Universidad Simón Bolívar
Inteligencia Artificial II
Prof. Ivette Carolina Martínez
Alumnos:
Carlos Martinez 11-10584
Yerson Roa 11-10876
Antonio Scaramazza 11-10957

Proyecto 3
Clustering

Resumen

La técnica de clustering es un tipo de análisis que busca encontrar grupos de elementos u objetos tal que las relaciones entre un grupo de objetos sea similar entre estos y que no tengan relación, sean diferentes, con los elementos u objetos en otros grupos. Expresado en términos de variabilidad hablamos de minimizar la variabilidad dentro de los grupos para al mismo tiempo maximizar la variabilidad entre los distintos grupos.

Este proyecto consta de entrenar y probar con el algoritmo de clustering k-means:

1. Los clusters obtenidos para 3 géneros de la especie Iris dadas las medidas de sus pétalos y sépalos.
2. Dada una imagen evaluar con distintos valores de k los resultados obtenidos.

Implementación

El algoritmo de *k-means* es ampliamente conocido y su pseudocódigo puede ser encontrado de manera simple y consistente en muchas páginas webs especializadas. Su implementación se realizó utilizando el lenguaje de programación Python y las librerías Numpy y Matplotlib.

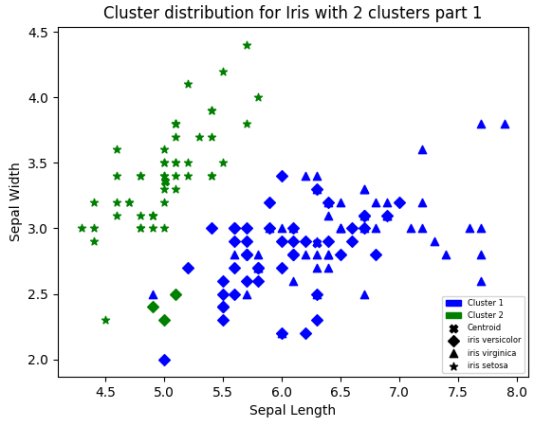
Para la actividad 2, se tomo el conjunto original de entrenamiento y se eliminó la etiqueta que identifica el género de la flor, la cual consistía de una cadena de caracteres. Se evaluó en la actividad el hecho de que si el index estaba en cierto rango pertenece a un género específico de flor.

- 0 - 49 cuando es identificada como “Iris Setosa”
- 50 - 99 cuando es identificada como “Iris Versicolour”
- 100 en adelante cuando es identificada como “Iris Virginica”

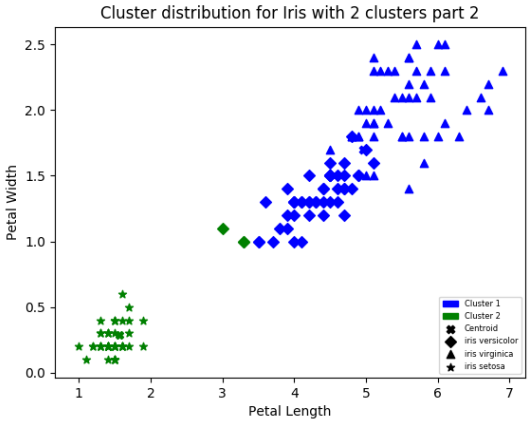
Resultados

Actividad 2:

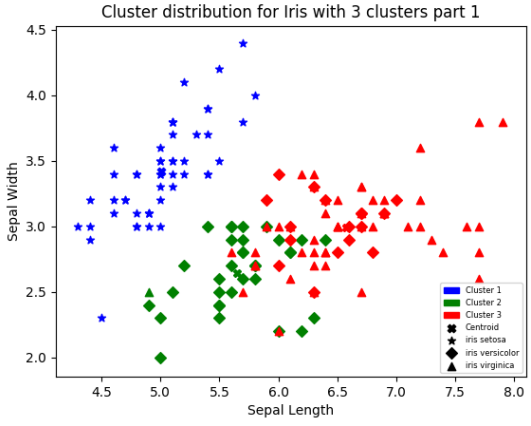
Al aplicar el algoritmo de k-means sobre los datos de *iris.txt* esperamos que retorne, en el caso ideal, clusters que contengan cada especie (3) de iris presente en el archivo, y que a medida que se aumenta el número de clusters encuentre subgrupos similares dentro de las mismas especies. Sin embargo, debido a que 2 de estas especies tienen características similares (*Iris Versicolour* e Iris Virginica) los clusters tienden a generar subgrupos que contienen instancias de ambas especies como se muestra en las gráficas 5,7 y 8 .



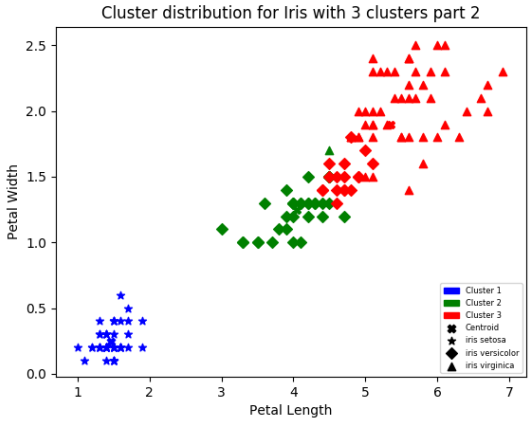
Gráfica 1



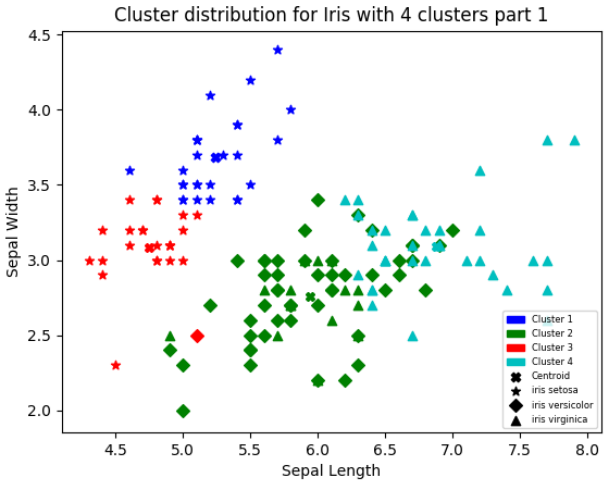
Gráfica 2



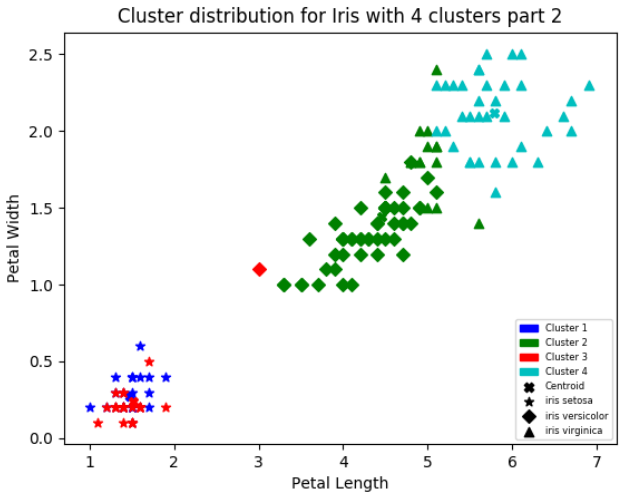
Gráfica 3



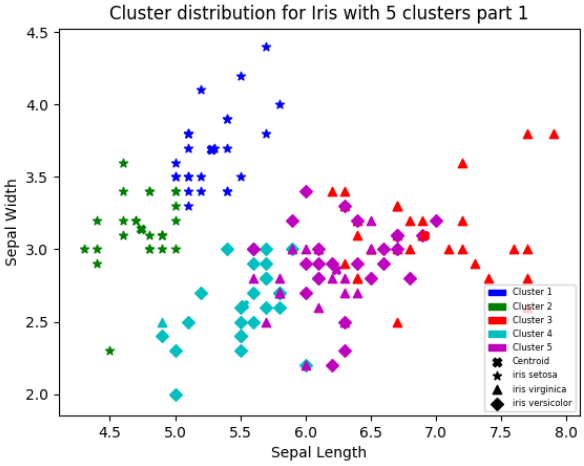
Gráfica 4



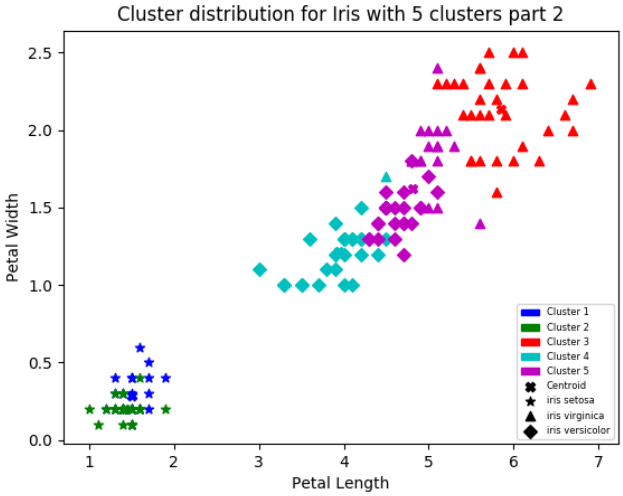
Gráfica 5



Gráfica 6



Gráfica 7



Gráfica 8

En los casos más simples para k={2,3} podemos ver que el algoritmo reconoce correctamente a las *Iris Setosa* al agruparlas todas bajo un mismo cluster, ya que sus características son bastantes definidas con respecto al resto (esto es fácilmente apreciable en la forma globular que posee su distribución en las gráficas). Para las dos especies restantes y con k=3 podemos ver que la separación no es del todo correcta como se aprecia en la Tabla 2. Esto (como se mencionó en el párrafo anterior) se debe a la gran similitud entre ellas, lo que genera dificultades en su correcta agrupación.

Iris	Cluster 1	Cluster 2
Iris Setosa	50	0
Iris Versicolour	3	47
Iris Virginica	0	50

Tabla 1. Distribución de las instancias para k=2

Iris	Cluster 1	Cluster 2	Cluster 3
Iris Setosa	50	0	0
Iris Versicolour	0	48	2
Iris Virginica	0	11	39

Tabla 2. Distribución de las instancias para k=3

Actividad 3:

La actividad 3 se realizó con el objetivo de ejemplificar los diversos usos del algoritmo k-means; en este caso se utilizó para comprimir una imagen de tal forma que está solo este compuesta por los k colores asociados a los centroides. Las imágenes resultantes de la compresión para k={2, 4, 8, 16, 32, 64, 128} se pueden apreciar en la carpeta adjunta a este documento.

Conclusiones

Como se pudo apreciar, el algoritmo de *k-means* puede utilizarse tanto para la clasificación de datos como la compresión de imágenes con cierto nivel de éxito. Sin embargo, también se pudo verificar que el desempeño del mismo depende de la distribución de los datos, los valores iniciales y del número de iteraciones que se le den al algoritmo.

Referencias

- *Machine Learning* . (2017). *Openclassroom.stanford.edu*. Retrieved 17 March 2017, from http://openclassroom.stanford.edu/MainFolder/DocumentPage.php?course=MachineLearning&doc=exercises/ex9/ex9.html
- *k-means clustering | Wikiwand*. (2017). *Wikiwand*. Retrieved 17 March 2017, from https://www.wikiwand.com/en/K-means_clustering