Instrucciones: En esta guía revisaremos los contenidos básicos de clusters y del algoritmo de $k-$means.

$k-$means desde cero

Usaremos el algoritmo de $K-$medias y el cuadrado de la distancia euclídea para agrupar los siguientes 8 datos en $K = 3$ clusters.

$x_1 = (2, 8), x_2 = (2, 5), x_3 = (1, 2), x_4 = (5, 8), x_5 = (7, 3), x_6 = (6, 4), x_7 = (8, 4), x_8 = (4, 7)$

Sugerencia: graficar los puntos.

Matriz de similitud: distancia euclídea

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
$x_{(1)}$	0	3.0000	6.0828	3.0000	7.0711	5.6569	7.2111	2.2361
x_2	3.0000	0	3.1623	4.2426	5.3852	4.1231	6.0828	2.8284
x_3	6.0828	3.1623	0	7.2111	6.0828	5.3852	7.2801	5.8310
x_4	3.0000	4.2426	7.2111	0	5.3852	4.1231	5.0000	1.4142
x_5	7.0711	5.3852	6.0828	5.3852	0	1.4142	1.4142	5.0000
x_6	5.6569	4.1231	5.3852	4.1231	1.4142	0	2.0000	3.6056
x_7	7.2111	6.0828	7.2801	5.0000	1.4142	2.0000	0	5.0000
x_8	2.2361	2.8284	5.8310	1.4142	5.0000	3.6056	5.0000	0

1. Asumamos que los puntos x_2, x_5 y x_6 son elegidos como centros iniciales. Realizar un paso del algoritmo y reportar los nuevos centros.
2. Reportar el valor de $W(\mathcal{C})$ antes y después de la iteración.
3. Representar, en un mismo archivo (gráfico), los plots coloreados según la pertenencia de los puntos a los clusters, antes y después de la iteración.

Aplicación

Los exoplanetas son planetas fuera del Sistema Solar. El primero de este tipo fue descubierto en 1995 por Mayor y Queloz (1995). El planeta, similar en masa a Júpiter, se encontró orbitando una estrella relativamente ordinaria, 51 Pegasus. En el período intermedio se han descubierto más de cien exoplanetas, casi todos detectados indirectamente, utilizando la influencia gravitacional que ejercen sobre sus estrellas centrales asociadas. Las propiedades de los exoplanetas encontradas hasta ahora parecen desafiar la teoría del desarrollo planetario construida para los planetas del sistema solar.

Los exoplanetas no se parecen en nada a los nueve planetas locales que conocemos tan bien. Un primer paso en el proceso de comprensión de los exoplanetas podría

ser tratar de clasificarlos con respecto a sus propiedades conocidas y este será el objetivo en este ejercicio. Los datos del archivo *HSAUR2::planets* contienen la masa (en Júpiter masa, mass), el período (en días terrestres, period) y la excentricidad (eccen) de los exoplanetas descubiertos hasta octubre de 2002.

4. Cuando las variables están en escalas muy diferentes, se necesitará usar alguna forma de estandarización. Explorar el comando *scatterplot3d* para realizar un gráfico tridimensional usando las observaciones escaladas (**scale**).

5. Aplicar, a los datos escalados, el comando *kmeans* usando 4 centros y obtener el valor de $W(\mathcal{C})$. Mediante el comando *scatterplot3d* inspeccione visualmente los grupos generados.

6. Utilizar el método Elbow para identificar la cantidad de clusters. ¿Cuántos grupos le sugiere?

7. ¿Cuáles son sus conclusiones?

Bonus 1

De las transparencias de clase

8. Probar la igualdad (1).

9. A partir de la igualdad (2) ¿se puede explicar cómo se comporta $W(\mathcal{C})$ a medida que crece la cantidad de grupos?

Bonus 2

K−means: Probaremos las igualdad (2) y (3) de las transparencias de clase

10. Dados $Y_1, \ldots, Y_N$, probar que

$$\frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2 = \frac{1}{2}\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(Y_i - Y_j\right)^2$$

11. supongamos que tenemos las observaciones $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ y consideremos los K clusters $\mathcal{C}_1, \ldots, \mathcal{C}_K$ tales que

    (I) $\#\{\mathcal{C}_j\} > 0$

(II) $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset$

(III) $\cup_{i=1}^{K} \mathcal{C}_i = \{\mathbf{x}_1 \ldots, \mathbf{x}_n\}$

y además $n_k = \#\mathcal{C}_k$. Llamemos al centro del grupo $\mathcal{C}_k$, $\bar{\mathbf{x}}_k = (\bar{x}_{1k}, \ldots, \bar{x}_{pk})$, donde

$$\overline{\mathbf{x}}_k = \frac{1}{n_k} \sum_{\ell : \mathbf{x}_\ell \in \mathcal{C}_k} \mathbf{x}_\ell \, .$$

Usando 11., probar que

$$\frac{1}{n_k} \sum_{i,j : \mathbf{x}_i, \mathbf{x}_j \in \mathcal{C}_k} \|\mathbf{x}_i - \mathbf{x}_j\|^2 = 2 \sum_{i : \mathbf{x}_i \in \mathcal{C}_k} \|\mathbf{x}_i - \overline{\mathbf{x}}_k\|^2$$

y

$$W(\mathcal{C}) = \sum_{k=1}^{K} \sum_{i : \mathbf{x}_i \in \mathcal{C}_k} \|\mathbf{x}_i - \overline{\mathbf{x}}_k\|^2 \, .$$