Ficha de Trabalho 5: Técnicas de Agrupamento de Dados (*Clustering*)

-Resolução-

***Objetivo**: Pretende-se promover a aquisição de conhecimentos e desenvolvimento de competências relativas aos **fundamentos** de algumas técnicas utilizadas para agrupar dados (Clustering & Data Mining)*

1) Considere dois pontos representados por $x_1=(1,1)$ e $x_2=(3,3)$. Determine a distância entre estes dois pontos utilizando as seguintes medidas: Euclidiana, Pombalina (ou *City Block*), Chebychev, Minkowski (utilizando a raiz quadrada) (T/**P**)

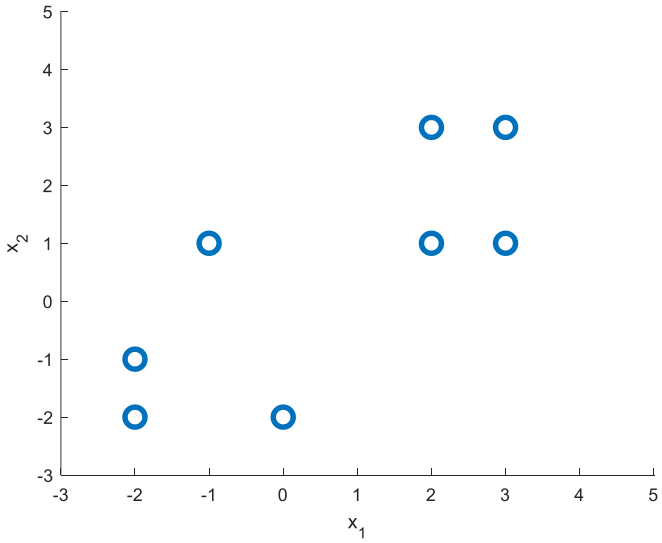
   Sugestão: utilize a função pdist do Matlab.

2) Considere o seguinte conjunto de dados, representado na Tabela 1:

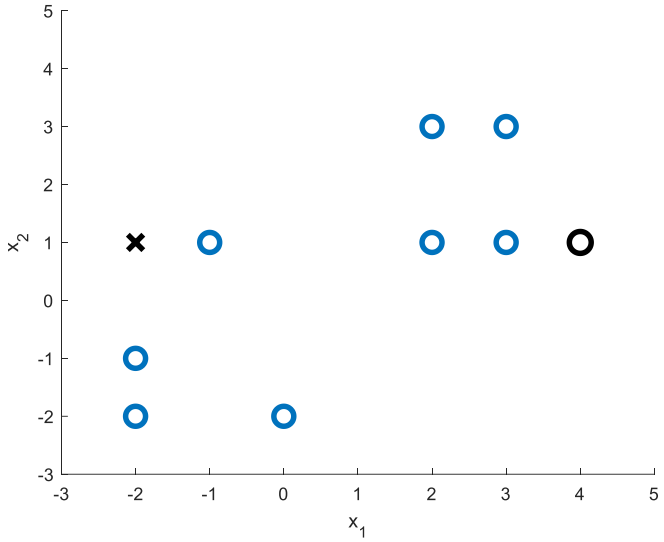
Tabela 1: Conjunto de Dados 1

# Amostra	(x_1, x_2)
1	(-2,-2)
2	(3,3)
3	(-1,1)
4	(3,1)
5	(-2,-1)
6	(2, 3)
7	(0,-2)
8	(2,1)

   i) Represente os pontos num gráfico (T/P).
   R:



   ii) Considere agora que se pretende agrupar os dados apresentados na Tabela 1 em dois grupos (clusters). Assume-se os dois centroides iniciais apresentados na Tabela 1. Represente os centroides no gráfico anterior utilizando um símbolo diferente das amostras.
   R:



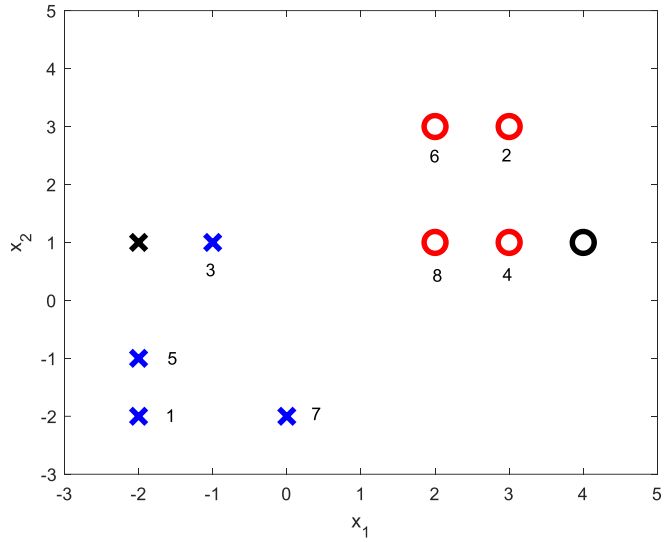
iii) Complete a Tabela 2 calculando a distância Euclidiana entre os pontos da amostra e os dois centroides.

iv) Com base na minimização das distâncias calculadas classifique os pontos no cluster C1 ou C2.

Tabela 2: Conjunto de Dados 1 com centroides iniciais

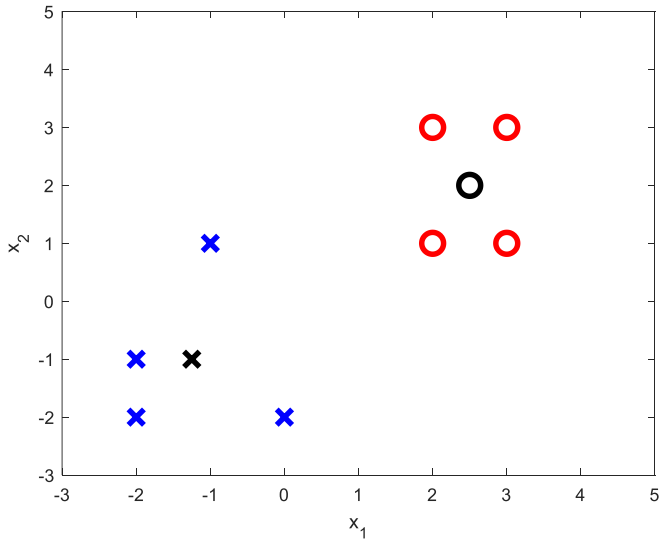
	Centroides Iniciais	***c1*** (-2 1)	***c2*** (4,1)	***Clusters***
# Amostra	***(x1, x2)***	***dist***	***dist***	***C1 ou C2?***
1	(-2,-2)	3.00	6.71	C1
2	(3,3)	5.39	2.24	C2
3	(-1,1)	1.00	5.00	C1
4	(3,1)	5.00	1.00	C2
5	(-2,-1)	2.00	6.32	C1
6	(2, 3)	4.47	2.83	C2
7	(0,-2)	3.61	5.00	C1
8	(2,1)	4.00	2.00	C2
	Centroides Novos	***c1*** (-1.25,-1)	***c2*** (2.5,2.0)	
	SSE	6,75	4,50	

v) Represente um gráfico diferenciando os pontos de acordo com o cluster a que pertencem.



vi) Determine os novos valores para os dois centroides conforme o agrupamento feito.

vii) Represente a nova localização dos centroides.



viii) Determine a soma dos erros quadráticos (SSE) para os dois clusters.

ix) Repita os cálculos para a nova iteração e preencha a Tabela 3

Tabela 3: Conjunto de Dados 1 ao fim de uma iteração

	Centroides Iniciais	***c1*** (-1.25,-1)	***c2*** (2.5,2.0)	***Clusters***
# Amostra	***(x1, x2)***	***dist***	***dist***	***C1 ou C2?***
1	(-2,-2)	1.25	6.02	C1
2	(3,3)	5.84	1.12	C2
3	(-1,1)	2.02	3.64	C1
4	(3,1)	4.70	1.12	C2
5	(-2,-1)	0.75	5.41	C1
6	(2, 3)	5.15	1.12	C2

7	(0,-2)	1,60	4.72	C1
8	(2,1)	3.82	1.12	C2
	Centroides	c_1	c_2	
	Novos	(-1.25,-1)	(2.5,2.0)	
	SSE	2.19	1.25	

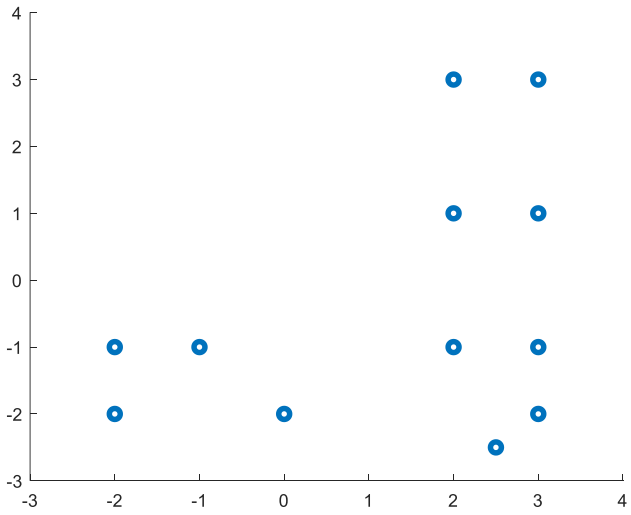
3) Considere o seguinte conjunto de dados com duas dimensões, representado na Tabela 4:

Tabela 4: Conjunto de Dados 2

# Amostra	(x_1, x_2)
1	(-2,-2)
2	(-1,1)
3	(-2,-1)
4	(0,-2)
5	(2,1)
6	(2, 3)
7	(3,1)
8	(3,3)
9	(3,-2)
10	(3,-1)
11	(2,-1)
12	(2.5,-2.5)

i) Represente os pontos num gráfico (T/P).

R:



ii) Aplique o algoritmo k-médias (k-means) com três clusters e confirme que o resultado obtido está de acordo com a seguinte figura e Tabela 4.

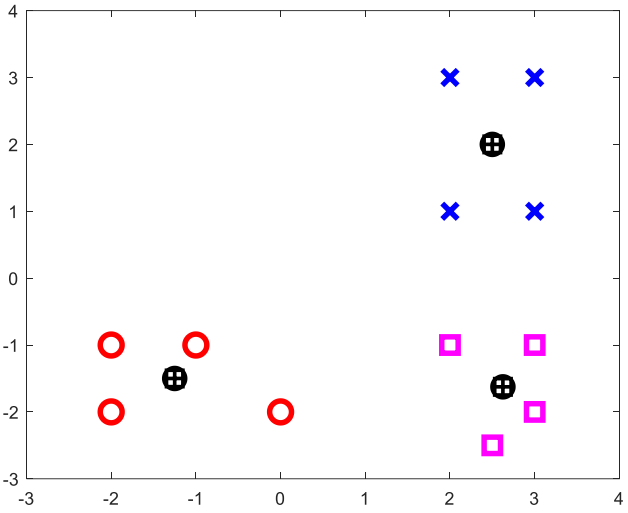


Tabela 5: Resultado do k-means para os dados da Tabela 4

# Cluster	***(c_1, c_2)***	***(x_1, x_2)***
1	(-1.25,-1.5)	(-2,-2)
		(-1,1)
		(-2,-1)
		(0,-2)
2	(2.5,2.0)	(2,1)
		(2,3)
		(3,1)
		(3,3)
3	(2.63,-1.63)	(3,-2)
		(3,-1)
		(2,-1)
		(2.5,-2.5)

iii) Calcule o valor da métrica Silhueta para a primeira amostra do primeiro cluster (-2,2) e para a primeira amostra do terceiro cluster e confirme se os valores obtidos são $S_{1,1}$=0.8945 e $S_{3,1}$= 0.933.

R:

a_1=2.33 ; $b_{1,2}$=37.5; $b_{1,3}$=22.125

b_{11}=min($b_{1,2}$, $b_{1,3}$)= min(37.5, 22.125)=22.125

max(a_1,b_{11})= max(2,33, 22.125)=22.125

$$S_{11} = \frac{b_{11} - a_1}{\max(a1, b11)} = \frac{22.125 - 2.33}{22.125} = 0.8945$$