CSL 603-Machine Learning

Lab 4 Report

K-Means Clustering and PCA

Keshav Garg 2015csb1016

Pankaj Verma 2015csb1022

Introduction

The aim of this lab is to experiment with clustering and dimensionality reduction techniques on MNIST hand written digits' dataset. We implemented the k means clustering using the inbuilt kmeans Matlab function. First we did pre processing on the input data by subtracting the mean values of corresponding column vectors. For task b, we used PCA to reduce the dimensionality of the digit images. Number of components required is 191 so that construction error is below 0.1. We created a new data set of these low dimensions and then again performed the k-means clustering on this new dataset. Observations are noted down and is as shown below.

Observations

Task A

1. K=10

Confusion Matrix-

Predicted Labels (Digits)	Actual Label (Digits)									
		0	1	2	3	4	5	6	7	8	9
	0	494	84	28	30	111	41	46	60	20	2
	1	0	332	10	14	1	31	2	6	4	1
	2	2	33	404	0	285	9	0	141	10	24
	3	2	11	19	289	40	10	182	14	242	2
	4	0	0	0	0	0	0	0	0	0	0
	5	0	13	3	9	13	400	0	3	2	10
	6	0	3	5	157	38	0	267	14	219	1
	7	2	17	30	1	6	0	1	261	1	1
	8	0	0	0	0	0	0	0	0	0	0
	9	0	7	1	0	6	9	2	1	2	459

Classification accuracy =(C/N)*100= 58.12%

C=No. of instances correctly labelled

N=Total number of instances

The reason for this low accuracy is because some digits are not even predicted. Some digits have majority count in more than one clusters. Thus classifying labels wrongly.

2. K=15

Confusion Matrix-

Predicted Labels (Digits)	Actual Label (Digits)									
	0	1	2	3	4	5	6	7	8	9
0	489	77	22	10	6	28	23	32	6	1
1	0	317	14	3	1	4	1	2	0	0
2	0	22	259	0	82	0	1	32	3	1
3	1	20	3	362	16	43	23	18	181	3
4	1	9	9	15	199	23	2	14	3	15
5	0	16	1	3	10	390	0	2	2	7
6	0	6	3	17	0	0	360	2	130	1
7	3	22	158	0	113	2	0	378	6	5
8	6	3	25	90	29	0	90	16	166	2
9	0	8	6	0	44	10	0	4	3	465

Classification accuracy =(C/N)*100= 67.7%

The no of misclassified instances decreases on increasing the clusters. As the number of cluster increases, old cluster gets separated in multiple clusters, thus all the digits have majority in at least one of the clusters.

3. K=5

Confusion Matrix-

Predicted Labels (Digits)	Actual Label (Digits)									
		0	1	2	3	4	5	6	7	8	9
	0	495	92	61	41	163	68	65	178	58	3
	1	0	0	0	0	0	0	0	0	0	0
	2	3	54	406	0	249	9	0	260	11	37
	3	0	0	0	0	0	0	0	0	0	0
	4	0	0	0	0	0	0	0	0	0	0
	5	0	339	7	33	14	410	3	23	7	29
	6	2	11	23	426	65	6	429	37	422	7
	7	0	0	0	0	0	0	0	0	0	0
	8	0	0	0	0	0	0	0	0	0	0
	9	0	4	3	0	9	7	3	2	2	424

Classification accuracy =(C/N)*100= 43.28%

As the number of clusters decreases, many of the digits are not predicted. Many clusters get combined which resulted in very less classification accuracy.

Task B

No. of principal components required = 191

No. of Principal Components	Reconstruction error
20	6.804467
40	3.375556
60	1.910678
80	1.168239
100	0.744528
120	0.484770
140	0.317008
160	0.204503
180	0.129305
191	0.098684

Variations of the data with 2 or 3 components i.e in 2D and 3D has been captured and the images are included in the folder task b.

For 2D, following images are recorded:



Please refer to the folder as image quality is not good here.

Task C

1. K=10

Confusion Matrix-

Predicted Labels (Digits)	Actual Label (Digits)									
	0	1	2	3	4	5	6	7	8	9
0	496	80	37	27	102	41	38	63	16	2
1	0	325	15	3	1	4	1	3	0	4
2	1	31	294	0	146	2	1	127	10	27
3	2	22	9	171	23	46	25	21	123	6
4	0	0	0	0	0	0	0	0	0	0
5	0	17	2	9	13	393	0	2	2	21
6	0	7	5	95	12	0	419	6	137	2
7	1	13	130	0	152	8	0	241	1	42
8	0	1	7	195	46	0	16	37	209	2
9	0	4	1	0	5	6	0	0	2	394

Classification accuracy =(C/N)*100= 58.84%

2. K=15

Confusion Matrix-

Predicted Labels (Digits)	Actual Label (Digits)									
	0	1	2	3	4	5	6	7	8	9
0	493	45	51	31	20	32	55	53	27	1
1	1	388	15	4	2	2	2	3	1	1
2	1	9	281	0	152	3	0	97	8	13
3	1	10	13	165	23	5	22	27	126	2
4	1	9	9	7	201	10	2	15	2	13
5	0	6	2	11	13	442	0	3	2	16
6	0	9	4	91	0	0	405	4	133	2
7	3	16	114	0	31	0	1	274	1	3
8	0	1	8	191	38	0	12	21	198	2
9	0	7	3	0	20	6	1	3	2	447

Classification accuracy =(C/N)*100= 65.88%

3. K=5

Confusion Matrix-

Predicted Labels (Digits)	Actual Label (Digits)									
	0	1	2	3	4	5	6	7	8	9
0	495	92	61	41	163	68	65	178	58	3
1	0	0	0	0	0	0	0	0	0	0
2	3	54	406	0	249	9	0	260	11	37
3	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0
5	0	339	7	33	14	410	3	23	7	29
6	2	11	23	426	65	6	429	37	422	7
7	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0
9	0	4	3	0	9	7	3	2	2	424

Classification accuracy =(C/N)*100= 43.28%

Accuracy decreases as we took data in lower dimension.