CS564 Foundations of Machine Learning
Assignment: 2

Mukuntha N S (1601CS27)
R J Srivatsa (1601CS35)
Shikhar Jaiswal (1601CS44)

August 27, 2019

1 Assignment Description

In the computational biology field, identifying the correct set of clusters helps to analyze the characteristics of the genes that helps to understand the unknown gene characteristics. Also, finding the correct cluster centers is one of the important steps for getting good results. We try to find the correct number of clusters for the **B-Chronic Lymphocytic Leukemia(BCLL) Dataset** .

2 Procedure

Installation

Install the following dependencies either using pip or through conda in a Python 3.5+ environment:
• jupyter
• pandas
• sklearn
• matplotlib

```
python3 -m pip install jupyter pandas sklearn matplotlib
```

or alternatively,

```
conda install -c anaconda jupyter pandas sklearn matplotlib
```

Running The Notebook

To run the program, head to the directory $Assignment2/Q1$. Please note that the **dataset should be present in the same directory** as the jupyter notebook. Use the following command to run the notebook:

```
jupyter notebook Q1.ipynb
```

3 Discussion

The primary objective of the assignment is to understand the change of silhouette score along with cluster centers while running K-Means algorithm. The following sub-sections contain the explanation for individual code snippets. For a detailed look at the output, please refer the notebook and the individual files provided.

Notebook Code

Pre-processing and Visualization

Import the Python dependencies, and check the data samples and their values.

```
# Import the required libraries.
import re
import math
import random
import collections
import pandas as pd
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
import matplotlib.pyplot as plt


df_list = []

# Function for loading the data into a pandas dataframe.
with open('./preprocessed_BCLL.txt') as f:
    for line in f:
        line = line.strip()
        columns = re.split('\s+', line, maxsplit = 23)

        for i, column in enumerate(columns):
            if i >= 2:
                try:
                    columns[i] = float(column)
                except:
                    columns[1] = columns[1] + ' ' + columns[2]
                    del(columns[2])

                    for j, new_column in enumerate(columns):
                        if j >= 2:
                            columns[j] = float(new_column)

        df_list.append(columns)

data = pd.DataFrame(df_list)

print(data)
```

Part A

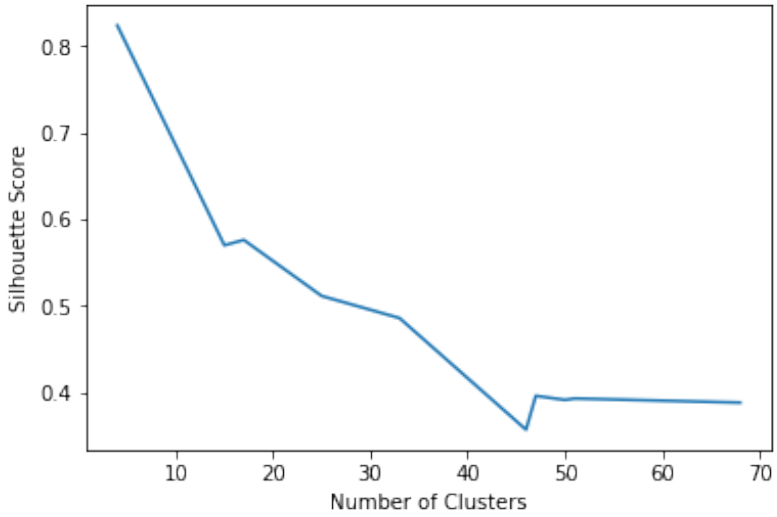
We run the K-Means algorithm 10 times where each time the number of clusters is randomly chosen between 2 to sqrt(N), where N represents the number of data points in the dataset. For each iteration, the Silhouette score is reported. Finally, we plot a graph, to understand the change of Silhouette score along with the number of cluster centers.

```
# Get all the numeric data from the original dataframe.
X = data._get_numeric_data().reset_index(drop = True)
# Normalize the dataframe.
X = X.apply(lambda column: (column - column.mean()) / column.std(), axis = 0)

SSE = {}
max_silhouette_score = -2
max_silhouette_score_cluster = 0

# Run the loop 10 times.
for k in range(1, 11):
    clusters = random.randint(2, int(math.sqrt(X.shape[0])))

    # Check for the previously seen values.
    if clusters in SSE.keys():
        clusters = random.randint(2, int(math.sqrt(X.shape[0])))

    # Get the labels for each sample.
    labels = KMeans(n_clusters = clusters, random_state = 42).fit_predict(X)
    # Get the silhouette score for each sample.
    score = silhouette_score(X, labels, metric = 'euclidean', sample_size = None,
                             random_state = 42)

    # Update the maximum score seen, and the cluster number which gives that score.
    if score > max_silhouette_score:
        max_silhouette_score = score
        max_silhouette_score_cluster = clusters

    # Store the number of clusters which give a particular score.
    SSE[clusters] = score
    print('Iteration: ', k, ' Number of Clusters: ', clusters, ' Silhouette Score: ', score)

# Order the cluster numbers.
sorted_SSE = sorted(SSE.items(), key = lambda kv: kv[0])
sorted_SSE = collections.OrderedDict(sorted_SSE)
# Plot the graph for silhouette score vs number of clusters.
plt.plot(list(sorted_SSE.keys()), list(sorted_SSE.values()))
plt.xlabel('Number of Clusters')
plt.ylabel('Silhouette Score')
```

```
Output:
Iteration:  1  Number of Clusters:  50  Silhouette Score:  0.3917516736746493
```

```
Iteration:  2  Number of Clusters:  4  Silhouette Score:  0.8238531718751286
Iteration:  3  Number of Clusters:  17  Silhouette Score:  0.5762882703193228
Iteration:  4  Number of Clusters:  47  Silhouette Score:  0.3964541504823024
Iteration:  5  Number of Clusters:  51  Silhouette Score:  0.39339718944920465
Iteration:  6  Number of Clusters:  25  Silhouette Score:  0.5116888578964914
Iteration:  7  Number of Clusters:  33  Silhouette Score:  0.48617397028658754
Iteration:  8  Number of Clusters:  68  Silhouette Score:  0.3886428959878027
Iteration:  9  Number of Clusters:  15  Silhouette Score:  0.5701544907503854
Iteration:  10  Number of Clusters:  46  Silhouette Score:  0.3574171449212913
```

Plot



Part B

For the iteration, with the highest Silhouette score, we show the gene names for each cluster in individual files, where each file represents a particular cluster.

```python
print('Maximum Silhouette Score: ', max_silhouette_score)
print('Number of Clusters: ', max_silhouette_score_cluster)

# Get the labels for each sample for the case where we see the maximum silhouette score.
labels = KMeans(n_clusters = max_silhouette_score_cluster, random_state = 42).fit_predict(X)

# Append the labels to the respective sample.
data['23'] = labels

# Output the clusters to individual files.
for label in set(labels.tolist()):
    cluster_data = data.loc[data['23'] == label, 1]
    cluster_data.to_csv('Cluster_' + str(label + 1) + '.csv', header = False, index = False)
```

```
Output:
Maximum Silhouette Score:  0.8238531718751286
Number of Clusters:  4
```

Additionally please refer $Q1/Cluster\_1.csv$ and similar files for the output.