SMAI Assignment – 3

Submitted by: Ayush Kumar Dwivedi (2018802002)

Answer to Q-1-2-: Dimensionality reduction using PCA

```
PCA Started !

Number of eigen values selected to maintain threshold at 10% is: 14

PCA Reduced Data

[[-2.50946520e+00  9.52118043e-01  8.05755168e-02 ...  8.50312294e-02
  -4.78761528e-01  3.80057517e-01]
 [ 1.19325993e+00 -5.59820431e+00  2.70795585e-01 ...  6.43676015e-02
   4.30117879e-01  2.64015116e-01]
 [-2.44710014e+00  9.08939709e-01  6.03192180e-02 ...  5.04823140e-02
  -7.54360905e-01  4.29615250e-01]
 ...
 [-2.14676535e-01 -3.09877590e-01 -1.37272986e+01 ...  1.72825356e+01
  -1.61696242e+00  1.79863104e+00]
 [-3.46663135e+00  5.46446031e-01 -7.44319096e-02 ...  3.55261207e-01
   4.57109427e-01 -7.13572892e-01]
 [ 4.01832089e+00  1.12480793e+00 -6.01524597e-02 ... -3.89962407e-03
   2.80523132e-01  2.18831703e-01]]

PCA Completed !
```

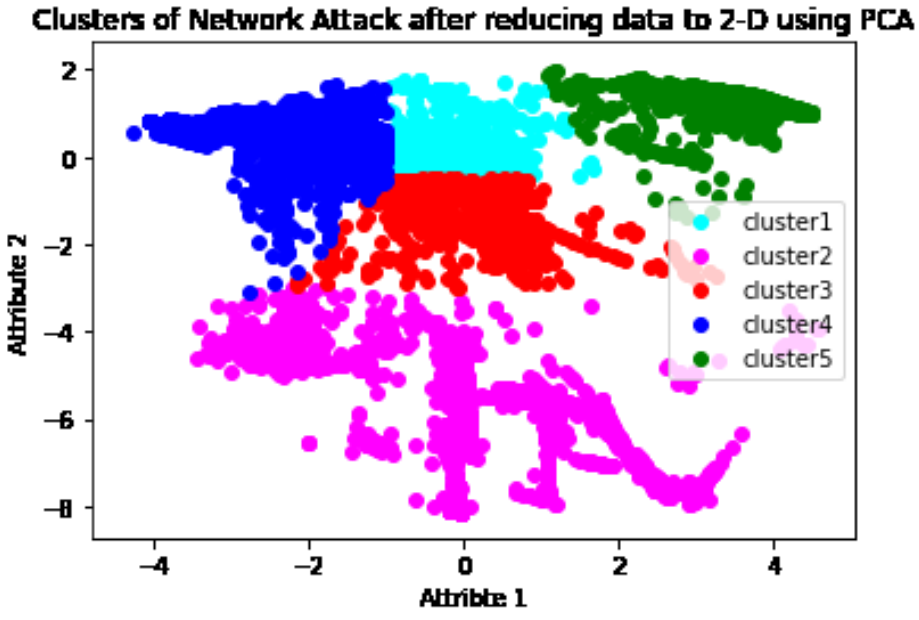
Answer to Q-1-2-: PCA for Categorical Values

Doing Clustering by selecting reduced number of dimentions in PCA as per threshold of 10%

```
Number of eigen values selected to maintain threshold at 10% is: 14

PCA Reduced Data

[[-2.50946520e+00  9.52118043e-01  8.05755168e-02 ...  8.50312294e-02
  -4.78761528e-01  3.80057517e-01]
 [ 1.19325993e+00 -5.59820431e+00  2.70795585e-01 ...  6.43676015e-02
   4.30117879e-01  2.64015116e-01]
 [-2.44710014e+00  9.08939709e-01  6.03192180e-02 ...  5.04823140e-02
  -7.54360905e-01  4.29615250e-01]
 ...
 [-2.14676535e-01 -3.09877590e-01 -1.37272986e+01 ...  1.72825356e+01
  -1.61696242e+00  1.79863104e+00]
 [-3.46663135e+00  5.46446031e-01 -7.44319096e-02 ...  3.55261207e-01
   4.57109427e-01 -7.13572892e-01]
 [ 4.01832089e+00  1.12480793e+00 -6.01524597e-02 ... -3.89962407e-03
   2.80523132e-01  2.18831703e-01]]

Clustering Started !!!

Initial Centroids Set to:
[[30 47 19 47 35 22 47 47 36 33 18 28 42 20]
 [22  1 53 10 19 36 22 39 25 29 20  4 41 40]
 [34 42 35 12 26 51 20 31 49  8 24  9 46 36]
 [28 25 20  7 36 15 31 14  3 16 34 47 53 51]
 [ 9 30 28 42 29 23 30 15 34 53 19 35 15 45]]

Distance moved by Centroids in next interation
[8.03514793e+09 1.10539586e+02 8.03514794e+09 1.29058126e+02
 2.05073158e+02]

Distance moved by Centroids in next interation
[ 0.          0.          0.         12.68857754  0.        ]

Distance moved by Centroids in next interation
[0.         0.         0.         6.70820393 0.        ]

Distance moved by Centroids in next interation
[0. 0. 0. 0. 0.]

Clustering Completed !!!


Purity while reducing data as per threshold:  0.5346427714217138
```

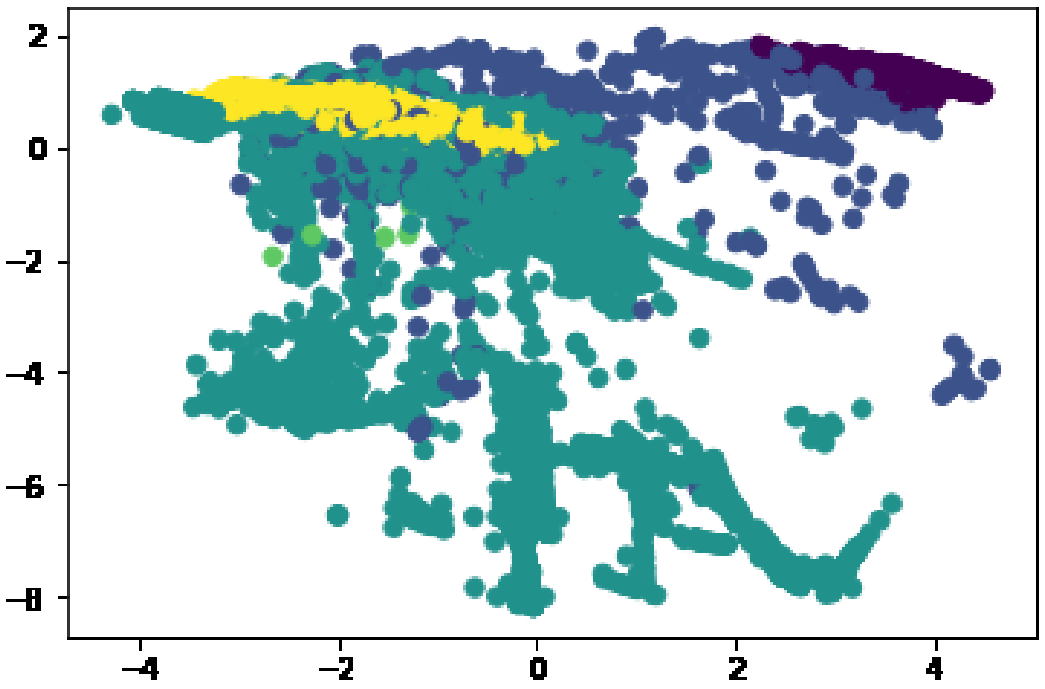
Doing Clustering by selecting reduced number of dimentions in PCA as 2 for getting plots

```
Number of eigen values selected to maintain threshold at 10% is: 2

PCA Reduced Data

[[-2.5094652   0.95211804]
 [ 1.19325993 -5.59820431]
 [-2.44710014  0.90893971]
 ...
 [-0.21467653 -0.30987759]
 [-3.46663135  0.54644603]
 [ 4.01832089  1.12480793]]

Clustering Started !!!

Initial Centroids Set to:
[[3 3]
 [1 0]
 [1 1]
 [3 1]
 [3 3]]

Distance moved by Centroids in next interation
[3.03700050e+09 2.23606798e+00 3.16227766e+00 0.00000000e+00
 3.03700050e+09]

Distance moved by Centroids in next interation
[0. 2. 1. 0. 0.]

Distance moved by Centroids in next interation
[0. 1. 0. 0. 0.]

Distance moved by Centroids in next interation
[0. 0. 0. 0. 0.]

Clustering Completed !!!

Purity while reducing data as per threshold:  0.837747019761581
```



Answer to Q-1-3: PCA for Categorical Values

```
PCA Started !

Number of eigen values selected to maintain threshold at 10% is: 14

PCA Reduced Data

[[-2.50946520e+00  9.52118043e-01  8.05755168e-02 ...  8.50312294e-02
  -4.78761528e-01  3.80057517e-01]
 [ 1.19325993e+00 -5.59820431e+00  2.70795585e-01 ...  6.43676015e-02
   4.30117879e-01  2.64015116e-01]
 [-2.44710014e+00  9.08939709e-01  6.03192180e-02 ...  5.04823140e-02
  -7.54360905e-01  4.29615250e-01]
 ...
 [-2.14676535e-01 -3.09877590e-01 -1.37272986e+01 ...  1.72825356e+01
  -1.61696242e+00  1.79863104e+00]
 [-3.46663135e+00  5.46446031e-01 -7.44319096e-02 ...  3.55261207e-01
   4.57109427e-01 -7.13572892e-01]
 [ 4.01832089e+00  1.12480793e+00 -6.01524597e-02 ... -3.89962407e-03
   2.80523132e-01  2.18831703e-01]]

PCA Completed !
GMM Started !!!

GMM Completed !!!


Purity while reducing data as per threshold:  0.7955036402912233
```



Answer to Q-1-5: PCA for Categorical Values

PCA can be applied in case of categorical data, but with a suitable non-linear transformation for each variable. Actually while trying to do PCA on categorical values, the difficulty comes in being able to represent distance between variable categories and individuals in feature space. There is a popular algorithm called as CATPCA which can be used. It converts categorical values to numerical ones using optimal scaling.