Determinación del número óptimo de clústeres

La determinación del número óptimo de clusters es un paso crucial en la tarea de agrupamiento o clusterización.

Cada método tiene sus ventajas y desventajas, y la elección depende de las características de los datos y del algoritmo. En muchos casos, es útil combinar múltiples métodos para obtener una estimación más robusta del número óptimo de clusters.

Éstos son algunos de los métodos más comunes para estimar el número adecuado de clusters:

1. Método del Codo (Elbow Method)

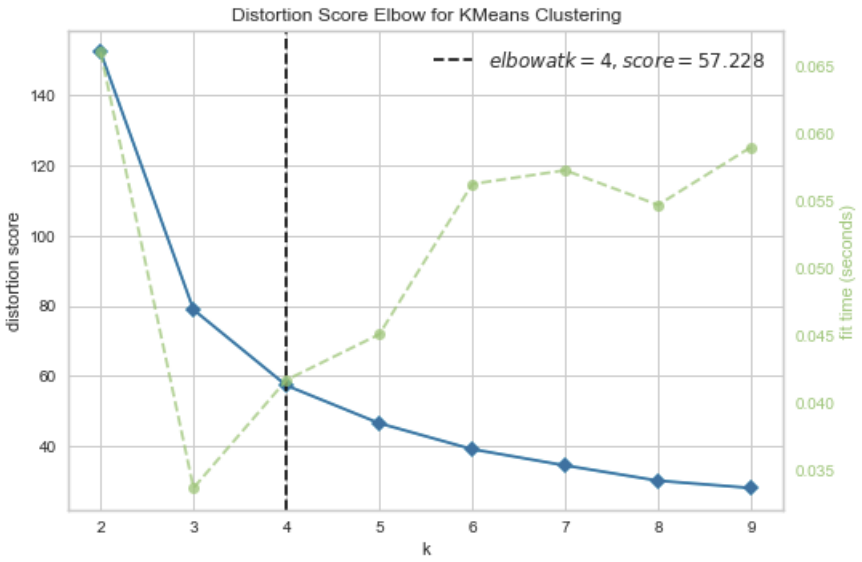
Este método implica graficar la suma de las distancias cuadráticas de los puntos al centro de cada cluster (Within-Cluster-Sum of Squares, WCSS, también conocido como *inercia*) contra el número de clusters.

La idea es que mientras más clusters se utilicen, menor será la inercia dentro de cada grupo, porque cada cluster contendrá menos puntos y, por ende, las distancias dentro del cluster serán menores. Sin embargo, al ser una distancia elevada al cuadrado, a medida que se aumenta el número de clusters, la reducción en el WCSS comienza a disminuir.

Pasos en el método del codo:

1. **Aplicar clustering para diferentes k:** Ejecutar el algoritmo de clustering (por ejemplo, K-means) para diferentes valores de k (número de clusters), típicamente en un rango de 1 a 10 o un rango similar.
2. **Calcular el WCSS:** Para cada k, calcular el WCSS de la siguiente forma:

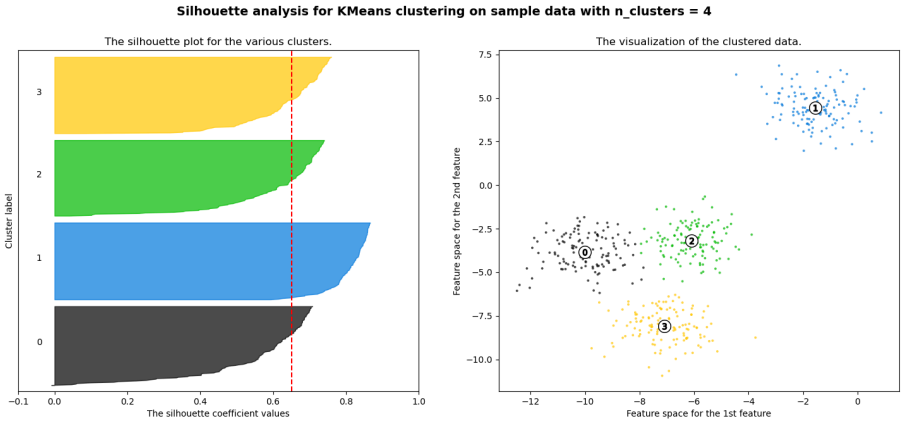
   $WCSS = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$
   donde:
   - C_i es el conjunto de puntos en el cluster i,
   - μ_i es el centroide del cluster i,
   - x es un punto en el cluster i,
   - $\|\cdot\|$ es la norma euclidiana.
3. **Graficar WCSS contra k:** Crear una gráfica donde el eje x represente el número de clusters (k) y el eje y sea el WCSS.
4. **Identificar el codo:** El punto donde la gráfica comienza a mostrar una reducción menos pronunciada en el WCSS se llama "codo" y suele indicar el número óptimo de clusters.



2. Índice de Silueta (Silhouette Score)

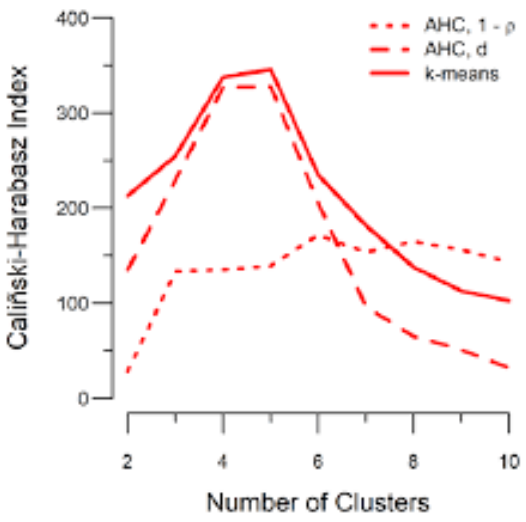
El índice de silueta evalúa la calidad de un agrupamiento basándose en la cohesión y la separación. La puntuación de silueta para un punto es una combinación de la distancia media a los puntos de su propio cluster y la distancia media a los puntos del cluster más cercano.

1. Ejecutar el algoritmo de agrupamiento con diferentes números de clusters.
2. Calcular la puntuación de silueta para cada k.
3. Graficar las puntuaciones frente al número de clusters.
4. El valor máximo de la puntuación indica el número óptimo de clusters.



3. Índice de Calinski-Harabasz (Calinski-Harabasz Index)

El índice de Calinski-Harabasz, también conocido como **criterio de varianza entre grupos**, mide la relación entre la dispersión **entre clusters** (BSS) y la dispersión **interna** (WCSS). Un valor más alto indica que los clusters están bien separados y son internamente compactos, lo que representa un buen agrupamiento.



Fórmula:

$CH(k) = \frac{B_k/(k-1)}{W_k/(n-k)}$

donde:

- B_k = *Between-Cluster Sum of Squares* (variabilidad entre clusters),
- W_k = *Within-Cluster Sum of Squares* (variabilidad dentro de los clusters),
- n= número total de muestras,
- k = número de clusters.

Pasos para utilizarlo:

1. Ejecutar el algoritmo de agrupamiento para distintos valores de kk.
2. Calcular el índice de Calinski-Harabasz para cada kk.
3. Graficar el valor del índice frente al número de clusters.
4. Seleccionar como número óptimo de clusters el valor de kk que **maximiza** el índice.

Ventajas:

- Computacionalmente eficiente.
- Interpretable directamente.
- Utiliza la misma información de dispersión ya considerada en otros métodos (como el WCSS/BSS del método del codo).

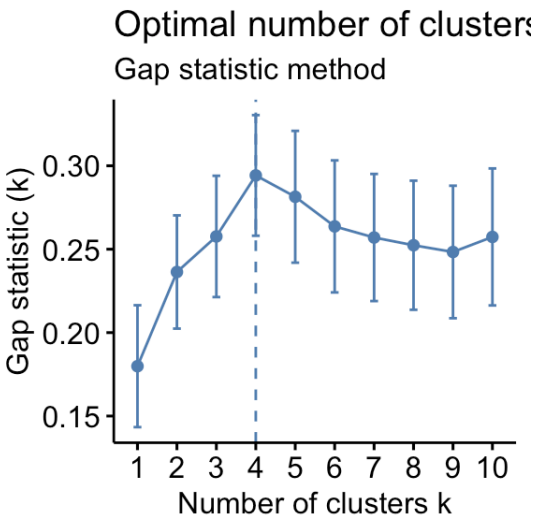
> Este índice es particularmente útil cuando ya se ha calculado WCSS y BSS, y puede integrarse de forma natural con análisis previos basados en varianza.

4. Método de la Brecha Estadística (Gap Statistic Method)

Este método compara la WCSS del agrupamiento observado con la WCSS esperada en una referencia generada aleatoriamente:

1. Ejecutar el algoritmo de agrupamiento para diferentes números de clusters.

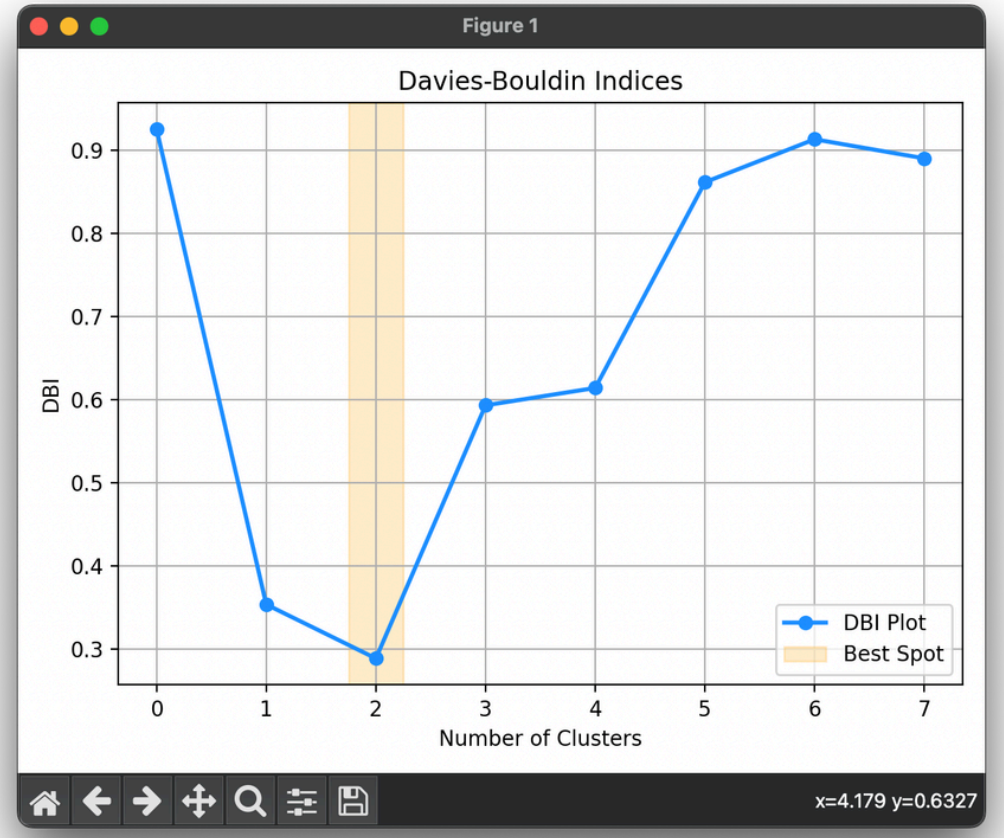
2. Generar un conjunto de datos de referencia mediante una distribución aleatoria.
3. Calcular el WCSS para ambos conjuntos (real y de referencia).
4. Comparar las diferencias entre la WCSS real y la de referencia para cada k.
5. El número óptimo de clusters es el valor k donde esta diferencia es máxima.



5. Criterio de Davies-Bouldin

Este índice mide la relación entre la distancia entre los clusters y el tamaño de los clusters mismos:

1. Ejecutar el algoritmo de agrupamiento para diferentes números de clusters.
2. Calcular el índice Davies-Bouldin para cada k.
3. Graficar los índices frente al número de clusters.
4. El valor mínimo indica el número óptimo de clusters.



6. Información de Criterio Bayesiano (BIC) / Información de Criterio de Akaike (AIC)

Estos métodos se aplican comúnmente en Modelos de Mezclas Gaussianas:

1. Ejecutar el algoritmo de mezclas gaussianas para diferentes números de clusters.
2. Calcular el BIC / AIC para cada k.
3. Graficar los valores frente al número de clusters.
4. El valor mínimo indica el número óptimo de clusters.

7. Índice Xie-Beni

Este índice mide la relación entre la dispersión interna de los clusters y la distancia mínima entre los centroides:

1. Ejecutar el algoritmo de agrupamiento para diferentes números de clusters.
2. Calcular el índice Xie-Beni para cada k.
3. Graficar los índices frente al número de clusters.
4. El valor mínimo indica el número óptimo de clusters.

8. Validación Cruzada

Para métodos basados en modelos (como mezclas gaussianas), se puede usar validación cruzada para determinar el k óptimo:

1. Dividir el conjunto de datos en entrenamiento y validación.
2. Aplicar el algoritmo para diferentes k en el conjunto de entrenamiento.
3. Evaluar el modelo en el conjunto de validación.
4. Seleccionar el k que maximice el rendimiento en el conjunto de validación.

9. Evaluación Visual (Métodos de Proyección)

A veces, una evaluación visual puede ayudar a determinar el número correcto de clusters:

1. Aplicar técnicas de reducción de dimensionalidad como PCA o t-SNE.
2. Graficar los datos en 2D o 3D.
3. Contar visualmente el número de grupos discernibles.

Alternativas a WCSS: BSS y TSS

El Between-Cluster Sum of Squares (BSS) representa la variabilidad entre los clusters, es decir, mide la distancia entre los centroides de diferentes clusters. Es una métrica complementaria al Within-Cluster Sum of Squares (WCSS), que mide la variabilidad dentro de los clusters. Ambas juntas permiten comprender la calidad del agrupamiento.

Para entender cómo se relaciona el BSS con el WCSS y cómo estos reflejan la calidad del agrupamiento, es útil introducir el concepto de suma total de cuadrados (Total Sum of Squares , TSS).

- **TSS (Total Sum of Squares):** Es la suma total de las distancias cuadradas entre cada punto y el centroide global (media de todos los puntos). Refleja la variabilidad total de los datos.

  $TSS = \sum_{x \in X} \|x - \bar{x}\|^2$
  donde:
  - X es el conjunto de todos los puntos,
  - $\bar{x}$ es el centroide global (media).
- **BSS (Between-Cluster Sum of Squares):** Es la suma de las distancias cuadradas entre cada centroide de cluster y el centroide global, ponderada por el número de puntos en cada cluster. Refleja la variabilidad entre los clusters.

  $BSS = \sum_{i=1}^{k} |C_i| \|\mu_i - \bar{x}\|^2$
  donde:
  - k es el número de clusters,
  - C_i es el conjunto de puntos en el cluster \(i\),
  - $|C_i|$ es el número de puntos en el cluster \(i\),
  - μ_i es el centroide del cluster i,
  - $\bar{x}$ es el centroide global.

Relación numérica entre TSS, WCSS y BSS

Existe una relación matemática entre estas tres métricas:

$$TSS = WCSS + BSS$$

Esto significa que la variabilidad total (TSS) se puede descomponer en la variabilidad dentro de los clusters (WCSS) y la variabilidad entre los clusters (BSS).

Interpretación en el Método del Codo

En el método del codo, cuando se grafica el WCSS contra el número de clusters k, se busca el punto donde la disminución en el WCSS se desacelera, lo que indica que la inclusión de más clusters no mejora significativamente la compacidad de los clusters.

- **WCSS**: Un WCSS bajo indica clusters compactos.
- **BSS**: Un BSS alto indica clusters bien separados.

Objetivo del clustering:

- Minimizar el WCSS (maximizar la compacidad).
- Maximizar el BSS (maximizar la separación).

Ejemplo numérico

Supongamos que tenemos un conjunto de datos con tres clusters y calculamos las siguientes métricas:

- **Cluster 1:**
  - Centroide: (1, 1)
  - Número de puntos: 10
  - WCSS1 = 50
- **Cluster 2:**
  - Centroide: (5, 5)
  - Número de puntos: 15
  - WCSS2 = 40
- **Cluster 3:**
  - Centroide: (9, 9)
  - Número de puntos: 20
  - WCSS3 = 60
- **Centroide Global:**
  (6, 6)
- **Cluster 1:**
  - Centroide: (1, 1)
  - Número de puntos: 10
  - Distancia al centroide global:

$$\|(1,1) - (6,6)\| = \sqrt{(1-6)^2 + (1-6)^2} = \sqrt{25+25} = \sqrt{50} \approx 7.071$$

- Contribución al BSS:

$$10 \cdot (7.071)^2 = 10 \cdot 50 = 500$$

- **Cluster 2:**
  - Centroide: \((5, 5)\)
  - Número de puntos: 15
  - Distancia al centroide global:

$$\|(5,5) - (6,6)\| = \sqrt{(5-6)^2 + (5-6)^2} = \sqrt{1+1} = \sqrt{2} \approx 1.414$$

  - Contribución al BSS:

$$15 \cdot (1.414)^2 = 15 \cdot 2 = 30$$

- **Cluster 3:**
  - Centroide: \((9, 9)\)
  - Número de puntos: 20
  - Distancia al centroide global:

$$\|(9,9) - (6,6)\| = \sqrt{(9-6)^2 + (9-6)^2} = \sqrt{9+9} = \sqrt{18} \approx 4.243$$

  - Contribución al BSS:

$$20 \cdot (4.243)^2 = 20 \cdot 18 = 360$$

- **BSS Total:**
  BSS = 500 + 30 + 360 = 890

Relación entre TSS, WCSS y BSS

Sabemos que:

$$TSS = WCSS + BSS$$

Dado que calculamos previamente el WCSS como 150, podemos encontrar el TSS:

$$TSS = 150 + 890 = 1040$$

Interpretación

- **WCSS (Within-Cluster Sum of Squares):** Representa la variabilidad dentro de los clusters; en este caso, es 150. Un valor bajo indica clusters compactos.
- **BSS (Between-Cluster Sum of Squares):** Representa la variabilidad entre los clusters (distancias entre los centroides); en este caso, es 890. Un valor alto indica clusters bien separados.

Método del Codo

El método del codo implica graficar el WCSS frente a diferentes valores de k y encontrar el punto donde la reducción en el WCSS comienza a disminuir significativamente. Este punto indica el número óptimo de clusters.

En términos de BSS, un valor más alto generalmente indica una mejor separación entre los clusters. Al combinar tanto el WCSS como el BSS, se puede determinar no solo la compacidad interna sino también la separación entre las agrupaciones.

Resumen

- **WCSS (Within-Cluster Sum of Squares):** Mide la compacidad interna de los clusters.
- **BSS (Between-Cluster Sum of Squares):** Mide la separación entre los clusters.
- **TSS (Total Sum of Squares):** Representa la variabilidad total de los datos y se descompone en WCSS + BSS.

El objetivo del clustering es minimizar el WCSS (maximizar la compacidad) y maximizar el BSS (maximizar la separación).