



Instituto Politécnico Nacional

Escuela Superior de Cómputo

Profesor: García Floriano Andrés

Alumnos:

Hernández Jiménez Erick Yael

Patiño Flores Samuel

Robert Garayzar Arturo

5BV1

Ejercicio 1: K-Means

Índice

| | |
|---|----------|
| 1. Introducción | 2 |
| 2. Desarrollo | 2 |
| 2.1. K-means | 2 |
| 2.2. Post K-means | 2 |
| 2.2.1. Clusters | 2 |
| 2.2.2. Gráfica | 3 |
| 3. Conclusiones | 3 |
| 3.1. Hernández Jiménez Erick Yael | 3 |
| 3.2. Patiño Flores Samuel | 3 |
| 3.3. Robert Garayzar Arturo | 3 |

1. Introducción

En el desarrollo de esta práctica llevamos a cabo el método de aprendizaje no supervisado K-means, el cual consiste en agrupar los datos con un número K. Originalmente, teniendo ese número K, se eligen K elementos del dataset aleatoriamente y se agrupan los demás datos calculando la distancia más corta a los puntos seleccionados. Una vez agrupados todos los elementos, se prueba con otros K elementos del dataset, cambiando continuamente hasta que se repitan los grupos seleccionados, terminando así el clustering.

2. Desarrollo

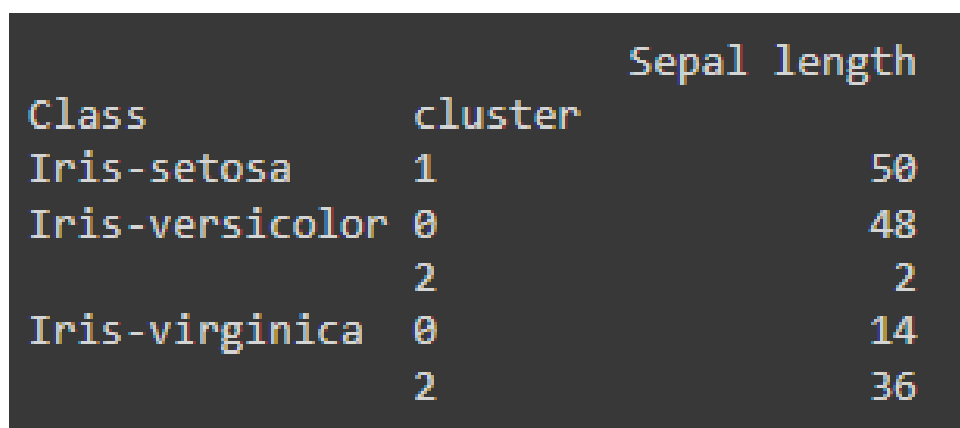
2.1. K-means

En el desarrollo de nuestra práctica, al aplicar el K-means, la función de K-means nos da a elegir entre dos procesos de ejecución para desarrollar el algoritmo: el primero es de manera aleatoria, como ya se había mencionado, y el segundo es el K-means++, una variante que utiliza distribución probabilística, lo que hace más eficiente el proceso. A este proceso se le llama *Greedy K-means++*.

2.2. Post K-means

2.2.1. Clusters

Una vez realizado el proceso de K-means, podemos ver un resumen de nuestro dataset, mostrando lo siguiente:

A terminal window with a dark background and light-colored text. It displays a table with three columns: 'Class', 'cluster', and 'Sepal length'. The data is as follows:

| Class | cluster | Sepal length |
|-----------------|---------|--------------|
| Iris-setosa | 1 | 50 |
| Iris-versicolor | 0 | 48 |
| | 2 | 2 |
| Iris-virginica | 0 | 14 |
| | 2 | 36 |

Figura 1: Información de clusters

Como podemos ver, el resultado para nuestro cluster 0 fue una combinación de 48 datos de *Iris-versicolor* y 14 datos de *Iris-virginica*. Para nuestro cluster 1, únicamente resultaron datos de *Iris-setosa*, y por último, el cluster 2 resultó con 2 datos de *Iris-versicolor* y 36 datos de *Iris-virginica*.

2.2.2. Gráfica

Para graficar nuestro resultado, utilizamos el Análisis de Componentes Principales (PCA) para reducir la dimensionalidad y poder mostrar nuestros datos en una gráfica de 2 dimensiones. El resultado es la siguiente gráfica:

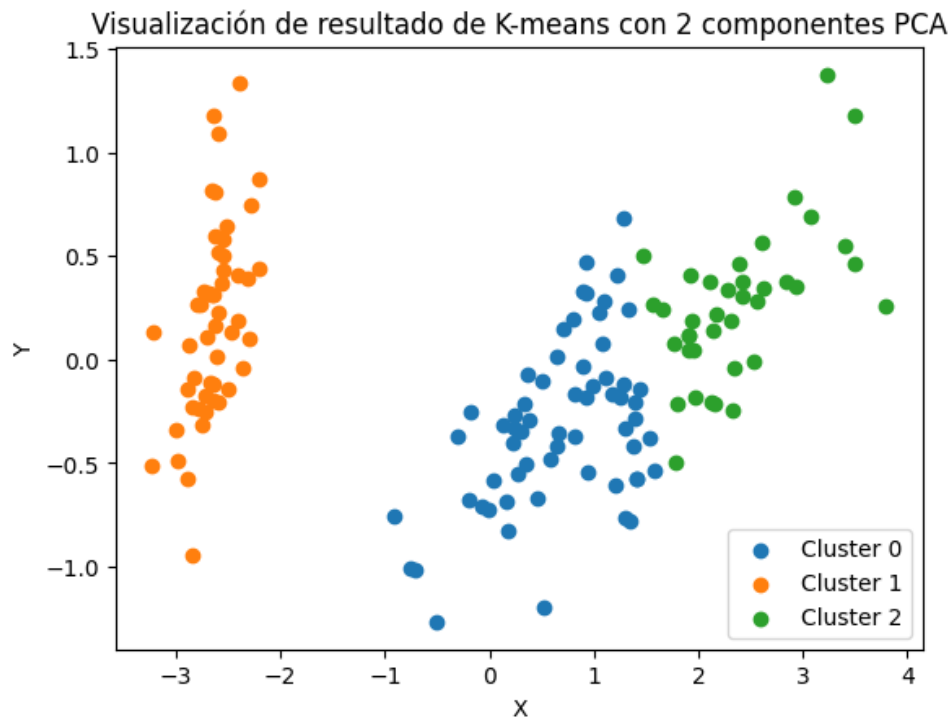


Figura 2: Análisis de Componentes Principales

3. Conclusiones

3.1. Hernández Jiménez Erick Yael

Con este ejemplo ejecutado, podemos ver que el método de K-means, concretamente el Greedy K-means, logra una asociación muy acertada y casi exacta de los datos, en este caso, del tamaño del sépalo de las clases 3 flores y su respectivo tipo, indicando entonces una estrecha relación entre la longitud de estas y la especie a la que corresponden.

3.2. Patiño Flores Samuel

En esta práctica pusimos en uso el profiling para poder analizar el dataset seleccionado. Asimismo, nos apoyamos de la información recabada con el profiling para el uso del PCA con la información de columnas correlacionadas. Se destacó el uso eficiente de *Greedy K-means* y la diferencia con su versión estándar.

3.3. Robert Garayzar Arturo

En conclusión sobre el algoritmo K-means destaca que se utilizó el profiling para analizar el conjunto de datos y con esta información, se implementó el Análisis de Componentes Principales para reducir la dimensionalidad y visualizar los datos. Además, se resalta la eficiencia de la versión Greedy K-means, mostrando una clara mejora en el rendimiento y en la precisión del agrupamiento.