**CSC 8850 – Homework Assignment 3**

**Sneha Gavini** – 002484406 - 11/01/2019

**Documentation for MDS,PCA & K-Means**

**1) Script for the conversion of DNA sequences from the dataset into numerical format.**

Firstly the data file is read and divided into name of the sequence list and DNA sequence list.

# Read in the file and create sequence list and dna list

def read\_file(file\_name):

sequence\_list = []

dna\_list = []

dna\_string\_list = []

with open(file\_name, 'r') as f:

count = 0

for i in f:

if count % 2 == 0:

sequence\_list.append(i[1:-1])

if count % 2 == 1:

dna\_list.append(list(i[:-1]))

dna\_string\_list.append(i[:-1])

count += 1

return sequence\_list, dna\_list, dna\_string\_list

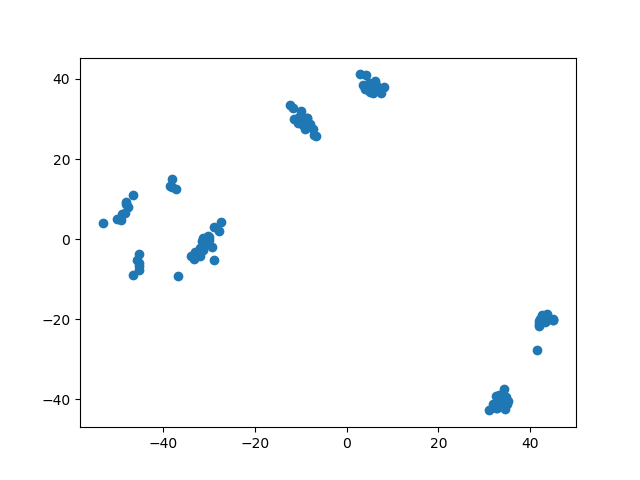
**Justification:**

At the maximum we have four possible classes in each column and these classes are ‘a’, ‘c’, ‘g’, ‘t’. Here each column is categorical and we can assign each of these characters or classes a numeric value where each number represents the character. We cannot give any numeric value as these might affect the calculations for example:

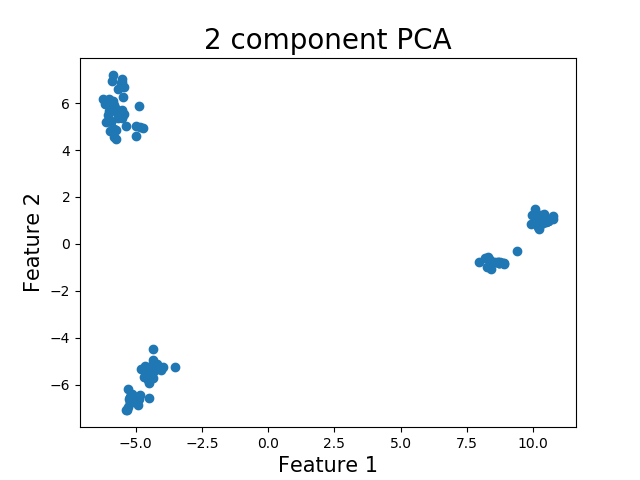
If we give the numbers 1, 100, 1000, 10000 to the classes 1 becomes insignificant.

Hence the numbers should not to be too widespread. Here each of the character has been given 1,2,3,4.

**2) MDS - Multi Dimensional Scaling:**



**PCA – Principal Component Analysis**



By observing the sequence names we can deduce that there are five clusters.

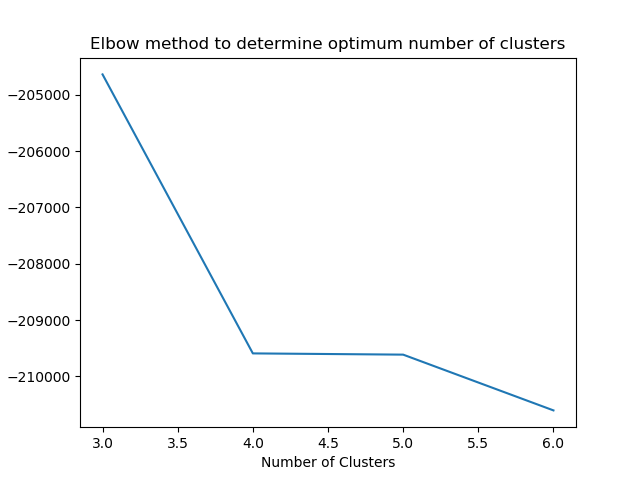
* AA8
* AC123
* AC121
* AN12
* AA20

In the 2 component PCA we can see that the five clusters have been detected, where as in the MDS the number of clusters visually is more than 5. So PCA is better than MDS.

**3)** In order to find out the best value for the k in K-means clustering, elbow method is used, where we start the number of clusters as 1 and increase the clusters by one. The process is repeated until we get a saturation point such that the sum of squares of the distances remain same.

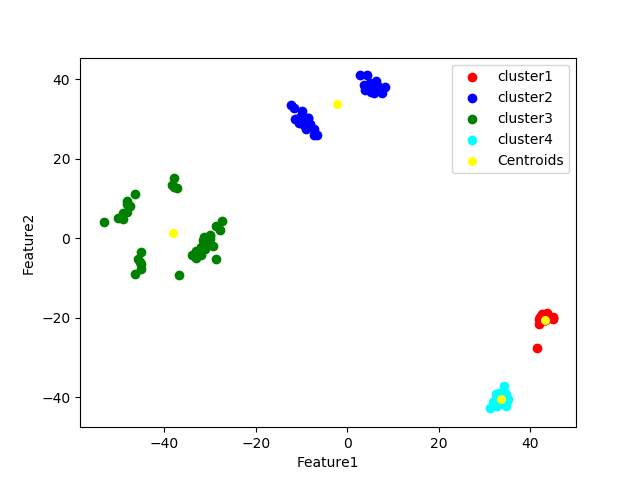
For this problem the number of clusters are started from 3 and got all the sum of squares till 6 and these are shown in the figure below.

**MDS – ELBOW GRAPH**

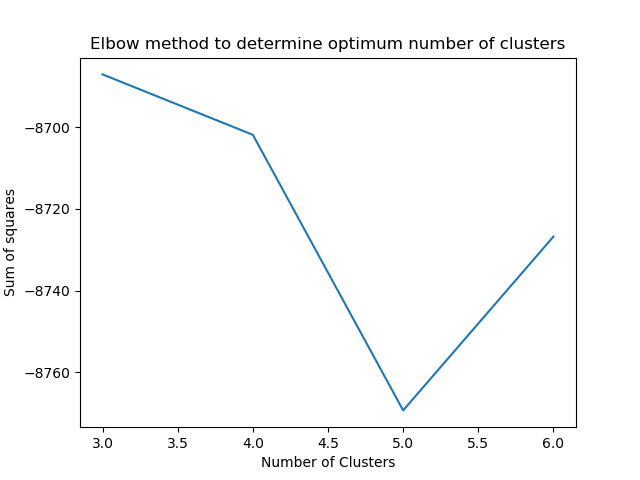


from the above graph we can see that the saturation point is reached at number of clusters = 4. The graph for the four number of clusters for MDS is as shown below.

**MDS – Clusters**

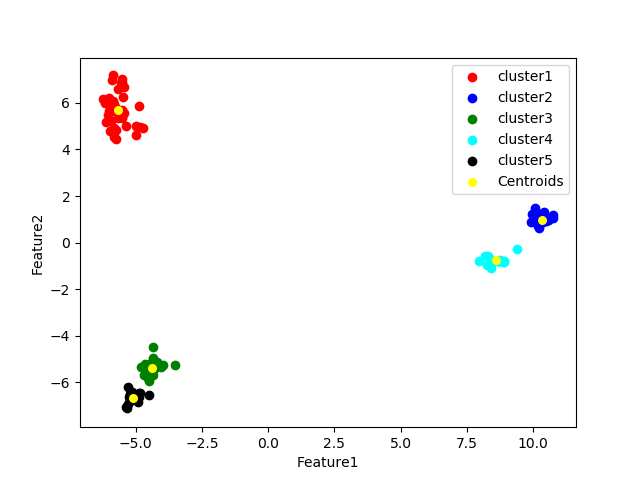


**PCA – ELBOW GRAPH**



from the above graph we can see that the saturation point or the minimum sum of squares is reached at number of clusters = 5. The graph for the five number of clusters for PCA is as shown below.

**PCA – Clusters**



From the two cluster graphs we can conclude that the K-Means over **PCA gives more meaningful clusters than K-Means over the MDS.**