2. EJERCICIOS: Clustering

1. **Coste computacional:** Dado un conjunto de N instancias, se desea obtener la agrupación formada por k clusters (sub-conjuntos) que sea óptima según un criterio de proximidad dado.

   a) ¿Cuántas agrupaciones posibles hay? es decir ¿cuántas formas distintas hay de obtener k sub-conjuntos?

   b) Dado que deseamos buscar la mejor de las agrupaciones, si el criterio de proximidad tuviera un coste del orden $\mathcal{O}(N) \approx N^2$ ¿cuál sería el coste de un algoritmo que implemente una búsqueda exhaustiva?

   c) ¿Recomendarías el uso de heurísticos de búsqueda?

2. $\star$ **Algoritmo:** Escribir, en pseudo-código, el algoritmo de clustering aglomerativo empleando *complete-link* como distancia inter-grupal.

3. **Clustering jerárquico con ayuda de Weka:**

   a) Descargar *mtcars* dataset

   b) Realizar clustering jerárquico de los datos mediante Weka (u otras herramientas)

   c) ¿Cómo se ha calculado la distancia entre instancias?

   d) ¿Cómo se ha calculado la distancia inter-grupal?

   e) Visualizar dendograma

   f) Interpretar el dendograma. ¿Qué coches que pertenecen al mismo cluster? (define los clusters de modo que sean separables según una distancia umbral dada)

4. **Clustering jerárquico:** La tabla 1 muestra la matriz de distancias 5 instancias: $m_{i,j} = d(x_i, x_j)$

	x_1	x_2	x_3	x_4	x_5
x_1	0	2.3	3.4	1.2	3.7
x_2		0	2.6	1.8	4.6
x_3			0	4.2	0.7
x_4				0	4.4
x_5					0

Tabla 1: Matriz de distancias entre instancias

Se pide:

a) Clustering jerárquico con distancia single-link.

b) Clustering jerárquico con distancia complete-link.

5. Disponemos de un conjunto de 5 instancias en un espacio bidimensional $X_1 \times X_2$

	X_1	X_2
$\mathbf{x_1}$	1	1
$\mathbf{x_2}$	2	4
$\mathbf{x_3}$	3	2
$\mathbf{x_4}$	3	5
$\mathbf{x_5}$	4	4

Instancias

	$\mathbf{x_1}$	$\mathbf{x_2}$	$\mathbf{x_3}$	$\mathbf{x_4}$	$\mathbf{x_5}$
$\mathbf{x_1}$	0	3.16	2.24	4.47	4.24
$\mathbf{x_2}$		0	2.24	1.41	2.00
$\mathbf{x_3}$			0	3.00	2.24
$\mathbf{x_4}$				0	1.41
$\mathbf{x_5}$					0

Matriz de distancias

Tabla 2: Conjunto de instancias y distancia euclídea entre instancias

Se pide representar las instancias en el espacio de atributos (el plano $X_1 \times X_2$).

6. **k-means clustering:** para las instancias de la tabla 2 y distancia euclidea

a) k=2 con centroides iniciales $\mathbf{m_1} = \mathbf{x_1}$ y $\mathbf{m_2} = \mathbf{x_2}$

b) k=2 con centroides iniciales $\mathbf{m_1} = \mathbf{x_4}$ y $\mathbf{m_2} = \mathbf{x_5}$

c) k=3 con centroides iniciales $\mathbf{m_1} = \mathbf{x_2}$, $\mathbf{m_2} = \mathbf{x_4}$ y $\mathbf{m_3} = \mathbf{x_5}$

7. **Clustering jerárquico aglomerativo:** para las instancias de la tabla 2 y distancia inter-grupal *single link*. A la vista del dendograma resultante:

a) ¿Cómo se agrupan las instancias en k=2 clusters?

b) ¿Cómo se agrupan las instancias en k=3 clusters?

8. **Clustering jerárquico aglomerativo:** para las instancias de la tabla 2 y distancia inter-grupal *complete link*. Construye el dendograma.

9. **k-means clustering:** para las instancias de la tabla 2 con

   a) k=2 con centroides iniciales $\mathbf{m_1} = \mathbf{x_1}$ y $\mathbf{m_2} = \mathbf{x_2}$

   b) k=2 con centroides iniciales $\mathbf{m_1} = \mathbf{x_4}$ y $\mathbf{m_2} = \mathbf{x_5}$

   c) k=3 con centroides iniciales $\mathbf{m_1} = \mathbf{x_2}$, $\mathbf{m_2} = \mathbf{x_4}$ y $\mathbf{m_3} = \mathbf{x_5}$

10. **Evaluación:** Supongamos que, en relación a los datos de la tabla 2, dos algoritmos de clustering han devuelto las siguientes , en relación a los datos de la tabla 2:

$$P_1 = \{\mathcal{G}_1, \mathcal{G}_2\} \quad \text{tal que} \quad \mathcal{G}_1 = \{\mathbf{x_1}\}, \quad \mathcal{G}_2 = \{\mathbf{x_2}, \mathbf{x_3}, \mathbf{x_4}, \mathbf{x_5}\} \qquad (1)$$

$$P_2 = \{\mathcal{G}'_1, \mathcal{G}'_2\} \quad \text{tal que} \quad \mathcal{G}'_1 = \{\mathbf{x_1}, \mathbf{x_3}\}, \quad \mathcal{G}'_2 = \{\mathbf{x_2}, \mathbf{x_4}, \mathbf{x_5}\} \qquad (2)$$

   Se pide:

   a) Según *Sum of Squared Error* ¿qué partición tiene mayor cohesión?

   b) Calcula la anchura media global de Silhouette para cada partición.

11. ⋆ **Dendograma:** La tabla 3 muestra un conjunto, $\mathcal{X}$, de 5 datos caracterizados con 3 atributos. Ejemplo: los atributos para la instancia $\mathbf{x}_4$ son $(x_{4,1}, x_{4,2}, x_{4,3}) = (3, 10, 1)$

	Atributos		
	X_1	X_2	X_3
$\mathbf{x}_1$	2	4	6
$\mathbf{x}_2$	3	5	7
$\mathbf{x}_3$	1	1	4
$\mathbf{x}_4$	3	10	1
$\mathbf{x}_5$	3	9	2

Tabla 3: $\mathcal{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_5\}$

Se pide representar gráficamente el agrupamiento de los datos mediante clustering jerárquico aglomerativo empleando como distancia inter-grupal *average-link* y la métrica (3) para la distancia entre instancias.

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sum_{r=1}^{3} (x_{i,r} - x_{j,r})^2 \tag{3}$$

12. Disponemos de un conjunto de 5 instancias en un espacio bidimensional $X_1 \times X_2$

	X1	X2
1	2	0
2	4	4
3	1	1
4	2	4
5	2	2
6	2	3
7	3	4
8	3	3

Instancias

	1	2	3	4	5	6	7	8
1	0	4.47	1.41	4.00	2.00	3.00	4.12	3.16
2		0	4.24	2.00	2.83	2.24	1.00	1.41
3			0	3.16	1.41	2.24	3.61	2.83
4				0	2.00	1.00	1.00	1.41
5					0			
6						0		
7							0	
8								0

Matriz de distancias

Tabla 4: Conjunto de instancias y distancia euclídea entre instancias

Se pide:

a) Completar la matriz de distancia euclídea entre instancias.

b) Se pide representar las instancias en el espacio de atributos.

13. ⋆ **k-means clustering:** para las instancias de la tabla 4 con:

a) k=2 con centroides iniciales $\mathbf{m_1} = \mathbf{x_1}$ y $\mathbf{m_2} = \mathbf{x_2}$

b) k=2 con centroides iniciales $\mathbf{m_1} = \mathbf{x_4}$ y $\mathbf{m_2} = \mathbf{x_5}$

c) k=3 con centroides iniciales $\mathbf{m_1} = \mathbf{x_2}$, $\mathbf{m_2} = \mathbf{x_4}$ y $\mathbf{m_3} = \mathbf{x_7}$

14. ⋆ **Evaluación:** Supongamos que, en relación a los datos de la tabla 4, dos algoritmos de clustering han devuelto las siguientes particiones:

$$P_1 = \{\mathcal{G}_1, \mathcal{G}_2\} \quad \text{tal que} \quad \mathcal{G}_1 = \{\mathbf{x_1}, \mathbf{x_3}\}, \quad \mathcal{G}_2 = \{\mathbf{x_2}, \mathbf{x_4}, \mathbf{x_5}, \mathbf{x_6}, \mathbf{x_7}, \mathbf{x_8}\} \tag{4}$$

$$P_2 = \{\mathcal{G}'_1, \mathcal{G}'_2\} \quad \text{tal que} \quad \mathcal{G}'_1 = \{\mathbf{x_1}, \mathbf{x_3}, \mathbf{x_5}\}, \quad \mathcal{G}'_2 = \{\mathbf{x_2}, \mathbf{x_4}, \mathbf{x_6}, \mathbf{x_7}, \mathbf{x_8}\} \tag{5}$$

Se pide:

a) Según *Sum of Squared Error* ¿qué partición tiene mayor cohesión?

b) Calcula la anchura media global de Silhouette para cada partición.