ESCUELA POLITÉCNICA NACIONAL

REDES NEURONALES

TAREA 7

David Fabián Cevallos Salas

2023-08-17

APRENDIZAJE NO SUPERVISADO: K-MEANS

SCRIPT DE MÉTRICAS PARA ANALIZAR DESEMPEÑO DE ALGORITMOS DE APRENDIZAJE NO SUPERVISADO

En la presente tarea analizaremos el dataset fisheriris mediante el algoritmo de aprendizaje no supervisado k-means.

Para ello utilizaremos únicamente dos descriptores: ancho de pétalo (en centímetros) y largo de pétalo (en centímetros).

El dataset cuenta con un total de 150 observaciones y no presenta valores atípicos ni faltantes.

1. Gráfica Ancho de pétalo vs Longitud de pétalo

En primer lugar, generamos el gráfico bidimensional (*scatter plot*) correspondiente al ancho de pétalo y longitud de pétalo, el cual se presenta en la Figura 1.

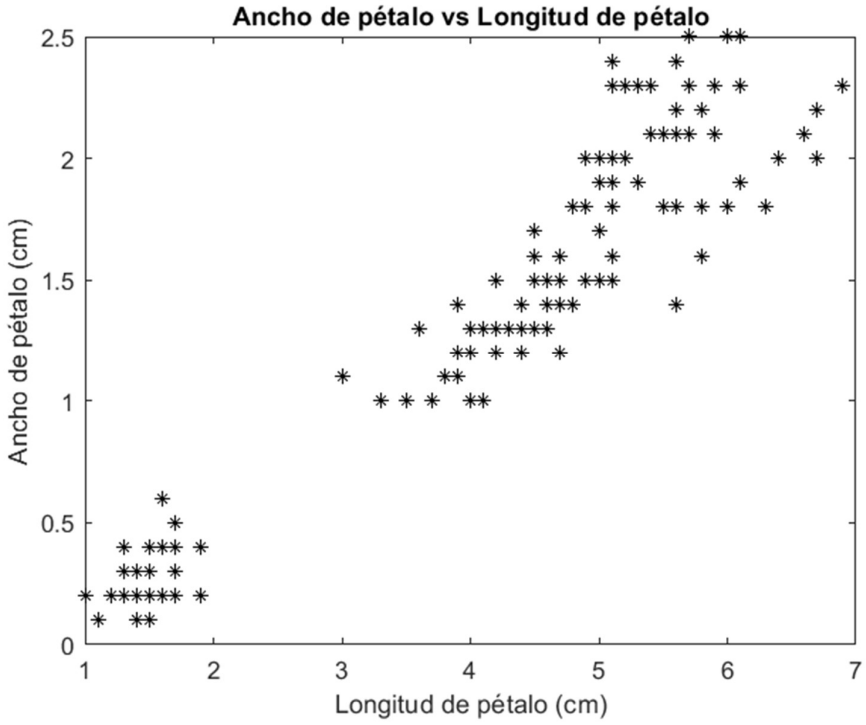

Figura 1 Scatter plot Ancho de pétalo vs Longitud de pétalo

Como se puede observar, la información disponible permite formar al menos dos grupos claramente identificables entre las 150 observaciones del dataset.

2. Valores de métricas obtenidas

En base a la información visualizada, se procedió a aplicar el algoritmo k-means sobre los datos con un valor de k igual a 3 (es decir, con un total de 3 clusters).

La Figura 2 presenta el resultado obtenido donde se han pintado de verde, rojo y azul cada una de las observaciones asignadas a un determinado cluster.

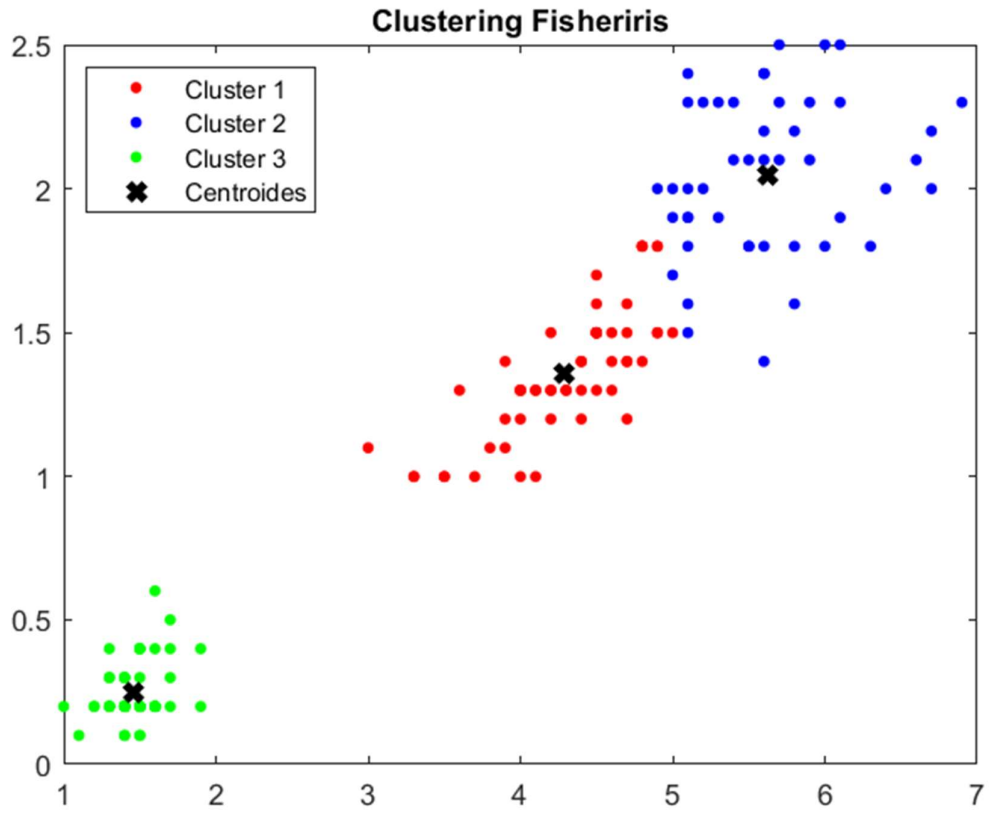

Figura 2 Clustering Fisheriris

La obtención de las métricas se obtuvo en base a la construcción de un script.

El script implementado de nombre `MetricasNoSupervisado` cuenta con los parámetros de entrada y salida que se detallan en la Tabla 1.

La Tabla 2 detalla los valores de métricas resultantes de la ejecución del script.

Tabla 1 Parámetros de entrada y salida de script de métricas

Parámetros de entrada	
X	Matriz de datos con observaciones (filas) y descriptores (columnas)
idx	Número de cluster asignado a cada observación a través de algoritmo no supervisado
Parámetros de salida	
SSW	Valor de métrica *Sum of Squared Within* (SSW)
SSB	Valor de métrica *Sum of Squared Between* (SSB)
WB	Valor de métrica *WB-Index*
SIL	Valor de métrica *Silhouette*

Tabla 2 Valores de métricas resultantes

Métrica	**Valor**
SSW	0,3689
SSB	1,6588
WB-index	0,6672
SIL	0,8058

3. Diagrama de codo

Finalmente, el diagrama de codo de la Figura 3 fue generado en donde se puede visualizar como decrece el valor de la función objetivo conforme el valor de k (número de clusters) se incrementa. En este caso, el valor de k varía entre 1 y 10.

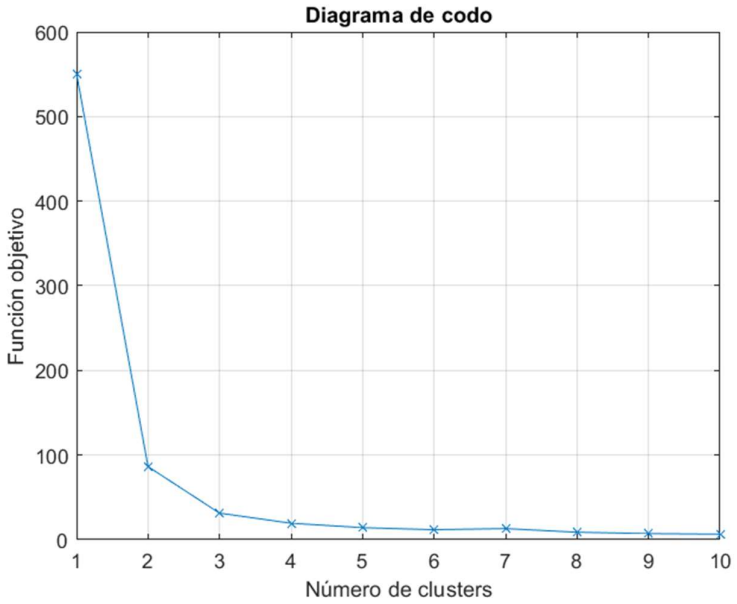

Figura 3 Diagrama de codo

4. Conclusión

Al culminar la presente actividad se puede concluir que la obtención de métricas es sumamente importante para la realización de una validación interna, es decir, para determinar qué tan bien logra clusterizar el algoritmo de aprendizaje no supervisado (en este caso k-means) la información disponible en base a algún patrón.

Por lo tanto, lo que se busca en este caso es obtener valores de SSW bajos (máxima cohesión dentro de cada cluster) y valores de SSB altos (máxima separación entre clusters) que involucren a la vez un valor de WB-index lo más bajo posible. De forma similar, se busca obtener un valor de métrica *Silhouette* lo más cercana a 1.

En lo referente al diagrama de codo se puede concluir que para un determinado umbral de k la función objetivo ya no decrecerá sustancialmente su valor, por lo que dicho diagrama es de utilidad para determinar este valor máximo.