

Agrupamiento K Means

Lectura Sugerida

Capítulo 10 de
Introduction to Statistical Learning
Gareth James

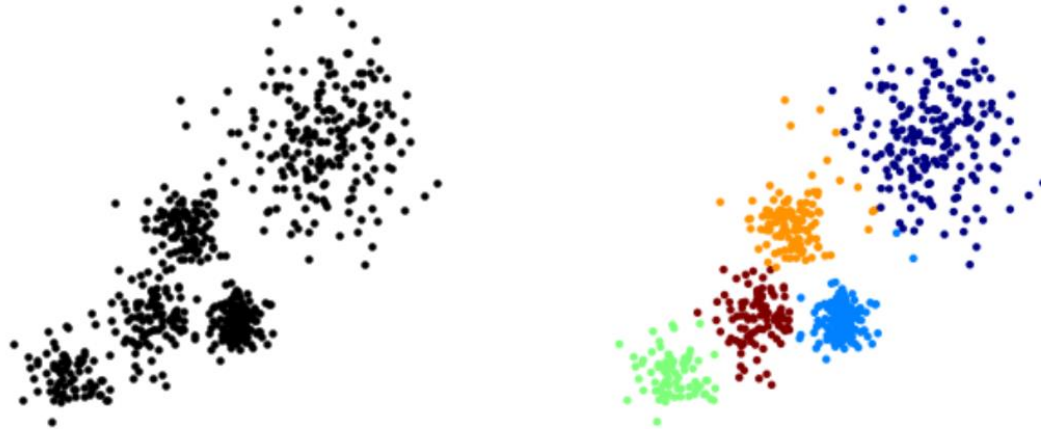
K Means

El agrupamiento K Means es un algoritmo de aprendizaje no supervisado que intentará agrupar grupos similares en sus datos. Entonces, ¿cómo se ve un problema típico de agrupamiento?

- Clústeres de Documentos similares
- Clústeres de Clientes basados en Características
- Segmentación de mercado
- Identificar grupos físicos similares.

K Means

- El objetivo general es dividir los datos en grupos distintos de modo que las observaciones dentro de cada grupo sean similares

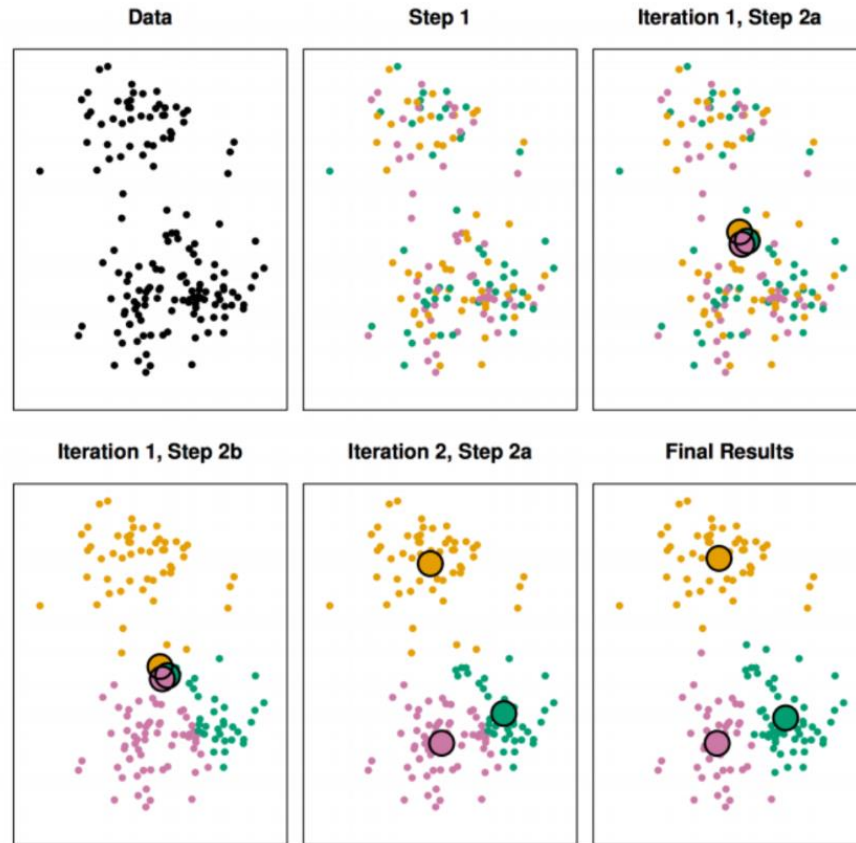


K Means

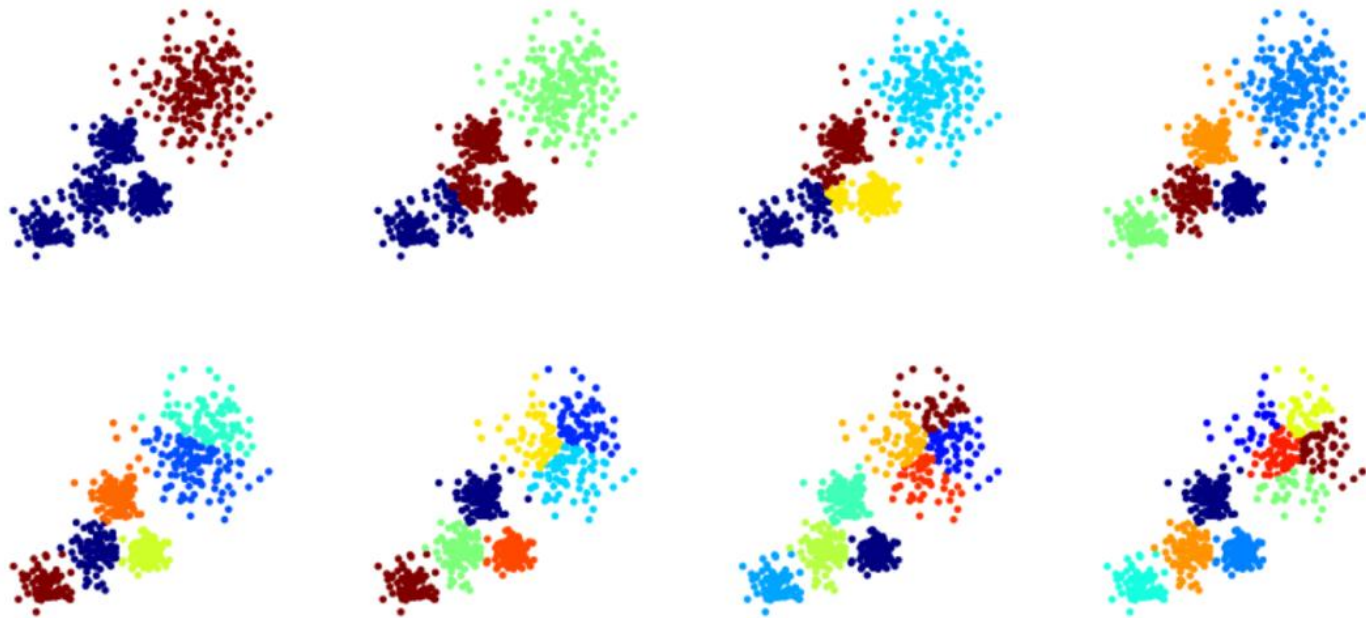
El Algoritmo K Means

- Elija un número de grupos "K"
- Asigna aleatoriamente cada punto a un grupo
- Hasta que los grupos dejen de cambiar, repita lo siguiente::
 - Para cada grupo, calcule el centroide del grupo tomando el vector promedio de puntos en el grupo
 - Asigne cada punto de datos al grupo para el cual el centroide es el más cercano

K Means



Elegir un valor de K



Elegir un valor de K

- No hay una respuesta fácil para elegir un "mejor" valor de K
- Una forma es el método del codo

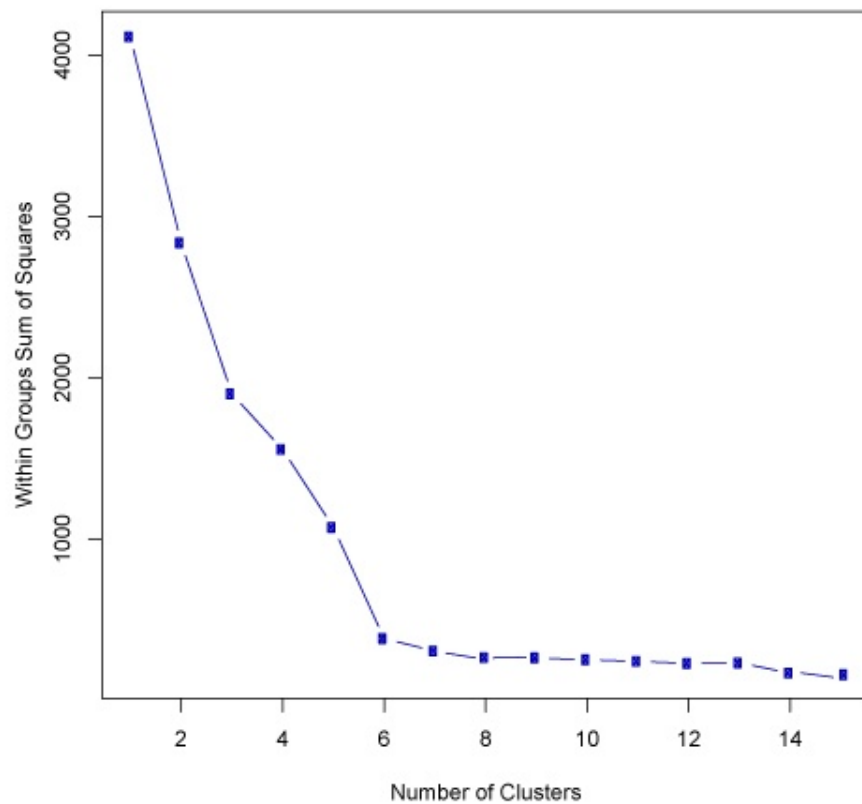
En primer lugar, calcule la suma de los errores al cuadrado (SSE) para algunos valores de k (por ejemplo, 2, 4, 6, 8, etc.).

El SSE se define como la suma de la distancia cuadrada entre cada miembro del grupo y su centroide.

Elegir un valor de K

- Si traza k contra el SSE, verá que el error disminuye a medida que k aumenta; esto se debe a que cuando aumenta el número de grupos, deberían ser más pequeños, por lo que la distorsión también es menor.
- La idea del método del codo es elegir la k en la que el SSE disminuye bruscamente.
- Esto produce un "efecto de codo" en el gráfico, como se puede ver en la siguiente imagen:

Elegir un valor de K



Elegir un valor de K

- ¡Pero no tome esto como una regla estricta al elegir un valor K!
- Mucho depende más del contexto de la situación exacta (conocimiento del dominio)