MACHINE LEARNING LAB 4

24th November, 2017
Prepared by
Sujit Rai,2017CSM1006
Manish Singh,2017CSM1003

ABSTRACT

Clustering is one of the important streams in data mining useful for discovering groups and identifying interesting distributions in the underlying data.This assignment aims in analyzing results of k-means algorithms on the MNIST dataset. The analysis has been done by changing the number of clusters and observing the kinds of misclassifications. Further we have applied the Principal Component Analysis on MNIST dataset to reduce the dimensions of the dataset to have a reconstruction error of 0.1 and further applied k-means on the transformed MNIST dataset.Results have been analysed as a confusion matrix which tells us the number of digits belonging to the correct clusters as well as the wrong clusters.

Question 1 (a)

Using a k-means clustering implementation of your choice, perform k-means clustering on the MNIST hand written digits' dataset. Indicate in the report the source of your k-means clustering implementation. Perform clustering with k=10. Suppose we were to label each cluster with the most frequently occurring digit, what is the classification accuracy? What are the kinds of misclassifications (confusion matrix)? Suppose we were to increase k to 15, some of the digits which were previously represented as a single cluster will split into 2 multiple clusters. Which are the digits that get split further? Now if we were to reduce k to 5, some of the clusters will be combined. Do your new clusters make any sense? For example, do you observe that clusters with digits 7 and 1 get combined? Discuss your observations

We have used the matlab inbuilt function for k-means implementation. We have taken the observation for 10 cluster and then we increased the number of clusters to 15. It has been observed that as we increase the number of clusters than some of the digits which were previously represented as a single cluster will split into 2 multiple clusters. We also

reduced the number of clusters to 5 and observed that some of the clusters which earlier were separate got combined. We have discussed our observations below:-

ACTUAL LABEL	PREDICTED LABEL									
	1	**6**	**4**	**5**	**0**	**3**	**6**	**2**	**9**	**7**
0	0	4	1	0	491	2	0	1	0	1
1	324	3	11	14	85	17	3	24	4	15
2	10	20	34	2	31	7	2	298	1	95
3	3	112	6	9	17	234	119	0	0	0
4	1	34	211	10	21	22	4	156	5	36
5	3	0	27	385	33	44	0	2	6	0
6	1	215	3	0	34	35	209	2	1	0
7	3	22	8	3	54	19	9	107	0	275
8	0	186	1	2	10	126	163	9	2	1
9	0	2	70	20	1	5	0	25	377	0

- The above results were obtained with k-means for k=10(clusters) and k-means was run for 500 iterations.
- The accuracy obtained is 60.3800% .
- The error was 39.62% .
- Total correct classifications are 3019 and misclassification were 1981.
- 491 digits for cluster 0 were identified correctly
- For cluster 0 there were 4 misclassification in cluster 6, 1 in cluster 4, 2 in cluster 3, 2 in cluster 1 and 1 in cluster 7.
- For cluster 9 there were 377 digits were correctly classified.
- For cluster 9 there were 2 misclassification in cluster 2, 70 in cluster 4, 20 in cluster 3, 1 in cluster 4, 5 in cluster 5 and 25 in cluster 7 .

ACTUAL LABEL	PREDICTED LABEL														
	2	2	3	5	4	6	0	3	9	1	9	7	8	0	6
0	2	0	0	0	1	0	232	0	0	0	0	0	2	261	2
1	18	8	0	14	13	4	47	18	0	329	7	9	3	28	2
2	164	231	0	3	11	3	28	2	1	8	0	24	22	1	2
3	0	0	202	3	15	0	7	151	0	2	0	0	99	9	12
4	120	90	4	10	195	0	5	9	3	1	5	3	54	0	1
5	3	1	0	362	45	0	29	43	4	4	4	0	0	5	0
6	0	0	7	0	2	187	17	19	0	0	0	1	44	18	205
7	111	24	6	2	15	0	34	13	1	3	1	243	38	5	4
8	7	2	67	1	3	4	6	110	0	0	2	1	168	3	126
9	12	9	0	11	16	1	1	4	199	0	244	1	2	0	0

- The results in table above were obtained with k means with k=15 for 500 iterations .
- The accuracy obtained was 67.46%.
- The error was 32.54%.
- Clusters for digit 0, digit 2, digit 3, digit 6 and digit 9 got split into two clusters.

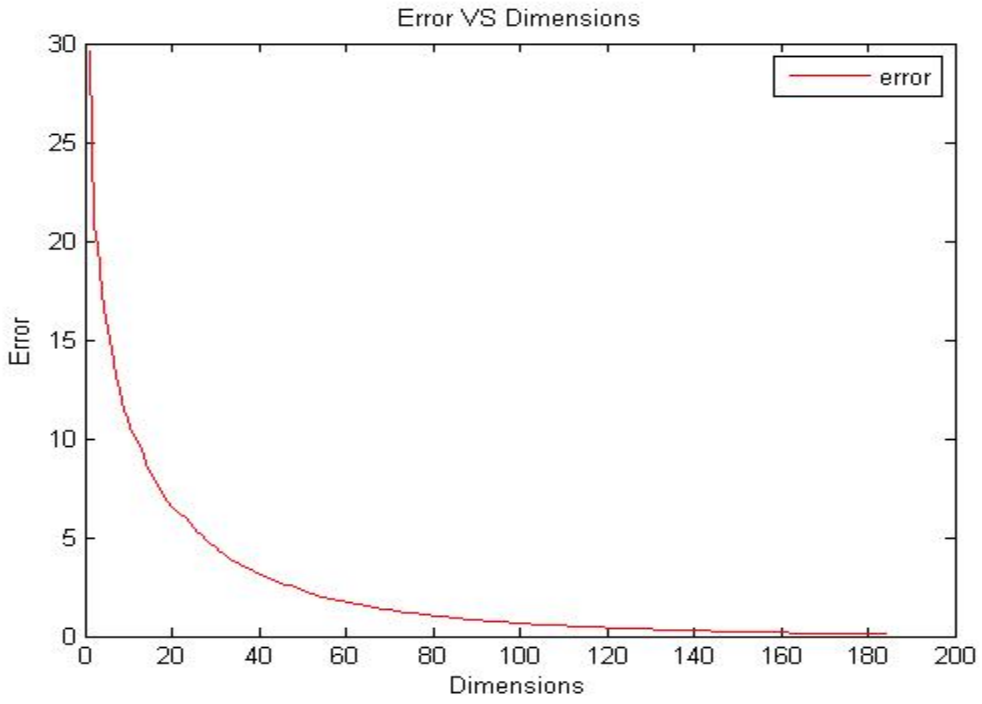
ACTUAL LABEL	PREDICTED LABEL				
	0	**6**	**5**	**2**	**9**
0	495	2	0	3	0
1	92	11	338	56	3
2	59	23	7	408	3
3	41	419	40	0	0
4	163	67	14	247	9
5	68	6	409	10	7
6	62	433	3	0	2
7	166	37	23	272	2
8	55	425	7	11	2
9	3	7	27	40	423

- The above table has been made by k-means algorithm for 500 iterations with 5 clusters.
- It is observed that clusters for digit 1 combined more with cluster for digit 5 followed by cluster for digit 1 and then with digit 2.
- The cluster for digit 3 combined with cluster for digit 6 followed by digit 0 and digit 5.
- The cluster for digit 7 combined with cluster for digit 2 followed by digit 0.
- The cluster for digit 8 combined with cluster for digit 6.
- The accuracy obtained by the clusters is 43.36% .
- The error for the clusters is 56.64%.
- There were 2168 correct classifications and 2832 misclassifications.

Question 1 (b)

Use PCA to reduce the dimensionality of the digit images. Select the number of components such that the residual reconstruction error is under 0.1. Visualize at least the two 2 or 3 components and try to interpret what kind of variation in the data are they capturing?

In this part we have reduced the dimensionality of the digit images by applying Principal Component Analysis such that the reconstruction error is 0.1 . We have plotted the graph of Number of Dimensions Vs the reconstruction error below:-



The graphs shows that as we are increasing the number of dimensions along which the data is projected the reconstruction error reduces rapidly. We have observed that as the number of dimensions along which we projected the MNIST data are 184 the reconstruction error obtained is 0.1 .Below we have given the table which shows us the reconstruction error for the different dimensions.

Number of Dimensions	Reconstruction Error
1	29.6301562073076
2	20.7902078721790
3	19.8008887388211
4	17.3943852040767
5	15.9696769929374
50	2.32071591558987
100	0.819376848900038
183	0.100423185712101
184	0.0982421985144112
400	Approximately zero($1.08*10^{-20}$)

Question 1 (c)

Perform k-means clustering on the data projected onto lower dimensions. Repeat the experiments that were conducted in part (a) on the low dimensional dataset. How does clustering in the low dimensional space compare to the original space?

We reduced the dimension of the MNIST dataset using PCA to have a reconstruction error of 0.1. The dimension was reduced to 184 for a reconstruction error of 0.1 . We have given the results in the table below:-

ACTUAL LABEL	PREDICTED LABEL				
	6	**9**	**2**	**5**	**0**
0	2	0	3	0	495
1	10	6	50	334	100
2	23	2	414	7	54
3	428	0	0	40	32
4	63	7	250	15	165
5	3	7	10	433	47
6	432	4	2	1	61
7	50	0	221	13	216
8	422	3	11	4	60
9	2	436	32	29	1

- The k-means was run for 500 iterations with k=10.
- Accuracy obtained is 44.2% .
- Misclassification error is 55.8%
- Clusters for digit 1 combined with cluster for digit 5 followed by cluster for digit 0 and than 2.
- Cluster for digit 3 got combined with cluster for digit 6 followed by cluster for digit 5 and than with digit 0
- Cluster for digit 4 got combined with cluster for digit 2 ,followed by cluster for digit 0 and than digit 6.
- Cluster for digit 7 got combined with cluster for digit 2 and than with digit 0 and than with digit 6
- Cluster for digit 8 got combined with cluster for digit 6 followed by cluster for digit 0.

ACTUAL LABEL	PREDICTED LABEL									
	7	**9**	**0**	**6**	**3**	**9**	**5**	**2**	**0**	**1**
0	1	0	270	0	1	0	0	1	227	0
1	14	0	38	3	14	9	19	25	57	321
2	125	1	2	4	16	1	2	294	42	13
3	0	2	12	160	284	0	7	0	28	7
4	123	5	0	16	44	12	13	145	141	1
5	7	6	2	0	11	3	404	2	58	7
6	0	1	24	271	160	1	0	0	42	1
7	271	3	13	20	27	1	3	112	47	3
8	2	2	5	221	230	2	2	9	26	1
9	12	206	0	1	1	253	11	12	2	2

The k-means algorithm was run for 500 iterations with k=10 and the results in above table was obtained.

The classification accuracy obtained is 56.02% .

The misclassification error is 45.98%.

ACTUAL LABEL	PREDICTED LABEL														
	0	9	6	3	0	3	6	7	4	8	5	1	2	5	2
0	234	0	0	1	261	0	0	1	1	0	0	0	0	0	2
1	51	3	2	12	34	0	3	10	9	4	5	323	13	15	16
2	27	1	2	3	2	0	2	33	10	12	1	8	239	2	158
3	7	0	12	160	12	209	0	0	5	85	5	2	0	2	1
4	8	6	0	12	0	6	0	7	198	33	4	0	100	5	121
5	18	5	0	4	5	0	0	0	10	0	270	0	2	183	3
6	13	0	195	18	22	3	172	0	3	72	0	2	0	0	0
7	26	0	2	12	8	7	0	253	11	35	1	3	25	1	116
8	5	2	115	120	3	60	3	0	2	176	1	1	2	1	9
9	1	381	0	2	0	0	2	0	47	2	23	0	19	2	21

The k-means algorithm was run for 500 iterations with k=15

The accuracy obtained is 68.24%.

The misclassification error is 31.76%.

OBSERVATIONS

It has been observed that as we decrease the dimension of the data using PCA the accuracy of clustering assignment decreases.But it was observed that even after decreasing the dimension the accuracy increased slightly for k=15. It may be due to noise in the data or random assignment of cluster centers.