

Inteligência Artificial

Técnicas de Agrupamento de Dados (Clustering)

Paulo Moura Oliveira

Departamento de Engenharias

Gabinete F2.15, ECT-1

UTAD

email: oliveira@utad.pt

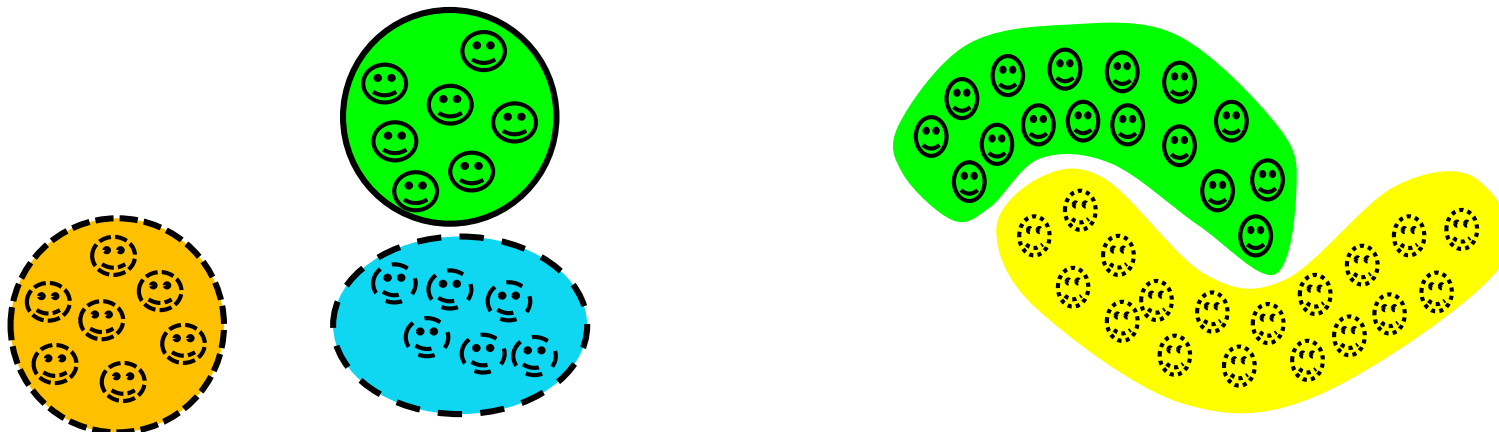
v: 2024

Em que consiste o *clustering*?

Clustering, consiste na organização (ou classificação) de um conjunto de dados em vários grupos a que se chamam *clusters*.

Como se faz o *clustering*?

Utiliza-se um dado **critério de similaridade** para agrupar os dados similares no mesmo grupo (ou de **critério de dissemelhança** para os distinguir dos outros grupos).



Medidas de Proximidade

Há muitas formas de determinar a distância entre dois pontos. Das mais conhecidas temos:

Distância Euclidiana

$$\text{dist}(x_i, x_j) = \sqrt{\sum_{k=1}^d (x_{ik} - x_{jk})^2}$$

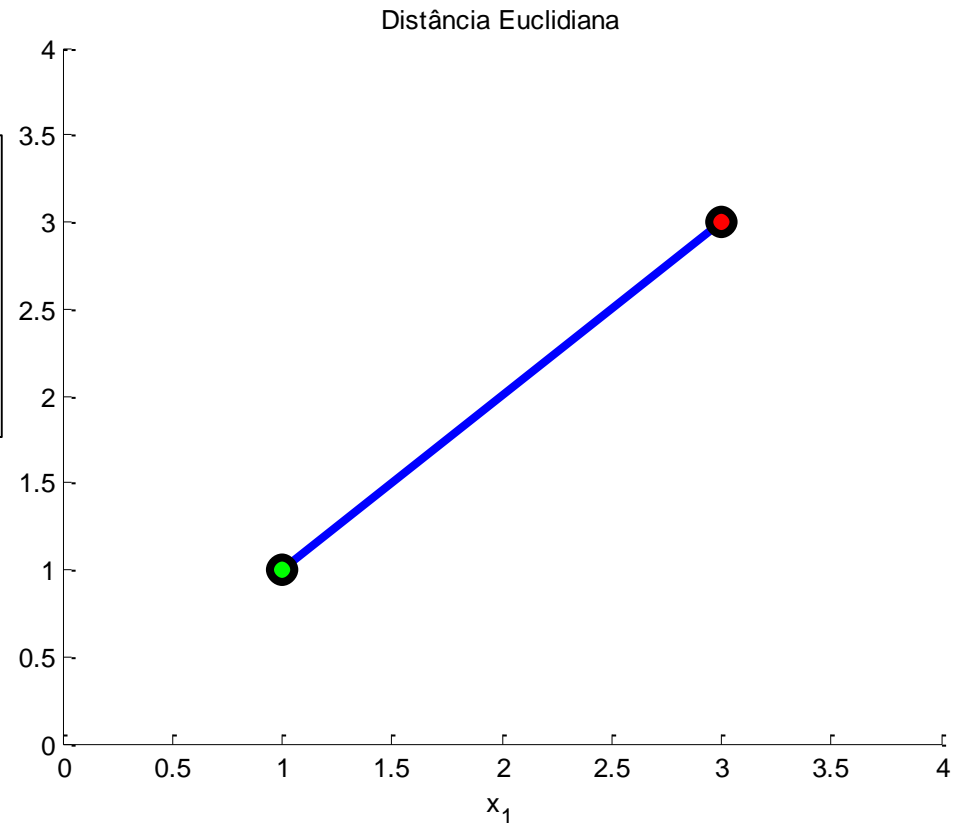
d: dimensão

Exemplo:

$d=2$

$x_1 = (1, 1)$; $x_2 = (3, 3)$

dist= 2.83



Medidas de Proximidade

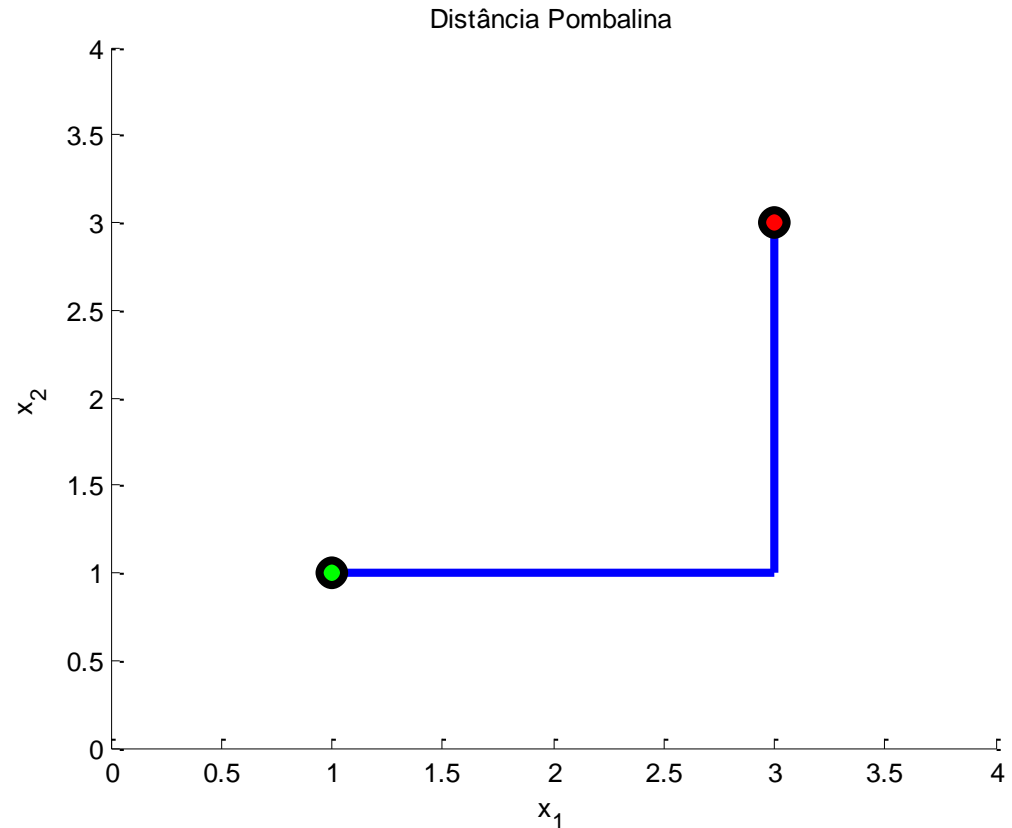
Distância Pombalina (conhecida como Manhattan ou *CityBlock*)

$$\text{dist}(x_i, x_j) = \sum_{k=1}^d |x_{ik} - x_{jk}|$$

Exemplo:

$x_1 = (1, 1)$; $x_2 = (3, 3)$

$d = 4$



Medidas de Proximidade

Distância Chebychev

$$\text{dist}(x_i, x_j) = \max_d |x_{id} - x_{jd}|$$

Exemplo:

$$x_1 = (1, 1) ; x_2 = (3, 3)$$

$$d = 2$$

Distância Minkowski

$$\text{dist}(x_i, x_j) = \sqrt[p]{\sum_{k=1}^d (x_{ik} - x_{jk})^p}$$

Exemplo:

$$p = 2$$

$$x_1 = (1, 1) ; x_2 = (3, 3)$$

$$d = 2.83$$

$p = 2$, Igual à
Euclidiana

Exemplo:

$$p = 5$$

$$x_1 = (1, 1) ; x_2 = (3, 3)$$

$$d = 2.297$$

Como Avaliar os Clusters (Agrupamentos)?

- ✓ **Coesão Intra-cluster:** avalia a proximidade dos seus pontos ao centróide do cluster

Uma medida muito utilizada é o Somatório do Erro Quadrático (SSE):

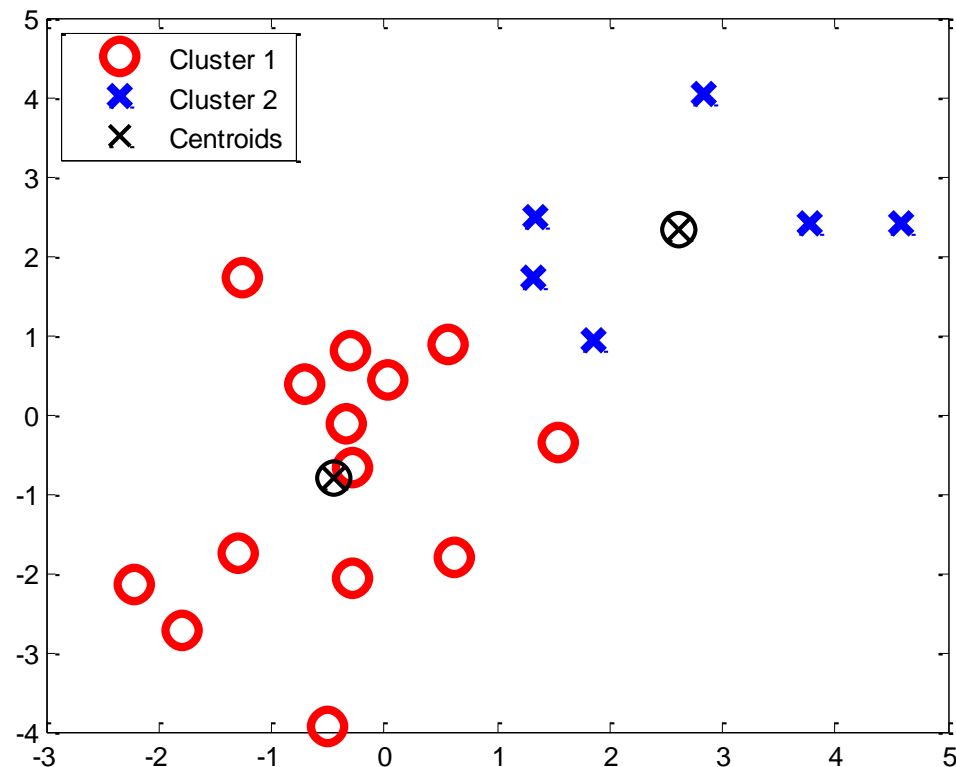
$$SSE = \sum_{r=1}^d dist^2(x_{ir} - x_{cr})$$

c: centróide do cluster

SSE_1= 47.0279

SSE_2= 14.3666

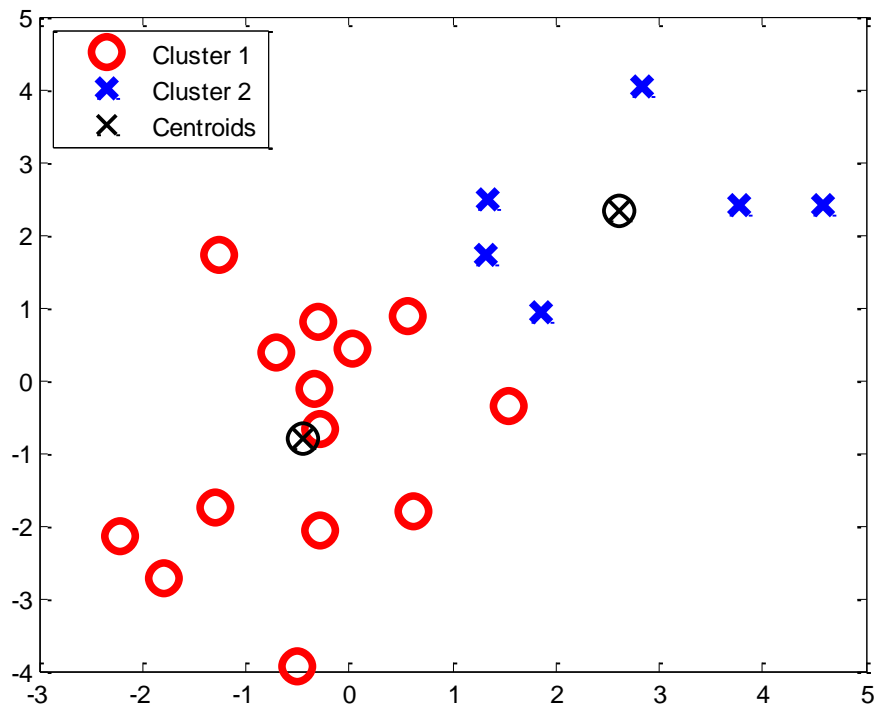
SSE= 61.3945



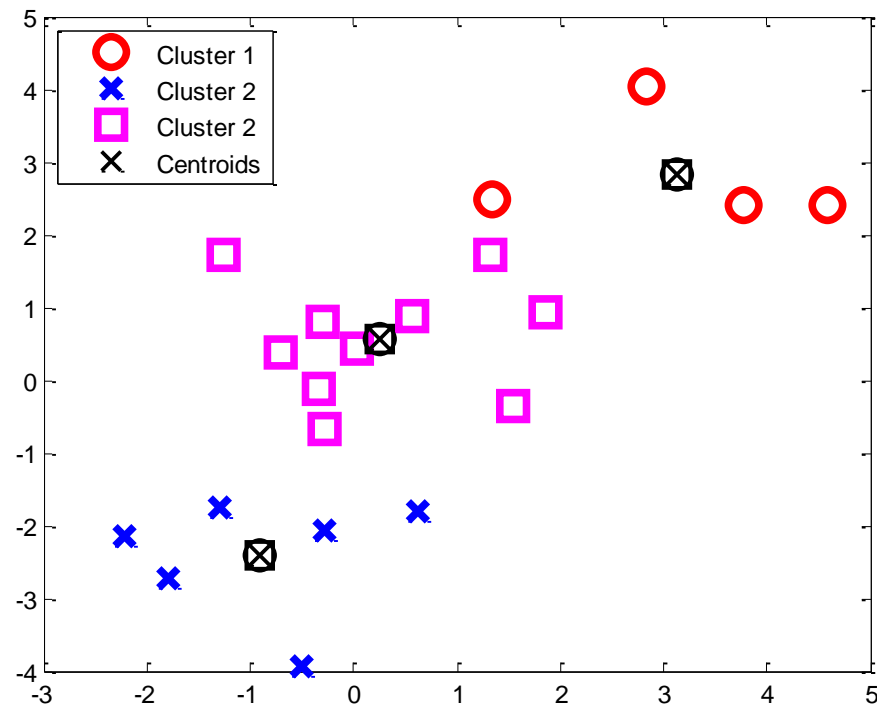
- ✓ **Separação Inter-cluster:** avalia a separação dos centroides dos vários clusters.

Qual o número de Clusters?

- ✓ Consideremos o mesmo exemplo do diapositivo anterior utilizando o k-means. Em vez de 2 clusters vamos agora considerar 3:



SSE_1= 47.0279
SSE_2= 14.3666
SSE= 61.3945



SSE_1= 7.7042
SSE_2= 8.9528
SSE_3= 15.4975
SSE= 32.1545

Qual o número de Clusters?

✓ Um procedimento possível é o seguinte:

1. Definir um **número fixo** de clusters
2. Executar o método de *clustering* e obter o melhor resultado para uma dada função de custo (função objetivo).
3. Voltar a 1 e **aumentar (ou diminuir) o número de clusters**

Quais as técnicas de *Clustering*?

- ✓ Existem várias taxonomias de técnicas de *clustering* que podem ser encontradas na literatura. Uma classificação comum usa três grupos:

1. Hierárquicas (*Hierarchical*)
2. Particionais (*Partitional*)
3. *Bayesianas* (*Bayesian*)



k-Means

- ✓ Como o algoritmo k-means é um dos mais utilizados vamos começar por esta técnica.

O que é?

Técnica de *clustering* que particiona um conjunto de dados em k clusters.

- ✓ Cada cluster tem um centro (centroide)
- ✓ O número de clusters, k , é especificado pelo utilizador.

Algoritmo k-means

Selecionar (ou Gerar) k -centros (centroides iniciais)

while(!(critério de paragem))

Atribuir cada amostra de dados ao cluster cujo centroide está mais próximo.

Recalcular os centroides utilizando os clusters atuais

end while

Critério de Paragem

- ✓ Alguns critérios que podem ser utilizados para parar o ciclo do k-means:
 1. Um número pré-definido de iterações;
 2. Variação dos centroides abaixo de um limiar mínimo;
 3. Variação dos pontos nos clusters menor que um valor baixo pré-definido;
 4. Soma do erro quadrático abaixo que um valor baixo pré-definido.

The diagram shows the Sum of Squared Errors (SSE) formula for k-means clustering. The formula is
$$SSE = \sum_{j=1}^k \sum_{\substack{r=1 \\ x_i \in C_j}}^d dist^2(x_{ir} - x_{jr})$$
. Annotations include: a red box labeled "k- clusters" with an arrow pointing to the summation index k ; a red box labeled "Para todas as dimensões de x" with an arrow pointing to the summation index d ; and a red box labeled "Elementos de cada cluster, j" with an arrow pointing to the inner summation condition $x_i \in C_j$.

k- clusters

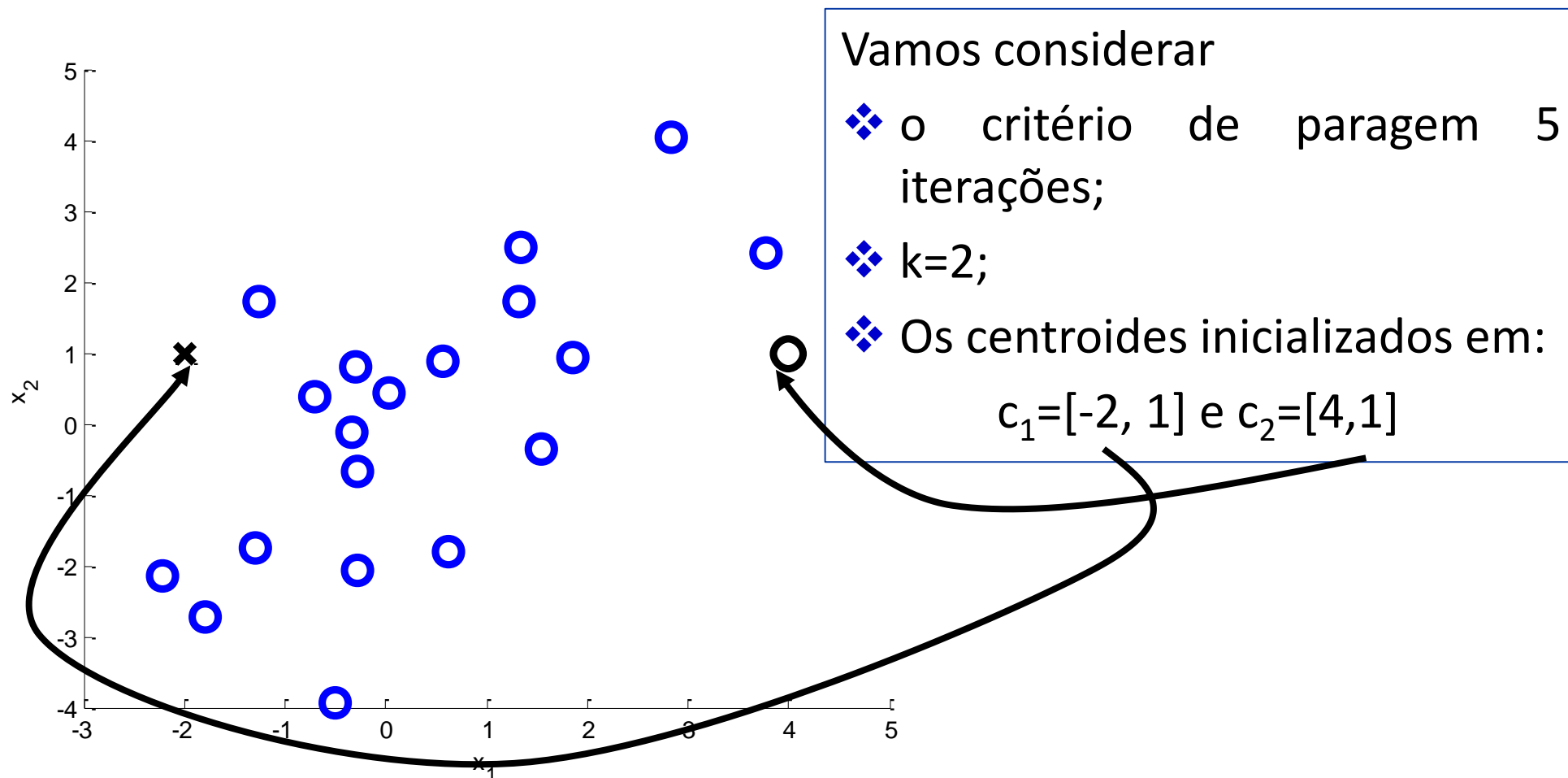
Para todas as dimensões de x

$SSE = \sum_{j=1}^k \sum_{\substack{r=1 \\ x_i \in C_j}}^d dist^2(x_{ir} - x_{jr})$

Elementos de cada cluster, j

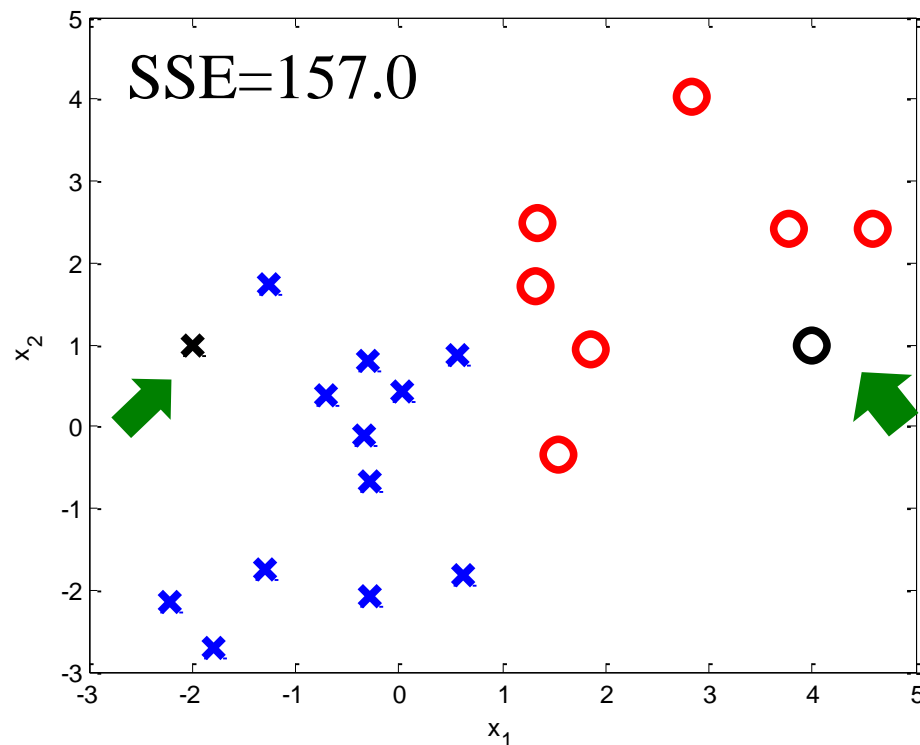
Exemplo 1

- ✓ Configuração inicial de um conjunto com 20 pontos.



Exemplo 1

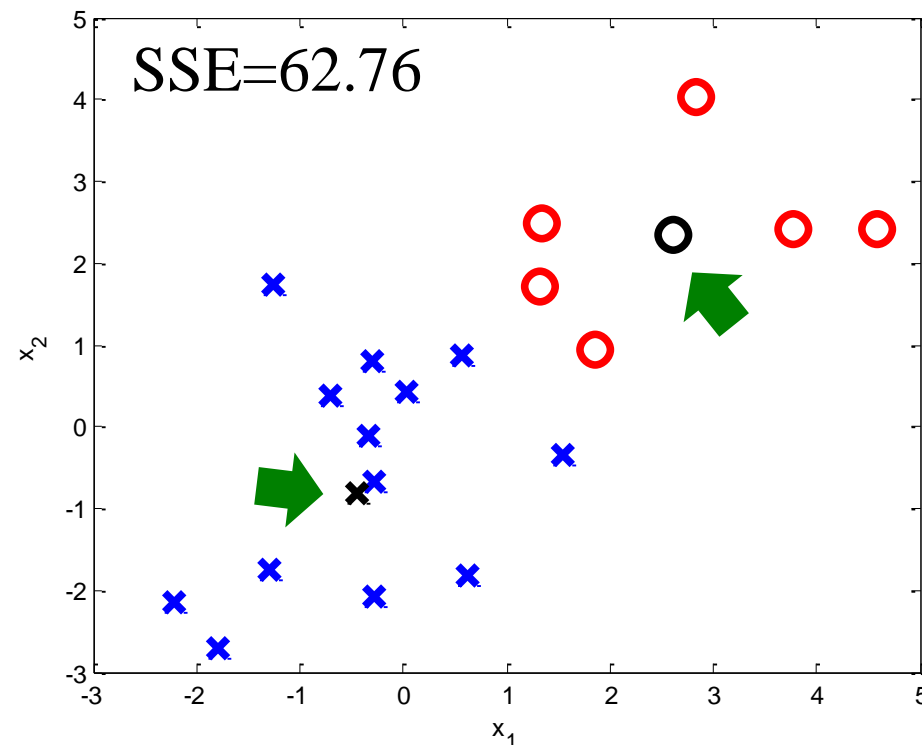
✓ Iteração 1:



$$c_1 = [-2, 1]$$

$$c_2 = [4, 1]$$

✓ Iteração 2:

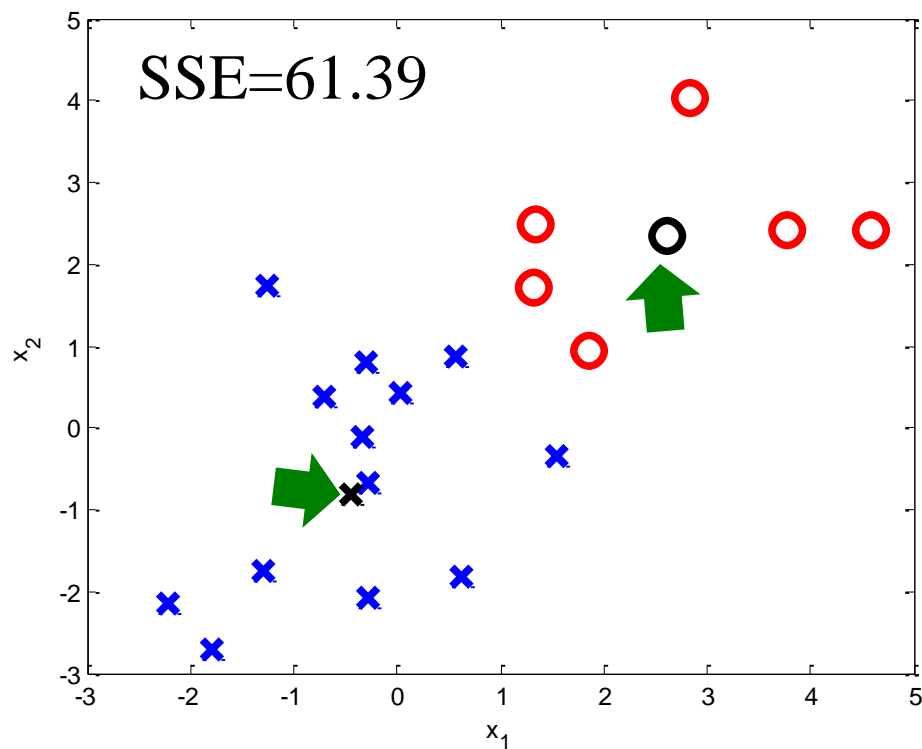


$$c_1 = [-0.5950, -0.8475]$$

$$c_2 = [2.4633, 1.9504]$$

Exemplo 1

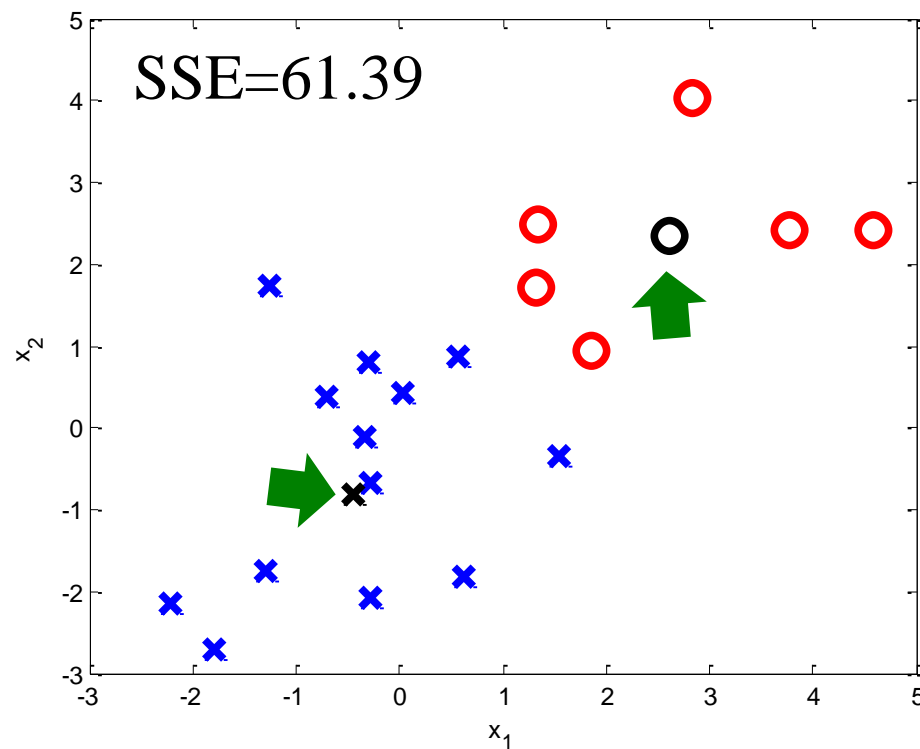
✓ Iteração 3:



$$c_1 = [-0.4427, -0.8120]$$

$$c_2 = [2.6175, 2.3338]$$

✓ Iteração 4:

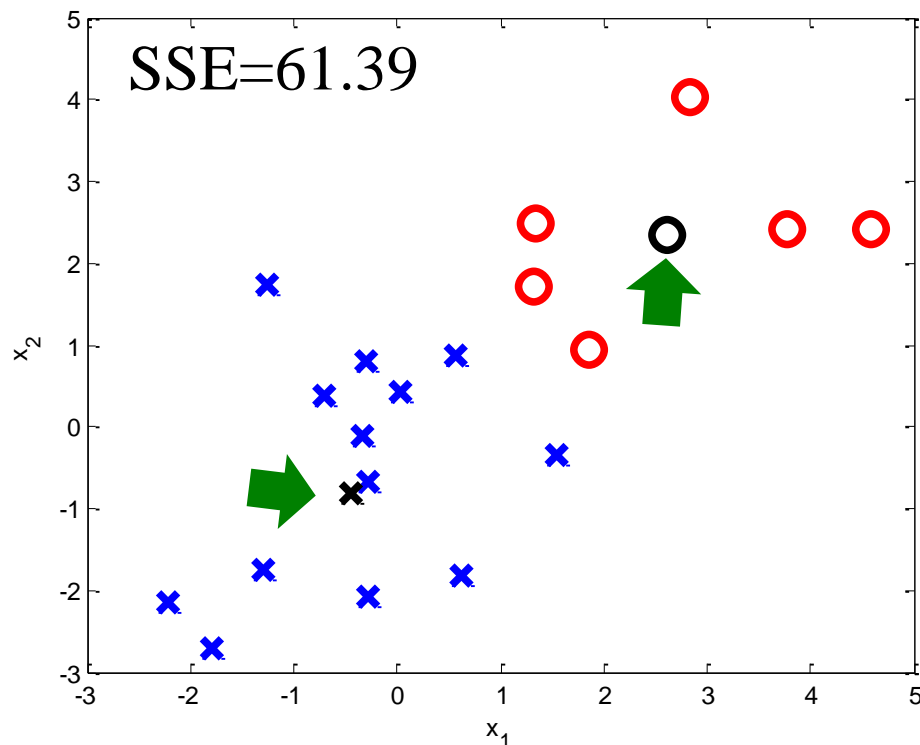


$$c_1 = [-0.4427, -0.8120]$$

$$c_2 = [2.6175, 2.3338]$$

Exemplo 1

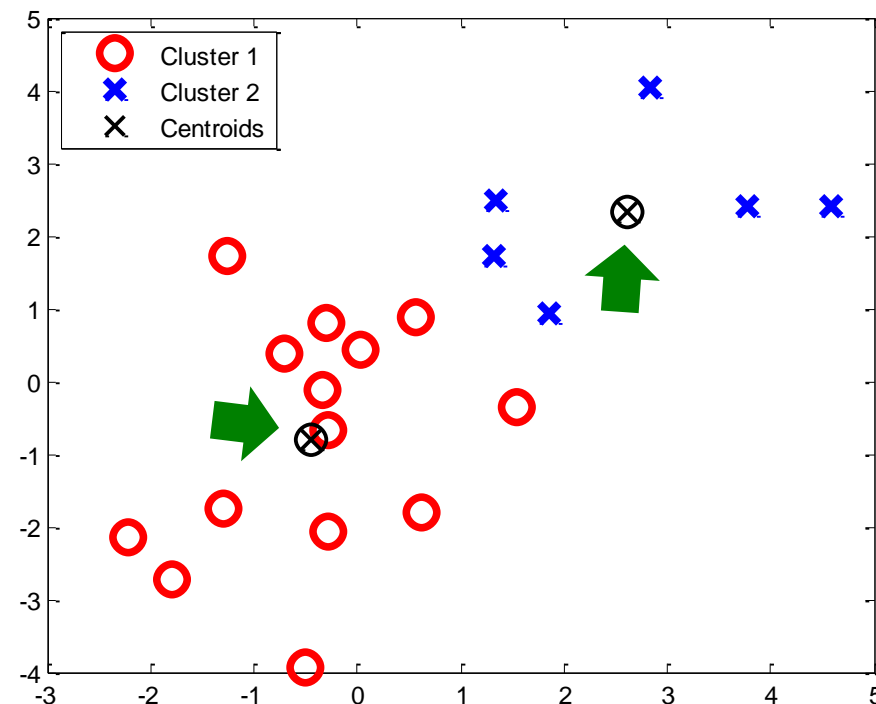
✓ Iteração 5:



$$c_1 = [-0.4427, -0.8120]$$

$$c_2 = [2.6175, 2.3338]$$

✓ Utilizando a função do Matlab (kmeans):



$$c_1 = [-0.4427, -0.8120]$$

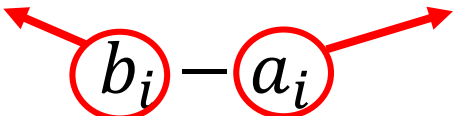
$$c_2 = [2.6175, 2.3338]$$

✓ Neste caso deu o mesmo resultado.

- ✓ Uma forma de avaliar a qualidade do *clustering* é utilizando o critério da **silhueta**, cujo valor pode ser determinado para o ponto i :

Mínimo das médias das distâncias do ponto i aos outros pontos dos outros clusters.

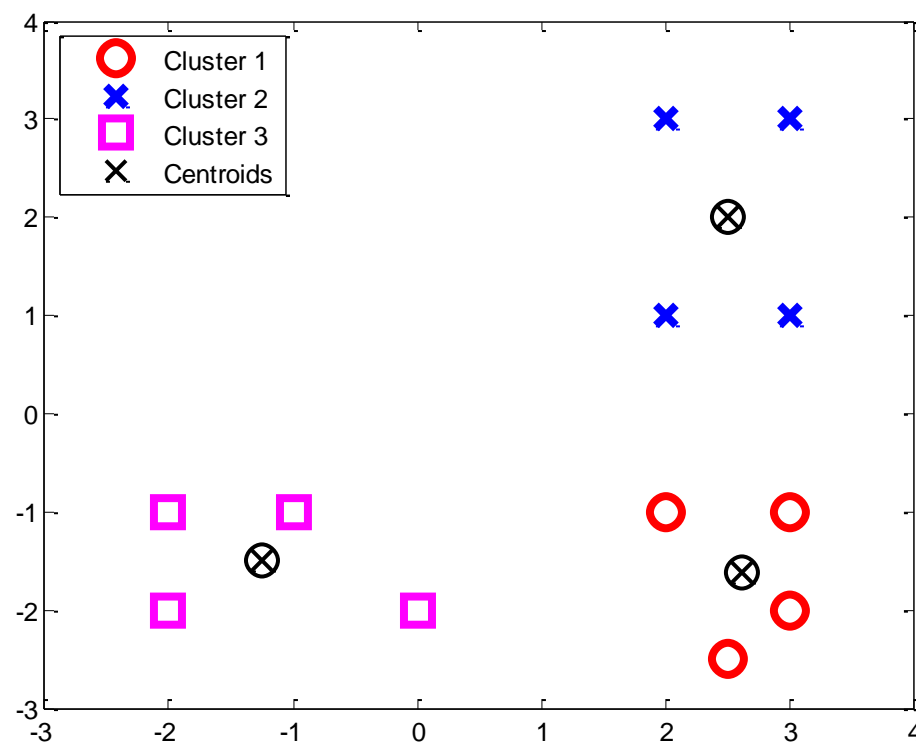
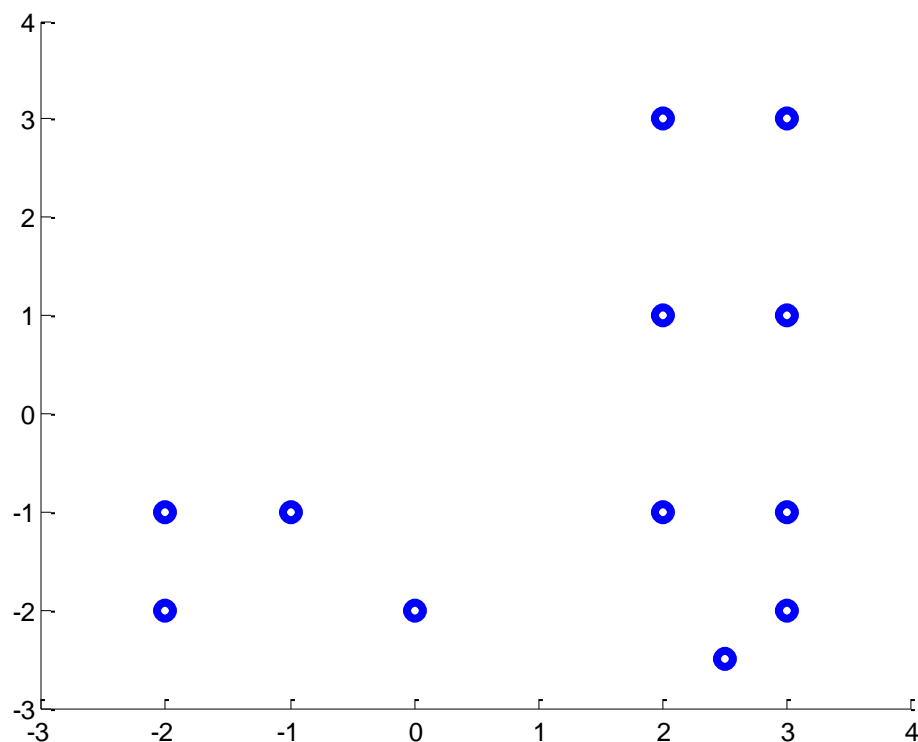
Média das distâncias do ponto i aos outros pontos do mesmo cluster.


$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

- ✓ Se:
 - Os valores de s_i podem variar entre $[-1$ e $1]$;
 - Se a maioria valores de s_i estiverem próximos de 1 , indica que o ***clustering* é bom**;
 - Se a muitos valores de s_i forem baixos ou próximos de -1 , indica que o ***clustering* é mau** (precisa de mais ou menos clusters)

Exemplo:

- ✓ Considere-se a seguinte representação inicial de dados com o respetivo agrupamento com o *k-means*:



Exemplo:

Separação:
Inter-cluster

$$d_{i1} = 25$$

$$d_{i2} = 17$$

$$d_{i3} = 26$$

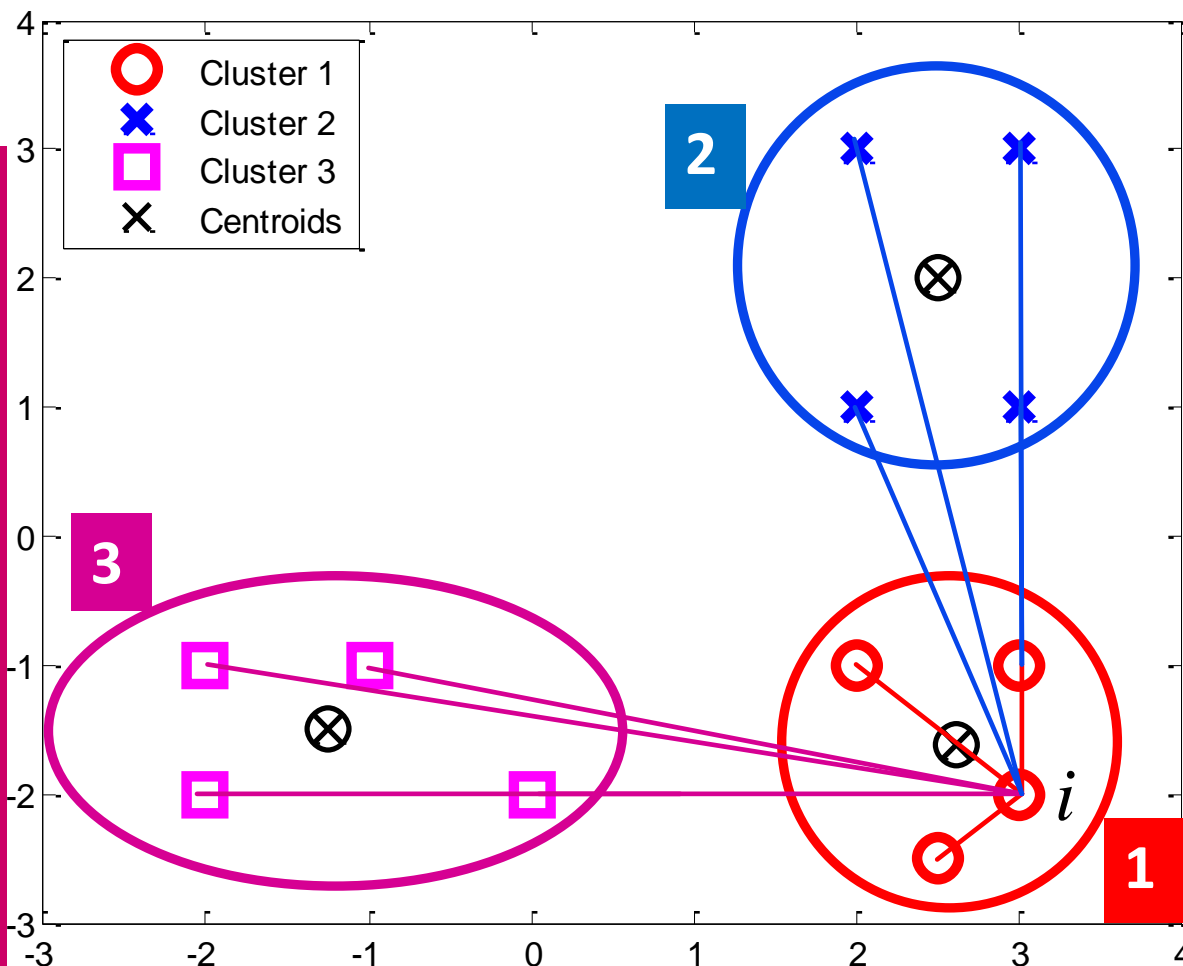
$$d_{i4} = 9$$

$$b_{i3} = (d_{i1} + d_{i2} + d_{i3} + d_{i4}) / 4 =$$

$$\underline{19.25}$$

$$b_i = \min(b_{i2}, b_{i3}) = 17.5$$

$$\max(a_i, b_i) = 17.5$$



Separação:
Inter-cluster

$$d_{i1} = 10$$

$$d_{i2} = 26$$

$$d_{i3} = 9$$

$$d_{i4} = 25$$

$$b_{i2} = 17.5$$

Coesão:

Intra-cluster

$$d_{i1} = 1$$

$$d_{i2} = 2$$

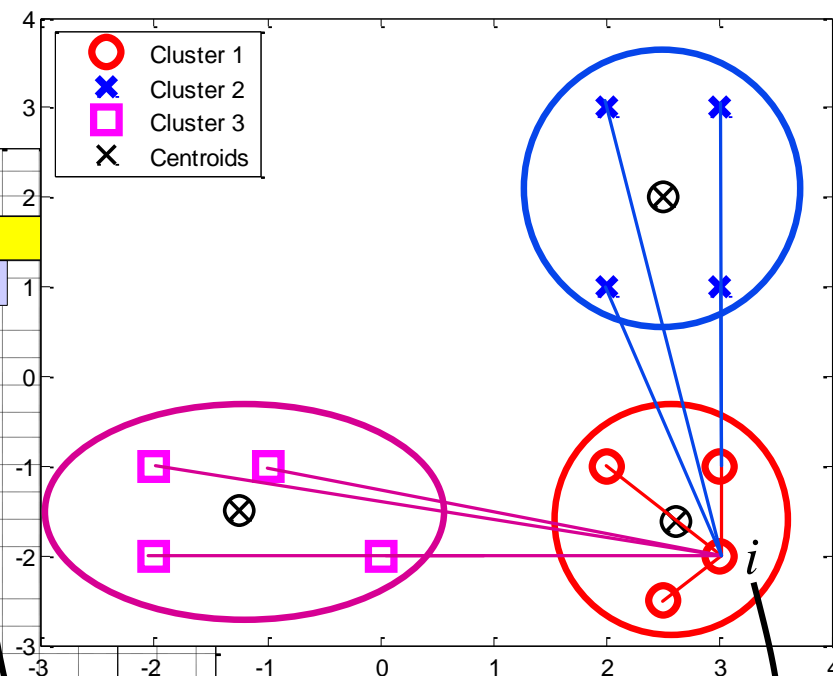
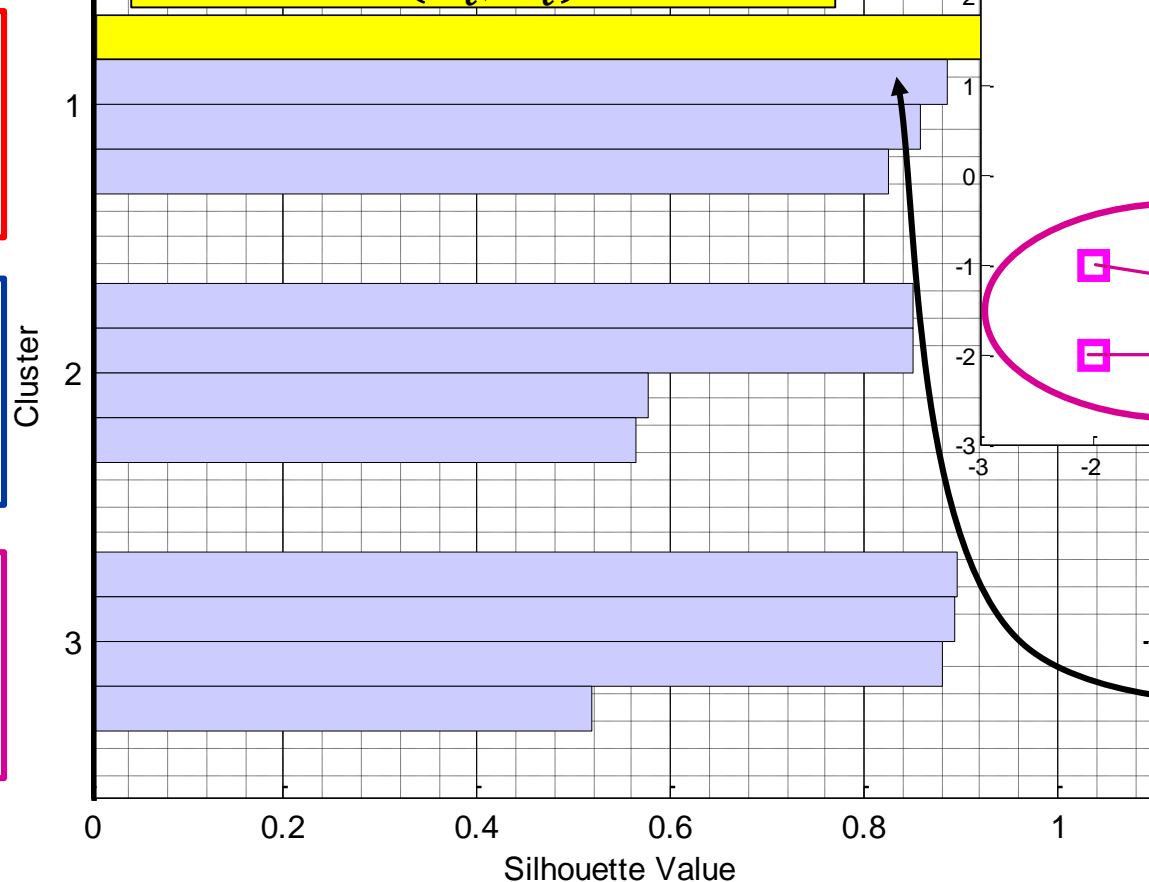
$$d_{i3} = 0.5$$

$$a_i = 1.1667$$

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} = 0.933$$

Exemplo:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} = 0.933$$



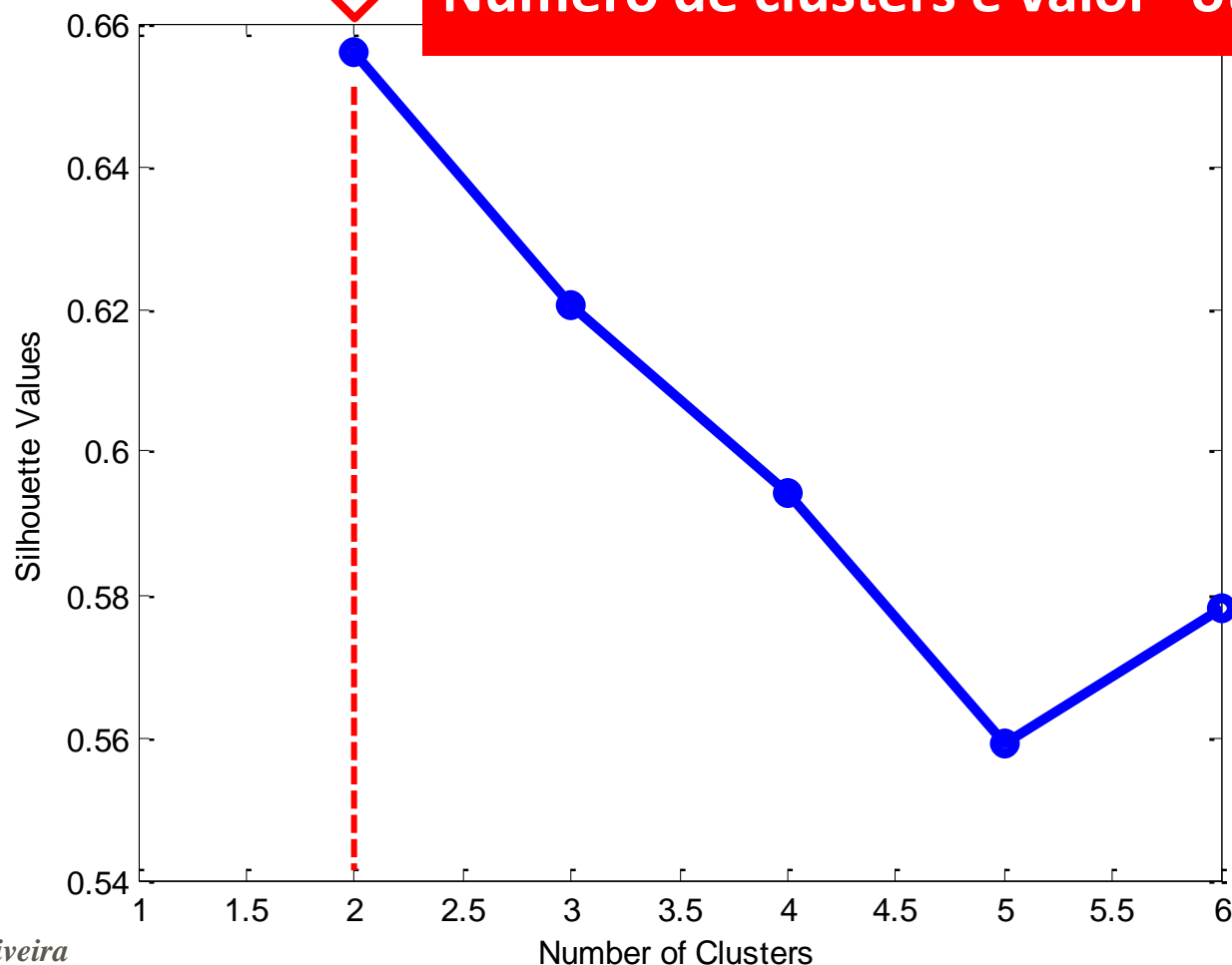
Como o valor de S_i está próximo de 1, indica que está bem classificado neste grupo.

Exemplo 1

- ✓ Podemos tentar vários números de clusters e ver qual deles dá a média dos valores silhueta menores:

[1	2	3	4	5	6]
[NaN	0.6559	0.6204	0.5944	0.5591	0.5780]

Número de clusters e valor “ótimo”

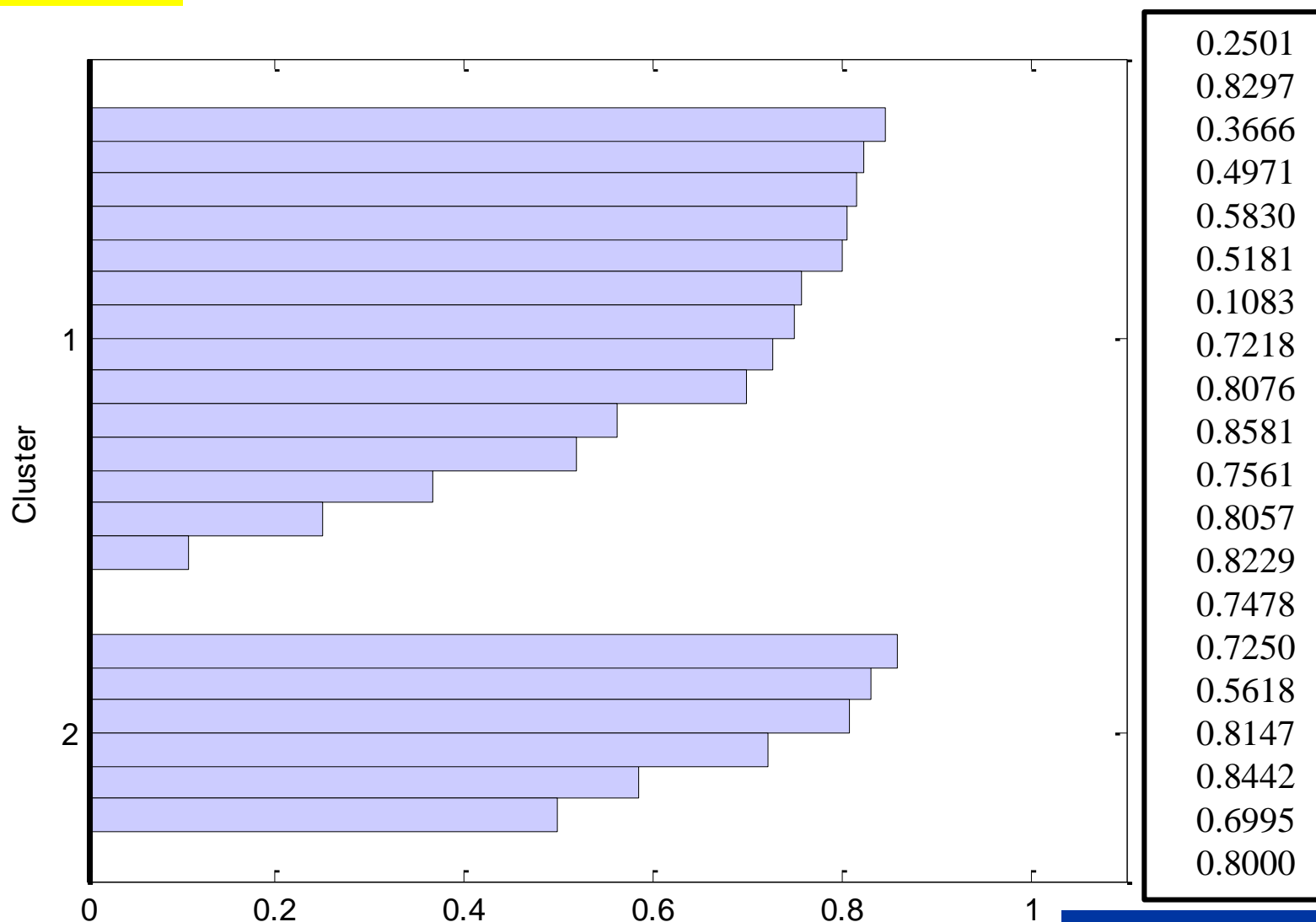


Nota:

As distâncias foram calculadas utilizando o quadrado da distância Euclidiana.

Exemplo 1

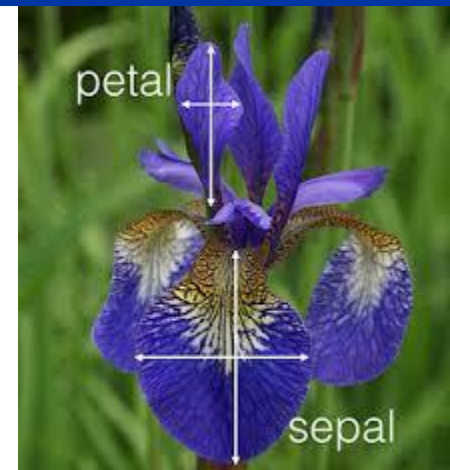
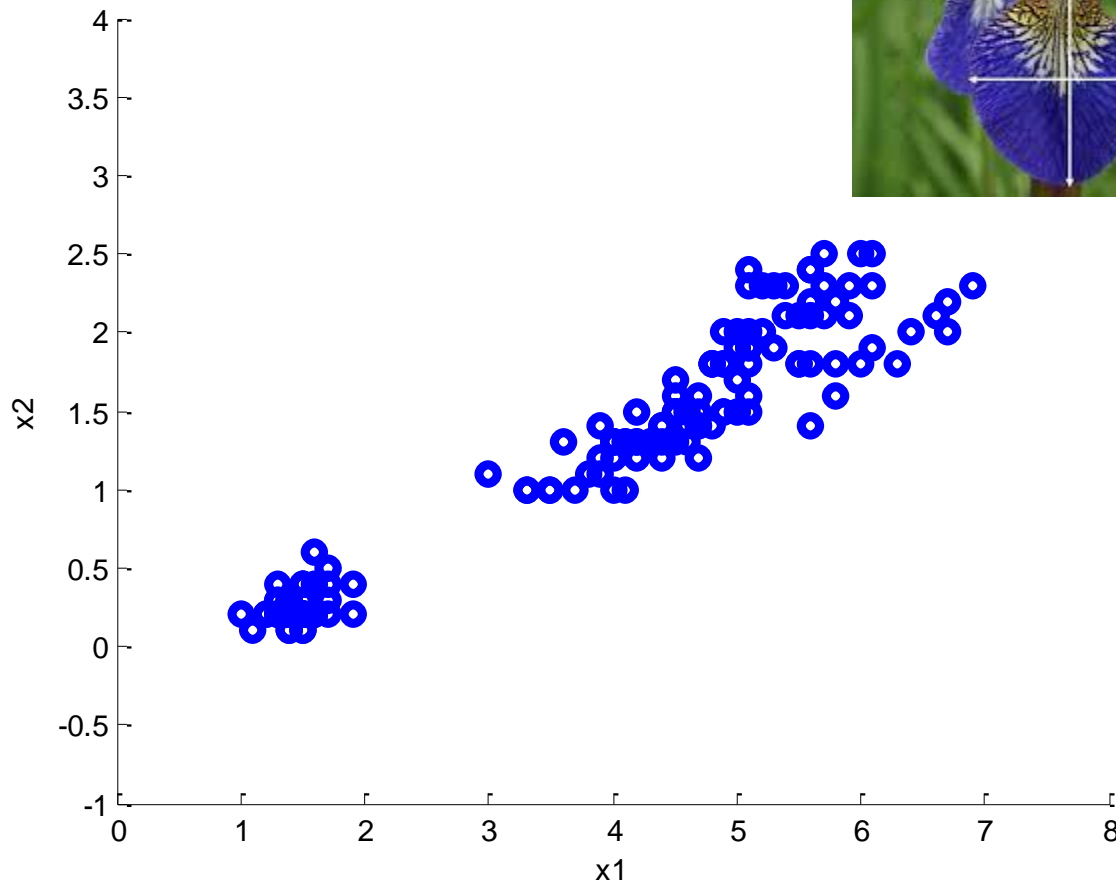
✓ Para o este caso o gráfico dos valores Silhueta é o seguinte:



Clustering, Exemplo 1:

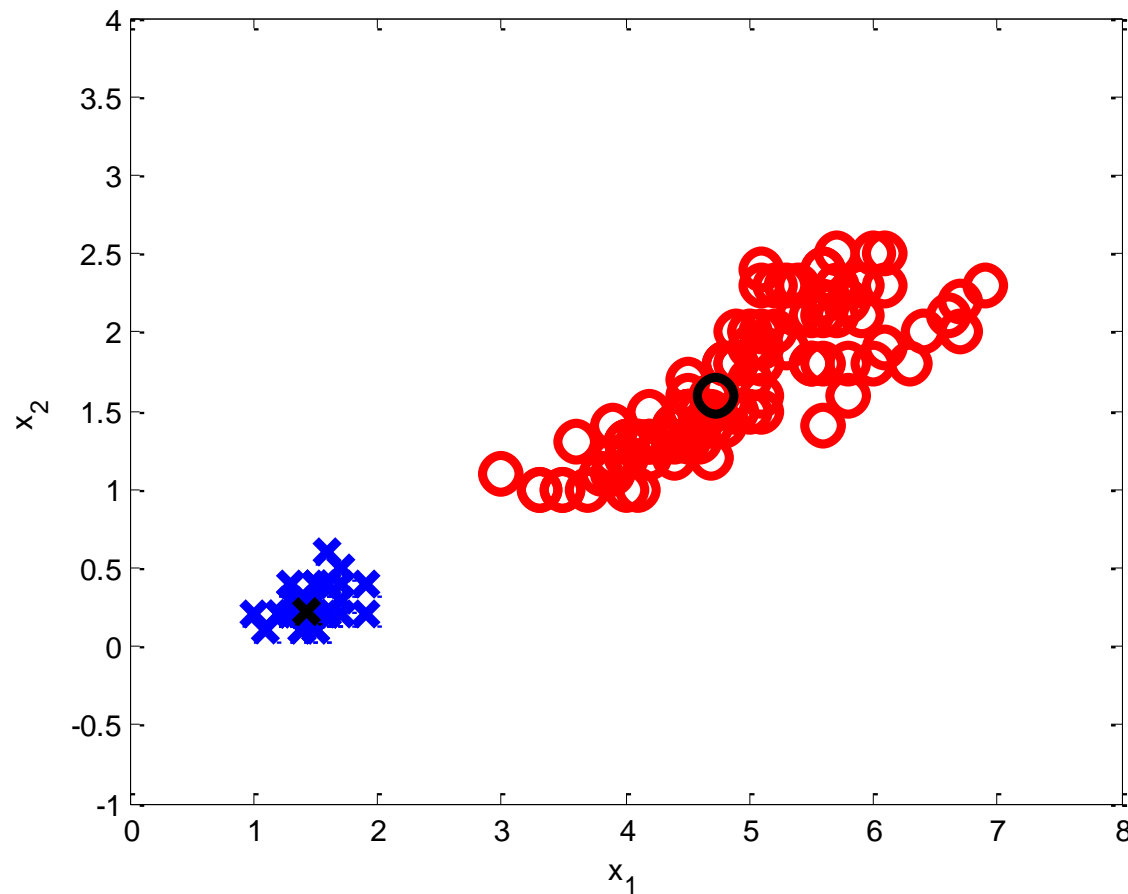
- ✓ Conjunto de dados relativos a flores (Iris).
- ✓ Vamos considerar dois dos atributos relativos às dimensões das pétalas (dum total de 4)

Largura da
Pétala (cm)



Comprimento da Pétala (cm)

✓ Configuração final:



'o' - Dados

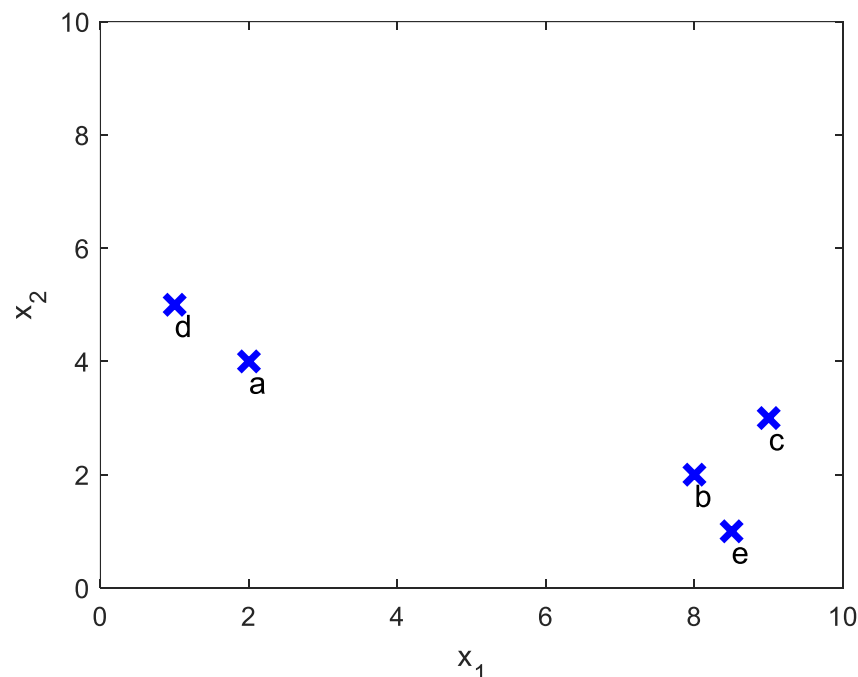
'x' - centróide c1

'o' - centróide c2

Clustering, Exemplo 2:

✓ Considere-se o seguinte exemplo didático (Adaptado de Peter Tryfus, 1997).

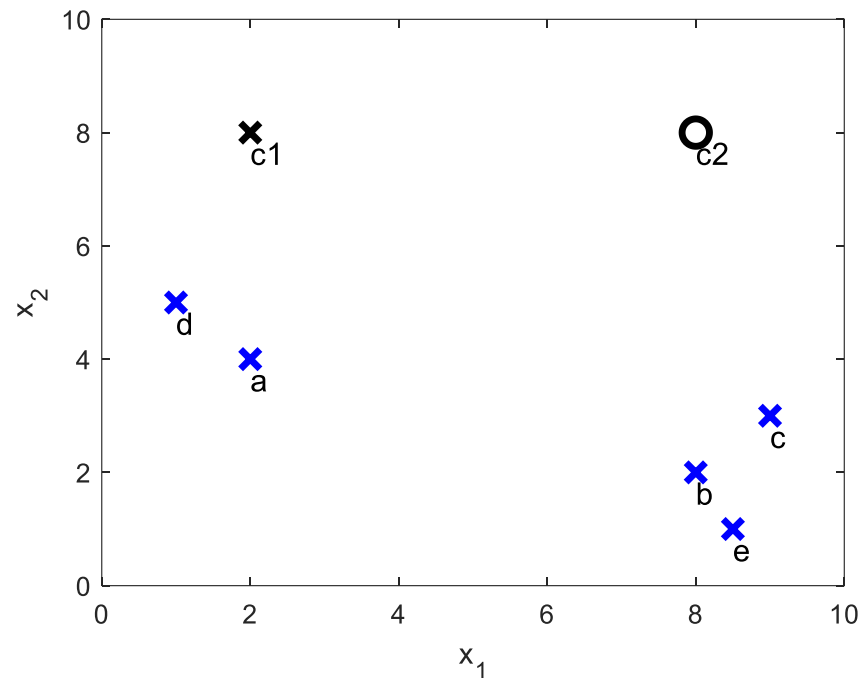
Amostra	x1	x2
a	2	4
b	8	2
c	9	3
d	1	5
e	8,5	1



Clustering, Exemplo 2:

- ✓ Vamos aplicar o algoritmo **k-means** com dois centroides iniciais em $c1=(2,8)$ e $c2=(8,8)$

Amostra	x1	x2
a	2	4
b	8	2
c	9	3
d	1	5
e	8,5	1



Clustering, Exemplo 2:

- ✓ Na primeira iteração do **k-means** obtemos $c1=(1.5,4.5)$ e $c2=(8.5,2)$ e o seguinte agrupamento :

Amostra	x1	x2
a	2	4
b	8	2
c	9	3
d	1	5
e	8,5	1

