Kiana Vang
X500: vang0951
CSCI 5521

HW2 Q3 Written Responses

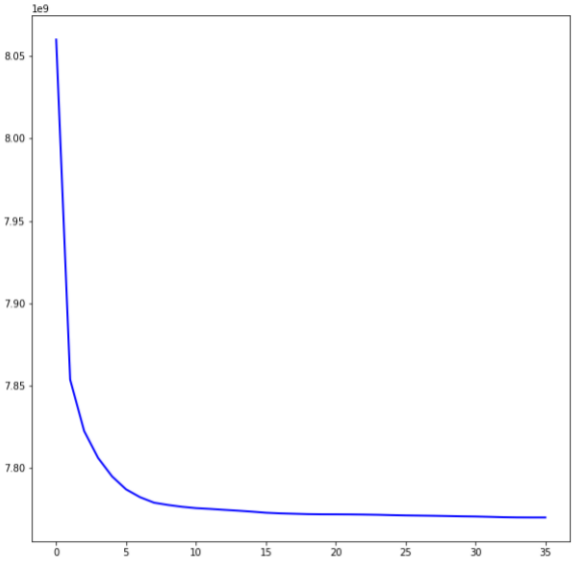
1) Raw Data

```
Using raw data converged in 36 iteration (8.49 seconds)
Contingency Matrix:
[[ 74. 438.  12.]
 [ 74. 487.  14.]
 [  1.  22. 325.]
 [  3.  21. 381.]
 [ 14.  24. 260.]
 [834.   8.   8.]]
##################
```

In the ideal case, we want majority of samples in a cluster to belong to a single class. According to the raw contingency matrix:

Samples of Class 0 mostly belong to Cluster 6 (left column)
Samples of Class 8 mostly belong to Clusters 1 and 2 (middle column)
Samples of Class 9 mostly belong to Clusters 3, 4, and 5 (right column)

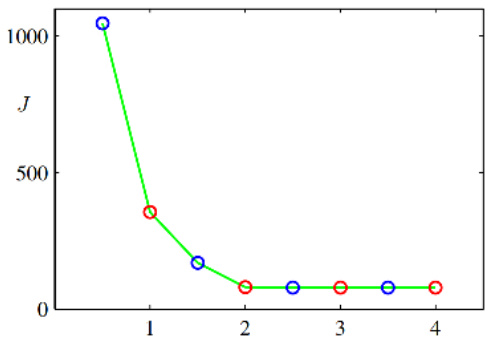
History of Reconstruction Error (J) for Raw-Data:



The reconstruction error represents the sum of the squares of the distances of each data point to its assigned vector (u_k). The goal is to minimize J as much as possible while iterating through E-step and M-step until the algorithm converges.

The shape of this plot is what I expected. Its shape is very similar to the example explained in Bishop's Ch9 text:

Figure 9.2 Plot of the cost function J given by (9.1) after each E step (blue points) and M step (red points) of the K-means algorithm for the example shown in Figure 9.1. The algorithm has converged after the third M step, and the final EM cycle produces no changes in either the assignments or the prototype vectors.



2) Low-Dimensionality obtained from PCA

```
Project data into 73 dimensions with PCA converged in 33 iteration (6.29 seconds)
Contingency Matrix:
[[ 75. 435.   7.]
 [ 74. 489.  14.]
 [  1.  27. 381.]
 [  2.  15. 326.]
 [ 14.  24. 264.]
 [834.  10.   8.]]
#################
```

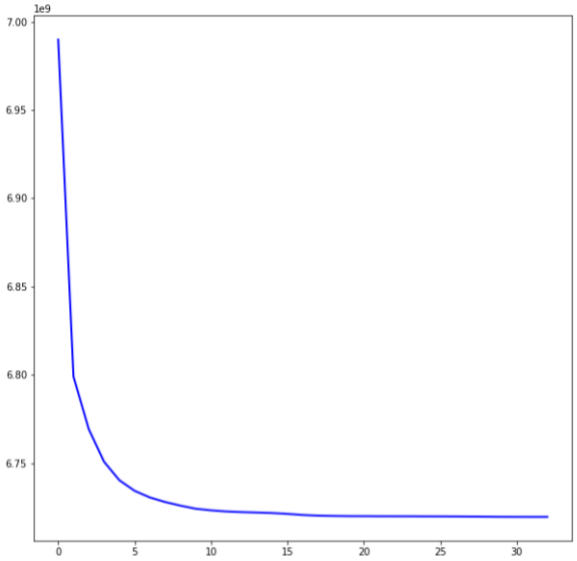
After running this algorithm successfully, a total of 73 dimensions is necessary to capture 90% of the variance. The algorithm converged in 33 iterations.

According to this contingency matrix:

Samples of Class 0 mostly belong to Cluster 6 (left column)
Samples of Class 8 mostly belong to Clusters 1 and 2 (middle column)
Samples of Class 9 mostly belong to Clusters 3, 4, and 5 (right column)

From this observation, this contingency matrix looks very similar to the raw-data version with some slight differences. In terms of number of iterations, this algorithm converged a bit faster than raw-data version. PCA helped with clustering, but only very slightly from this experiment.

History of Reconstruction Error (J) for PCA – 73 dimensions:



3) First Principal Component Only

```
Project data into 1 dimension with PCA converged in 71 iteration (14.13 seconds)
Contingency Matrix:
[[247.  44.  29.]
 [ 65. 228. 106.]
 [ 13. 566. 259.]
 [  2. 162. 603.]
 [379.   0.   3.]
 [294.   0.   0.]]
```

It is observed that using only the first principal component produced poorer results. Looking at contingency matrix results, it's hard to distinguish samples of class 8 and 9 belonging to cluster 3 or 4. The total number of iterations is also a lot higher than the previous experiments.

History of Reconstruction Error (J) for PCA – 1st PCA only: