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

***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 para confirmar os valores.

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).

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

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

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

Tabela 2: Conjunto de Dados 1 com centroides iniciais

	Centroides Iniciais	c_1 (-2 1)	c_2 (4,1)	***Clusters***
# Amostra	(x_1, x_2)	***dist***	***dist***	***C1 ou C2?***
1	(-2,-2)			
2	(3,3)			
3	(-1,1)			
4	(3,1)			
5	(-2,-1)			
6	(2, 3)			
7	(0,-2)			

8	(2,1)			
	Centroides Novos	c_1	c_2	
	SSE			

v) Represente um gráfico diferenciando os pontos de acordo com o cluster a que pertencem. (T/P)
vi) Determine os novos valores para os dois centroides conforme o agrupamento feito. (T/P)
vii) Represente a nova localização dos centroides. (T/P)
viii) Determine a soma dos erros quadráticos (SSE) para os dois clusters. (T/P)
ix) Repita os cálculos para a nova iteração e preencha a Tabela 3. (T/P)

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

	Centroides	c_1	c_2		
		(-1.25,-1)	(2.5,2.0)		***Clusters***
# Amostra	(x_1, x_2)	***dist***	***dist***		***C1 ou C2?***
1	(-2,-2)				
2	(3,3)				
3	(-1,1)				
4	(3,1)				
5	(-2,-1)				
6	(2, 3)				
7	(0,-2)				
8	(2,1)				
	Centroides Novos	c_1	c_2		
	SSE				

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)

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 5. (T/P)

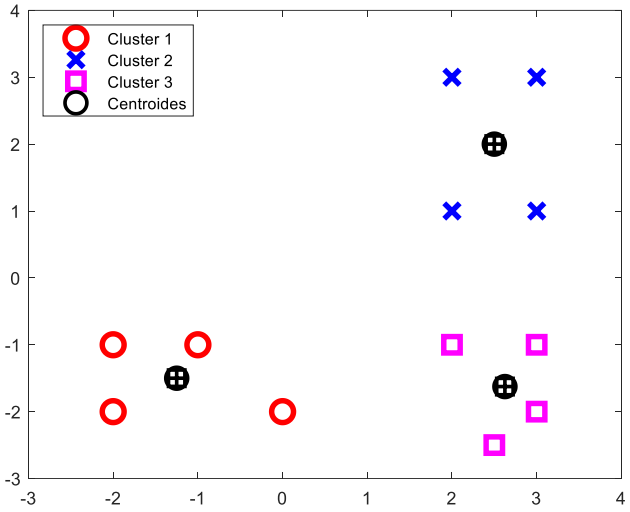

Tabela 5: Resultado do k-means para a os dados da Tabela 4

# 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. (T/P)