**El algoritmo k-means**

*K-means* es un algoritmo de clasificación no supervisada (clusterización) que agrupa objetos en *k* grupos basándose en sus características. El agrupamiento se realiza minimizando la suma de distancias entre cada objeto y el centroide de su grupo o cluster. Se suele usar la distancia cuadrática.

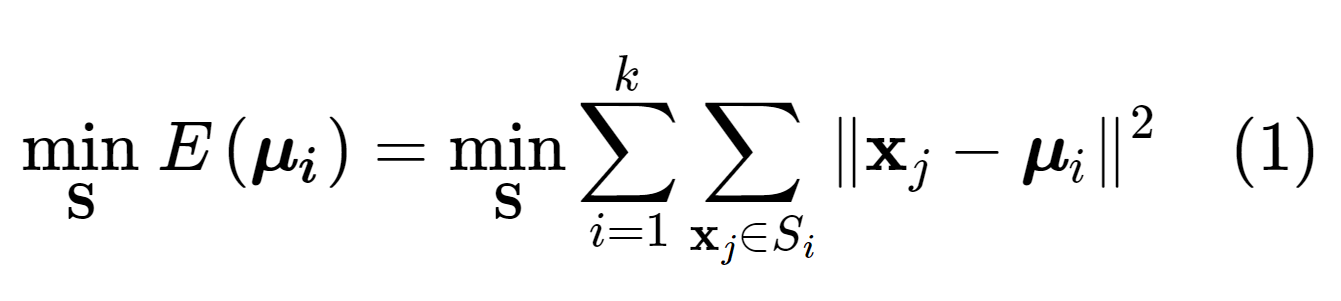
El algoritmo consta de tres pasos:

1. **Inicialización:** una vez escogido el número de grupos, *k*, se establecen *k* centroides en el espacio de los datos, por ejemplo, escogiéndolos aleatoriamente.
2. **Asignación objetos a los centroides:** cada objeto de los datos es asignado a su centroide más cercano.
3. **Actualización centroides:** se actualiza la posición del centroide de cada grupo tomando como nuevo centroide la posición del promedio de los objetos pertenecientes a dicho grupo.

Se repiten los pasos 2 y 3 hasta que los centroides no se mueven, o se mueven por debajo de una distancia umbral en cada paso.

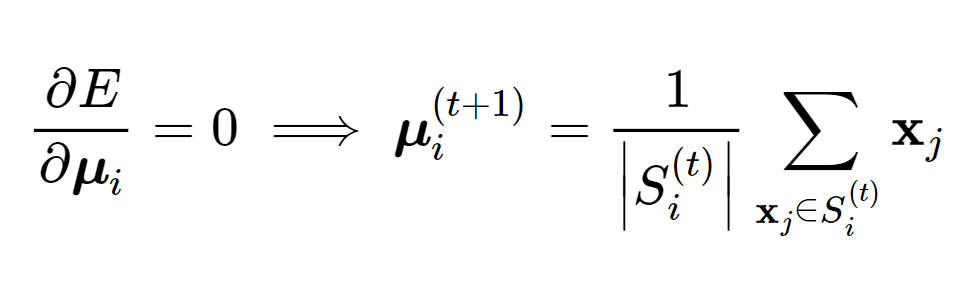
El algoritmo *k-means* resuelve un **problema de optimización**, siendo la función a optimizar (minimizar) la suma de las distancias cuadráticas de cada objeto al centroide de su cluster.

Los objetos se representan con vectores reales de dd dimensiones (x1,x2,…,xn)(x1,x2,…,xn) y el algoritmo *k-means* construye kk grupos donde se minimiza la suma de distancias de los objetos, dentro de cada grupo S={S1,S2,…,Sk}S={S1,S2,…,Sk}, a su centroide. El problema se puede formular de la siguiente forma:



donde SS es el conjunto de datos cuyos elementos son los objetos xjxj representados por vectores, donde cada uno de sus elementos representa una característica o atributo. Tendremos kk grupos o clusters con su correspondiente centroide μiμi.

En cada actualización de los centroides, desde el punto de vista matemático, imponemos la condición necesaria de extremo a la función E(μi)E(μi) que, para la función cuadrática (1)(1) es



y se toma el promedio de los elementos de cada grupo como nuevo centroide.

Las principales ventajas del método *k-means* son que es un método sencillo y rápido. Pero es necesario decidir el valor de kk y el resultado final depende de la inicialización de los centroides. En principio no converge al mínimo global sino a un mínimo local.