K Means Algorithm - Theory

El algoritmo K Means es un tipo de algoritmo de aprendizaje NO supervisado que intenta agrupar grupos de clusters similares juntos dado un set de datos.

Ejemplos de típicos problemas de clustering

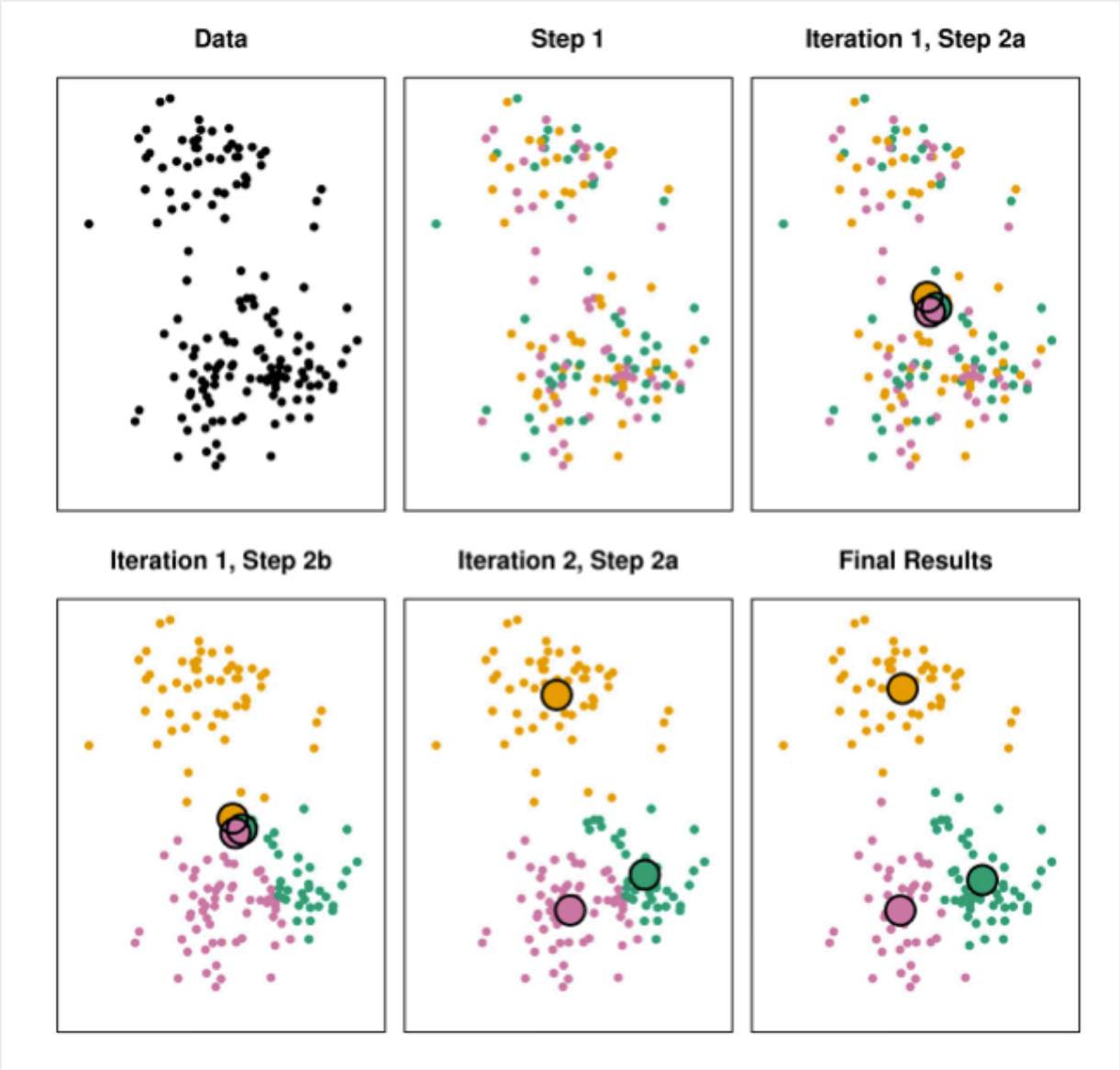
- Clusterizar documentos similares
- Clusterizar clientes basado en features
- Segmentación de mercado
- Identificar grupos psicológicos similares

El objetivo sobre todo es tratar de dividir los datos en distintos segmentos de grupos de tal manera que las observaciones (data points) compartan características similares a las similes que se encuentren en el mismo grupo.

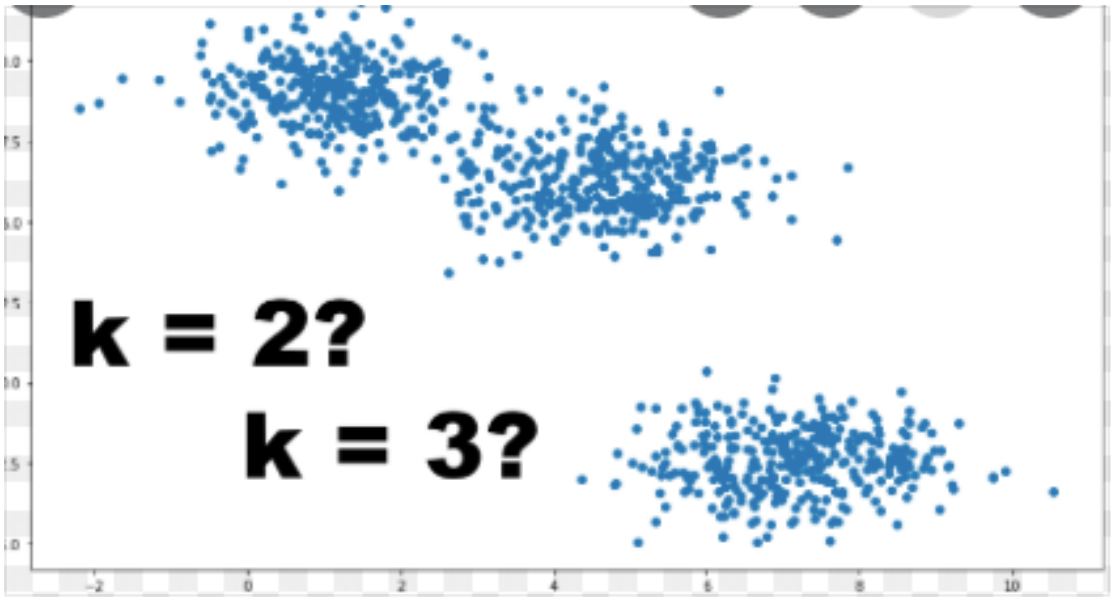
¿Cómo funciona el algoritmo en la práctica?

1. Se escoge un número de clusters esperados (K),
2. Aleatoriamente se asigna a cada punto un clúster,
3. Hasta que los clusters dejan de cambiar, repita lo siguiente:
    1. Por cada cluster, calcular el centroide de ese cluster tomando la media del vector de puntos en el cluster,
    2. Asignar a cada observación al cluster en donde el centroide es más

cercano.



Problema al escoger K

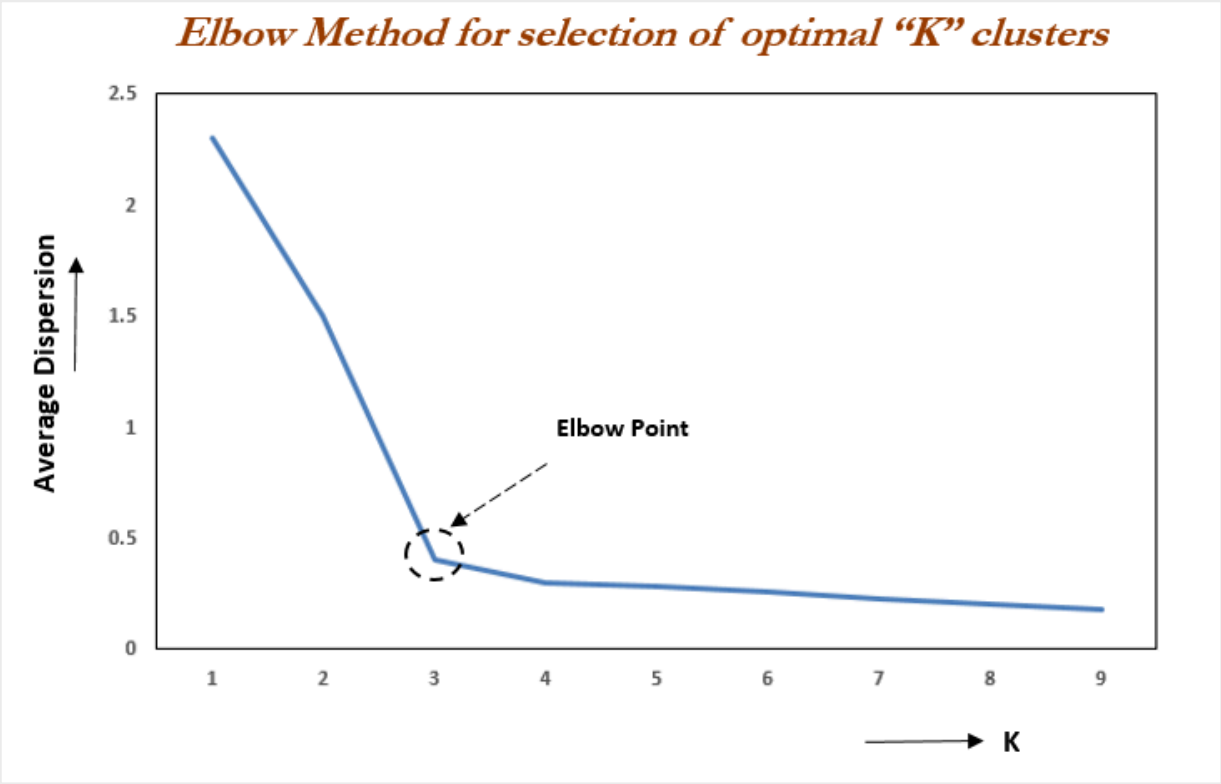


Escoger un valor adecuado para K no es trivial, una manera de hacerlo es el método del codo.

Pimrero que todo, se calcula la suma del valor cuadrado (SSE = Sum of Squared Error) para algunos valores de K (i.e 2,4,6,8, etc).
El SSEes definido como la suma de la distancia cuadrada entre cada miembro del cluster y su centroide

$$\sum_{i=1}^{n} \sum_{j=1}^{n} (x(j) - u(i))^2$$

Si se grafica el SSE, se verá que el error decrese como K se hace más grande, esto es porque cuando el número de clusters aumenta, debiesen hacerse más pequeños, por lo que la distorción es más pequeña.
La idea del métrodo del codo es escoger el valor de K donde veamos que el SSE decrese abdruptamente.


Elbow Method for selection of optimal "K" clusters
Average Dispersion
2.5
2
1.5
1
0.5
0
1
2
3
4
5
6
7
8
9
K
Elbow Point