

Problema 1

Verifica la igualdad que vimos en la clase:

$$\frac{1}{2} \sum_{k=1}^K \sum_{i:g(i)=k} \sum_{j:g(j)=k} \|x_i - x_j\|^2 = \sum_{k=1}^K N_k \sum_{i:g(i)=k} \|x_i - \mu_k\|^2 \text{ con } \mu_k = \text{promedio}\{x_i : g(i) = k\}$$

donde N_k es el número de elementos en cluster k . Puedes limitarte al caso cuando $x \in R$.

Sol.

Para aligerar la notación tomaremos $\sum_{i:g(i)=k} = \sum_i$ y lo mismo para j . Comenzamos enfocandonos en la suma sobre j de lado izquierdo

$$\begin{aligned} \sum_j \|x_i - x_j\|^2 &= \sum_j |x_i - x_j|^2 = \sum_j |x_i - \mu_k - x_j + \mu_k|^2 = \sum_j ((x_i - \mu_k) - (x_j - \mu_k))^2 \\ &= \sum_j [(x_i - \mu_k)^2 - 2(x_i - \mu_k)(x_j - \mu_k) + (x_j - \mu_k)^2] \end{aligned}$$

Donde usamos que, al estar en R , la norma es el valor absoluto y la propiedad del valor absoluto $|a|^2 = (a)^2$. Ahora, tomando en cuenta que x_i y μ_k no dependen del índice j , se cumple que $\sum_j (x_i - \mu_k)^2 = N_k (x_i - \mu_k)^2$, por lo que:

$$= N_k (x_i - \mu_k)^2 - 2(x_i - \mu_k) \sum_j (x_j - \mu_k) + \sum_j (x_j - \mu_k)^2$$

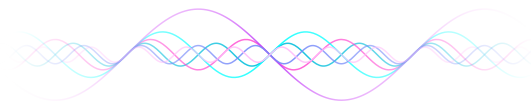
Ahora, tomamos la suma sobre i

$$\begin{aligned} \sum_i \sum_j \|x_i - x_j\|^2 &= N_k \sum_i (x_i - \mu_k)^2 - 2 \sum_i (x_i - \mu_k) \sum_j (x_j - \mu_k) + N_k \sum_j (x_j - \mu_k)^2 \\ &= 2N_k \sum_i (x_i - \mu_k)^2 - 2 \sum_i (x_i - \mu_k) \sum_j (x_j - \mu_k) \end{aligned}$$

Pero $\sum (x_i - \mu_k) = k \left(N_k \left(\frac{\sum_i x_i}{N_k} \right) - N_k \mu_k \right) = (N_k \mu_k - N_k \mu_k) = 0$

De modo que:

$$\sum_i \sum_j \|x_i - x_j\|^2 = 2N_k \sum_i (x_i - \mu_k)^2 + 0$$



Ahora, tomamos la suma sobre k

$$\sum_{k=1}^K \sum_i \sum_j \|x_i - x_j\|^2 = \sum_{k=1}^K 2N_k \sum_i \|x_i - \mu_k\|^2$$

Y, finalmente dividiendo entre 2:

$$\frac{1}{2} \sum_{k=1}^K \sum_i \sum_j \|x_i - x_j\|^2 = \sum_{k=1}^K N_k \sum_i \|x_i - \mu_k\|^2$$

Problema 2

Sea la fórmula de **average linkage** que se usa para un **Algoritmo Jerarquico Aglomerativo**

$$d(C_i, C_j) = \frac{1}{|C_i| \cdot |C_j|} \sum_{x \in C_i} \sum_{y \in C_j} d(x, y). \quad (1)$$

donde $|C_i|$ y $|C_j|$ representan la cardinalidad de los clusters C_i y C_j respectivamente, y $d(x, y)$ una medida de distancia entre x y y .

En cada paso, los clusters mas cercanos C_i y C_j se combinan en un nuevo cluster $C_i \cup C_j$. Muestra que la distancia del cluster $C_i \cup C_j$ a otro cluster C_k se puede calcular mediante la fórmula recursiva:

$$d(C_i \cup C_j, C_k) = \frac{|C_i| \cdot d(C_i, C_k) + |C_j| \cdot d(C_j, C_k)}{|C_i| + |C_j|}$$

Sol.

Nuevamente, para aligerar la notación haremos un pequeño cambio: $\sum_{x \in C_i} = \sum_i$ y lo mismo para j y k . Partimos del lado derecho de la expresión anterior y aplicamos la ecuación 1.

$$\begin{aligned} \frac{|C_i| \cdot d(C_i, C_k) + |C_j| \cdot d(C_j, C_k)}{|C_i| + |C_j|} &= \frac{1}{|C_i| + |C_j|} \left[\left(\frac{|C_i|}{|C_i| \cdot |C_k|} \right) \sum_i \sum_k d(x, y) + \left(\frac{|C_j|}{|C_j| \cdot |C_k|} \right) \sum_j \sum_k d(x, y) \right] \\ &= \frac{1}{|C_i \cup C_j|} \left[\frac{1}{|C_k|} \sum_i \sum_k d(x, y) + \frac{1}{|C_k|} \sum_j \sum_k d(x, y) \right] \\ &= \frac{1}{|C_i \cup C_j| \cdot |C_k|} \left[\sum_i \sum_k d(x, y) + \sum_j \sum_k d(x, y) \right] \\ &= \frac{1}{|C_i \cup C_j| \cdot |C_k|} \left[\sum_{i,j} \sum_k d(x, y) \right] = d(C_i \cup C_j, C_k) \end{aligned}$$