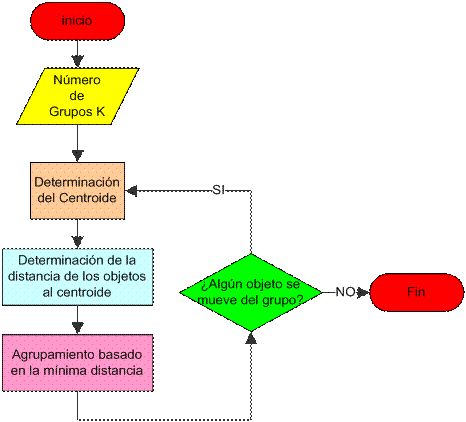
# 3.1 - diferentes pasos (sin matemática) -> pasos y ejemplificación ¿Y después del agrupamiento qué? (tenemos los grupos ya hechos y los puntos divididos en grupos y están clasificados)

Los pasos para llevar a cabo el procedimiento de agrupamiento k-means, se describe en la próxima imagen:

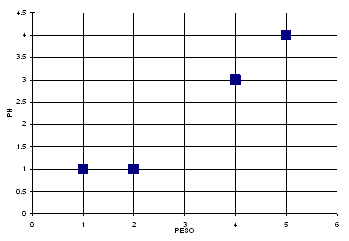


A continuación, vamos a explicar este procedimiento con un ejemplo práctico.

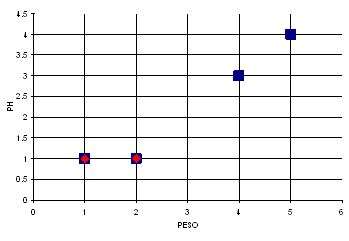
Supongamos que tenemos cuatro medicinas con dos atributos (peso e índice PH) como se muestra a continuación:

|  |  |  |
| --- | --- | --- |
| MEDICINA | PESO | ÍNDICE PH |
| A | 1 | 1 |
| B | 2 | 1 |
| C | 4 | 3 |
| D | 5 | 4 |

Cada medicina puede ser representada como un punto en un espacio de coordenadas:



Colocamos los centroides en las mismas coordenadas que las medicinas A y B:



Calculamos la matriz de distancias euclidianas, donde cada fila representa un centroide, y cada columna la distancia de cada punto a dicho centroide:

Ejemplo Numerico

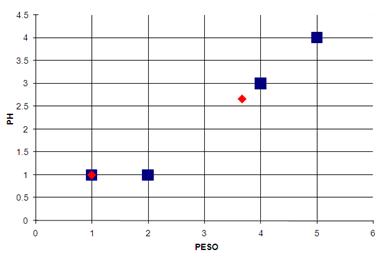
A partir de esto, representamos con una matriz de 0 y 1 que punto quedaría asociado a que grupo, quedando de la siguiente forma:

Ejemplo Numerico

Observando que el grupo 1 tiene asociado únicamente 1 punto, no cambiaría su centroide, pero en el caso del grupo 2 si que cambiaría el centroide ya que tiene 3 puntos asociados, por lo tanto, el siguiente paso es calcula dicho cambio de centroide:

Ejemplo Numerico

Quedando de la siguiente forma:



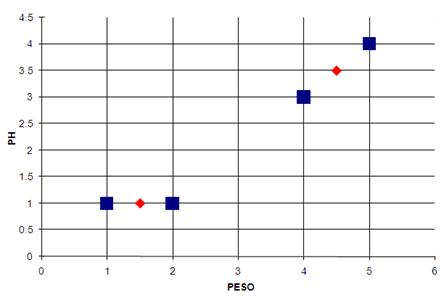
Con esto finalizaríamos la primera iteración, repetimos los pasos para la siguiente iteración:

Ejemplo Numerico

Ejemplo Numerico

Ejemplo Numerico

Obteniendo el siguiente resultado:



Tras la siguiente iteración:

Ejemplo Numerico

Ejemplo Numerico

Observamos que los puntos siguen perteneciendo a los mismos grupos que en la iteración anterior, por lo tanto, los centroides se han estabilizado y no es necesario continuar realizando iteraciones.

# 3.2 - Heurística para decidir los centroides

En el proceso de agrupamiento k-means estándar, para determinar la posición de los centroides que vamos a utilizar, se utilizan dos alternativas prácticas, la primera de ellas consiste en seleccionas K elementos, de nuestra lista de elementos, de forma aleatoria, y que estos sean nuestros centroides, la otra alternativa consiste en seleccionas los primeros K elementos de nuestra lista, de forma secuencial, y que estos sean nuestros centroides.

# 3.3 - Variaciones

El mayor problema con el que nos encontramos, como hemos nombrado en el punto anterior, se trata de la elección de los centroides, que al ser de forma aleatoria es bastante probable que la convergencia del algoritmo tome más tiempo del que debería.

Por esto motivo han ido apareciendo distintas variantes del algoritmo a lo largo del tiempo, nosotros vamos a centrarnos en la variante llamada k-means++, que pretende mejorar la elección de los centroides, para ello utiliza el siguiente procedimiento:

1. Escoger un centro de entre los puntos de datos de forma aleatoria.
2. Para cada punto x, calcular D(x), que es la distancia entre x y el centro más cercano que ya ha sido seleccionado.
3. Escoger un nuevo punto como nuevo centro, utilizando una distribución de probabilidad ponderada donde un punto x es escogido con la probabilidad proporcional a D(x)2.
4. Repetir paso 2 y 3 hasta que se hayan seleccionado k centros.
5. Ahora que los centros iniciales han sido elegidos, continuar utilizando k-means estándar.

Este método produce una mejora considerable en el error final de k-means. Aunque la selección inicial en el algoritmo toma tiempo extra, k-means converge muy rápidamente después de la selección de centroides y por lo tanto este algoritmo reduce el tiempo de cálculo.