Clustering:

Algoritmo K-means

Kevin A. Vega Hernandez.
Universidad Tecnológica de Bolívar Ingeniería de
Sistemas Kevin.vega.h @hotmail.com

Abstract: This work was carried out in order to study the concept of Clustering using the K-means algorithm, having two datasets with which they will be essential to understand the functioning of this Clutering algorithm.

In the present document we will see tests carried out using this algorithm that we have previously worked on and will perform its respective analysis, drawing conclusions to the data thrown by the software.

Resumen: Este trabajo se realizó con el fin de estudiar el concepto de Clustering mediante el algoritmo K-means teniendo dos set de datos con los cuales serán esenciales para entender el funcionamiento de este algoritmo de Clutering.

En el presente documento veremos pruebas realizadas mediante este algoritmo que construiremos previamente y realizaremos su respectivo análisis, sacando conclusiones a los datos arrojados por el software.

Palabras clave: Clustering, K-means, algoritmo, set de datos, Centroides, Cluster.

I. INTRODUCCION.

El Clustering consiste en agrupar un conjunto de objetos los cuales no están etiquetados, en subconjuntos de objetos llamados clusters. El clustering es una técnica de minería de datos (data mining) dentro de la disciplina de la inteligencia artificial.

A. Justificación

En el presente documento realizaremos clustering a dos conjuntos de datos definidos como dataset1 y dataset2. Esto lo realizaremos mediante el algoritmo de K-means que es un método para dividir datos, en base a una serie de variables, un conjunto en un numero K de segmentos
(Clusters).

B. Objetivo general

Realizar pruebas en base al algoritmo K-mean y analizar los resultados arrojados de este, entendiendo el funcionamiento del K-menas y el concepto de Clustering.

II. ESTADO DEL ARTE

El Clustering consiste en agrupar un conjunto de objetos no etiquetados, en subconjuntos de objetos llamados Clusters. Cada Cluster está conformado por una colección de objetos que son similares o pueden llegar a ser similares entre sí, pero diferentes al respecto a otros Clusters.

El clustering es una técnica de minería de datos (data mining) dentro de la disciplina de la inteligencia artificial. Se enmarca dentro del aprendizaje no supervisado, esto significa entonces que para esta técnica solo disponemos de un conjunto de entrada de datos y no poseemos datos de salida.

Para el Clustering podemos hacer uso del método de agrupamiento K-means, un método utilizado en la minería de datos.

K-means es un método de agrupación, es un método para dividir datos, en base a una serie de
Variables, un conjunto en un numero K de segmentos (Clusters). El algoritmo estándar fue propuesto por Stuart Lloyd en 1957 para modulación por impulsos codificados.

K-means es un algoritmo de clasificación no supervisada (Cauterización) que agrupa objetos en *k* grupos basándose en sus características. El agrupamiento se realiza minimizando la suma de distancias entre cada objeto y el centroide de su grupo o cluster.

El algoritmo consta de tres pasos:

1. Inicialización: una vez escogido el número de grupos, k, se establecen kcentroides en el espacio de los datos, por ejemplo, escogiéndolos aleatoriamente.
2. Asignación objetos a los centroides: cada objeto de los datos es asignado a su centroide más cercano.
3. Actualización centroides: se actualiza la posición del centroide de cada grupo tomando como nuevo centroide la posición del promedio de los objetos pertenecientes a dicho grupo.

Se repiten los pasos 2 y 3 hasta que los centroides no se mueven, o se mueven por debajo de una distancia umbral en cada paso.

El algoritmo K-means resuelve un problema de optimización, siendo la función a optimizar (minimizar) la suma de las distancias cuadráticas de cada objeto al centroide de su cluster.

Las principales ventajas del método k-means son que es un método sencillo y rápido. Pero es necesario decidir el valor de k y el resultado final depende de la inicialización de los centroides. En principio no converge al mínimo global sino a un mínimo local.

III. DESARROLLO DEL PROYECTO

Para este trabajo se realizó la construcción del algoritmo K-means el cual se desarrolló en el lenguaje de programación PYTHON 3. Tomamos los dos sets de datos que son dataset1.txt y dataset2.txt, primero hicimos el análisis para dataset2.txt ya que el número de datos es menor a comparación de dataset1.

Definimos un conjunto de K en este caso 10 estableciendo los centroides en el espacio de los datos tomándolos aleatoriamente del conjunto de datos. Luego asignamos cada dato del conjunto a su K respectivo mediante la distancia más corta al centroide. Aparte, calculamos la función de costo que juega un papel fundamental. Cabe aclarar que lo anterior es un proceso iterativo. Del proceso anterior tomamos los K con menor valor de costos. Asignamos los datos a los centroides y actualizamos los centroides iterativamente. Todo lo anterior fue aplicado para dataset1.text

IV. RESULTADOS

Para encontrar el número óptimo de clusters se hizo uso del método del codo para los dos sets de datos y se obtuvieron las siguientes graficas:

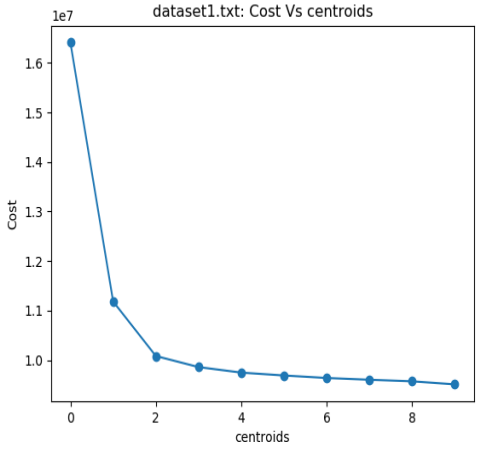

Figura1. Dataset1

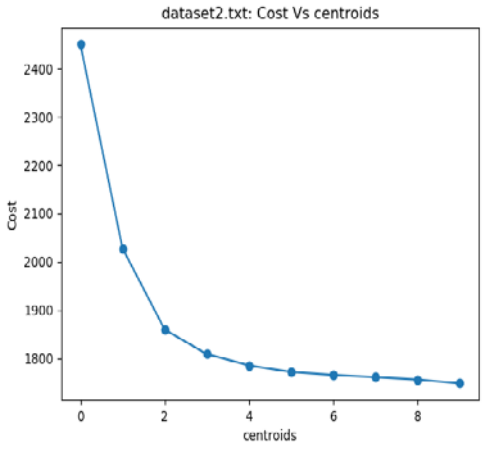

Figura2. Dataset2

Trazamos una línea recta desde el primer punto hasta el último y medimos cada punto de inercia hasta la línea. La línea de mayor longitud será nuestro número óptimo de K:

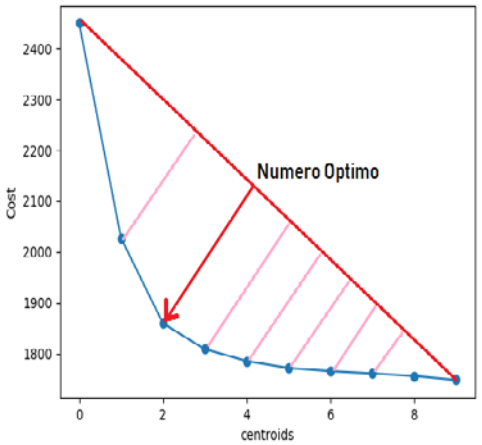

Figura.3 número óptimo

Teniendo entonces:

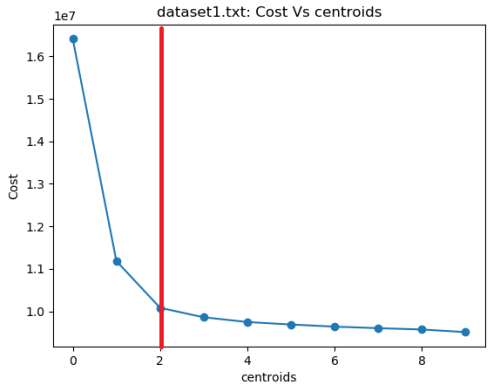

Figura4. Método del codo dataset1

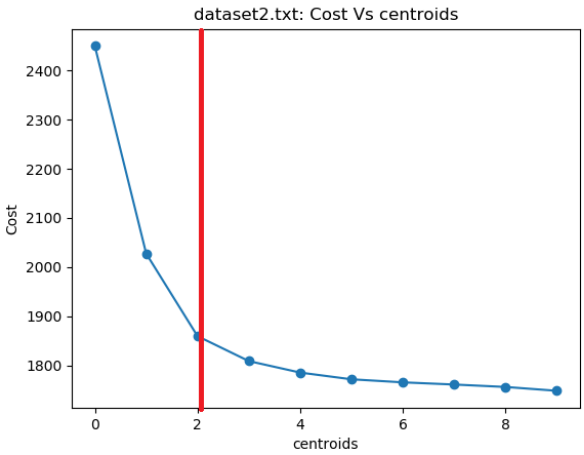

Figura5. Método del codo dataset2

De lo anterior podemos observar que para la figura4, el Koptimo es 2 al igual para la figura 5

V. CONCLUSION

Se pudo realizar satisfactoriamente el algoritmo de K-means aplicándolo a los dataset y teniendo un resultado satisfactorio al final, pudimos aprender acerca del clustering y todo lo que conlleva.

VI. REFERENCIAS

- k-means. (s.f.). Recuperado de https://es.wikipedia.org/wiki/K-means
- El algoritmo k-means aplicado a clasificación y procesamiento de imágenes. (s.f.). Recuperado de https://www.unioviedo.es/compnum/laboratorios_py/kmeans/kmeans.html