# CSL - 603
# LAB - 4 REPORT
# K-Means Clustering and PCA

Eeshaan Sharma
2015CSB1011
Shivam Mittal
2015CSB1032

November 2017

## 1 Introduction

The aim of this lab is to experiment with clustering and dimensionality reduction techniques. K-means clustering is implemented on hand written digits dataset and then Principal Component Analysis is used to project the original 400 dimensional data to lower dimensions.

## 2 Experiment 1

As a part of Experiment 1, K-means clustering is performed on the original dataset where each example is a 400 dimensional vector. The implementation used is the MATLAB inbuilt function to perform k-means clustering. $idx = kmeans(X, k)$ performs k-means clustering to partition the observations of the N X D data matrix X into k clusters, and returns an N X 1 vector (idx) containing cluster indices of each observation. By default, inbuilt kmeans function uses the squared Euclidean distance measure.//To measure accuracy each cluster is assigned the label of the most frequently occurring digit.

### 2.1 Observations

The number of cluster centres $k$ are varied and classification accuracy is measured. Confusion matrix is constructed to study the kinds of mis-classifications. The formula used for calculating accuracy is as follows -

$$Accuracy = \frac{N_C}{N} \tag{1}$$

$N_C$ is the number of examples that are classified correctly.

$N$ is the number of training examples.

### 2.1.1 k = 10

As a result of performing k-means clustering with a value of k = 10, different outputs are obtained on each run due to random initialization of the cluster centres which is taken care by the inbuilt function.

On a sample run the following cluster center labels are obtained when each cluster center is assigned the label of the most frequently occurring digit in it -

Cluster Center Labels - 4, 3, 1, 8, 1, 5, 2, 6, 0, 7

This shows that examples belonging to digit 1 are split into 2 separate clusters and there is no cluster which contains digit 9 as the most frequently occurring digit.

The accuracy obtained varies roughly from 56% to 58% in each run.

Average Accuracy obtained after 100 runs = 56.98%

Confusion Matrix obtained is as follows -

PREDICTED LABEL

| ACTUAL LABEL | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 394 | 3 | 0 | 26 | 8 | 27 | 16 | 0 | 26 | 0 |
| 1 | 0 | 490 | 0 | 1 | 1 | 1 | 0 | 0 | 8 | 0 |
| 2 | 4 | 79 | 326 | 28 | 11 | 18 | 14 | 4 | 16 | 0 |
| 3 | 1 | 52 | 11 | 258 | 14 | 12 | 2 | 5 | 145 | 0 |
| 4 | 0 | 36 | 7 | 0 | 293 | 48 | 9 | 107 | 0 | 0 |
| 5 | 5 | 20 | 0 | 141 | 37 | 182 | 10 | 3 | 102 | 0 |
| 6 | 8 | 51 | 12 | 2 | 10 | 31 | 383 | 0 | 3 | 0 |
| 7 | 1 | 66 | 1 | 0 | 151 | 12 | 1 | 268 | 0 | 0 |
| 8 | 0 | 75 | 7 | 99 | 13 | 36 | 3 | 18 | 249 | 0 |
| 9 | 2 | 38 | 1 | 7 | 233 | 21 | 1 | 196 | 1 | 0 |

Figure 1: Confusion Matrix for k = 10

The confusion matrix shows that the digits 0, 1 and 6 are being classified with high accuracy, which means they are discriminative enough in the pixel space (400 pixels).
A large number of examples which actually belong to digit 4 are predicted to belong to cluster of digit 7 and those belonging to digit 7 are predicted to

belong to cluster labelled with digit 4. This shows that the algorithm gets a little confused between the two digits and tends to misclassify examples belonging to them.
Also, similar errors are being made between digits 3 and 8. Also, classification for digit 5 is making mistakes and being classified as 3 and 8.
Majority examples belonging to digit 9 are predicted to belong to either cluster labelled with digit 4 or 7.
These observations can be attributed to the structural similarities between the digits, for example, 3 and 8 have common right side curvature, and 5 has lower curvature similar to 3 and 8. So, this gives rise to a possibility that the image corresponding to one digit, is more closer to other digits' cluster center than itself. It is perfectly understandable since the images do not have any high-level feature in which we perform clustering, we are performing clustering in the very low-level space of pixels, which are prone to rotation, scale and other variance.
Also, no cluster is being detected for the digit 9, this may be because in the pixel space in which we are performing clustering, there the distribution of images corresponding to digit 9 have high variance, they are not close to each other, and hence no cluster center is being allotted the label 9, because they couldn't contribute as the majority vote. Also it is perfectly plausible that the distribution of one class may not follow the spherical assumptions that k-means assume.

#### 2.1.2 k = 15

As a result of performing k-means clustering with a value of k = 15, as in the previous case different outputs are obtained on each run. Since the number of clusters exceeds the actual number of types of digits in the dataset (which is 10, 0-9), it is observed that some of the digits which earlier formed a single cluster when k = 10 have now split into multiple clusters.

This can be seen from the following observation where on a sample run the following cluster center labels are obtained when each cluster center is assigned the label of the most frequently occurring digit in it -

Cluster Center Labels - 0,0,1,2,3,3,4,4,5,5,6,7,7,8,8

This shows that examples belonging to digit 0,3,4,5,7 and 8 have split further into 2 separate clusters and as observed in majority of cases there is still no cluster which contains digit 9 as the most frequently occurring digit..

The accuracy obtained varies roughly around 67% in each run. This shows that when the number of cluster centers increase the accuracy increases.

Average Accuracy obtained after 100 runs = 67.14%

Confusion Matrix obtained is as follows -

PREDICTED LABEL

| ACTUAL LABEL | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 407 | 0 | 0 | 6 | 3 | 72 | 7 | 1 | 4 | 0 |
| 1 | 0 | 490 | 0 | 0 | 0 | 2 | 0 | 0 | 8 | 0 |
| 2 | 4 | 70 | 330 | 19 | 16 | 22 | 11 | 5 | 23 | 0 |
| 3 | 1 | 21 | 6 | 340 | 6 | 21 | 1 | 7 | 97 | 0 |
| 4 | 0 | 13 | 1 | 0 | 398 | 31 | 5 | 51 | 1 | 0 |
| 5 | 3 | 4 | 1 | 84 | 32 | 323 | 7 | 0 | 46 | 0 |
| 6 | 5 | 5 | 5 | 1 | 37 | 124 | 323 | 0 | 0 | 0 |
| 7 | 1 | 32 | 2 | 0 | 35 | 3 | 0 | 427 | 0 | 0 |
| 8 | 2 | 31 | 3 | 53 | 18 | 21 | 1 | 2 | 369 | 0 |
| 9 | 2 | 13 | 0 | 5 | 317 | 6 | 1 | 147 | 9 | 0 |

Figure 2: Confusion Matrix for k = 15

The confusion matrix shows that examples belonging to digits 0 - 8 have been in majority clustered correctly, thus they belong to cluster centers with the correct label. Only examples belonging to digit 9 have been mis-classified and are majorly predicted to belong to cluster belonging to digit 4 and some are predicted to belong to cluster labeled with digit 7.
These observations in the errors being made can be explained the explanation we gave above for k=10, that two letters are being confused because of structural similarity, and error being made because of the distribution of some specific class data.
Also, we observe that the accuracy obtained has increased when we increase the cluster centers from 10 to 15, this is because the radius of the clusters in the 10 cluster scenario were greater than the radius off the clusters in the 15 cluster scenario. There might be cases that instances in 10 cluster scenario were close to the boundary of a cluster, but still being classified as a part of the cluster because of the large radius, this error will be reduced as a result of increasing the number of cluster centers. Also, since increasing the number of cluster centers, decrease the cluster radius, this imposes more stronger similarity to be clustered together, hence reducing the error.

### 2.1.3 k = 5

As a result of performing k-means clustering with a value of k = 5, as in the previous case different outputs are obtained on each run. Since the number of clusters are less than the actual number of types of digits in the dataset (which is 10, 0-9), it is obvious that only 5 classes would be predicted now, corresponding to the labels assigned to the cluster centers.

This can be seen from the following observation where on a sample run the following cluster center labels are obtained when each cluster center is assigned

the label of the most frequently occurring digit in it -

Cluster Center Labels - 3, 7, 0, 1, 6

The accuracy obtained varies roughly around 43% in each run. This shows that when the number of cluster centers decrease the accuracy decreases.

Average Accuracy obtained after 100 runs = 43.20%

Confusion Matrix obtained is as follows -

| | | PREDICTED LABEL | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| ACTUAL LABEL | 0 | 423 | 3 | 0 | 40 | 0 | 0 | 27 | 7 | 0 | 0 |
| | 1 | 0 | 495 | 0 | 3 | 0 | 0 | 0 | 2 | 0 | 0 |
| | 2 | 3 | 92 | 0 | 56 | 0 | 0 | 338 | 11 | 0 | 0 |
| | 3 | 3 | 59 | 0 | 408 | 0 | 0 | 7 | 23 | 0 | 0 |
| | 4 | 0 | 41 | 0 | 0 | 0 | 0 | 34 | 425 | 0 | 0 |
| | 5 | 9 | 163 | 0 | 247 | 0 | 0 | 14 | 67 | 0 | 0 |
| | 6 | 7 | 69 | 0 | 10 | 0 | 0 | 409 | 5 | 0 | 0 |
| | 7 | 2 | 64 | 0 | 0 | 0 | 0 | 3 | 431 | 0 | 0 |
| | 8 | 2 | 166 | 0 | 272 | 0 | 0 | 23 | 37 | 0 | 0 |
| | 9 | 2 | 56 | 0 | 11 | 0 | 0 | 7 | 424 | 0 | 0 |

Figure 3: Confusion Matrix for k = 5

The accuracy obtained is lower than the previous runs of the algorithm, this is obvious as only 5 classes will be predicted now, and the maximum accuracy can be 50%.

As number of clusters have reduced, some of the clusters have combined.
Do the new clusters make any sense? Yes the new clusters make sense because in the case of 10 cluster centers, we had observed that 0, 1 and 6 had the best classification accuracy as mentioned in the observations above. So, these clusters labels being obtained in the case of reduced clusters make sense, as the distribution of the data of these classes have less variance giving rise to the majority vote which leads to the label being selected. Also, the classes 3, 5 and 8 were being confused, so these classes merged into one. Also 9, 4 and 7 were being confused, so these classes merged into one. This can also be seen from the confusion matrix that the label predicted for actual clas 4,7 and 9 are majorily 7. And same is the case with 3, 5 and 8.
Do you observe that clusters with digits 7 and 1 get combined? No, the digits 7 and 1 have different cluster centers. This is expected because the confusion between 7 and 1 was minimal, and is still minimal as seen from the confusion matrices.

# 3 Experiment 2

In this experiment, we used PCA to reduce the dimensionality of the digit images. Principal Component Analysis (PCA) is the general name for a technique which uses sophisticated underlying mathematical principles to transforms a

number of possibly correlated variables into a smaller number of variables called principal components. In general terms, PCA uses a vector space transform to reduce the dimensionality of large data sets. Using mathematical projection, the original data set, which may have involved many variables, can often be interpreted in just a few variables (the principal components). It is therefore often the case that an examination of the reduced dimension data set will allow the the user to spot trends, patterns and outliers in the data, far more easily than would have been possible without performing the principal component analysis. (Parts of the PCA theory taken from here :
http://www.dsc.ufcg.edu ta.br/ hmg/disciplinas/
posgraduacao/rn-copin-2014.3/material/SignalProcPCA.pdf) In class, we had mathematically derived the procedure to do PCA, and we found that the transformation matrix to give highest variance would consist of eigenvectors with largest eigenvalues. We calculate the residual reconstruction error by re-projecting the transformed data to the original space. The re-construction can be achieved by just multiplying the transformed data with the transformation matrix U and adding the mean of the original data. We used mean square error to calculate the reconstruction error.

### 3.1 Number of components (dimesion of new spaace) vs reconstruction error

The table and graph which shows the reconstruction error as a function of the number of components chosen in PCA is shown.

Table 1: Reconstruction error

| Number of components | Reconstruction error |
|---|---|
| 10 | 10.780197 |
| 30 | 3.375556 |
| 50 | 2.516359 |
| 100 | 0.744528 |
| 150 | 0.255230 |
| 200 | 0.0784978 |
| 250 | 0.016222 |
| 300 | 0.001365 |
| 350 | 0.000010 |
| 400 | 0.000000 |

We can see that the reconstruction error decreases as the number of components taken in PCA increases, this is because as more number of components are taken in PCA, the amount of variance keeps on increasing, and more of the variations in PCA are being captured. As more and more information is being incorporated, the reconstruction error (measure of information loss) is being minimized. We observe that the error becomes 0 when we take 400 principal components (all the dimensions), because all the information and the variance

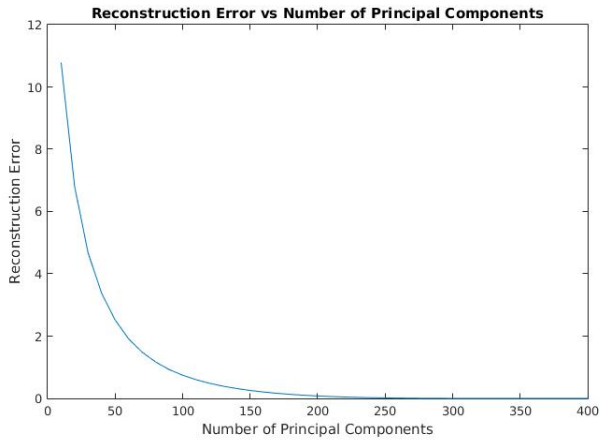


Figure 4: Reconstruction error

gets captured.

### 3.1.1 Number of components such that the residual error is under 0.1

If we take 191 components, then the residual error drops below 0.1. This shows a very important observation, it shows that if we taken all the 400 components then we get 0 reconstruction error. But if we take 191 components which is significantly less than 400, then we get minimal reconstruction error, i.e, most of the information is being captured in the 191 components being taken. Hence experiments can be run on this lower dimension space instead of the original space which may contains very very less extra information than the reduced dimension space, but at the cost of so many increased dimensions. Because of these results, PCA is one of the best things that came out of linear algebra.

### 3.1.2 What do the components represent?

Figure 5: The 3 principal components (with highest eigenvalues)

The principal components are basically those directions in which the variance for the data is maximum, so the components which we show in the form of images

will have bright values at those spot corresponding to whose pixel values, there is maximum variance in all the data.
From the above image, we can see the pixels where maximum variance in the data is there, this is somewhat related to the structural form of the digits.

# 4 Experiment 3

As a part of Experiment 3, K-means clustering is performed on the data projected onto lower dimensions. For the purpose of observations each example is now projected onto 191 principal components as reconstruction error is less than 0.1 when data is projected onto 191 components. The implementation used to implement k-means clustering is the same as that of Experiment 1.

## 4.1 Observations

The number of cluster centres $k$ are varied and classification accuracy is measured. Confusion matrix is constructed to study the kinds of mis-classifications. The formula used for calculating accuracy is the same as used in Experiment 1

$$Accuracy = \frac{N_C}{N} \tag{2}$$

$N_C$ is the number of examples that are classified correctly.

$N$ is the number of training examples.

### 4.1.1 k = 10

As a result of performing k-means clustering with a value of k = 10, different outputs are obtained on each run due to random initialization of the cluster centres which is taken care by the inbuilt function.

On a sample run the following cluster center labels are obtained when each cluster center is assigned the label of the most frequently occurring digit in it -

Cluster Center Labels - 5,3,1,4,8,7,6,0,7,2

This shows that examples belonging to digit 7 are split into 2 separate clusters and there is no cluster which contains digit 9 as the most frequently occurring digit.

The accuracy obtained varies roughly from 55% to 57% in each run, which is pretty close to the accuracy obtained in Experiment 1 with k = 10. Thus accuracy does not change much after reducing the dimensions of the data, as data PCA projects data onto principal components that are uncorrelated, and are ordered by the fraction of the total information each retains, thus information

loss is minimum and with 191 principal components the reconstruction error is also less than 0.1

Average Accuracy obtained after 100 runs = 56.2%

Confusion Matrix obtained is as follows -

| | | PREDICTED LABEL | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| ACTUAL LABEL | 0 | 392 | 1 | 0 | 27 | 5 | 27 | 17 | 3 | 28 | 0 |
| | 1 | 0 | 489 | 0 | 2 | 0 | 3 | 0 | 5 | 1 | 0 |
| | 2 | 3 | 79 | 323 | 25 | 14 | 17 | 14 | 6 | 19 | 0 |
| | 3 | 1 | 29 | 11 | 247 | 5 | 11 | 2 | 23 | 171 | 0 |
| | 4 | 0 | 12 | 1 | 0 | 238 | 22 | 10 | 217 | 0 | 0 |
| | 5 | 5 | 7 | 0 | 138 | 17 | 185 | 10 | 37 | 101 | 0 |
| | 6 | 7 | 29 | 4 | 3 | 43 | 29 | 382 | 0 | 3 | 0 |
| | 7 | 1 | 33 | 1 | 2 | 28 | 9 | 0 | 426 | 0 | 0 |
| | 8 | 0 | 54 | 2 | 102 | 20 | 44 | 3 | 42 | 233 | 0 |
| | 9 | 2 | 8 | 0 | 10 | 127 | 8 | 2 | 342 | 1 | 0 |

Figure 6: Confusion Matrix for k = 10

Similar to the results obtained in Experiment 1, the confusion matrix shows that the digits 0, 1,6 and 7 are being classified with high accuracy, which means they are discriminative enough even in the reduced pixel space (191 pixels).
As observed in Experiment 1, a large number of examples which actually belong to digit 4 are still predicted to belong to cluster of digit 7 but the same error is not repeated for those those belonging to digit 7.
Similar errors between the digits 3 and 8 are still prevelant and similar error in the classification for digit 5 still exists as it is being classified as 3 and 8.
Again on similar lines majority examples belonging to digit 9 are predicted to belong to either cluster labelled with digit 4 or 7.
These observations can be attributed to the structural similarities between the digits as was the case in Experiment 1, Even in the lower dimensional space it is still the case that images do not have any high-level feature in which we perform clustering, we are performing clustering in the very low-level space of pixels, which are prone to rotation, scale and other variance.
Also, again since no cluster is being detected for the digit 9, this may be because even in the lower dimensional pixel space in which we are performing clustering, there the distribution of images corresponding to digit 9 have high variance, they are not close to each other, and hence no cluster center is being allotted the label 9, because they couldn't contribute as the majority vote. It can be seen that even in the lower dimensional space it is perfectly plausible that the distribution of one class may not follow the spherical assumptions that k-means assume.

#### 4.1.2 k = 15

As a result of performing k-means clustering with a value of k = 15, as in the previous case different outputs are obtained on each run. Similar to the case of performing clustering with k = 15 as in Experiment 1, since the number of

clusters exceeds the actual number of types of digits in the dataset (which is 10, 0-9), it is observed that some of the digits which earlier formed a single cluster when k = 10 have now split into multiple clusters.

This can be seen from the following observation where on a sample run the following cluster center labels are obtained when each cluster center is assigned the label of the most frequently occurring digit in it -

Cluster Center Labels - 0,0,1,1,2,3,3,4,4,5,6,7,7,8,9

This shows that examples belonging to digit 0,1,3,4 and 7 have split further into 2 separate clusters. This also shows that all digits are now majority occurring digits in one of the clusters and thus the error which was made in Experiment 1 with k = 15 where no cluster center was assigned to digit 9 is not repeated as using PCA the data is now projected onto lower dimensions where each dimension retains maximum information

The accuracy obtained varies roughly around 67% in each run. This is roughly similar to one obtained in Experiment 1 and shows that when the number of cluster centers increase the accuracy increases.

Average Accuracy obtained after 100 runs = 67.04%

Confusion Matrix obtained is as follows -

| | | PREDICTED LABEL | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| ACTUAL LABEL | 0 | 438 | 1 | 0 | 21 | 4 | 15 | 14 | 1 | 4 | 2 |
| | 1 | 0 | 494 | 0 | 2 | 1 | 1 | 0 | 0 | 2 | 0 |
| | 2 | 6 | 77 | 325 | 24 | 22 | 10 | 16 | 6 | 11 | 3 |
| | 3 | 1 | 33 | 8 | 387 | 7 | 12 | 3 | 3 | 32 | 14 |
| | 4 | 0 | 21 | 3 | 0 | 370 | 11 | 3 | 16 | 0 | 76 |
| | 5 | 8 | 9 | 1 | 209 | 20 | 198 | 10 | 0 | 7 | 38 |
| | 6 | 8 | 37 | 4 | 4 | 43 | 23 | 381 | 0 | 0 | 0 |
| | 7 | 0 | 31 | 0 | 0 | 23 | 2 | 0 | 342 | 1 | 101 |
| | 8 | 2 | 39 | 3 | 137 | 17 | 16 | 2 | 3 | 251 | 30 |
| | 9 | 2 | 8 | 0 | 11 | 182 | 2 | 2 | 127 | 0 | 166 |

Figure 7: Confusion Matrix for k = 15

The confusion matrix shows that examples belonging to every digit, except digit 5 and 9 have been in majority clustered correctly, thus they belong to cluster centers with the correct label. Examples belonging to digit 5 haven been majorly predicted to belong to cluster with label of digit 3 and examples belonging to digit 9 have been mis-classified and are majorly predicted to belong to cluster belonging to digit 4 and some are predicted to belong to cluster labeled with digit 7.
These observations in the errors being made can be explained the explanation we gave above for k=10 above and also the explanation for k = 15 in Experiment 1, that two letters are being confused because of structural similarity, and error

being made because of the distribution of some specific class data.
Also, we observe that the accuracy obtained has increased when we increase the cluster centers from 10 to 15 even in the lower dimension space, this is because of similar reasoning as in Experiment 1 that the radius of the clusters in the 10 cluster scenario were greater than the radius off the clusters in the 15 cluster scenario. There might be cases that instances in 10 cluster scenario were close to the boundary of a cluster, but still being classified as a part of the cluster because of the large radius, this error will be reduced as a result of increasing the number of cluster centers. Also, since increasing the number of cluster centers, decrease the cluster radius, this imposes more stronger similarity to be clustered together, hence reducing the error.

#### 4.1.3 k = 5

As a result of performing k-means clustering with a value of k = 5, as in the previous case different outputs are obtained on each run. Similar to the case of performing clustering with k = 5 in Experiment 1, since the number of clusters are less than the actual number of types of digits in the dataset (which is 10, 0-9), it is obvious that only 5 classes would be predicted now, corresponding to the labels assigned to the cluster centers.

This can be seen from the following observation where on a sample run the following cluster center labels are obtained when each cluster center is assigned the label of the most frequently occurring digit in it -

Cluster Center Labels - 0,1,3,6,7

Similar to the result obtained in Experiment 1 with k = 5, the accuracy obtained varies roughly around 43% in each run. This shows that even in the lowere dimensional space when the number of cluster centers decrease the accuracy decreases.

Average Accuracy obtained after 100 runs = 43.20%

Confusion Matrix obtained is as follows -

| | | PREDICTED LABEL | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| ACTUAL LABEL | 0 | 423 | 3 | 0 | 40 | 0 | 0 | 27 | 7 | 0 | 0 |
| | 1 | 0 | 495 | 0 | 3 | 0 | 0 | 0 | 2 | 0 | 0 |
| | 2 | 3 | 92 | 0 | 56 | 0 | 0 | 338 | 11 | 0 | 0 |
| | 3 | 3 | 59 | 0 | 408 | 0 | 0 | 7 | 23 | 0 | 0 |
| | 4 | 0 | 41 | 0 | 0 | 0 | 0 | 40 | 419 | 0 | 0 |
| | 5 | 9 | 163 | 0 | 247 | 0 | 0 | 14 | 67 | 0 | 0 |
| | 6 | 7 | 68 | 0 | 10 | 0 | 0 | 409 | 6 | 0 | 0 |
| | 7 | 2 | 62 | 0 | 0 | 0 | 0 | 3 | 433 | 0 | 0 |
| | 8 | 2 | 166 | 0 | 272 | 0 | 0 | 23 | 37 | 0 | 0 |
| | 9 | 2 | 55 | 0 | 11 | 0 | 0 | 7 | 425 | 0 | 0 |

Figure 8: Confusion Matrix for k = 5

The accuracy obtained is lower than the previous runs of the algorithm with the data projected onto lower dimensional space. On similar lines to reasoning as in the case of Experiment 1 with k = 5, this is obvious as only 5 classes will be predicted now, and the maximum accuracy can be 50%.

As number of clusters have reduced, some of the clusters have combined.
Do the new clusters make any sense? Yes the new clusters make sense because similar to the explantion in Experiment 1 with k = 5, in the case of 10 cluster centers, in the lower dimensional space, we had observed that 0, 1,6 and 7 had the best classification accuracy as mentioned in the observations above. So, these clusters labels being obtained in the case of reduced clusters make sense, as the distribution of the data of these classes have less variance giving rise to the majority vote which leads to the label being selected. Also, the classes 3, 5 and 8 were being confused even in the lower dimensional space, so these classes merged into one. Also 9 and 4were being confused, so these classes merged into one. This can also be seen from the confusion matrix that the label predicted for actual class 4,7 and 9 are majorly 7. And same is the case with 3, 5 and 8.
Do you observe that clusters with digits 7 and 1 get combined? No, the digits 7 and 1 have different cluster centers. Similar to the explanation with k = 5 in Experiment 1.This is expected because the confusion between 7 and 1 was minimal, and is still minimal as seen from the confusion matrices.
Thus it is observed that a lot of similarity exists between the results that were obtained on performing k-means on the original dimensions and in the results obtained on performing k-means on the data projected onto 191 dimensions. This can be explained by the face that the principal components that PCA chooses are uncorrelated, and are ordered by the fraction of the total information each retains. Upon projecting the data to 191 dimensions it is observed that the reconstruction error is also less than 0.1 and thus very little information is lost and thus results obtained are almost similar.