Herramientas computacionales: el arte de la analítica

Entregable

Implementación de K-means

Presenta

Javier Valente Rodríguez	A01662693
Juan Francisco García Rodríguez	A01660981
Adrián Aguilar Sánchez	A01651592
César Andrés Ceballos Castillo	A01661893
Klaus Manuel Cedillo Arredondo	A01653257

Docente

Sergio Ruiz Loza

Ciudad de México

Mayo, 12 de 2022

1. Cargamos las librerías y el archivo en cuestión.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sb
from sklearn.cluster import KMeans
from sklearn.metrics import pairwise_distances_argmin_min

%matplotlib inline
from mpl_toolkits.mplot3d import Axes3D
plt.rcParams['figure.figsize'] = (16, 9)
plt.style.use('ggplot')
```

```
#Cargamos el archivo

dataframe = pd.read_csv(r"avocado.csv")
dataframe.head()
```

	Unnamed: 0	Date	AveragePrice	Total Volume	4046	4225	4770	Total Bags	Small Bags	Large Bags	XLarge Bags	type	year	region
0	0	2015-12-27	1.33	64236.62	1036.74	54454.85	48.16	8696.87	8603.62	93.25	0.0	conventional	2015	Albany
1	1	2015-12-20	1.35	54876.98	674.28	44638.81	58.33	9505.56	9408.07	97.49	0.0	conventional	2015	Albany
2	2	2015-12-13	0.93	118220.22	794.70	109149.67	130.50	8145.35	8042.21	103.14	0.0	conventional	2015	Albany
3	3	2015-12-06	1.08	78992.15	1132.00	71976.41	72.58	5811.16	5677.40	133.76	0.0	conventional	2015	Albany
4	4	2015-11-29	1.28	51039.60	941.48	43838.39	75.78	6183.95	5986.26	197.69	0.0	conventional	2015	Albany

2. Si determinas alguna variable que no sirve, elimínala y justifica las razones.

Consideramos que la variable llamada "*Unnamed: 0"* no tiene relevancia en el análisis del archivo, por lo que decimos hacer exclusión de la misma con el siguiente comando:

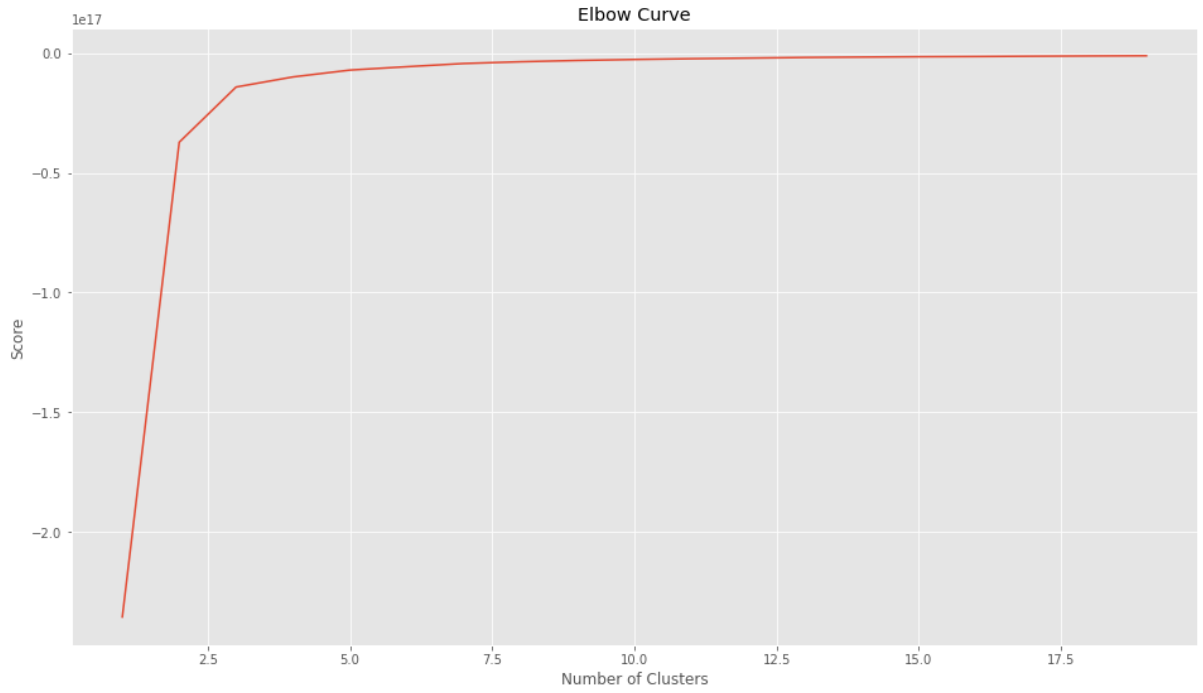
```
dataframeL=dataframe.drop(['Unnamed: 0'],axis=1)
dataframeL
```

3. Determina el valor de k.

De acuerdo con la información obtenida de la página provista en canvas (https://www.aprendemachinelearning.com/k-means-en-python-paso-a-paso/), no hay una única manera de determinar el valor de k. En este caso, se decidió implementar el código provisto en el ejemplo de la página, que sigue el método del codo para obtener dicho valor:

```
#Obtener valor de k (punto de codo)

Nc = range(1, 20)
kmeans = [KMeans(n_clusters=i) for i in Nc]
kmeans
score = [kmeans[i].fit(X).score(X) for i in range(len(kmeans))]
score
plt.plot(Nc,score)
plt.xlabel('Number of Clusters')
plt.ylabel('Score')
plt.title('Elbow Curve')
plt.show()
```

Y la gráfica obtenida se ve así:



Puesto que, de acuerdo con este método, el punto en que se observa un cambio brusco en la inercia en nuestra gráfica es cercano al tres, se decidió tomar a ese número como el óptimo número de clusters para nuestro data set. Entonces, el valor de k es:
K = 3

4. Utilizando scikitlearn calcula los centros de del algoritmo de k-means.

Ocupamos el código que se aprecia en la imagen para calcular los centros.

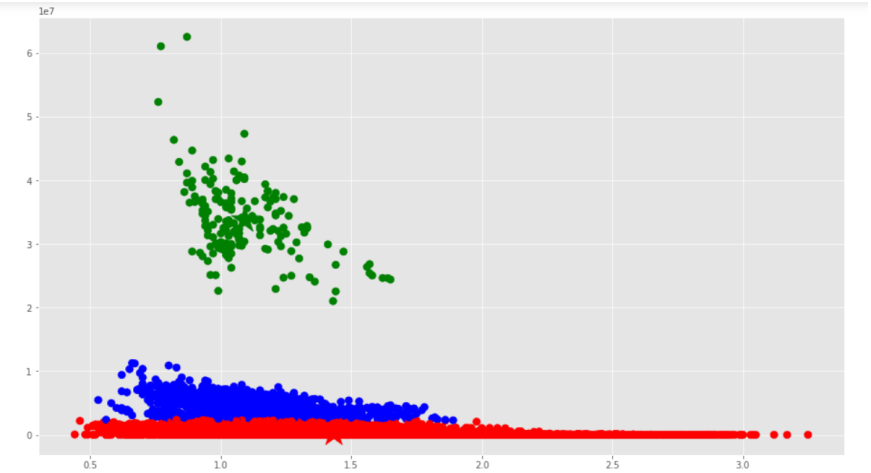
```
#Ejecutamos k-means

kmeans = KMeans(n_clusters=3).fit(X)
centroids = kmeans.cluster_centers_
print(centroids)
```

Dando como resultado:

```
[[1.43340863e+00 2.40087076e+05 7.20529553e+04]
 [1.09201183e+00 3.37350390e+07 9.19049275e+06]
 [1.09325653e+00 4.45007684e+06 1.23738713e+06]]
```

a) ¿Crees que estos centros puedan ser representativos de los datos?, ¿Por qué?

Puede que no todos los centros sean representativos de los datos. Por ejemplo, como se puede ver en la gráfica debajo, el centroide de los datos en verde, identificado como una estrella, podría considerarse a primera vista como un representativo de los datos, puesto que alrededor de él están acumulados gran parte de los datos. En cambio, en el caso de los datos en rojo, su centroide, también representado por una estrella, puede que no sea un representativo de los datos, puesto que éstos se encuentran mucho más dispersos con respecto a su centroide, a diferencia de los datos en verde.

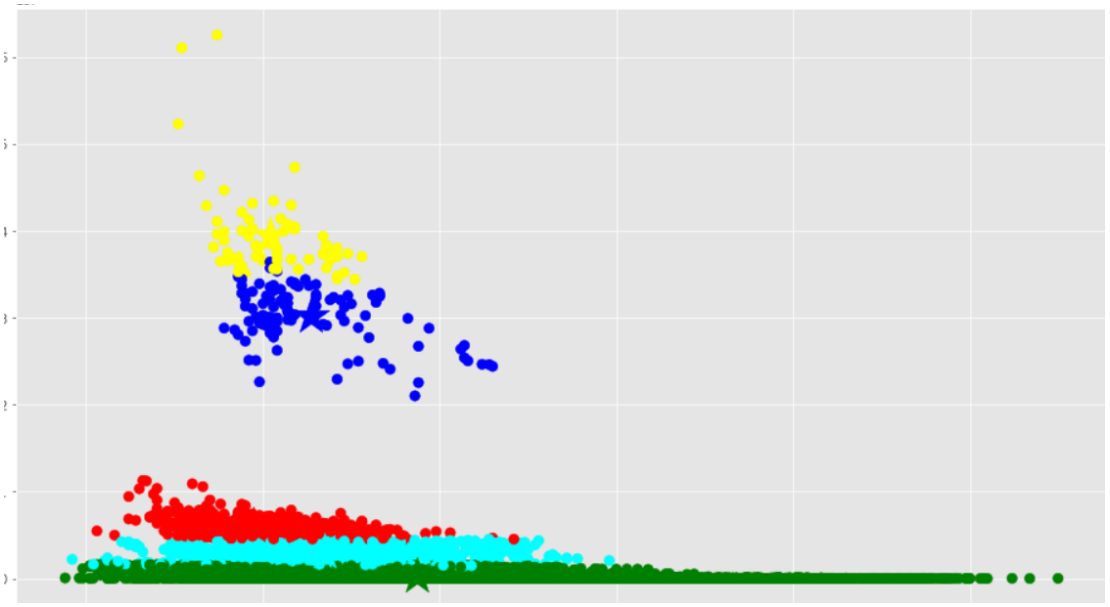


b) ¿Cómo obtuviste el valor de k a usar?

Como se explicó anteriormente, no hay un método único para obtener este valor. En este caso, se decidió hacer una gráfica con el código proporcionado en la página de referencia para aprender acerca de K-means en canvas, en la que el valor de k se determina al observar el punto de la gráfica en la que hay un cambio brusco. Este valor es el valor k, también llamado punto codo según este método, que indica cuál podría ser el número óptimo de clusters

c) ¿Los centros serían más representativos si usaras un valor más alto?, ¿Más bajo?

Creemos que el método del codo nos proveé del número adecuado de clusters para que los centros sean lo más representativo a los datos. Sin embargo, para apreciarlo gráficamente, se decidió hacer las mismas gráficas, pero cambiando el valor de k a 5. Por ejemplo, en la gráfica de abajo, en el caso de los datos en azul y amarillo sí se podría considerar que son más representivos los centroides, pero no se podría decir lo mismo respecto a los otros grupos de datos.



d) ¿Qué distancia tienen los centros entre sí?, ¿Hay alguno que esté muy cercano a otro?

Esos son los centros que se obtuvieron usando el valor de k ya mencionado, y también en la gráfica adjunta en el inciso a, se puede observar que en realidad todos los centros no son realmente lejanos; e incluso se puede observar que hay dos centros, representados por el color azul y rojo en la gráfica, que están relativamente cerca.

```
[[1.43340863e+00 2.40087076e+05 7.20529553e+04]
 [1.09201183e+00 3.37350390e+07 9.19049275e+06]
 [1.09325653e+00 4.45007684e+06 1.23738713e+06]]
```

e) ¿Qué pasaría con los centros si tuviéramos muchos outliers en el análisis de cajas y bigotes?

Su distancia sería mayor debido a que esto solo podría significar que los datos se encuentran en un mayor rango; es decir, están más dispersos.

f) ¿Qué puedes decir de los datos basándote en los centros?

Se puede concluir que los datos tiene propiedades similares, ya que el número de racimos (clusters) que nos arrojó el valor de k fue adecuado comparado con el número de variables que teníamos en los datos. Además, los centros estaban a una distancia corta, lo que afirma que la diferencia entre los datos no es grande; y esto conviene al momento de analizar los datos.