CSL7670 : Fundamentals of Machine Learning

Lab Report

Name:	**SOUHITYA KUNDU**
Roll Number:	**M20PH209**
Program:	**MSc-MTech(Physics & Materials Sc.)**

Chapter 1

Lab-9

1.1 Objective

The objective of this whole assignment is to learn about K-means clustering techniques

1.2 Problem-1

The main objective is to understand K-means clustering approach by proper application in a dataset. with two features, Feature - 1 and Feature - 2.

- First, I have computed using different values of K for the clustering technique ranging from 2 to 5 as 2,3,4,5
- Second, I have plotted the scatter plot for the various K values in the order by forming loop.

Solution 1:

```
#!/usr/bin/env python
# coding: utf-8

# In[26]:


import numpy as np
import pandas as pd


# In[27]:


df = pd.read_csv("DATA.csv")


# In[28]:


df.head(10)


# In[29]:


```

```
from sklearn.cluster import KMeans


# In[30]:


kmeans = KMeans(n_clusters = 5, random_state = 0,n_init="auto").fit(df)
kmeans.labels_


# In[31]:


kmeans.cluster_centers_


# In[32]:


import seaborn as sns
sns.scatterplot(df,x="Feature-1",y="Feature-2",hue=kmeans.labels_)


# In[33]:


# Try different values of k=2,3,4, and 5 and show clustering using appropriate colors in a scatter
# plot. Also, show cluster centers.
# BLACK DOT represents the cluster centers.

import matplotlib.pyplot as plt

cluster_vals= [2, 3, 4, 5]

for i in cluster_vals:
    kmeans = KMeans(n_clusters=i,random_state = 0,n_init="auto")
    kmeans.fit(df)

    cluster_centers = kmeans.cluster_centers_

    plt.scatter(df["Feature-1"], df["Feature-2"], c=kmeans.labels_, cmap='autumn')
    plt.scatter(cluster_centers[:, 0], cluster_centers[:, 1], c='black', marker='o',s=50)
    plt.title(f'K-Means Clustering (k={i})')
    plt.xlabel('Feature 1')
    plt.ylabel('Feature 2')
    plt.show()
    print(cluster_centers)



# In[ ]:
```

```





# In[ ]:
```

The K-means cluster based scatter plots are given below for different values of k such as 2,3,4,5. The Black Dots in the pictures represent the center of the corresponding neighborhood

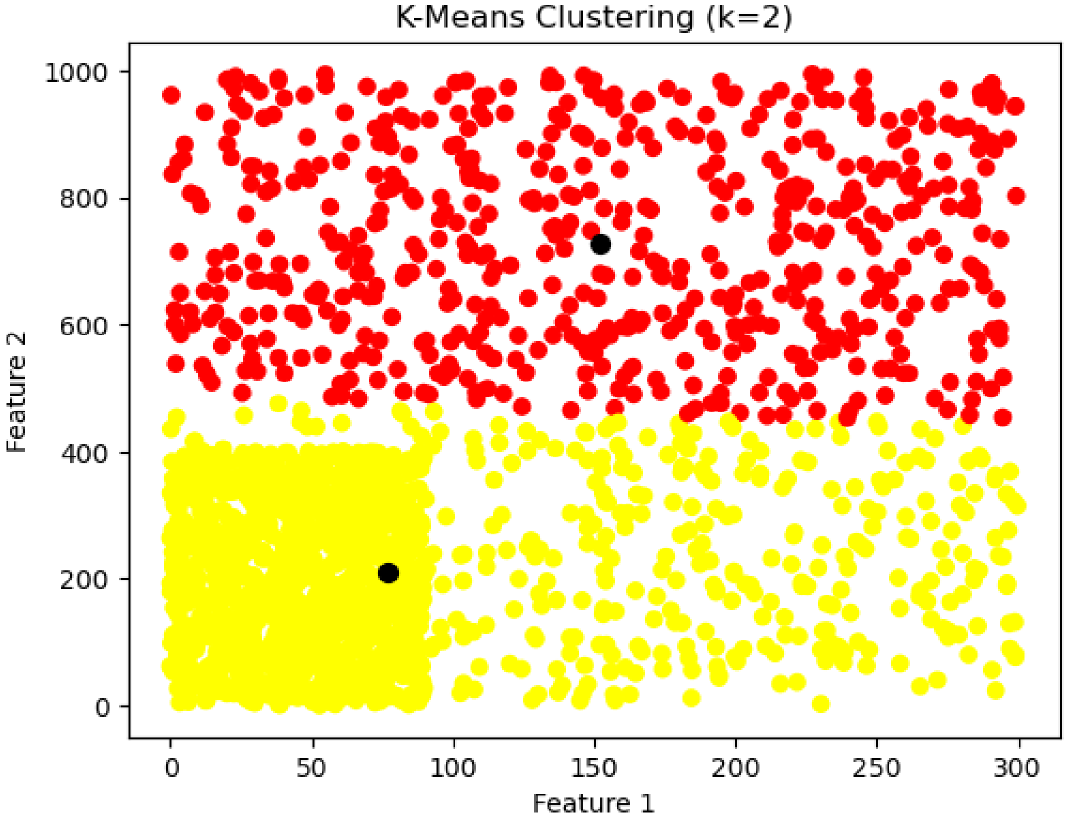

Figure 1.1: K-means cluster with k = 2.

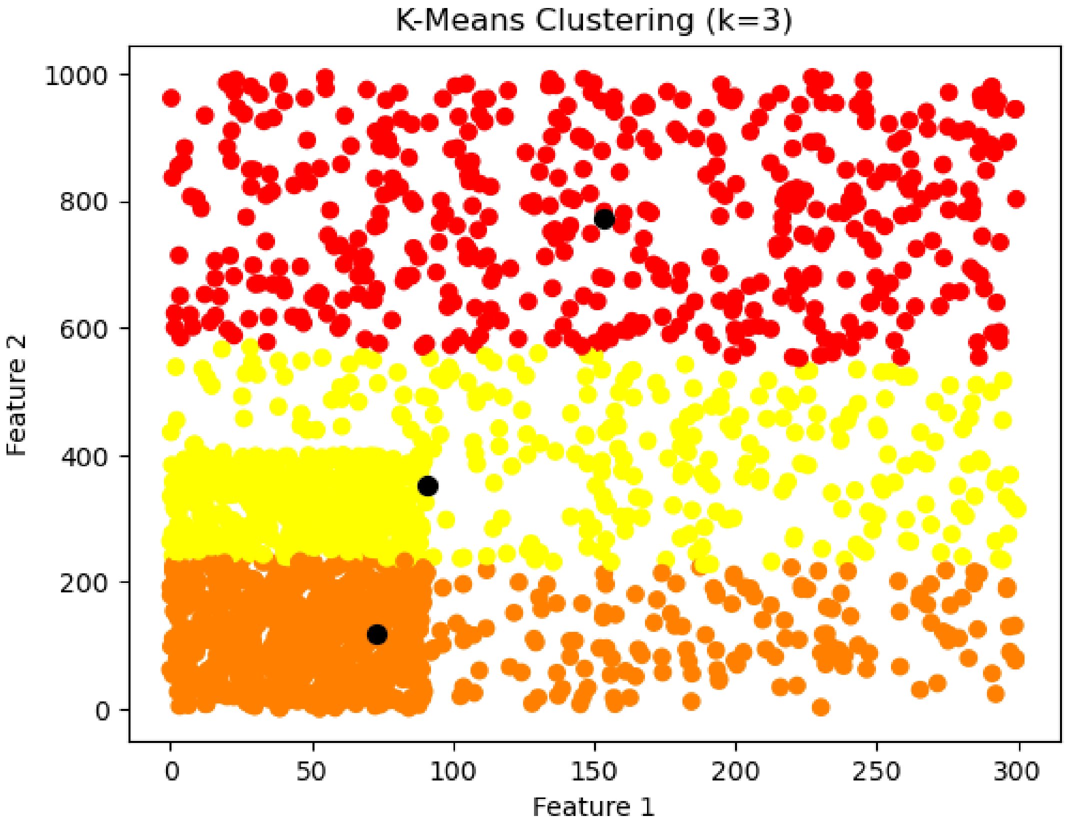

Figure 1.2: K-means cluster with k = 3.

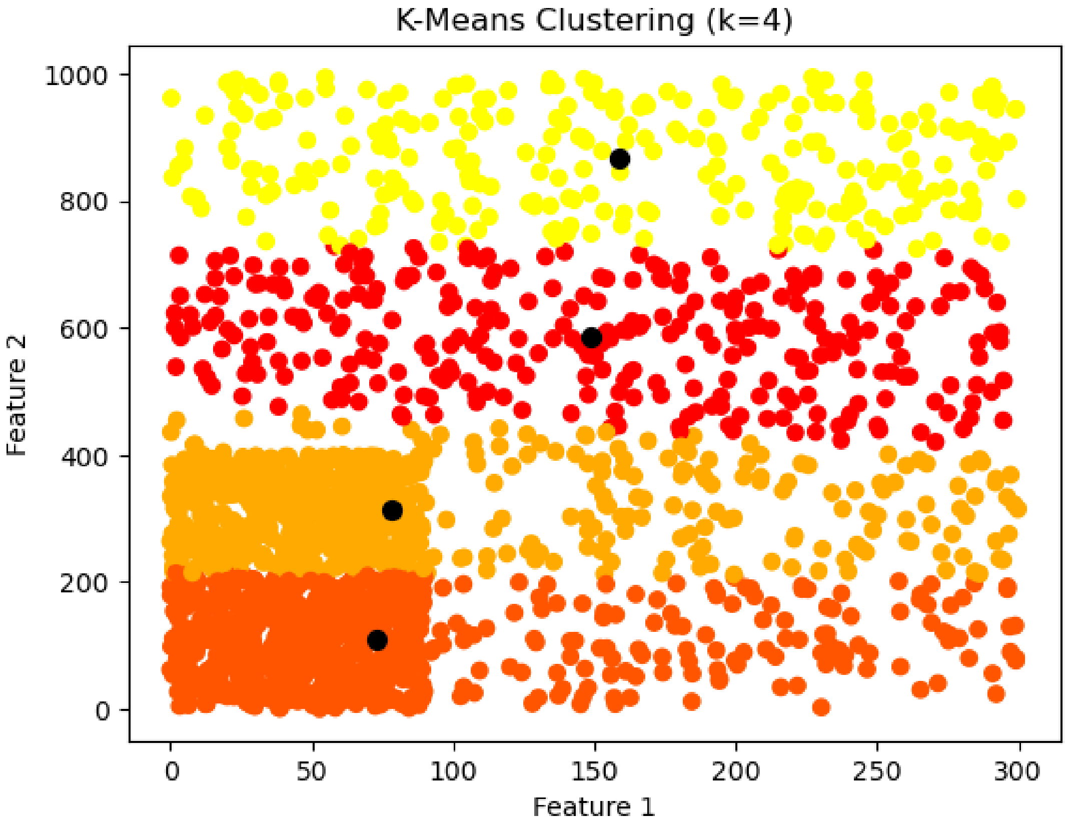

Figure 1.3: K-means cluster with k = 4.

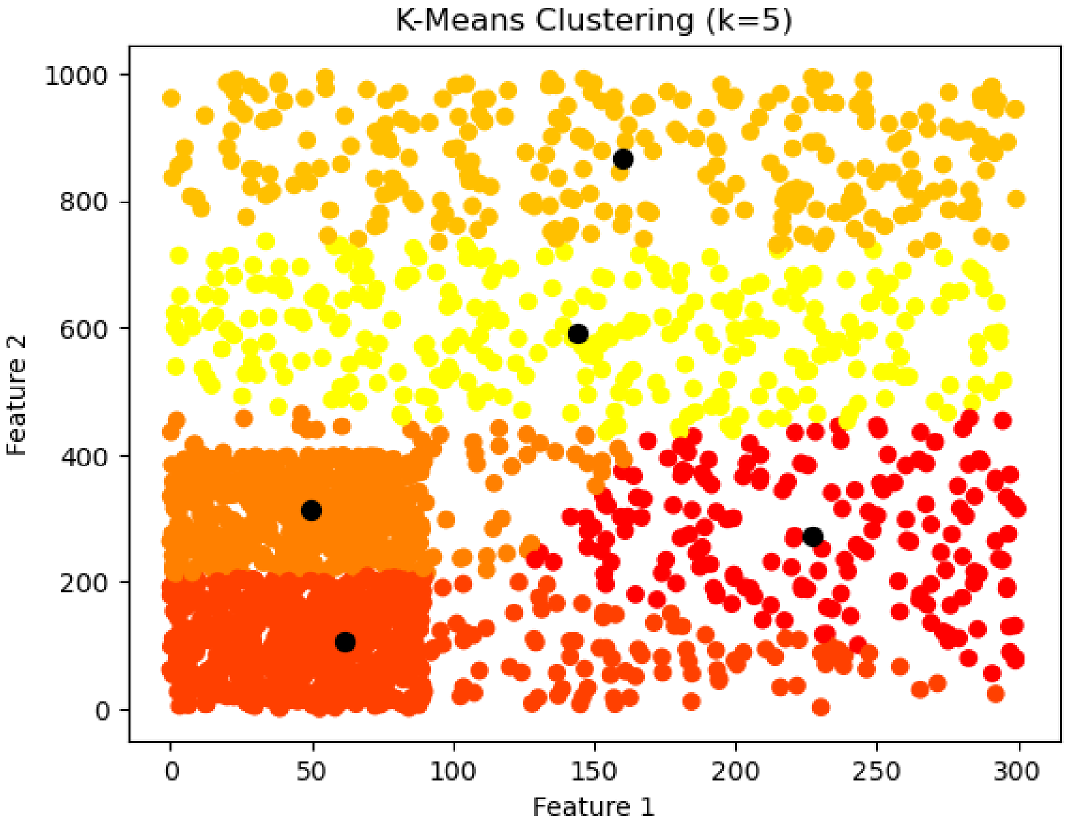

Figure 1.4: K-means cluster with k = 5.

1.3 Problem-2

The Wikipedia article document has been provided. I had to manually identify 5 keywords in these documents and use them to represent documents using Bag of Words(BOW). I have clustered them in 2 clusters and shared my observations.

Solution 2:

```
#!/usr/bin/env python
# coding: utf-8

# In[9]:


from sklearn.feature_extraction.text import CountVectorizer
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.cluster import KMeans

f1='C:/Users/user/FML␣+␣IMAGE_PROCESSING+AI_B/FML_Assignments/A_9/dataset/n1.txt'
f2='C:/Users/user/FML␣+␣IMAGE_PROCESSING+AI_B/FML_Assignments/A_9/dataset/n2.txt'
f3='C:/Users/user/FML␣+␣IMAGE_PROCESSING+AI_B/FML_Assignments/A_9/dataset/n3.txt'
f4='C:/Users/user/FML␣+␣IMAGE_PROCESSING+AI_B/FML_Assignments/A_9/dataset/n4.txt'
f5='C:/Users/user/FML␣+␣IMAGE_PROCESSING+AI_B/FML_Assignments/A_9/dataset/n5.txt'

paths=[f1,f2,f3,f4,f5]

output = []

for path in paths:
    with open(path, 'r', encoding='utf-8') as file:
        content = file.read()
        output.append(content.lower())

# The 5-Keywords

keywords = ["tendulkar", "politician", "australian", "minister", "economic"]

#BOW

countvectorizer = CountVectorizer(vocabulary=keywords)
X = countvectorizer.fit_transform(output)

kmeans = KMeans(n_clusters=2, random_state=0).fit(X)
print(kmeans.labels_)



```

```

# In[ ]:
```

CONCLUSION:

[1	**0**	**0**	**0**	**0**]

Based on the above result I have made two clusters and fitted them accordingly in 1 and 0 labels. The manually chosen keywords are tendulkar, politician, australian, minister, economic. Hence the conclusion observed as [**1—0—0—0—0**]