Evaluación de Agrupación

La evaluación en modelos de agrupación es un poco más compleja que en modelos de clasificación o regresión, ya que en la agrupación no hay etiquetas de clase conocidas de antemano para comparar directamente
con los clusters encontrados por el modelo. Métricas de evaluación comunes utilizadas en el contexto de modelos de agrupación:

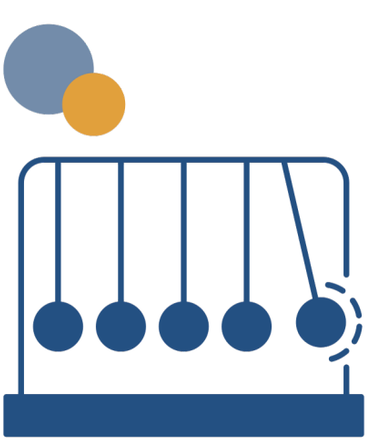
1. Índice de Silueta (Silhouette Score)

El índice de silueta es una métrica que proporciona una medida de cuán bien se han separado los clusters en un conjunto de datos. Se calcula para cada muestra y se basa en la distancia media dentro del cluster y la distancia media al cluster más cercano.
Un valor de silueta cercano a +1 indica que la muestra está bien dentro de su propio cluster y lejos de otros clusters, mientras que un valor cercano a 0 indica que la muestra está cerca del límite de decisión entre dos clusters, y un valor negativo indica que la muestra puede haber sido asignada incorrectamente a su cluster.

La puntuación de silueta se calcula como el promedio de las puntuaciones de silueta de todas las muestras.

2. Inercia

La inercia mide la suma de las distancias cuadradas de cada punto a su centroide más cercano. Cuanto menor sea la inercia, más compactos y bien definidos serán los clusters.
Sin embargo, la inercia por sí sola no proporciona una evaluación completa de los clusters, ya que no considera la dispersión entre los clusters.



3. Índice de Rand Ajustado (Adjusted Rand Index, ARI) y Mutual Information (MI)

Estas métricas son útiles cuando se disponen de etiquetas de clase verdaderas para los datos, aunque estas últimas no estén disponibles en muchos casos.

El ARI y MI comparan los clusters encontrados por el modelo con las etiquetas de clase reales y proporcionan una medida de la similitud entre ellos.

El ARI ajusta el índice de Rand para tener en cuenta el azar, mientras que la información mutua mide la cantidad de información compartida por las asignaciones de cluster y las etiquetas de clase.

4. Homogeneidad, Completitud y V-Measure

Estas métricas evalúan diferentes aspectos de la calidad de los clusters cuando se tienen etiquetas de clase verdaderas disponibles. Conozcamos cada una de ellas:

La **homogeneidad** mide la proporción de instancias de una clase que están contenidas en un solo cluster.

La **completitud** mide la proporción de instancias que pertenecen a una clase y que se agrupan en el mismo cluster.

La puntuación **V-measure** es la media armónica de homogeneidad y completitud, y proporciona una medida combinada de ambas.

Es importante tener en cuenta que la elección de la métrica de evaluación adecuada depende del contexto específico del problema y de las características del conjunto de datos. Además, en algunos casos puede ser útil utilizar múltiples métricas para obtener una evaluación más completa del rendimiento del modelo de agrupación.

Ejemplo

- En este ejemplo, primero generamos datos sintéticos utilizando la función make_blobs de scikit-learn. Luego, inicializamos y ajustamos un modelo de KMeans a los datos. Finalmente, calculamos varias métricas de evaluación, incluyendo el índice de silueta, la inercia, el índice de Rand ajustado y las métricas de homogeneidad, completitud y puntuación V-measure.
- Estas métricas proporcionan una evaluación completa del rendimiento del modelo de agrupación en los datos. Recuerda que puedes cambiar el modelo de agrupación y ajustar los parámetros según sea necesario para tu aplicación específica.

```
from sklearn.datasets import make_blobs
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score, calinski_harabasz_score,
adjusted_rand_score, homogeneity_completeness_v_measure

# Generar datos sintéticos para agrupación
X, y_true = make_blobs(n_samples=300, centers=4, cluster_std=0.60,
random_state=0)

# Inicializar y ajustar un modelo de KMeans
kmeans = KMeans(n_clusters=4, random_state=0)
kmeans.fit(X)

# Calcular el índice de silueta
silhouette_avg = silhouette_score(X, kmeans.labels_)
print("Silhouette Score:", silhouette_avg)

# Calcular la inercia
inertia = kmeans.inertia_
print("Inertia:", inertia)

# Calcular el índice de Rand ajustado
ari = adjusted_rand_score(y_true, kmeans.labels_)
print("Adjusted Rand Index:", ari)

# Calcular la homogeneidad, completitud y puntuación V-measure
homo, comp, v_measure = homogeneity_completeness_v_measure(y_true,
kmeans.labels_)
print("Homogeneity:", homo)
print("Completeness:", comp)
print("V-Measure:", v_measure)
```