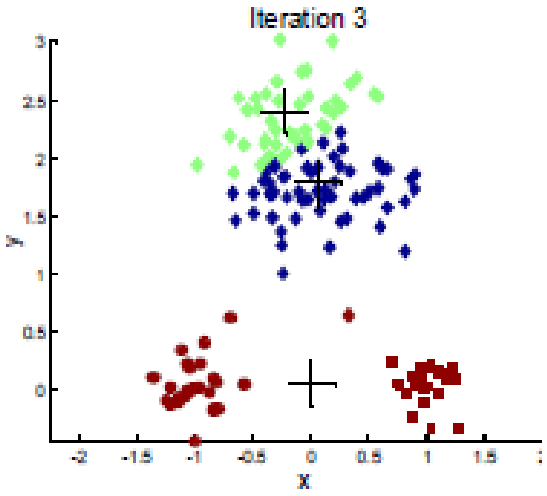
Importance of Choosing Initial Centroids …

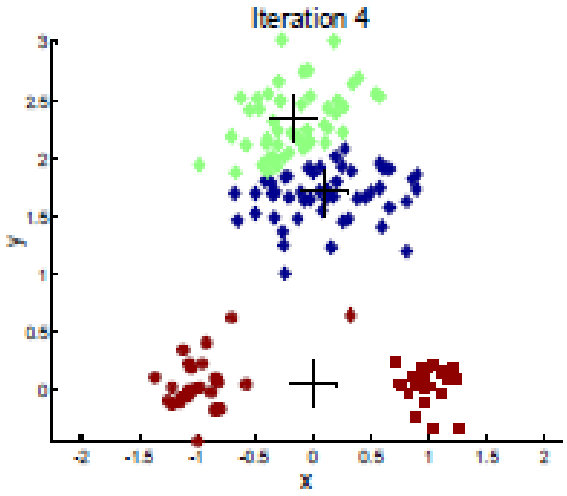


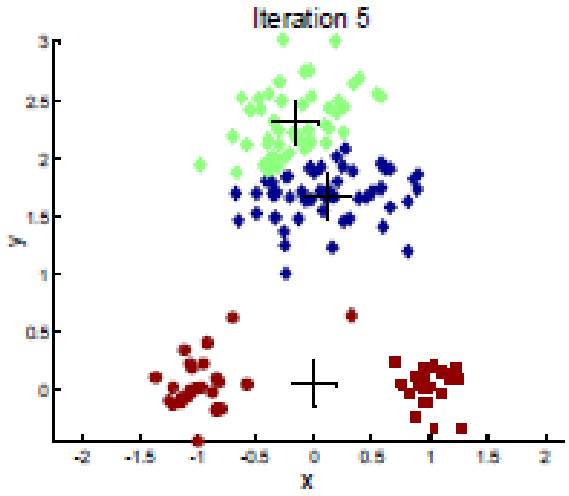
Importance of Choosing Initial Centroids …







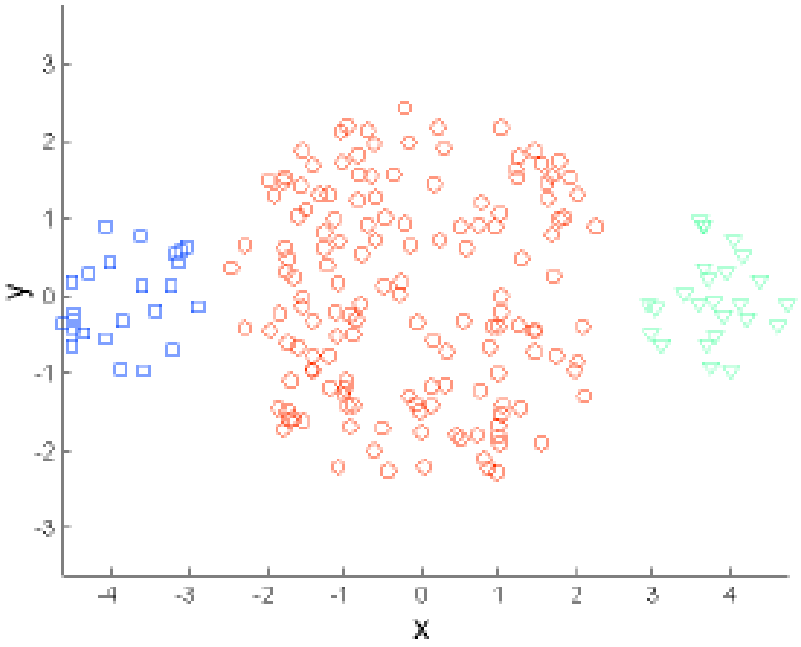




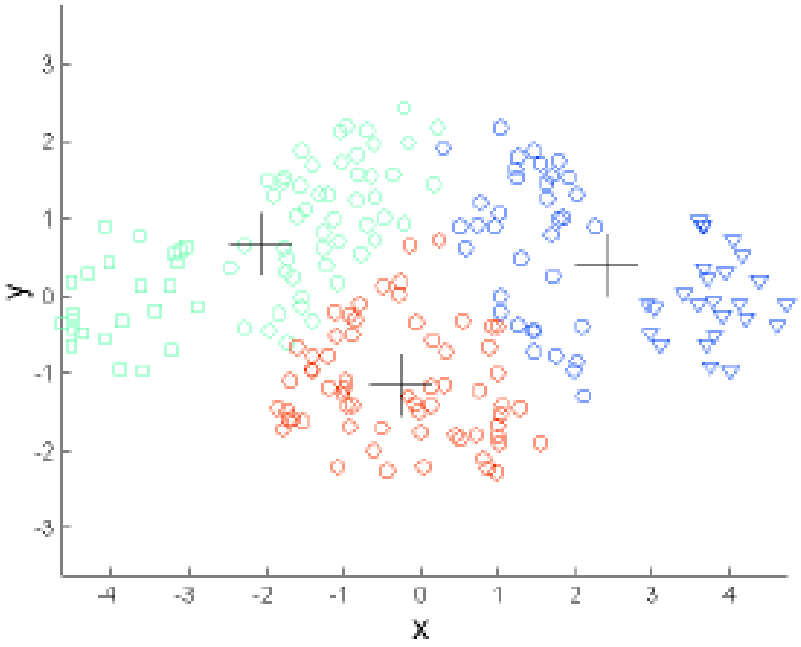
Discussão

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

Differing Sizes

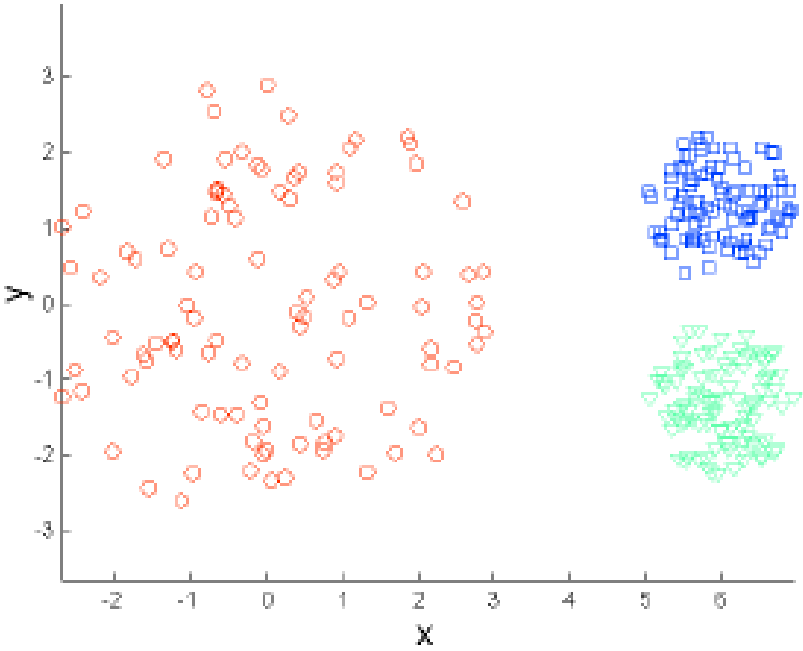


Original Points

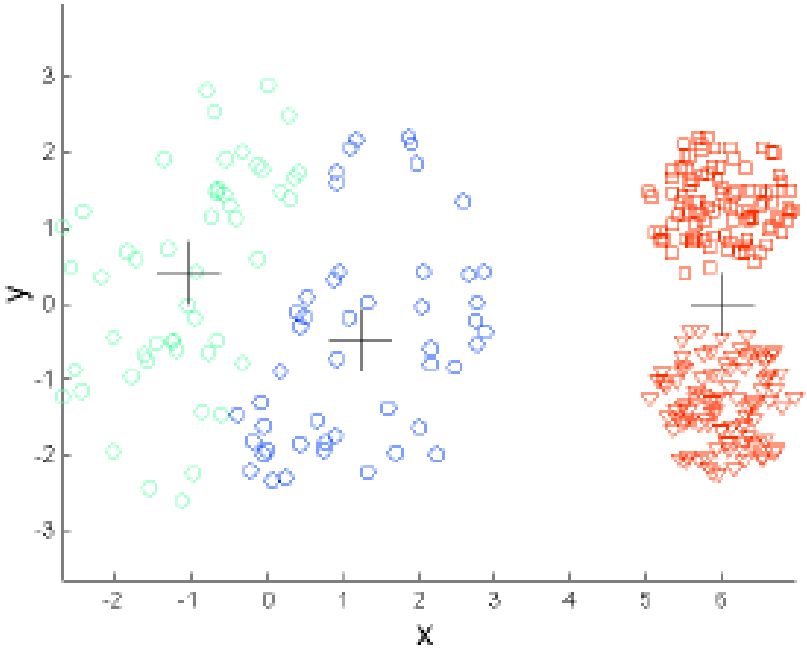


K-means (3 Clusters)

Differing Density



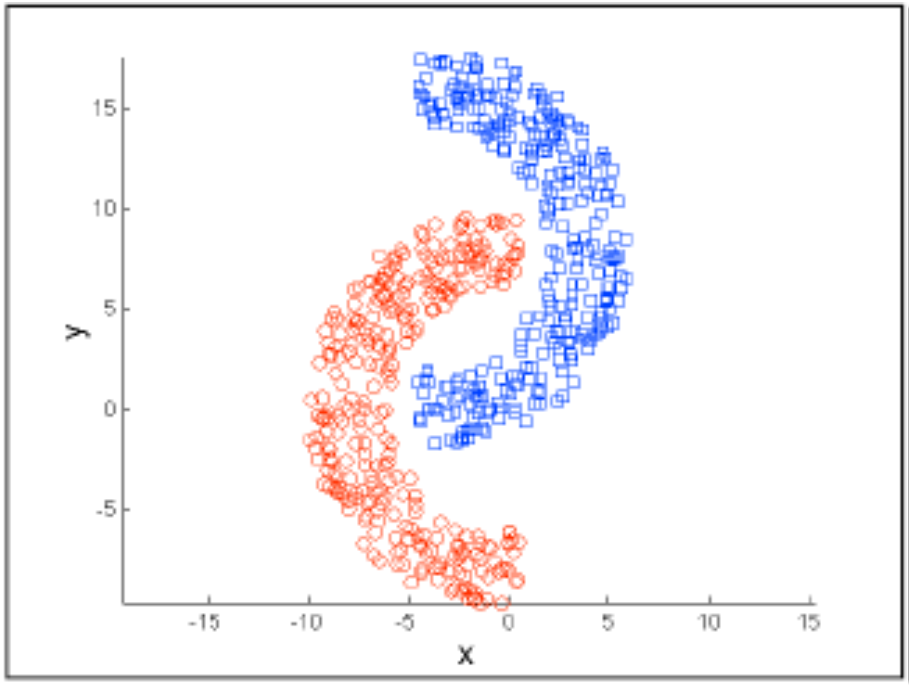
Original Points

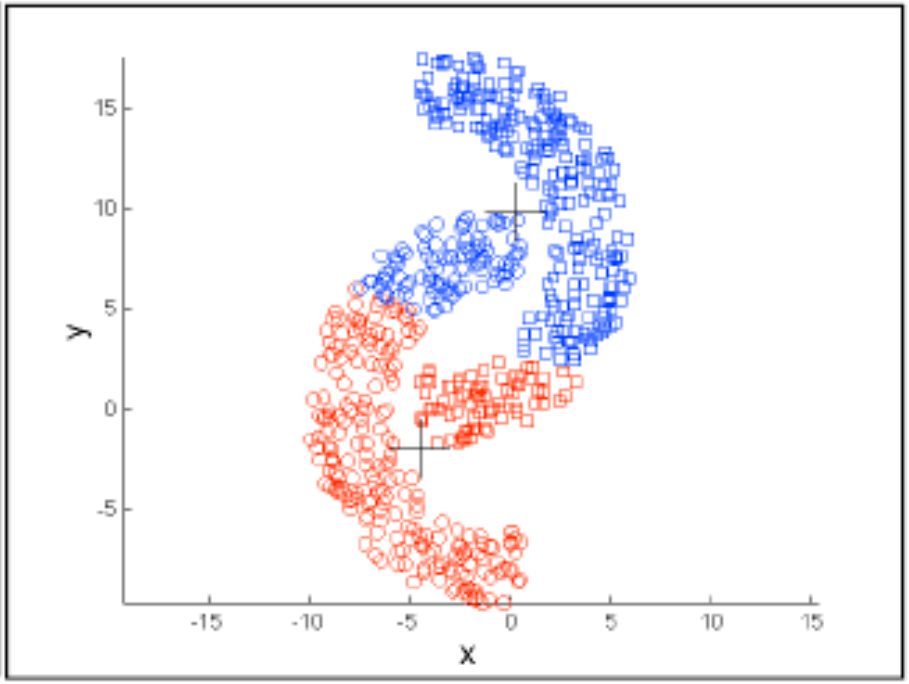


K-means (3 Clusters)

Formas Não-Globulares

- Ao contrário do que muitos pensam, o “problema” abaixo usualmente é de pouco interesse em data mining real
  - Grandes BDs (muitos objetos & atributos) e necessidade de interpretação dos resultados (e.g. segmentação de mercado...)

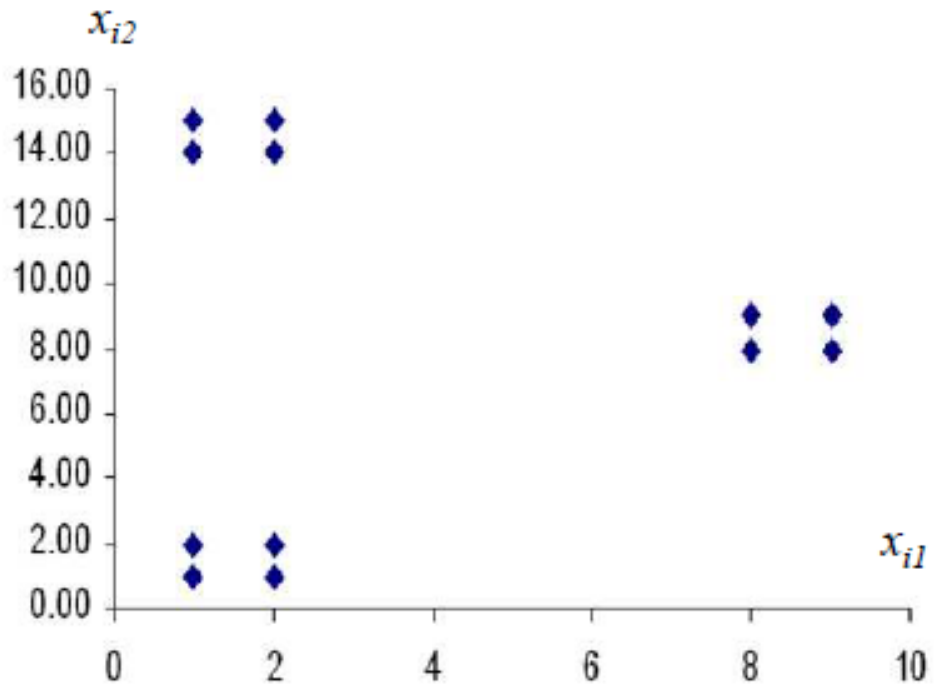




Exercício

- Executar k-means com k=3 nos dados acima a partir dos protótipos [6 6], [4 6] e [5 10] e outros a sua escolha

Objeto $\mathbf{x}_i$	x_{i1}	x_{i2}
1	1	2
2	2	1
3	1	1
4	2	2
5	**8**	**9**
6	**9**	**8**
7	**9**	**9**
8	**8**	**8**
9	1	15
10	2	15
11	1	14
12	2	14



Implementações Eficientes

- Desempenho pode ser “turbinado” de diferentes formas:
  - Heurísticas de Inicialização (e.g. PAM)
  - Estruturas de Dados (e.g. kd-trees)
  - Algoritmos, e.g.:
    - Atualização incremental dos centroides:
      - Cálculo de cada centroide só depende do número de objetos do cluster em questão, dos novos objetos atribuídos ao cluster, dos objetos que deixaram o cluster, e do valor anterior do centroide
      - Não demanda recalcular tudo novamente
      - Exercício: a partir da equação do cálculo do centroide, escrever a equação de atualização incremental descrita acima!

Resumo do k-means

- **Vantagens**
  - Simples e intuitivo
  - Possui complexidade computacional linear em todas as variáveis críticas (N, n, k)
  - Eficaz em muitos cenários de aplicação e produz resultados de interpretação relativamente 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 volumétricos / 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**: complexidade computacional do cálculo da mediana (ordenação) é maior que aquela do cálculo da média

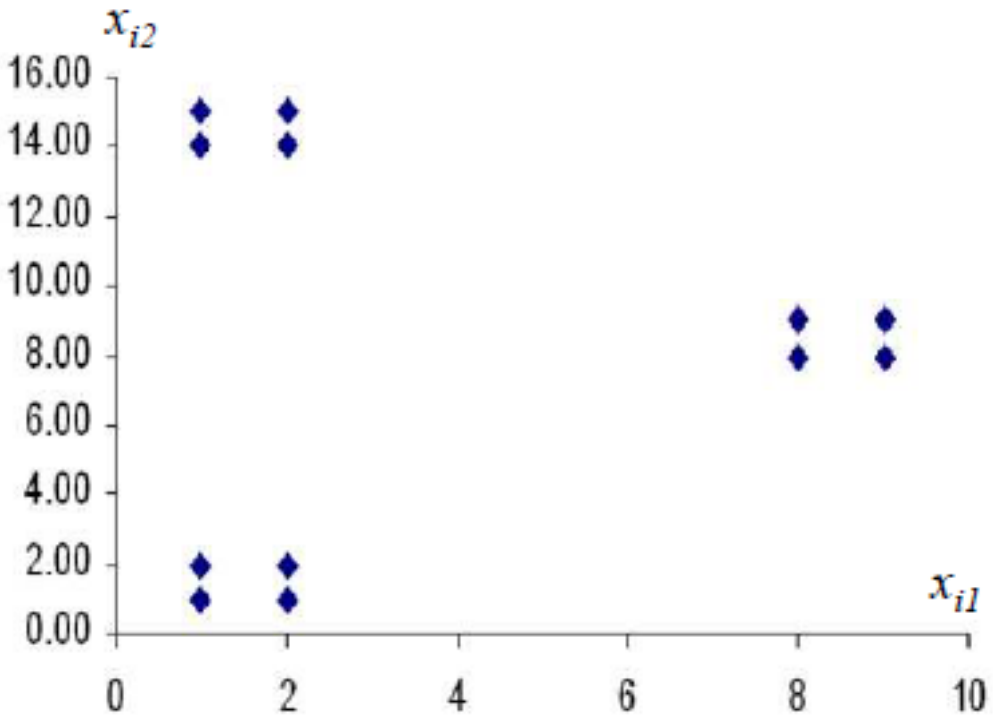
Algumas Variantes do k-means

- **K-medóides**: Substituir cada centroide 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**:
    - assim como o k-medianas, também é menos sensível a outliers
    - permite cálculo relacional
    - logo, pode ser aplicado a bases com atributos categóricos
  - **Desvantagem**: Complexidade quadrática com o no. de objetos N

Exercício

- Executar k-medóides com k=3 nos dados acima, com medóides iniciais dados pelos objetos 5, 6 e 8

Objeto x_i	x_{i1}	x_{i2}
1	1	2
2	2	1
3	1	1
4	2	2
5	**8**	**9**
6	**9**	**8**
7	**9**	**9**
8	**8**	**8**
9	1	15
10	2	15
11	1	14
12	2	14



Algumas Variantes do k-means

- K-means para Data Streams: usa o conceito de vizinhos mais próximos (K-NN)
  - Objetos são dinamicamente incorporados ao cluster mais próximo
  - Atualização do centroide do cluster pode ser incremental
  - Heurísticas podem ser usadas para criação ou remoção de clusters

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 entre $[k_{min},k_{max}]$
    - Aleatório: n_k 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

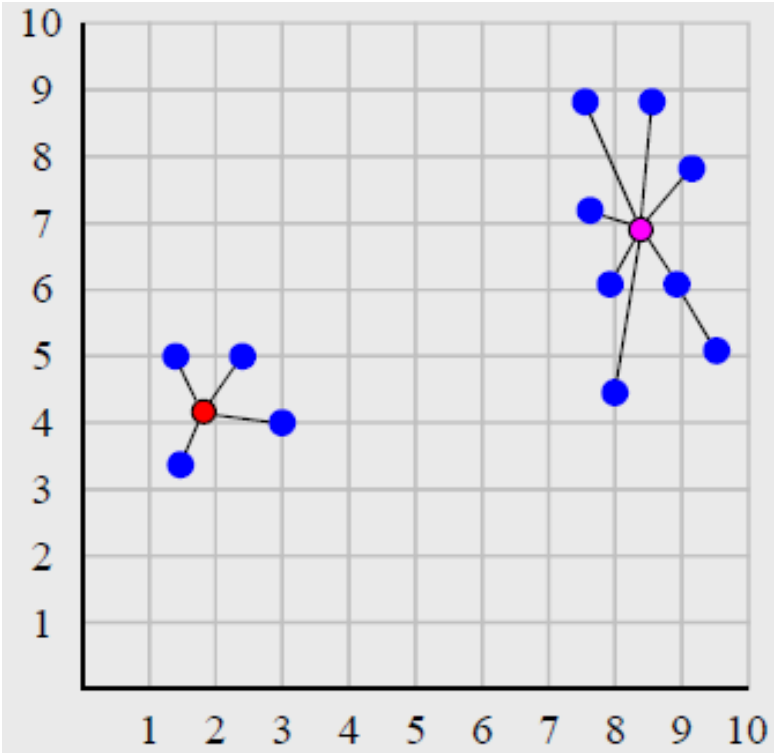
Questão...

- A própria função objetivo J do k-means não poderia ser utilizada como medida de qualidade 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, portanto variável...?
- Para responder, considere que as partições são geradas:
  - Por múltiplas execuções de k-means com k entre [k_{min}, k_{max}], ou
  - por outra variante que também estime o valor de k
    - X-means, k-means evolutivo,...

Questão...

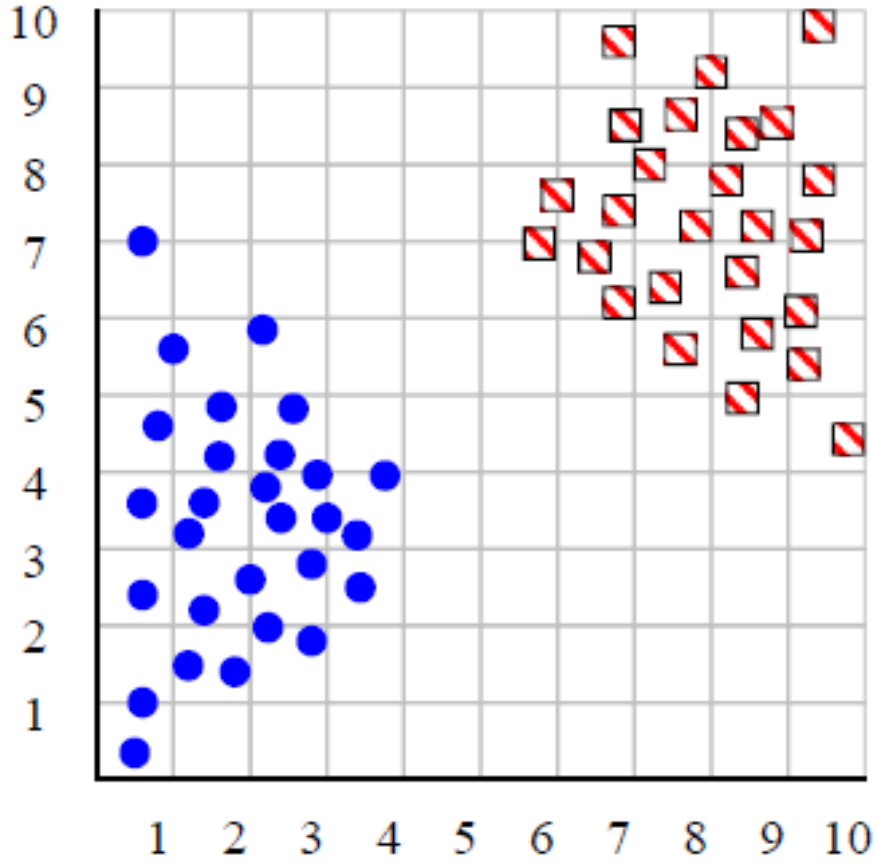
- Para tentar responder a questão anterior, vamos considerar o método de múltiplas execuções ordenadas de k-means, com uso da função objetivo J
  - Erro Quadrático:

$$J = \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in \mathbf{C}_i} d\left(\mathbf{x}_j, \overline{\mathbf{x}}_i\right)^2$$

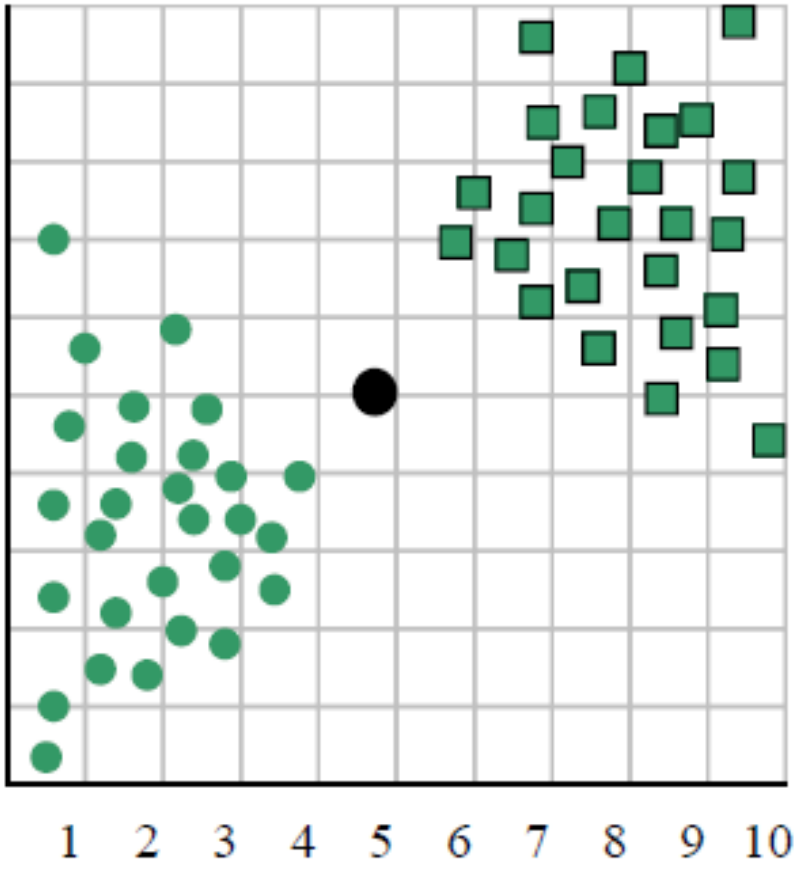


Questão...

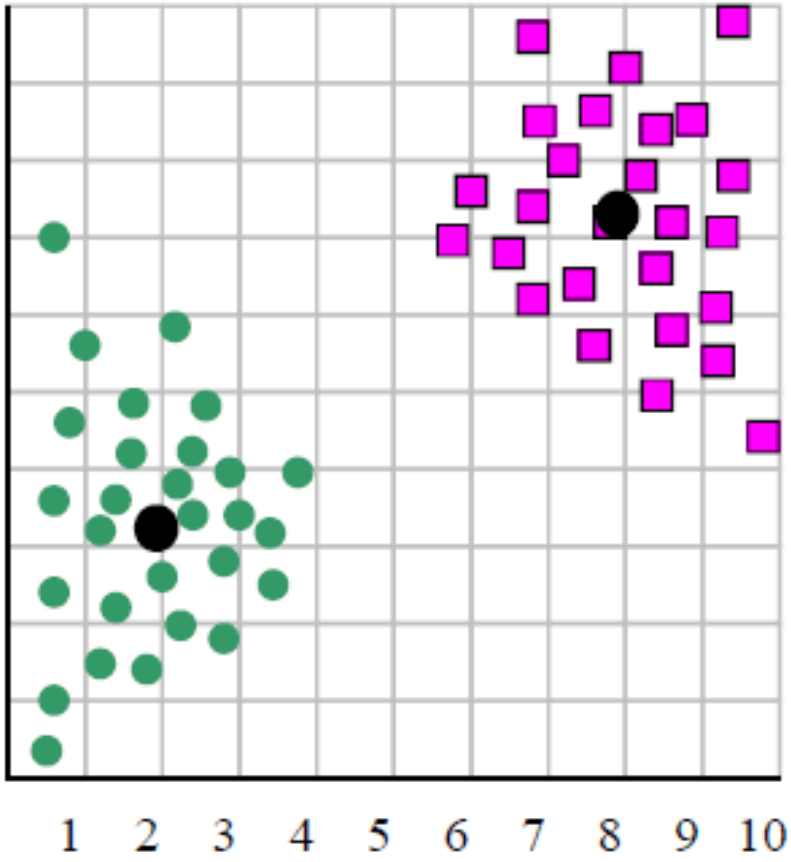
- Considere o seguinte exemplo:



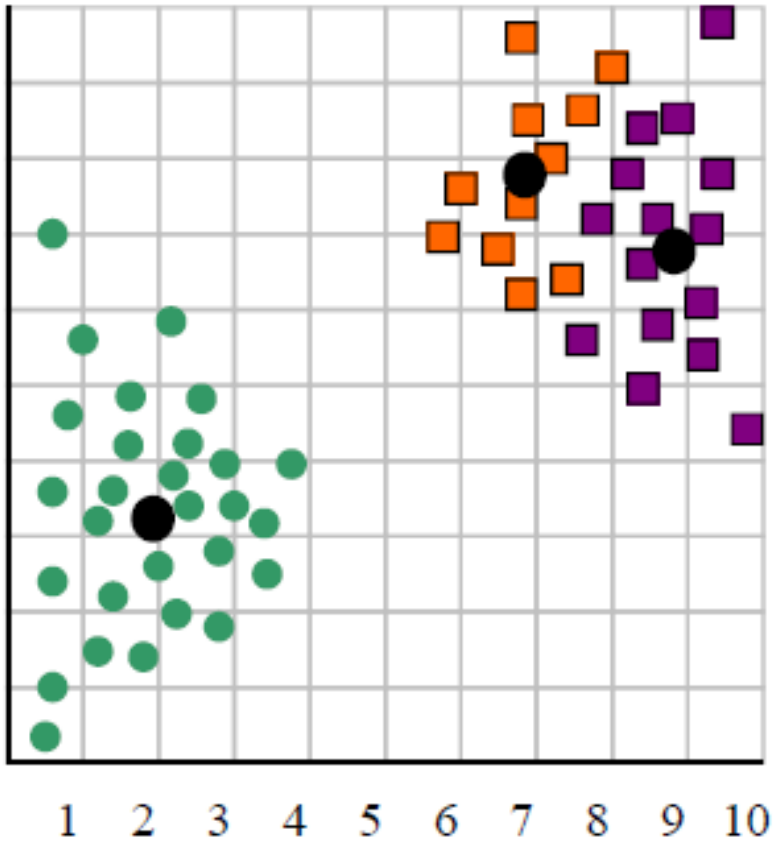
- Para k = 1, o valor da função objetivo é 873,0.



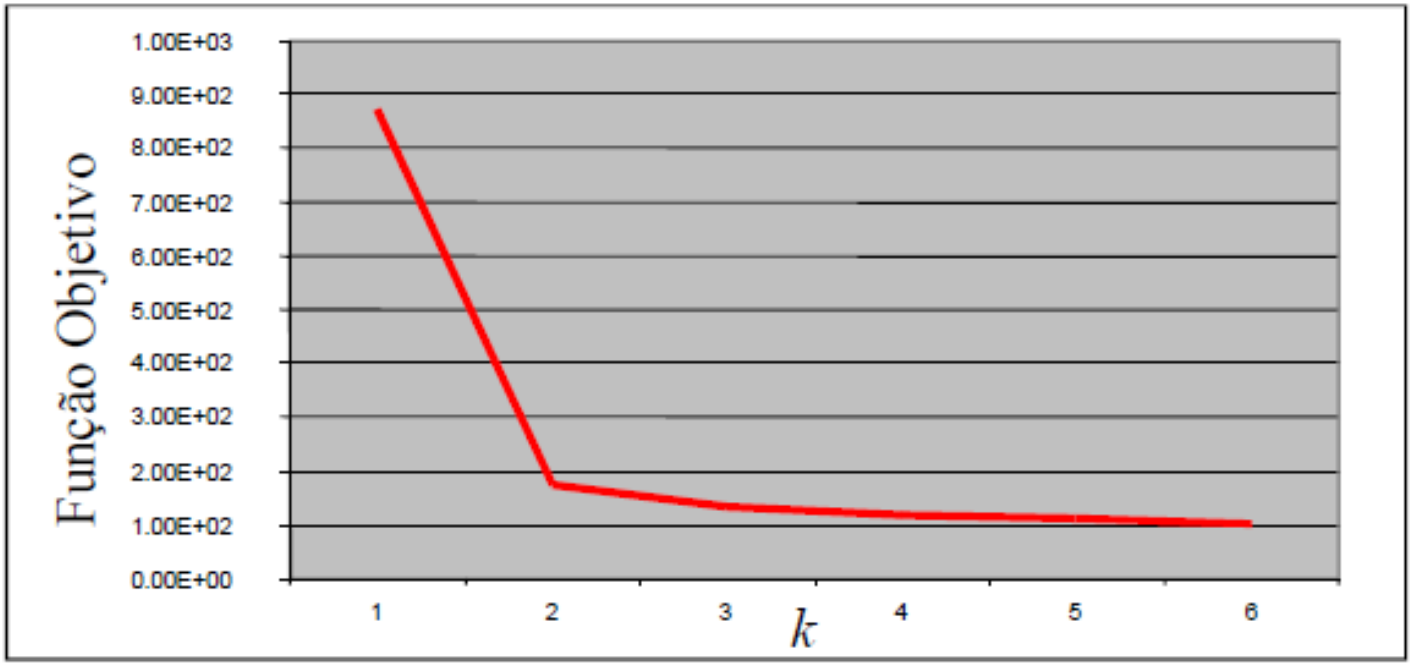
- Para k = 2, o valor da função objetivo é 173,1.



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



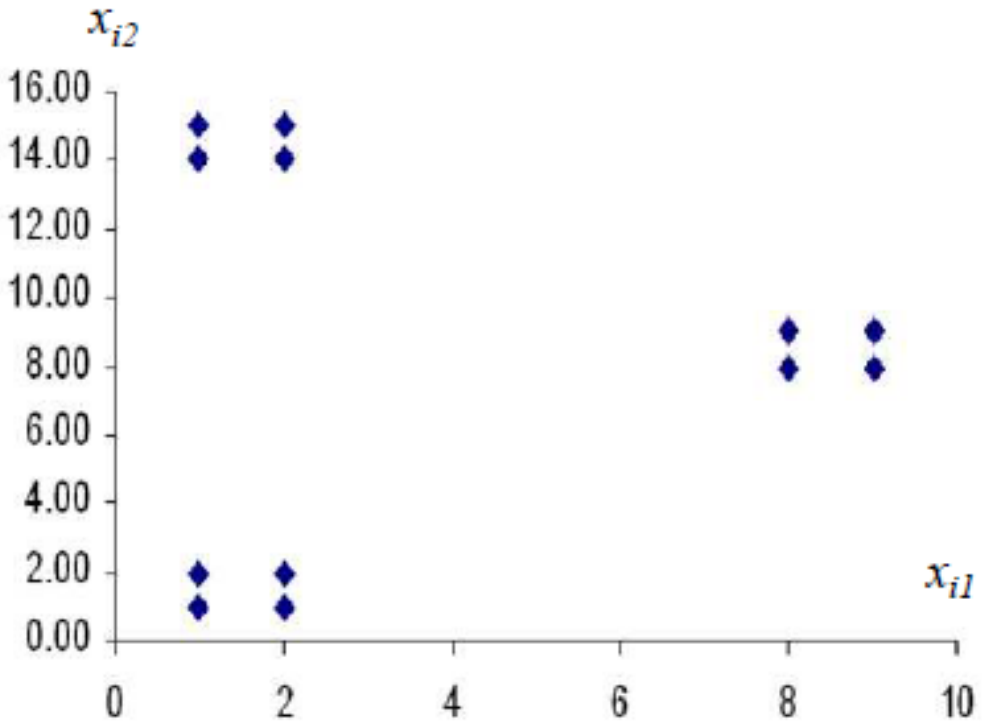
- Podemos então repetir este procedimento e plotar os valores da função objetivo J para k=1,…,6, ... e tentar identificar um “joelho” :



Exercício

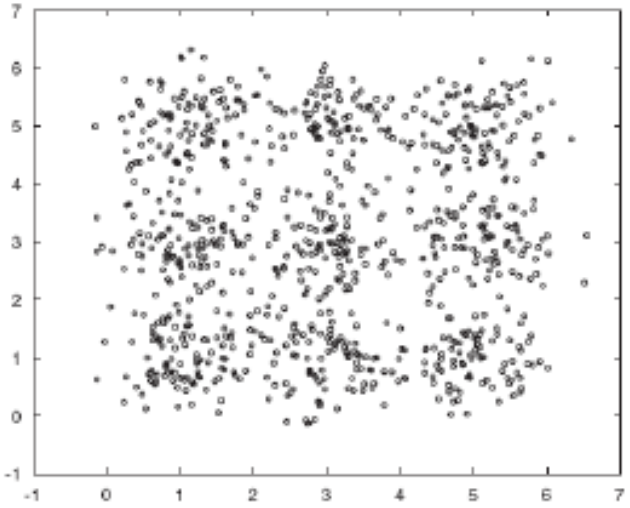
- Executar k-means com k=2 até k=5 nos dados acima e representar graficamente a função objetivo J em função de k

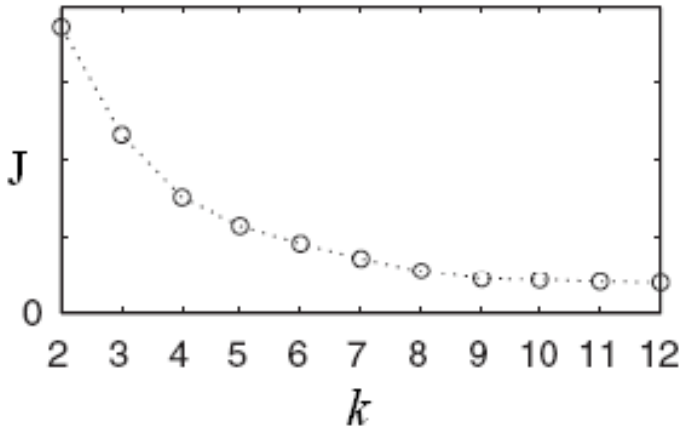
Objeto x_i	x_{i1}	x_{i2}
1	1	2
2	2	1
3	1	1
4	2	2
5	**8**	**9**
6	**9**	**8**
7	**9**	**9**
8	**8**	**8**
9	1	15
10	2	15
11	1	14
12	2	14



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,...?

Critérios de Validade Relativos

- Para avaliar relativamente a qualidade de diferentes partições, possivelmente com números distintos de grupos, faz-se necessário um tipo de índice:
  - Critério Relativo de Validade de Agrupamento
- Existem dezenas de tais critérios na literatura
- Estudos apontam alguns deles como superiores em algumas classes de problemas comuns na prática
- Para problemas em geral, no entanto, não há qualquer garantia que um dado critério será o mais apropriado
  - No free lunch !!!

Comment on Cluster Validity

- “The validation of clustering structures is the most difficult and frustrating part of cluster analysis.

- Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage.”
  - Jain and Dubes, Algorithms for Clustering Data, 1988

Critério da Largura de Silhueta

- SWC = Silhueta média sobre todos os objetos:

$$SWC = \frac{1}{N} \sum_{i-1}^{N} s(i)$$

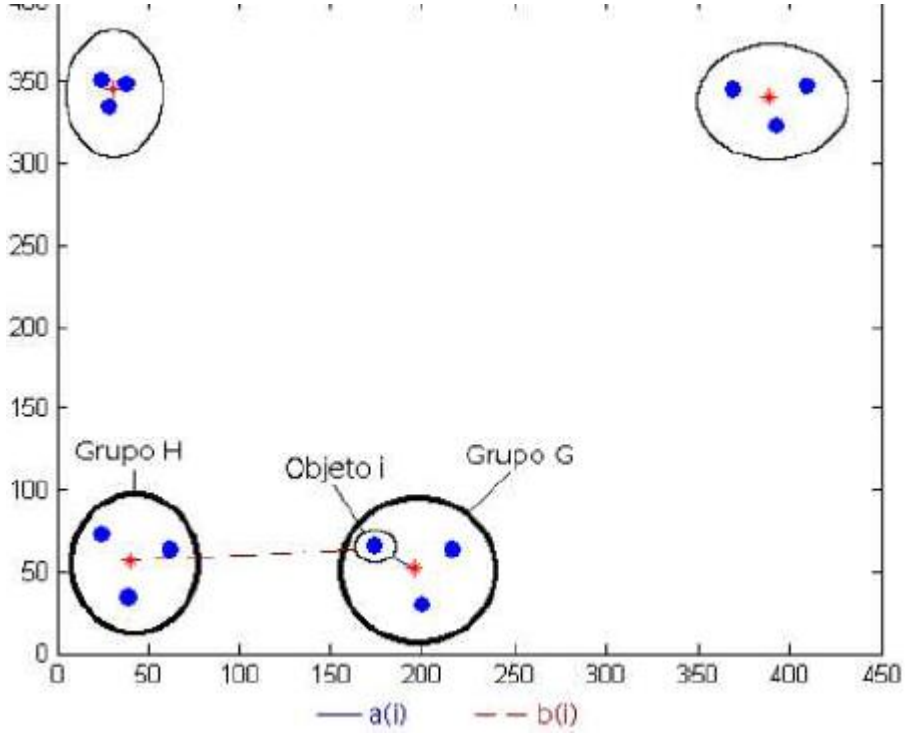
- Silhueta (i-ésimo objeto):
  - a(i): dissimilaridade média do i-ésimo objeto ao seu cluster
  - b(i): dissimilaridade média do i-ésimo objeto ao cluster vizinho mais próximo

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

- Silhueta Original: a(i) e b(i) são calculados como a distância média (Euclidiana, Mahalanobis, etc) do i-ésimo objeto a todos os demais objetos do cluster em questão. Complexidade O(N2)
  - SWC ϵ [–1,+1] ; s(i) := 0 para singletons

Critério da Largura de Silhueta

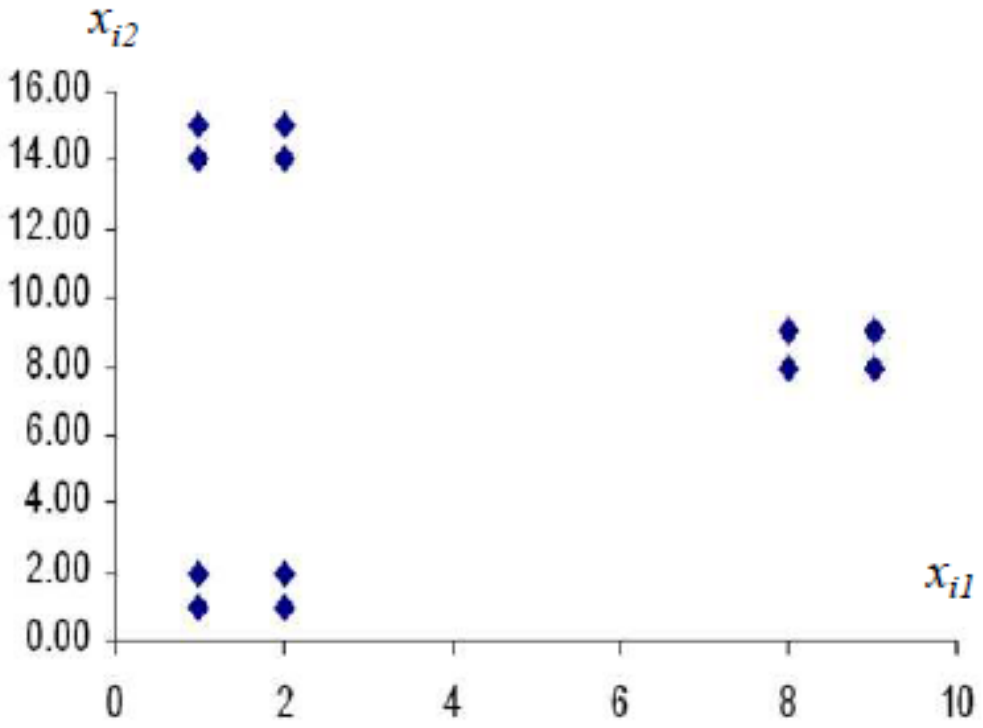
- Silhueta Simplificada
  - a(i) e b(i) são calculados como a distância do i-ésimo objeto ao centroide do cluster em questão. Complexidade O(N).



Exercício

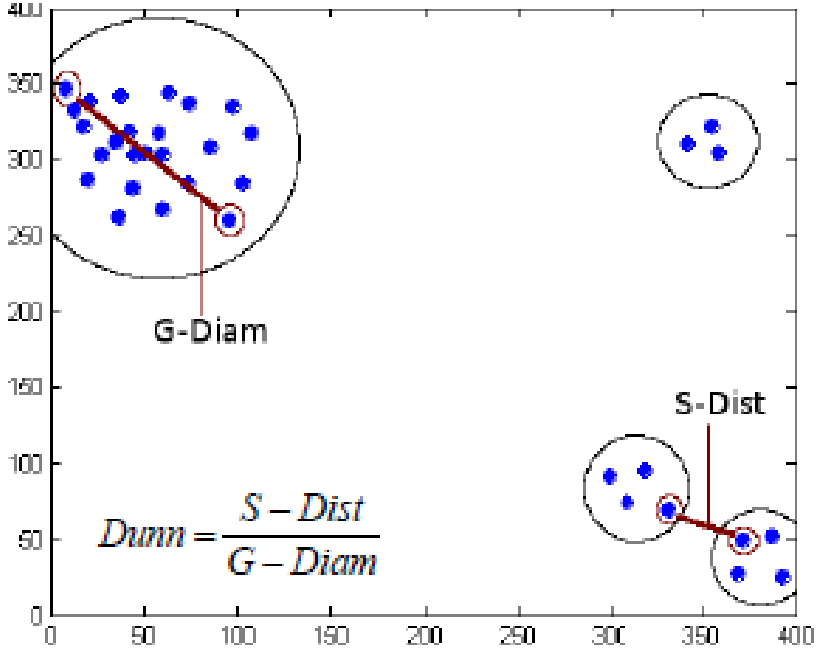
- Executar k-means nos dados acima e calcular as silhuetas (original e simplificada) das partições obtidas para vários k

Objeto $\mathbf{x}_i$	x_{i1}	x_{i2}
1	1	2
2	2	1
3	1	1
4	2	2
5	**8**	**9**
6	**9**	**8**
7	**9**	**9**
8	**8**	**8**
9	1	15
10	2	15
11	1	14
12	2	14



Muitos Outros Critérios...

- Variance Ratio Criterion (VRC)
  - também denominado Calinski-Harabaz
- Davies-Bouldin
- Índice de Dunn e Variantes
  - e muito mais...



Critérios de Validade Externos

- Embora o problema de clustering seja não supervisionado, em alguns cenários o resultado de agrupamento desejado pode ser conhecido. Por exemplo:
  - Reconhecimento visual dos clusters naturais (bases 2D, 3D)
  - Especialista de domínio
  - Bases geradas sinteticamente com distribuições conhecidas
    - Benchmark data sets
  - Bases de classificação sob a hipótese que classes são clusters

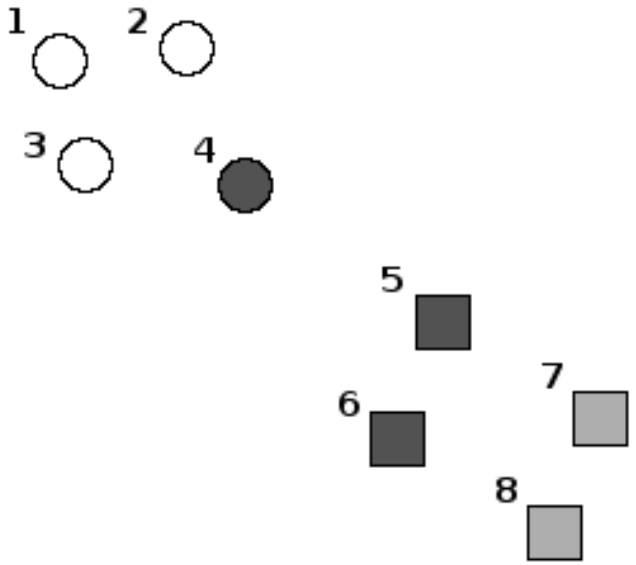
- Índices que medem o nível de compatibilidade entre uma partição obtida e uma partição de referência dos mesmos dados são denominados **critérios de validade externos**

Exemplo de Critério Externo de Validade de Agrupamento

- **Jaccard:**

$$J = \frac{a}{a+b+c}$$

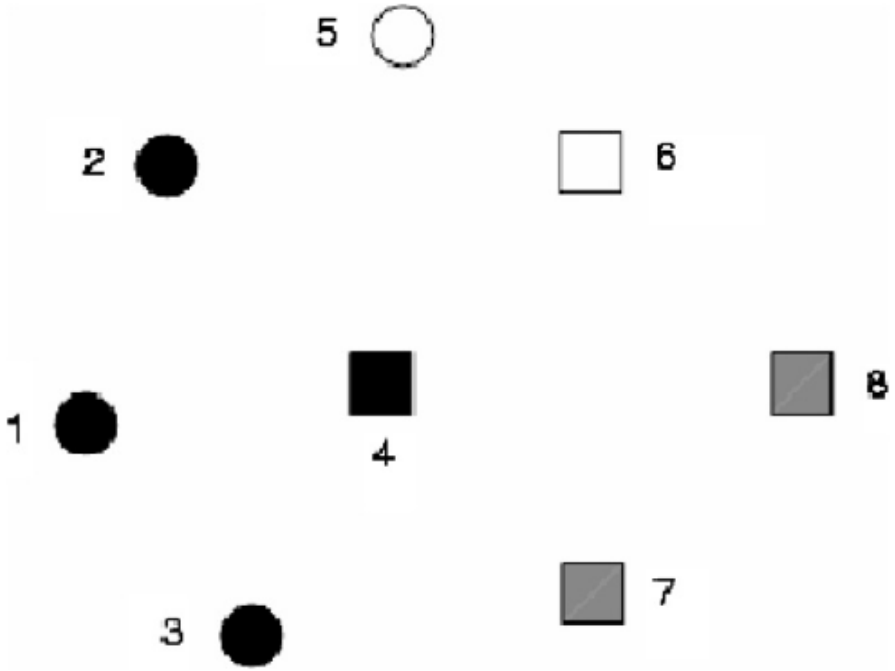
  - a: No. de pares que pertencem à mesma classe e ao mesmo cluster
  - b: No. de pares que pertencem à mesma classe e a clusters distintos
  - c: No. de pares que pertencem a classes distintas e ao mesmo cluster



- 2 Classes (Círculos e Quadrados)
- 3 Clusters (Preto, Branco e Cinza)
  - a = 5; b = 7; c = 2
  - J = 5/(5+7+2) = 0.3571

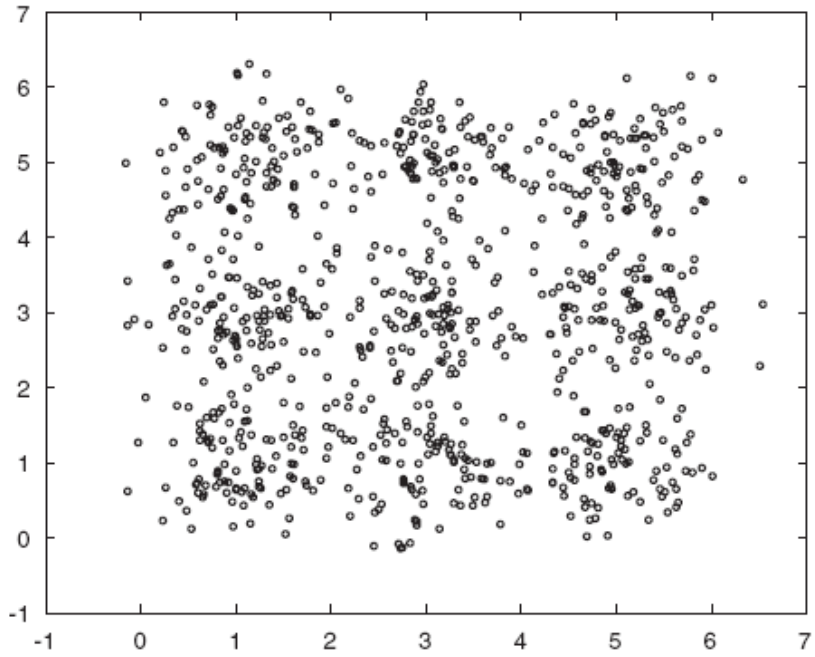
Exercício

- Calcule o valor do critério de Jaccard entre as duas partições (dos mesmos 8 objetos) ilustradas na figura abaixo. Uma das partições é representada por 3 cores, enquanto a outra é representada por duas formas geométricas.



Métodos de Partição (Com Sobreposição)

- Como uma estratégia do tipo partição sem sobreposição, k-means produz uma partição rígida da base de dados:
  - cada objeto pertence ou não a um determinado grupo
  - Usualmente refere-se a esse tipo de partição como Hard ou Crisp
- No entanto, muitos problemas envolvem grupos mal delineados, que não podem ser separados adequadamente dessa maneira
- Em outras palavras, existem situações nas quais os dados compreendem categorias que se sobrepõem umas às outras em diferentes níveis. Por ex.:



Métodos de Partição (Com Sobreposição)

- Métodos de agrupamento com sobreposição, denominados overlapping clustering algorithms em inglês, são concebidos para lidar com situações como esta. Geram partições de 3 tipos:
  - Soft: Objetos podem pertencer (de forma integral) a mais de um grupo
  - Fuzzy: Objetos pertencem a todos os grupos com diferentes graus de pertinência (possivelmente nulo)
  - Probabilísticas: Objetos possuem probabilidades de pertinência associadas a cada cluster

- Vamos discutir brevemente os representantes mais clássicos dos últimos dois tipos, que são os mais comumente utilizados

Agrupamento Fuzzy

- Fuzzy c-Means (FCM):

$$\min_{f_{ij},\,\mathbf{v}_i} J = \sum_{j=1}^{N}\sum_{i=1}^{c} f_{ij}^{m}\left\|\mathbf{x}_j - \mathbf{v}_i\right\|^2$$

$$s.a \quad 0 \le f_{ij} \le 1$$

$$\sum_{i=1}^{c} f_{ij} = 1 \qquad \forall j \in \{1, 2, \ldots, N\}$$

$$0 < \sum_{j=1}^{N} f_{ij} < N \quad \forall i \in \{1, 2, \ldots, c\}$$

$$\begin{cases} f_{ij} \Rightarrow \text{Pertinência do objeto } j \text{ ao grupo } i. \\ \mathbf{v}_i \in \Re^n \Rightarrow \text{Centróide do } i\text{-ésimo grupo.} \\ m > 1 \end{cases}$$

Fuzzy c-Means

- NOTAS:
  - O critério de custo J força que os dados mais próximos aos protótipos estejam associados (sejam multiplicados) às funções de pertinência maiores e vice-versa, de modo que os produtos sejam minimizados.
  - As restrições limitam os valores de pertinência ao intervalo unitário, evitam que esses sejam todos nulos ou unitários e forçam a consistência (“probabilística”) com relação à pertinência a grupos distintos.
  - Para valores maiores de m tem-se que os valores de pertinência mais próximos a 1, associados aos dados mais próximos aos protótipos, passam a ter maior importância no critério de custo.

Fuzzy c-Means

- Algoritmo FCM:

1 – *Selecionar os centros iniciais* $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_c$.

2 – *Calcular* f_{ij} : *Para cada* $j \in \{1, \cdots, N\}$, *se* $\|\mathbf{x}_j - \mathbf{v}_i\|^2 > 0$ *para* $i = 1, \cdots, c$ *então* :

$$f_{ij} = \left[\sum_{l=1}^{c} \left(\frac{\|\mathbf{x}_j - \mathbf{v}_i\|^2}{\|\mathbf{x}_j - \mathbf{v}_l\|^2} \right)^{\frac{1}{m-1}} \right]^{-1}$$

Se $\|\mathbf{x}_j - \mathbf{v}_i\|^2 = 0$ *para* $i \in I \subseteq \{1, \cdots, c\}$, *então definir* f_{ij} *para* $i \in I$ *como qquer nro real não negativo que satisfaça* $\sum_{i \in I} f_{ij} = 1$ *e definir* $f_{ij} = 0$ *para* $i \in \{1, \cdots, c\} - I$.

3 – *Atualizar os centros* :

$$\mathbf{v}_i = \frac{\sum_{j=1}^{N} f_{ij}^m \mathbf{x}_j}{\sum_{j=1}^{N} f_{ij}^m}$$

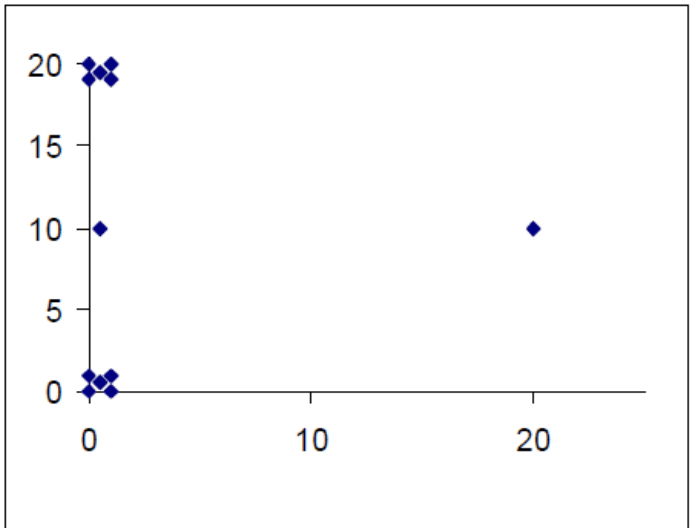
4 – *Parar em caso de convergência ou voltar ao passo* 2.

Fuzzy c-Means

- Valor ótimo de m desconhecido. Usualmente m = 2

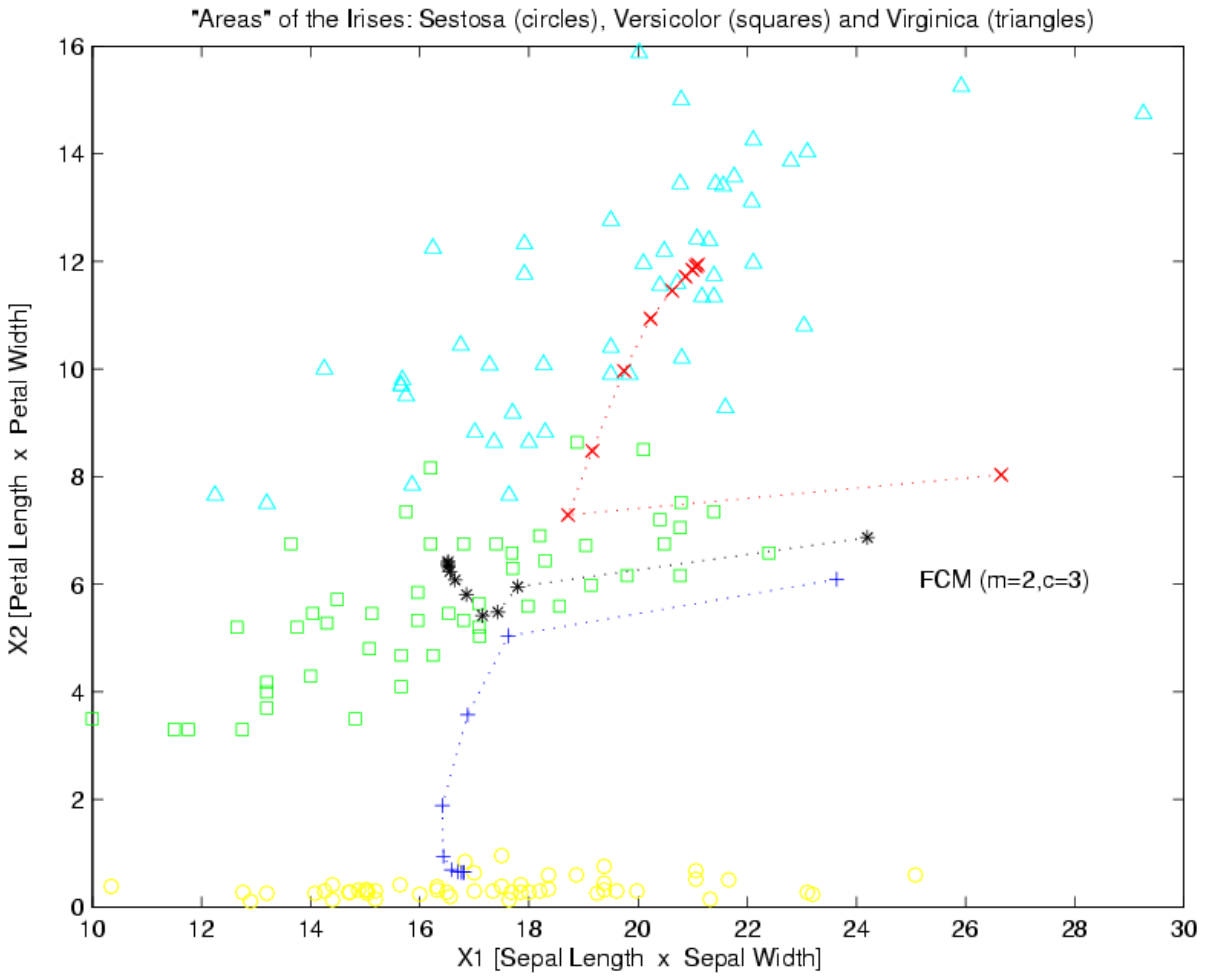
- Trata-se de uma extensão de k-means para o domínio fuzzy
  - Como tal ⇒ apenas garantia de convergência para soluções locais !
  - Ou seja, também é susceptível a mínimos locais da função objetivo J
    - depende da inicialização dos protótipos
    - esquemas de múltiplas execuções podem ser utilizados...

- Existem dezenas de variantes!
  - e.g. versão possibilística
    - considere a figura ao lado com c = 2
    - pertinência dos dois outliers...?

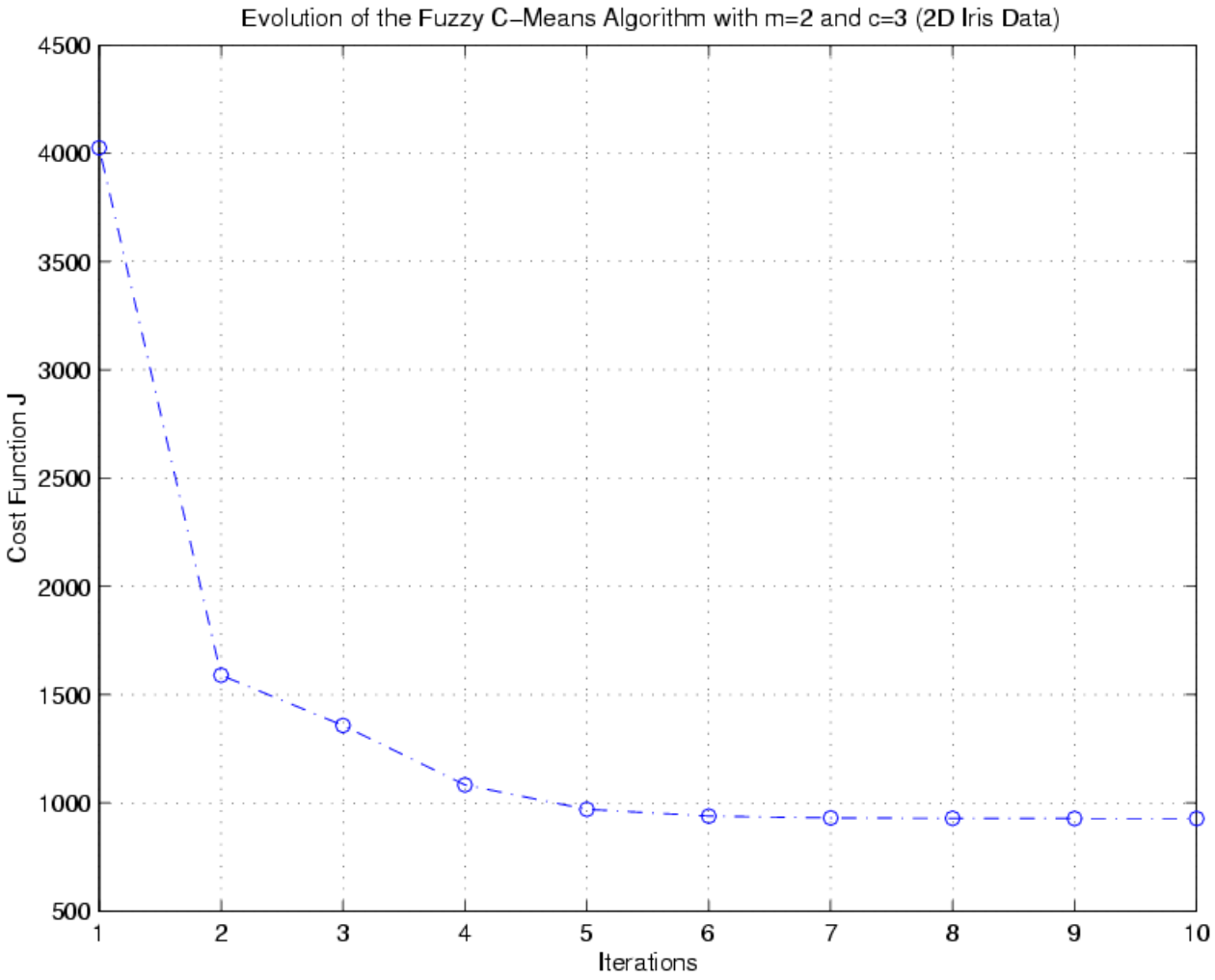


Fuzzy c-Means

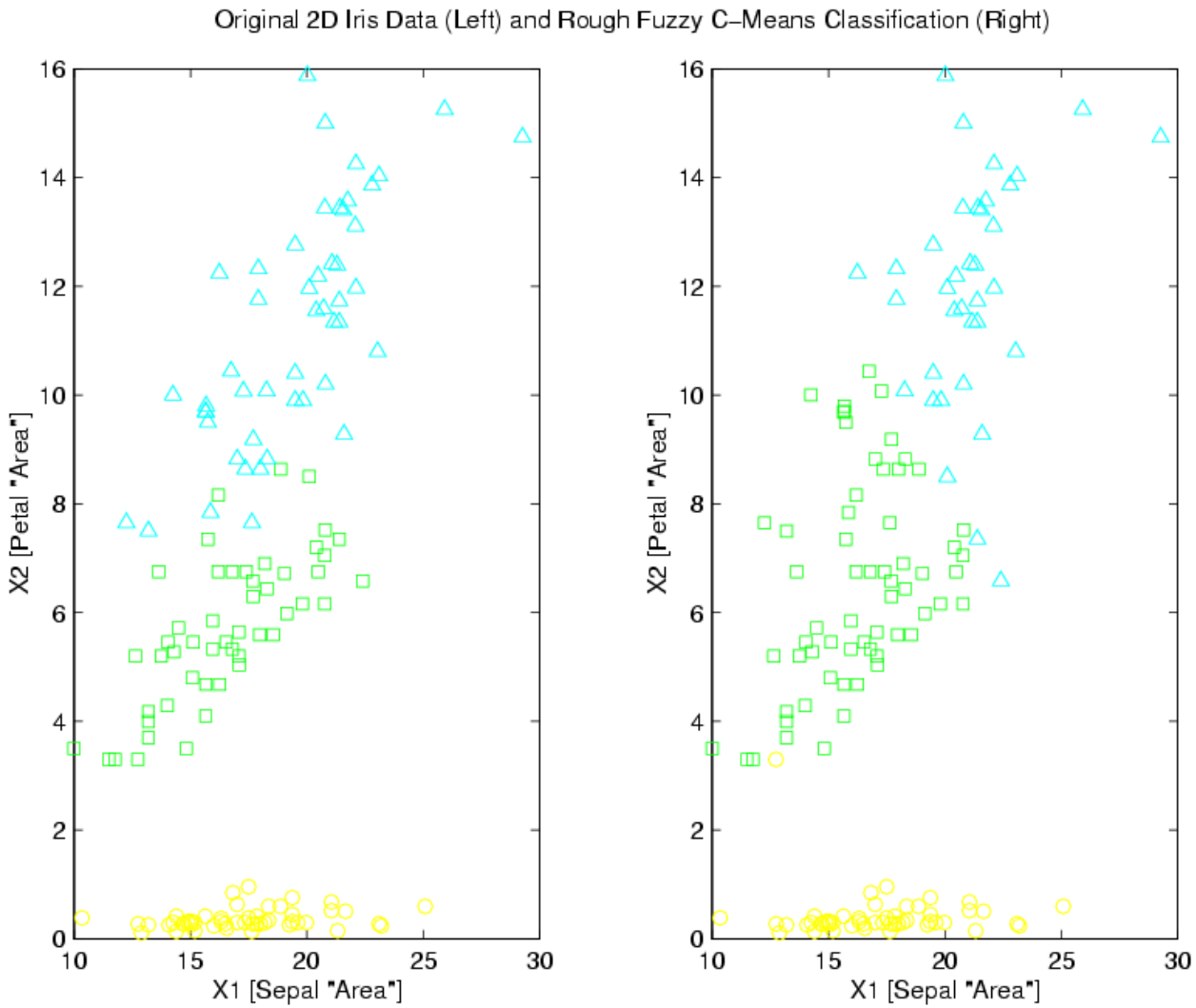
- Exemplo (“Iris” 2D):



Fuzzy c-Means



Fuzzy c-Means

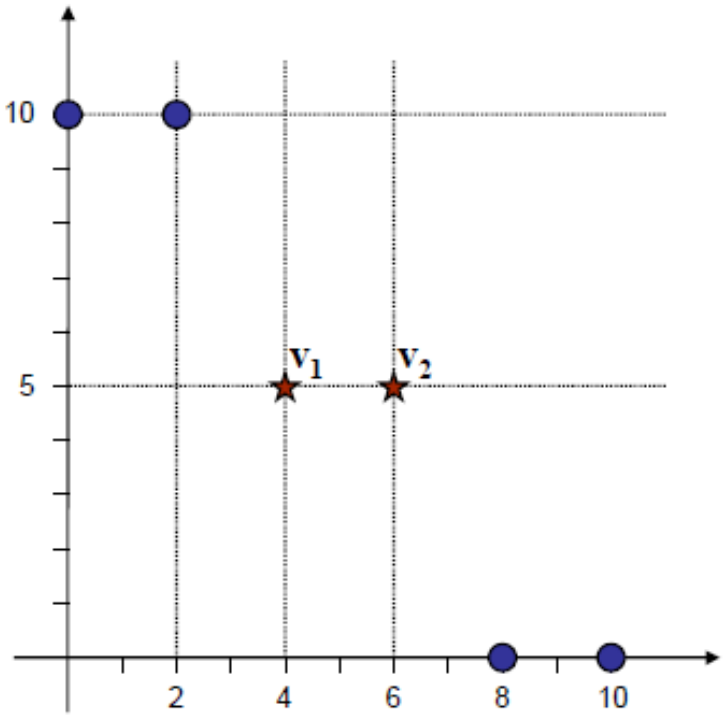


Agrupamento Fuzzy

- Existem versões fuzzy para todos os tipos de critérios de validade de agrupamento discutidos anteriormente:
  - Critérios **relativos** de validade de agrupamento fuzzy
    - Silhueta Fuzzy
    - e muitos outros...
  - Critérios **externos** de validade de agrupamento fuzzy
    - Jaccard Fuzzy
    - e muitos outros...

Exercício

- Agrupar os dados em azul na figura abaixo através do método FCM com 2 clusters, m=2 e centros iniciais assinalados em vermelho. Apresentar os centros dos grupos e a matriz de valores de pertinência para cada iteração.



Expectation Maximization (EM)

- O Algoritmo **EM (Expectation Maximization)** é um procedimento genérico para a modelagem **probabilística** de um conjunto de dados

- Basicamente, o algoritmo otimiza os parâmetros de uma função de distribuição de probabilidades de forma que esta represente os dados da forma mais verossímil possível
  - Maximização da Verossimilhança

- Modelo mais utilizado é aquele cuja função de distribuição de probabilidades é dada por uma **Mistura de Gaussianas**

EM – Mistura de Gaussianas

- O modelo de mistura de Gaussianas é dado pela seguinte função de densidade de probabilidade p:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

- **x** é um padrão (objeto)
- *N* é uma Gaussiana (da mesma dimensão dos padrões)
  - $\boldsymbol{\mu}_k$ é o centro da k-ésima Gaussiana (vetor da mesma dimensão de **x**)
  - $\boldsymbol{\Sigma}_k$ é a matriz de covariância da k-ésima Gaussiana
- K é o número de Gaussianas que compões a mistura

EM – Mistura de Gaussianas

- Dado um conjunto de N padrões $\mathbf{x}_n$ (n = 1, ..., N), o algoritmo opera em dois passos:
  - Passo E (Expectation)
  - Passo M (Maximization)

EM for Gaussian Mixtures

1. Initialize the means $\boldsymbol{\mu}_k$, covariances $\boldsymbol{\Sigma}_k$ and mixing coefficients π_k, and evaluate the initial value of the log likelihood.

2. **E step**. Evaluate the responsibilities using the current parameter values

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

continua...

EM – Mistura de Gaussianas

3 **M step.** Re-estimate the parameters using the current responsibilities

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n$$

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(\mathbf{x}_n-\boldsymbol{\mu}_k^{\text{new}})(\mathbf{x}_n-\boldsymbol{\mu}_k^{\text{new}})^{\mathrm{T}}$$

$$\pi_k^{\text{new}} = \frac{N_k}{N}$$

where

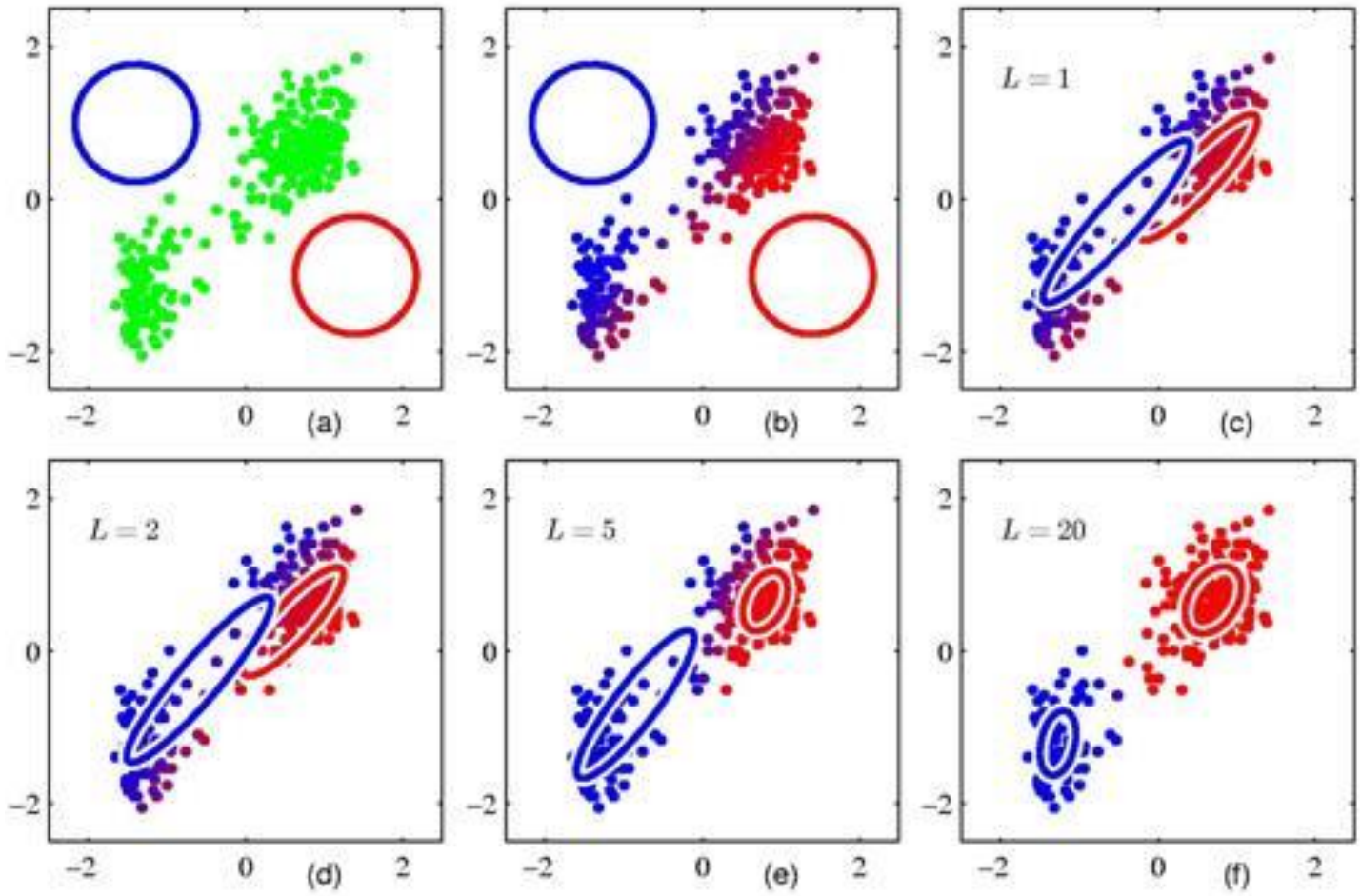
$$N_k = \sum_{n=1}^{N}\gamma(z_{nk}).$$

4. Evaluate the log likelihood

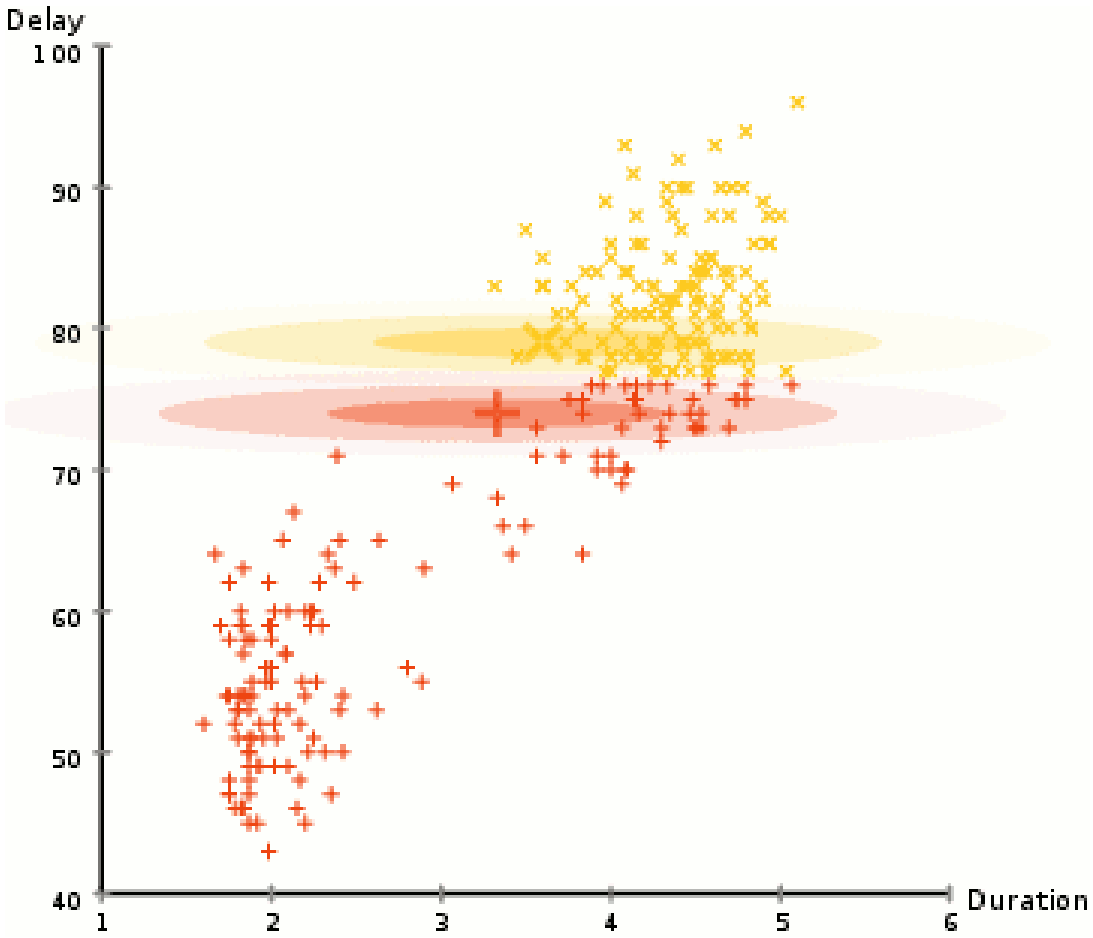
$$\ln p(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi}) = \sum_{n=1}^{N}\ln\left\{\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right\}$$

and check for convergence of either the parameters or the log likelihood. If the convergence criterion is not satisfied return to step 2.

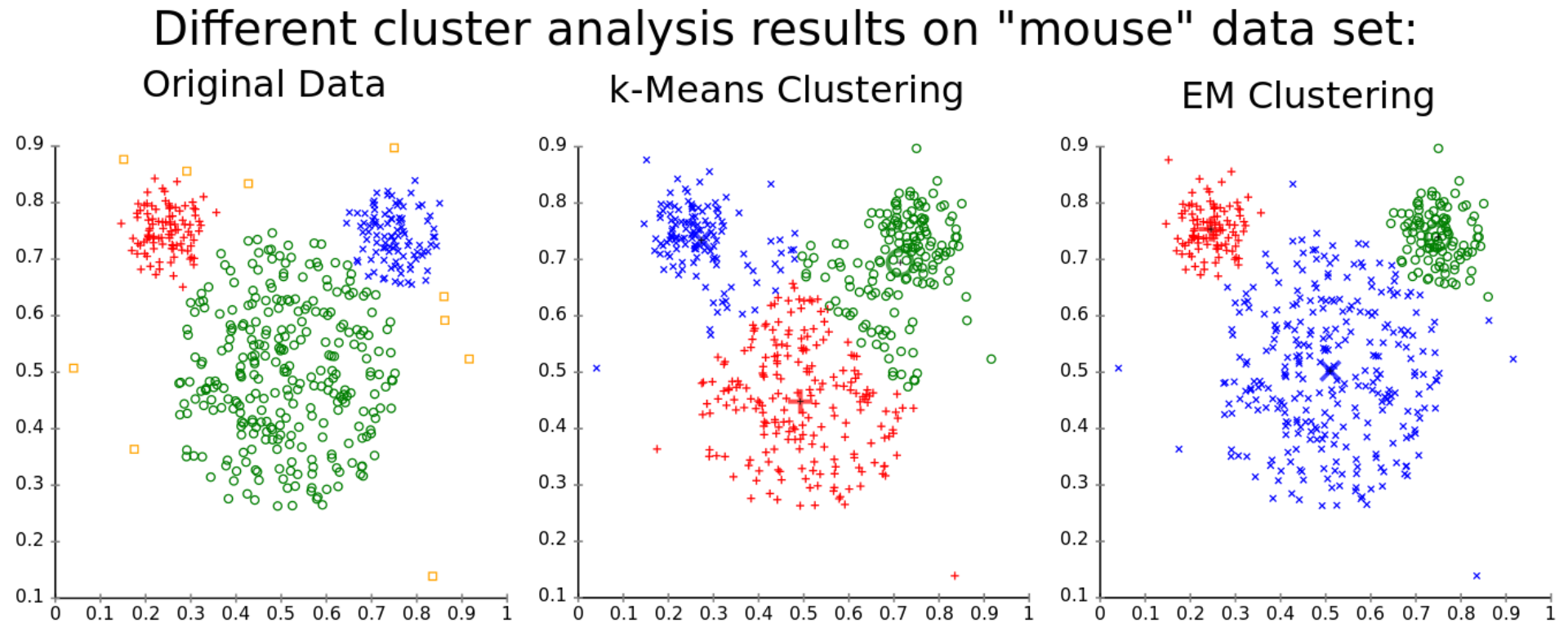
EM – Exemplo



EM – Exemplo



EM x k-Means


Different cluster analysis results on "mouse" data set:
Original Data
k-Means Clustering
EM Clustering
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

EM x k-Means

- EM produz informação muito mais rica sobre os dados
  - Probabilidades associadas a cada padrão / cluster
- Probabilidades produzidas por EM podem facilmente ser convertidas em uma partição rígida, caso desejado
  - Via projeção do maior valor
  - Essas partições podem representar clusters alongados, elipsoidais, com atributos correlacionados
- No entanto, todas as vantagens acima vêm com um elevado custo computacional associado...
  - Cálculo das Normais Multi-dimensionais N demanda as inversas das matrizes de covariância $\mathbf{\Sigma}_k$, que é O(n3)
- k-means é um caso particular de EM. Ambos estão sujeitos a mínimos locais

Alguns Outros Algoritmos de Agrupamento

- em Densidade / Grids:
  - DBSCAN, CLIQUE, DENCLUE,...
- Baseados em Hierarquias / Árvores:
  - CURE, Chameleon, BIRCH,...
- Baseados em Grafos:
  - CLICK, Chameleon, ROCK,...
- Baseados em Medóides:
  - PAM, CLARA, CLARANS,...
- Auto-Organizáveis:
  - SOM,...

Notas Finais

- O problema de normalização / padronização dos dados é mais complexo em clustering do que em outras tarefas de AM e DM (e.g. classificação)

- As técnicas disponíveis são essencialmente as mesmas, bem como o objetivo da aplicação destas

  - Por exemplo, evitar que atributos com escalas muito maiores do que outros dominem os cálculos de dissimilaridade e, portanto, induzam sozinhos a estrutura de clusters

- No entanto, a aplicação dessas técnicas pode distorcer totalmente a estrutura original dos dados em clusters !!!
  - É preciso mais cautela, experimentação e conhecimento de domínio para realizar pré-processamento de dados em clustering!

Notas Finais

- A observação anterior também é particularmente válida no que diz respeito à seleção de atributos
- Clusters podem ser bem definidos em um sub-conjunto de atributos mas não no conjunto completo ou em outro subconjunto!!!
- Áreas de pesquisa ativas relacionadas a esta questão:
  - Subspace Clustering
  - Seleção de atributos para clustering
  - Ponderação de atributos em clustering
  - Biclustering

Referências

- Jain, A. K. and Dubes, R. C., **Algorithms for Clustering Data**, Prentice Hall, 1988
- Everitt, B. S., Landau, S., and Leese, M., **Cluster Analysis**, Arnold, 4th Edition, 2001.
- Gan, G., Ma, C., and Wu, J., **Data Clustering: Theory, Algorithms and Applications**, ASA SIAM, 2007
- Bishop, C. M., **Pattern Recognition and Machine Learning**, Springer, 2006
- 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