Aprendizaje No Supervisado

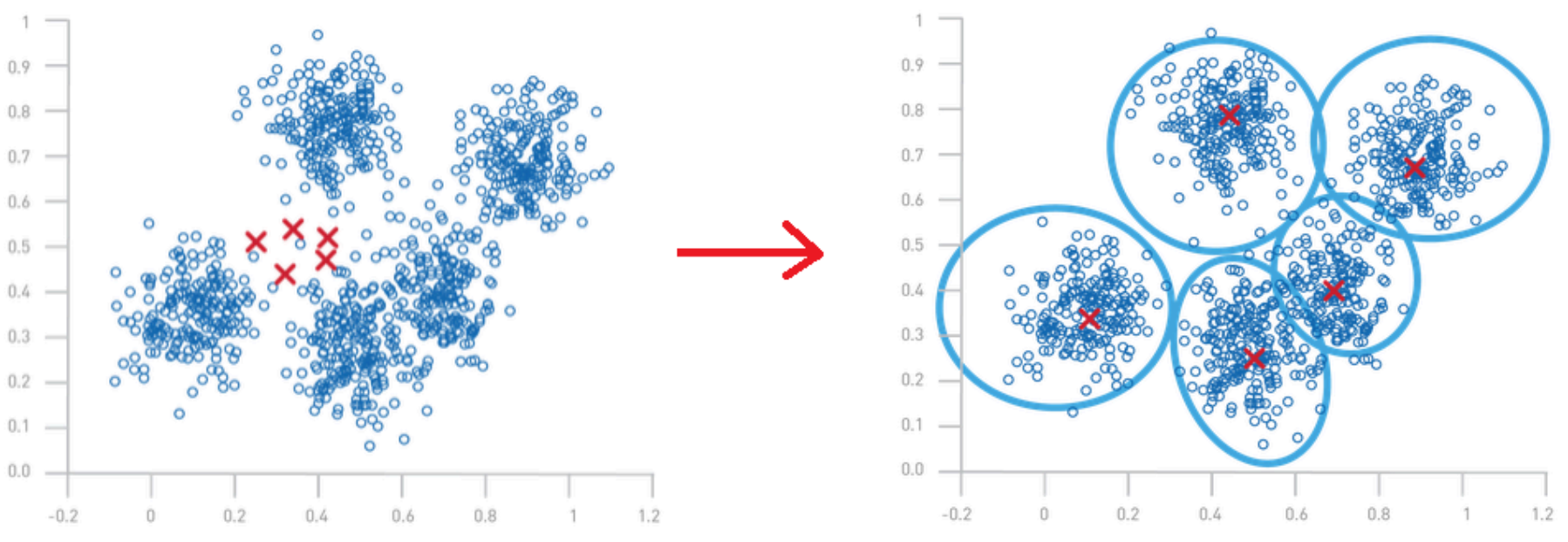
Análisis de Clustering o Agrupamientos

Tópicos

1. Algoritmo K-Means
2. Determinación del Número Óptimo de Grupos (clusters)

1. Algoritmo K-Means

Particiona n **observaciones** en k **grupos (clústeres)**, donde cada observación pertenece al grupo con el centroide más cercano. Es necesario especificar el número de grupos (k) de antemano.



Etapas del Algoritmo K-means

```
1. INICIALIZACIÓN:
   - Seleccionar k puntos aleatorios como centroides iniciales

2. ASIGNACIÓN:
   - Para cada punto de datos:
     • Calcular distancia a todos los centroides
     • Asignar el punto al grupo del centroide más cercano

3. ACTUALIZACIÓN:
   - Para cada grupo:
     • Calcular el nuevo centroide (media de todos los puntos del grupo)

4. REPETIR:
   - Repetir pasos 2 y 3 hasta que:
     • Los centroides no cambien significativamente, o
     • Se alcance el número máximo de iteraciones
```

Fórmula Matemática

La función objetivo de K-Means es minimizar la **inercia** (suma de distancias al cuadrado dentro de cada grupo):

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2$$

Donde:

- k = número de grupos
- C_i = conjunto de puntos en el grupo i
- μ_i = centroide del grupo i
- $||x - \mu_i||^2$ = distancia euclidiana al cuadrado

2. Determinación del Número Óptimo de Grupos

2.1 Método del Codo (Elbow Method)

Busca un equilibrio entre:

- **Pocos grupos**: Explicar insuficientemente la varianza
- **Muchos grupos**: Sobreajuste y poca generalización

Método:

1. Ejecutar K-means para diferentes valores de k (por ejemplo: de 2 a 10)
2. Calcular la **inercia** para cada k
3. Graficar inercia vs. k
4. Buscar el "codo", es decir, el punto donde la mejora se estabiliza.

2.2 Coeficiente de Silueta (Silhouette Score)

Mide qué tan bien está asignada cada muestra a su grupo, comparando:

- **Cohesión**: Qué tan cerca está de su propio grupo
- **Separación**: Qué tan lejos está del grupo más cercano

Fórmula para una muestra

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Donde:

- $a(i)$ = distancia media al resto de puntos en su mismo grupo
- $b(i)$ = distancia media al grupo más cercano

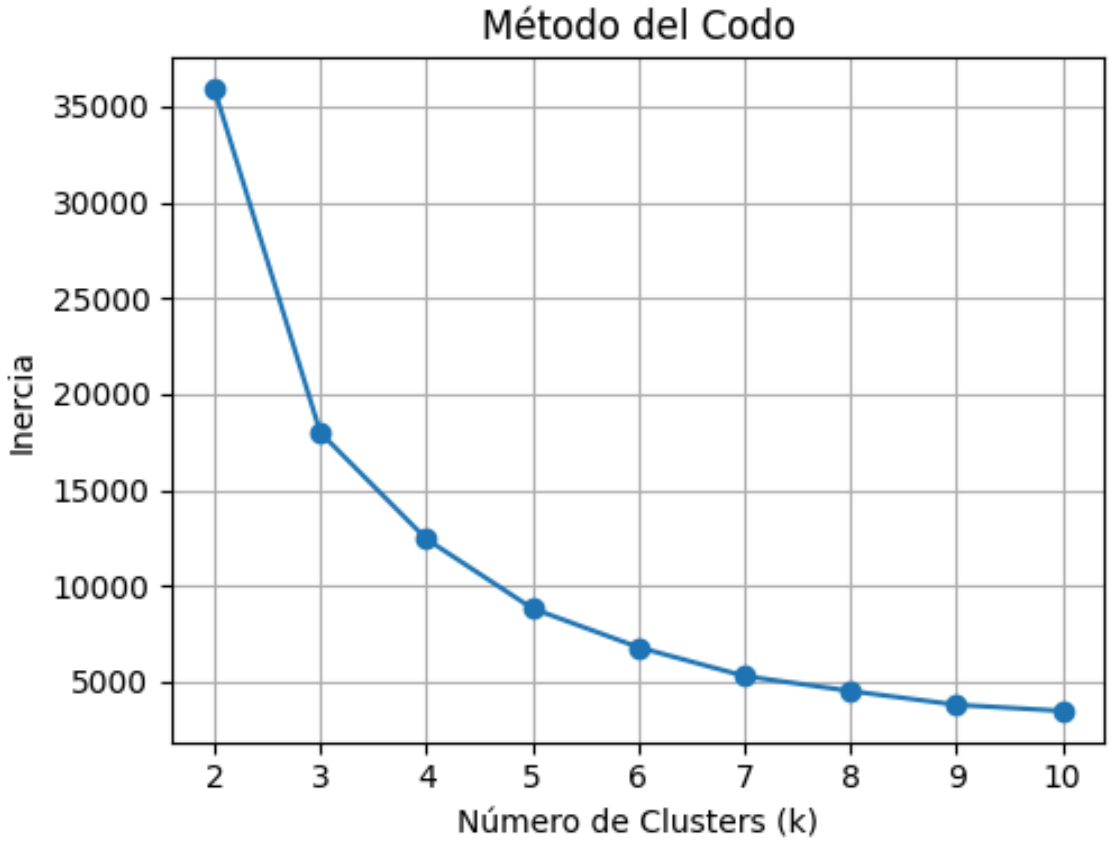
Interpretación

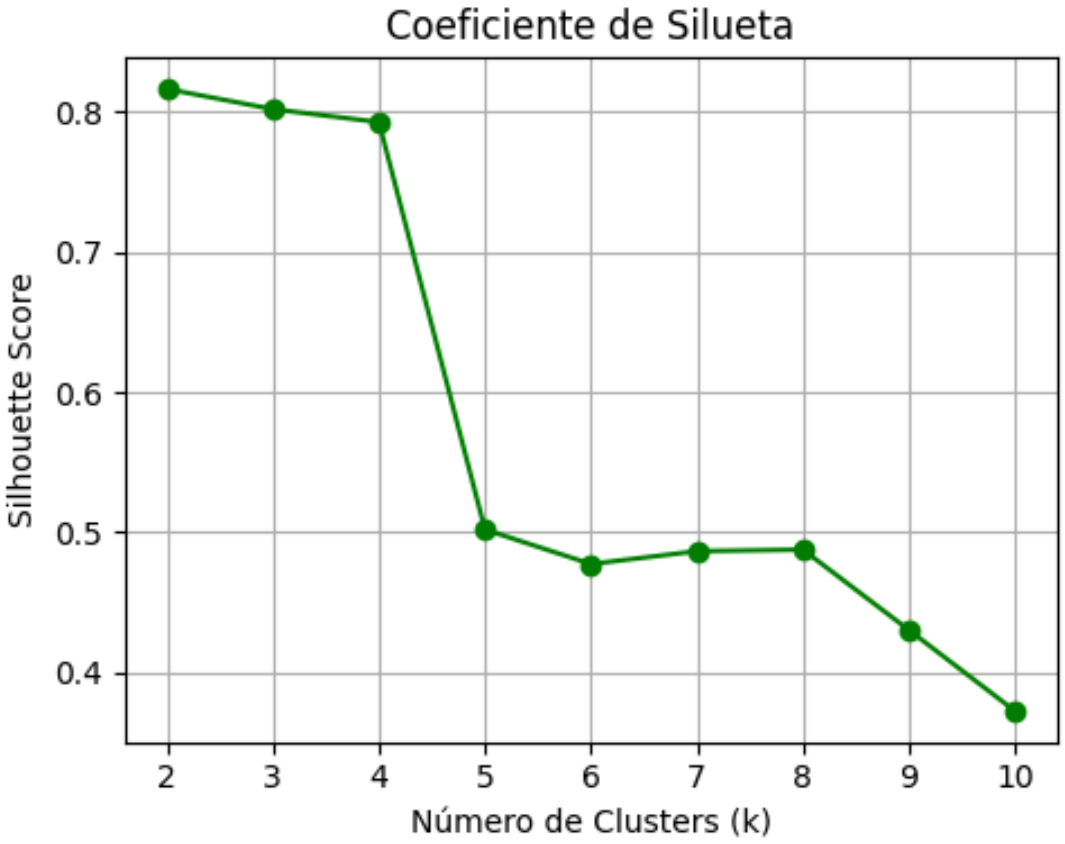
El coeficiente de silueta está en el rango $[-1, 1]$:

+1	Perfectamente asignado (lejos de otros grupos)
0	En la frontera entre dos grupos
-1	Probablemente mal asignado

Reglas generales:

0.7 - 1.0	Estructura fuerte
0.5 - 0.7	Estructura razonable
0.25 - 0.5	Estructura débil
< 0.25	Sin estructura sustancial


Método del Codo
Inercia
35000
30000
25000
20000
15000
10000
5000
2
3
4
5
6
7
8
9
10
Número de Clusters (k)


Coeficiente de Silueta
Silhouette Score
0.8
0.7
0.6
0.5
0.4
2
3
4
5
6
7
8
9
10
Número de Clusters (k)

Fin