# CSL 603 – Machine Learning

# Lab 4

# Clustering and Dimensionality Reduction

Jatin Goyal (2015csb1014)

Uneet Patel(2015csb1038)

# 1 . Introduction

The aim of this lab is to learn the concepts of clustering and dimensionality reduction techniques. In this, we were to classify the hand written digits' images correctly. To accomplish this task we performed k-means clustering technique on the given MNIST hand written digits' dataset. All the images are of dimension 20x20. Initially we processed the data on original input. For this, we vary the number of clusters and checked the accuracy on it. Then we performed PCA technique for dimensionality reduction and checked the accuracy on it.

# 2 . Pre – Processing

First of all we created the input matrix X from the given data.txt file. The dimension of this matrix is NxD (where N = 5000 , the number of instances and D = 400 , the dimensions). Then , output matrix Y was created from the given label.txt file. The dimension of this matrix is NxO (where O = 10 , the number of output classes (as there are 10 different digits possible)).

# 3 . Experiment – 1

To label each cluster with the most frequently occurring digit.

Each time when the experiment is run different output comes due to random initialization of the cluster centres.

The formulae used for calculating accuracy is :

$$\text{Accuracy} = M/N$$

Where M = Number of instances classified correctly

N = Total number of instances

**For K = 10**

Each column in this vector represents the label of the cluster.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 5 | 6 | 8 | 7 | 4 | 7 | 1 | 0 | 3 |

Classification Accuracy = 59.0200 %

On each run the accuracy is around 57-60 %. The reason behind this low accuracy is that some of the digits are not even labelled, as shown in above vector 9 is not labelled thus, all of those digits are misclassified. However, some of the digits have majority in more than one cluster so majority of those digits are correctly classified and therefore rest all the digits in that clusters are misclassified.

The average accuracy obtained on 100 runs is : 58.3200 %

The Confusion Matrix is as shown below.

| | | ACTUAL CLASS | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| | 0 | 392 | 0 | 3 | 1 | 0 | 5 | 6 | 1 | 0 | 2 |
| | 1 | 1 | 490 | 78 | 30 | 12 | 7 | 29 | 33 | 54 | 8 |
| | 2 | 0 | 0 | 323 | 10 | 2 | 0 | 5 | 1 | 2 | 0 |
| PREDICTED CLASS | 3 | 28 | 2 | 25 | 260 | 0 | 136 | 3 | 2 | 98 | 10 |
| | 4 | 5 | 0 | 14 | 5 | 240 | 22 | 42 | 28 | 20 | 129 |
| | 5 | 28 | 2 | 17 | 11 | 19 | 195 | 31 | 7 | 42 | 7 |
| | 6 | 17 | 0 | 14 | 2 | 8 | 11 | 383 | 0 | 3 | 2 |
| | 7 | 3 | 5 | 8 | 23 | 219 | 37 | 0 | 428 | 41 | 341 |
| | 8 | 26 | 1 | 18 | 158 | 0 | 87 | 1 | 0 | 240 | 1 |
| | 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Figure 1

As we can see that, instances corresponding to digit 9 are classified to either digit 4 or 7. Large number of instances belonging to digit 4 is classified as 7. This shows that the algorithm get confused and misclassify these instances. Almost all the Instances belonging to digit 1 are correctly classified.

**For K = 15**

Each column in this vector represents the label of the cluster.

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 7 | 3 | 6 | 6 | 4 | 0 | 0 | 2 | 1 | 1 | 8 | 9 | 8 | 5 |

Classification Accuracy = 67.4000 %

Average accuracy obtained is : 66.8200 %

As the number of clusters have increased and are more than the actual number of distinct digits. Some of the digits which were previously classified in single cluster are now divided in multiple clusters. The table is shown below:

| Digits | Number of clusters for K=10 | Number of clusters for K=15 |
|---|---|---|
| 6 | 1 | 2 |
| 3 | 1 | 2 |
| 0 | 1 | 2 |
| 8 | 1 | 2 |
| 1 | 1 | 2 |

On increasing the number of clusters some of the digits get split as the digits which were initially together are now partitioned but have the majority in the new clusters. But not all the digits get split and some of the digits are not even labelled. The number of misclassified instances decreases because some of the digits are now correctly classified and which were not labelled previously gets labelled. So, the classification accuracy increases.

The Confusion Matrix is:

| | | ACTUAL CLASS | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| | 0 | 446 | 0 | 6 | 1 | 0 | 8 | 8 | 2 | 2 | 2 |
| | 1 | 1 | 494 | 77 | 32 | 23 | 10 | 25 | 48 | 39 | 19 |
| | 2 | 0 | 0 | 304 | 6 | 4 | 1 | 0 | 1 | 4 | 1 |
| PREDICTED CLASS | 3 | 15 | 0 | 28 | 363 | 0 | 160 | 2 | 1 | 60 | 5 |
| | 4 | 0 | 0 | 6 | 0 | 271 | 4 | 0 | 6 | 8 | 69 |
| | 5 | 15 | 1 | 8 | 10 | 15 | 196 | 11 | 7 | 13 | 11 |
| | 6 | 15 | 0 | 48 | 3 | 13 | 6 | 450 | 1 | 2 | 10 |
| | 7 | 0 | 2 | 2 | 2 | 28 | 2 | 0 | 256 | 4 | 168 |
| | 8 | 6 | 1 | 16 | 70 | 0 | 93 | 3 | 1 | 364 | 9 |
| | 9 | 2 | 2 | 5 | 13 | 146 | 20 | 1 | 177 | 4 | 206 |

Figure 2

**For K=5**

Each column in this vector represents the label of the cluster.

| | | | | |
|---|---|---|---|---|
| 1 | 0 | 6 | 3 | 7 |

Classification Accuracy = 43.3600 %

Average accuracy obtained is : 44.6300 %

On decreasing the number of clusters the accuracy decreases by a huge amount. As, the number of clusters are less than the actual number of distinct digits so, majority of the instances are misclassified. Due to this some of the clusters get combined.

Yes the new clusters make sense because for 10 clusters as the digits 1, 0 and 6 have good classification. So, in the reduced clusters number these clusters have less variance and majority in that which leads to the selection of these labels. However some digits like 3, 5 and 8 get merge to single clusters because they do not have that good classification and the algorithm get confused on classifying them. Same case for digits 4, 7 and 9.

No the digits 7 and 1 didn't combined. This is because the digits 7 and 1 do not have much confusion and they have less overlapping. So, these digits remain separated.

The confusion matrix is :

| | | ACTUAL CLASS | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| | 0 | 424 | 0 | 3 | 3 | 0 | 9 | 7 | 2 | 2 | 2 |
| | 1 | 3 | 495 | 92 | 59 | 41 | 163 | 69 | 63 | 166 | 55 |
| | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| PREDICTED CLASS | 3 | 40 | 3 | 56 | 408 | 0 | 246 | 10 | 0 | 272 | 11 |
| | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 6 | 27 | 0 | 338 | 7 | 34 | 14 | 409 | 3 | 23 | 7 |
| | 7 | 6 | 2 | 11 | 23 | 425 | 68 | 5 | 432 | 37 | 425 |
| | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Figure 3

# Experiment – 2

In this experiment we used PCA to reduce the dimensions of the given images. In class, we mathematically derived the procedure to do PCA, and we found that the transformation matrix to give highest variance would consist of eigenvectors with largest eigen values. Here we calculated residual reconstruction error by re-projecting the data of transformed space to original space. The reconstruction method is mentioned in question PDF.

The observed variation of reconstruction error with the number of components is shown in table below:

| **Number of Components** | **Reconstruction Error** |
|---|---|
| 10 | 10.7802 |
| 20 | 6.8045 |
| 50 | 2.5164 |
| 80 | 1.1682 |
| 100 | 0.7445 |
| 120 | 0.4848 |
| 150 | 0.2552 |
| 180 | 0.1293 |
| 200 | 0.0785 |
| 250 | 0.0162 |
| 300 | 0.0014 |
| 350 | 9.7922e-6 |
| 400 | 0.000000 |

Here we can observe that as the number of components increases the reconstruction error keeps on decreasing and it is zero when we transformed for the same dimensions. This is because as the number of components keeps on decreasing more information loss occurs. Therefore, more error comes on lower dimensions.

On taking **191 components** the reconstruction error is under 0.1. This means that on taking only 191 components we don't loss much information however, the number of dimensions decreases significantly. So, we can run our experiments on the transformed data without losing much information.

Some of the images that are obtained in the new space .

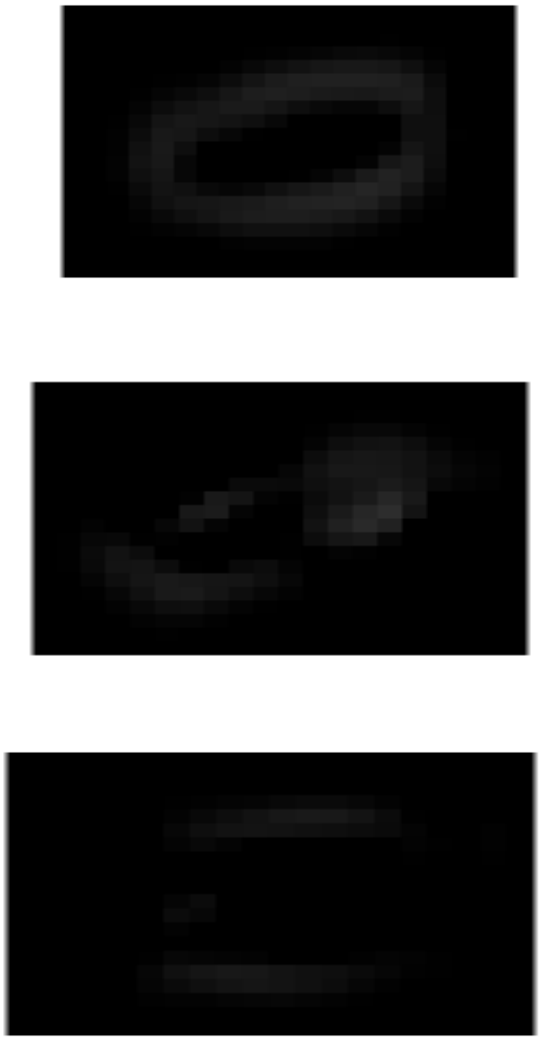

Figure 4 : The three principal components (with eigen values 1, 3, 5 respectively)

Scatter plots in 3-dimension of eigen vectors.

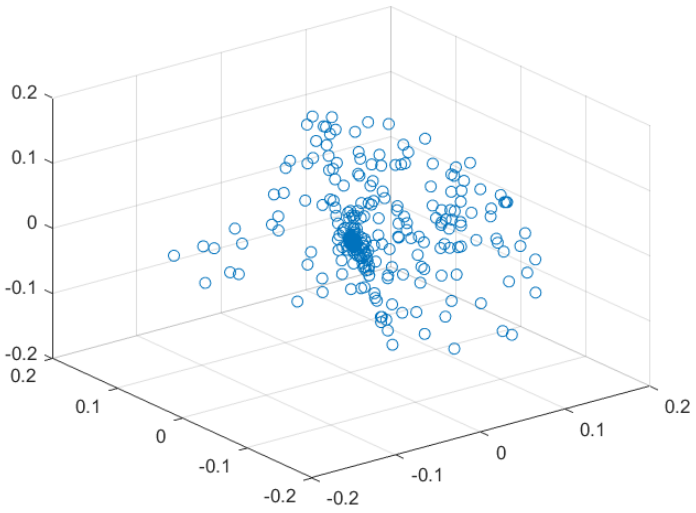


Figure 5 : Plot corresponding to eigen vector 1,2 and 3.

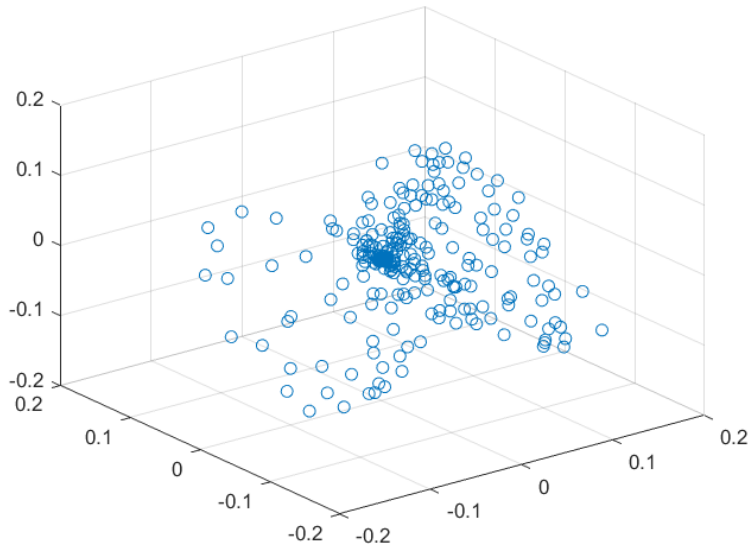


Figure 6 : Plot corresponding to eigen vector 1, 3 and 5.

# Experiment – 3

In this experiment we performed all the experiments of experiment – 1 but on the reduced dimension i.e. 191 components. We projected the original data onto 191 components using PCA with 0.1 reconstruction error.

**For K = 10**

Each column in this vector represents the label of the cluster.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 3 | 8 | 4 | 6 | 1 | 7 | 8 | 1 | 2 |

Classification Accuracy = 56.2400 %

Average Accuracy obtained is : 55.4300 %

The confusion matrix is :

| | | ACTUAL CLASS | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| | 0 | 394 | 0 | 3 | 1 | 0 | 4 | 7 | 1 | 0 | 2 |
| | 1 | 5 | 497 | 89 | 52 | 50 | 106 | 75 | 69 | 62 | 40 |
| | 2 | 4 | 0 | 327 | 12 | 7 | 1 | 6 | 2 | 4 | 1 |
| PREDICTED CLASS | 3 | 11 | 0 | 21 | 253 | 0 | 105 | 1 | 0 | 28 | 3 |
| | 4 | 6 | 0 | 15 | 9 | 296 | 21 | 11 | 155 | 3 | 227 |
| | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 6 | 17 | 0 | 16 | 1 | 10 | 10 | 386 | 1 | 2 | 2 |
| | 7 | 1 | 0 | 2 | 2 | 137 | 10 | 0 | 272 | 14 | 213 |
| | 8 | 62 | 3 | 27 | 170 | 0 | 243 | 14 | 0 | 387 | 12 |
| | 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Figure 7

As we can see that, majority of all classification remains same. However, some of the changes were observed like, 7 was having previously 2 clusters but now it has only one cluster. Now digit 8 is divided into 2 clusters which was having previously single cluster. However, instances corresponding to digit 9 are classified to either digit 4 or 7. Large

number of instances belonging to digit 4 is classified as 7. This shows that the algorithm get confused and misclassify these instances. Almost all the Instances belonging to digit 1 are correctly classified.

**For K = 15**

Each column in this vector represents the label of the cluster.

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 5 | 6 | 3 | 6 | 3 | 0 | 7 | 8 | 1 | 4 | 1 | 2 | 0 | 5 |

Classification Accuracy = 65.6800 %

Average Accuracy obtained is : 65.1700 %

The confusion matrix is :

| | | ACTUAL CLASS | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| | 0 | 413 | 0 | 3 | 1 | 0 | 5 | 3 | 2 | 2 | 3 |
| | 1 | 0 | 494 | 39 | 39 | 30 | 10 | 22 | 64 | 45 | 34 |
| | 2 | 1 | 1 | 393 | 6 | 8 | 1 | 7 | 4 | 4 | 4 |
| PREDICTED CLASS | 3 | 10 | 1 | 23 | 389 | 1 | 92 | 0 | 0 | 135 | 10 |
| | 4 | 1 | 0 | 9 | 11 | 273 | 29 | 4 | 146 | 7 | 221 |
| | 5 | 55 | 2 | 12 | 17 | 40 | 350 | 17 | 18 | 30 | 24 |
| | 6 | 18 | 0 | 5 | 2 | 16 | 7 | 447 | 0 | 2 | 2 |
| | 7 | 0 | 0 | 3 | 5 | 132 | 4 | 0 | 265 | 15 | 201 |
| | 8 | 2 | 2 | 13 | 30 | 0 | 2 | 0 | 1 | 260 | 1 |
| | 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Figure 8

As we can see that, majority of all classification remains same. However digit 5 was having only 1 cluster previously but now it has 2 clusters. On increasing the number of clusters some of the digits get split as the digits which were initially together are now partitioned

but have the majority in the new clusters. But not all the digits get split and some of the digits are not even labelled. The number of misclassified instances decreases because some of the digits are now correctly classified and which were not labelled previously gets labelled. So, the classification accuracy increases.

**For K = 5**

Each column in this vector represents the label of the cluster.

| | | | | |
|---|---|---|---|---|
| 7 | 3 | 0 | 6 | 1 |

Classification Accuracy = 43.2100 %

Average Accuracy obtained is : 43.7600 %

The confusion matrix is :

| | | ACTUAL CLASS | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| | 0 | 424 | 0 | 3 | 3 | 0 | 9 | 7 | 2 | 2 | 2 |
| | 1 | 3 | 495 | 92 | 59 | 41 | 163 | 69 | 63 | 166 | 55 |
| | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| PREDICTED CLASS | 3 | 40 | 3 | 56 | 408 | 0 | 246 | 10 | 0 | 272 | 11 |
| | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 6 | 27 | 0 | 338 | 7 | 34 | 14 | 409 | 3 | 23 | 7 |
| | 7 | 6 | 2 | 11 | 23 | 425 | 68 | 5 | 432 | 37 | 425 |
| | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Figure 9

In the reduced dimensions also on decreasing the number of clusters the accuracy decreases by a huge amount. As, the number of clusters are less than the actual number of

distinct digits so, majority of the instances are misclassified. Due to this some of the clusters get combined.

Yes the new clusters make sense because for 10 clusters as the digits 1, 0 and 6 have good classification. So, in the reduced clusters number these clusters have less variance and majority in that which leads to the selection of these labels. However some digits like 3, 5 and 8 get merge to single clusters because they do not have that good classification and the algorithm get confused on classifying them. Same case for digits 4, 7 and 9.

No the digits 7 and 1 didn't combined. This is because the digits 7 and 1 do not have much confusion and they have less overlapping. So, these digits remain separated.

On comparing the results obtained in this space and original space we found that the accuracy in this space decreases but not much. This is because we took 191 components which have only 0.1 construction error and therefore we didn't lose much information. But on the reduced dimensions the results were obtained more quickly and each iteration was processed faster.

**Thank you**