**Report –Lab5**

Submitted By- **Himanshu Tolani (2014CSb1015)**

**Jatin Garg (2014CSB1017)**

Part (1)

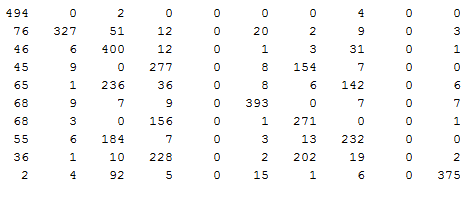
Source for K-means clustering is the matlab inbuilt function “kmeans(X,c)” where X is the dataset and c is the number of clusters. Since there are 10 classes and if we take 10 clusters it is not necessary that there will be one-one map of the classes observed from inbuilt function and the actual classes of digits from 0-9.

\*\*\*\*\*\*\*\*\*\*\* Note- Class label 10 means digit 0.

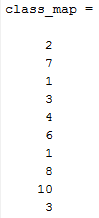
**For number of cluster 10**

Accuracy observed is around 53%. Confusion Matrix is given below

Predicted Label



Class Map that shows to which digit each class given by kmeans function maps to:



**Observations:**  The entries that are shown in the table above is for digits 1-9 the last row represents digit 0. The observation are as follows:

(a) digits ‘1’ in the first row is correctly classified with 494 instances being correctly predicted.

(b) digit ‘2’ in the second row has 327 instances as correctly classified and 76 instances classified as 1 and 51 instances classified as 3.

(c) digit ‘3’ in the next row has 400 instances as correctly classified and 46 classified as 1and 31 instances classified as 8.

(d) digit ‘4’ has 277 instances classified correctly and  **large number 154 instances classifying 4 as 7 .**

(e) digit ‘5’ has **236 instances classified as 3** and  **large number 142 instances classifying 5 as 8 . Hence digit 5 is highly misclassified.**

(f) digit ‘6’ has 393 instances being correctly classified.

(g) digit ‘7’ has 271 instances correctly classified and large number **156 instances classified as 4.**

(h) digit ‘8’ has 232 instances correctly classified while large number **184 instances classified as 3.**

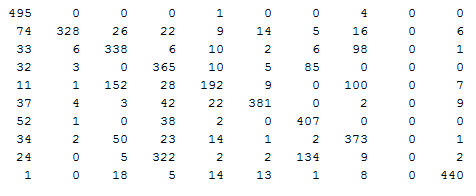
(i) digit ‘9’ **is highly misclassified as 228 instances are classified as 4 and 202 as 7.**

(j) digit ‘0’ has 375 instances classified correctly and 92 instances classifying it has 3.

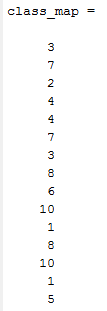
**For number of clusters 15**

Accuracy observed is around 65% and various a little bit between 63%-66% . Confusion Matrix is given below:

Predicted Label



Class Map that shows to which digit each class given by kmeans function maps to:



**Observations:**

From table for K=10 and K=15 we can observe that :

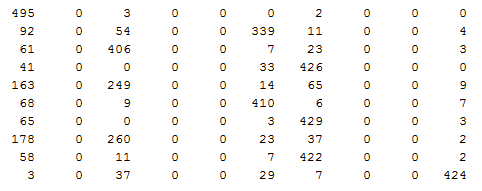
(a) The digit 8 which classified 156 times as 4 and 271 times as 7 for K=10 has now splitted and has 407 instances being classified as 7.

(b) The digit 7 which classified 184 times as 3 and 232 times as 8 for K=10 has now splitted and has 373 instances being classified as 8.

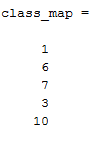
(c) The digit 5 which was highly misclassified as 3 and 8 for K=10 has now 192 instances being correctly labelled.

**For number of clusters 5**

Accuracy observed is around 43% .Confusion matrix is as follows



Class Map that shows to which digit each class given by kmeans function maps to:



**Observations:**

**(a) digit 2 merges with digit 6.**

**(b) digit 4 merges with 7.**

**(c) digit 5 merges into digit 1 and digit 3.**

**(d) digit 8 merges with digit 1 and digit 3.**

**(e) digit 9 merges with digit 7.**

**Part(ii) On applying PCA and using least squared error per instance as the reconstruction error by using L2 we were able to reduce the dimensions to 191 by taking 191 vectors corresponding to largest 191 eigen values .**

**Total Squared error=**

**Where X(i,j) is jth feature value for ith instance of the original data**

**X’(i,j) is the jth feature value of the ith instance for the reconstructed data.**

**Mean squared Error=**

**New Dimensions received for the data set use just 191 features instead of all 400 features.**

***Number of principal Component – 191.***

three_K=98_const.png three_K=120_const.png three_K=191_const.png three_K=217_const.png three_K=305_const.png three.png

K=98 K=120 K=191 K=217 K=305 K=400

seven_K=98_const.png seven_K=120_const.png seven_K=191_const.png seven_K=217_const.png seven_K=305_const.png seven.png

K=98 K=120 K=191 K=217 K=305 K=400

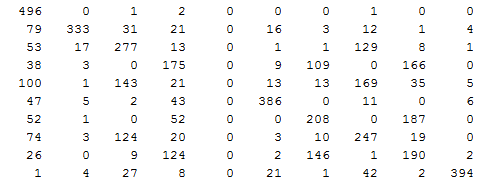
In the above images we visualise the digits 3 and 7 with the variation in the number of principal components we observe that as the number of principal components increase the images become more and more similar to the original images. The reason for the background highlighted images is that in the reconstructed image we have added the mean value which is responsible for the background highlighted .

mean.png Image observed using mean value of features.

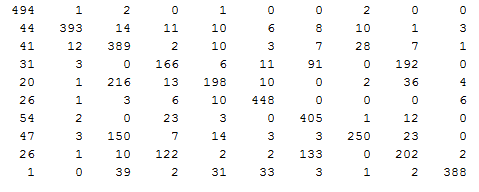
As the number of Principal component increases the sharp and fine features of the images can be observed more clearly.

Part (iii)

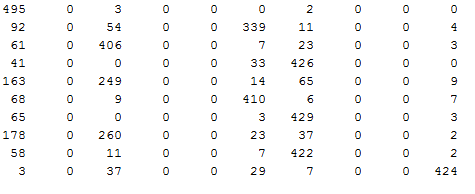
Confusion matrix for K=10



Confusion matrix for K=15



Confusion Matrix for K=5



|  |  |
| --- | --- |
| Value of number of clusters | Avg accuracy observed |
| 10 | 54.1 |
| 15 | 66.6 |
| 5 | 43.3 |

From the Confusion matrix observed above and results of accuracy on performing K means clustering with lower dimensional data the accuracy almost remains unchanged . But computationally more efficient than using all the 400 features. As PCA helps us in removing all the redundant features and includes feature that provide the maximum information.