Machine Learning Lab4 Report.

24.11.2017

Alok Kiran (2015CSB1006)
Kartik Vishwakarma(2017CSM1001)

Goals

- Measuring prediction accuracy of K-means clustering algorithm for varying number of clusters.
- Dimensionality reduction using principle component analysis.

K-Means clustering.

K-means clustering algorithm is used to label the data based on the Euclidian distance of data from the center of its cluster.

We performed K-means clustering for various number of clusters whose observations are given below.

I. Observations: here K = number of clusters we are considering.

For K = 10

Prediction accuracy is in between 50 % to 60 % for various run of the code for K-means.

And the average accuracy is about 53.14 %.

As We can see that the accuracy is not a constant value it is because every time K-means is picking the cluster centers randomly.

Here is the confusion matrix for K = 10

Actual Label

	0	1	2	3	4	5	6	7	8	9
0	493	0	2	2	1	0	0	2	0	0
1	83	332	29	12	13	9	4	15	0	3
2	28	7	398	14	13	1	4	34	0	1

3	23	13	1	305	13	7	137	1	0	0
4	10	0	224	37	199	11	5	8	0	6
5	33	31	6	11	25	388	0	0	0	6
6	45	2	0	174	5	0	272	1	0	1
7	43	6	156	10	13	3	14	255	0	0
8	17	4	11	245	6	2	212	1	0	2
9	1	0	51	7	42	18	0	1	0	380

Note : Leftmost column in every confusion matrix shows Predicted label.

From above table most of diagonal entries have greater value than other entry, which

Represents most of data goes to corresponding label cluster as in data.txt file given.

I.e. A true cluster is formed, But also Accuracy is fall between 50-60 with tell us forming cluster based on distance is not great idea.

For K = 15

In this case prediction accuracy is in between 61 % to 68 %.

And average accuracy is 64.86 %

We can see that the prediction accuracy is much better than the previous case because here we have more number of clusters in which data points can be classified accurately.

Confusion matrix for K = 15

Actual Label

	0	1	2	3	4	5	6	7	8	9
0	493	1	1	2	1	0	0	1	1	0
1	44	389	11	12	8	6	6	14	3	7
2	43	15	263	8	9	2	2	140	16	2
3	18	3	0	166	7	9	191	0	106	0
4	11	2	146	19	196	9	8	58	31	20
5	23	1	3	6	3	457	0	0	0	7
6	36	2	1	36	2	0	294	1	127	1
7	42	4	102	19	17	4	19	262	28	3
8	8	1	9	115	2	2	214	1	146	2
9	0	1	14	2	12	15	0	3	3	450

From above table most of diagonal entries have greater value than other entry, which

Represents most of data goes to corresponding label cluster

For K = 5

Accuracy is in between 39 % to 44 %. And the average accuracy is 42.16 %

Reasoning is same as above.

Confusion matrix for K = 5.

Actual Label

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

8	55	0	11	0	0	7	425	0	0	2
9	3	0	40	0	0	27	7	0	0	423

From above table most of diagonal entries have greater value than other entry, which

Represents most of data goes to corresponding label cluster

Question 2. PCA (Principle Component analysis)

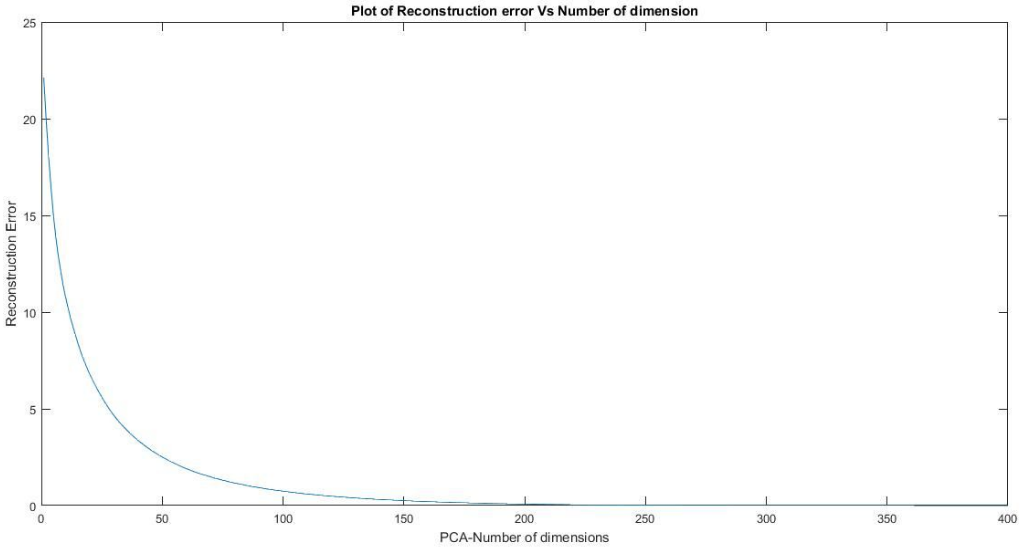
In the PCA we reduced the dimension of original dataset by projecting it onto lower dimension.Since the original data may contains so many dummy points which are not important.So using PCA we can get almost same accuracy on the dataset with information loss penalty.

Here are the observations for the Reconstruction errors obtained on some of the Principal components.

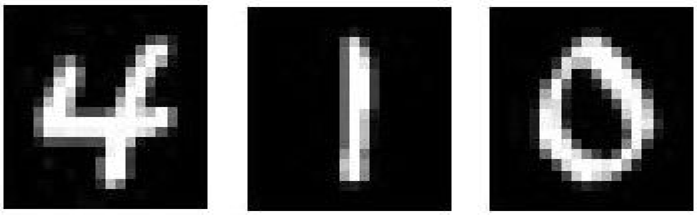
Serial Number	Reconstruction Error	PCA
1	0.0984	191
2	14.9936	5
3	3.37	40
4	0.2490	150
5	1.07762 e -28	400

We can observe that the error is continuously decreasing as the PCA increases because information loss is decreasing.

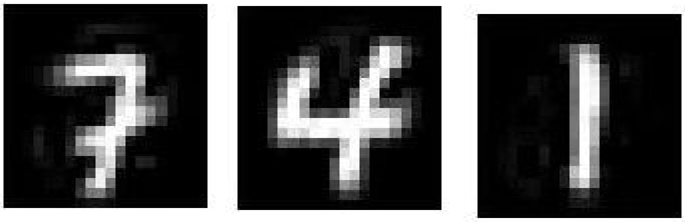
Here is the plot.



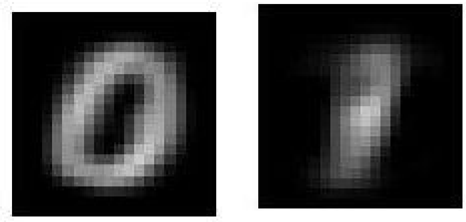
For 2 or 3 k(PCA) values and we are showing 2 or 3 images as follows;



For PCA=191



For PCA=50



For PCA=2

Question3.

Here we are repeating the same process done in question 1. Here we are taking the data projected on lower dimension.And then applying K-means on those data sets.

From Observation We can see with reducing Dimension we achieved less error(greater accuracy) than higher dimension.

I.e. Curse of Dimensionality is minimized .

With reducing dimension we able to get more data per dimension and greater accuracy.

K =10

For 5 runs Average accuracy is 55.912 %

Actual Label

	0	1	2	3	4	5	6	7	8	9
0	496	0	1	2	0	0	0	1	0	0
1	91	317	26	17	0	24	6	16	0	3
2	41	12	274	11	0	3	14	144	0	1
3	24	2	0	190	0	10	274	0	0	0
4	17	1	138	25	0	12	162	138	0	7
5	39	4	2	44	0	394	0	10	0	7
6	45	0	0	54	0	0	401	0	0	0
7	63	3	113	22	0	4	53	242	0	0
8	14	1	10	128	0	20	342	1	0	0
9	1	2	17	6	0	19	7	38	0	410

From above table most of diagonal entries have greater value than other entry, which Represents most of data goes to corresponding label cluster.

K = 15

For 5 runs of K-means Average accuracy is 67.216 %

Actual Label

	0	1	2	3	4	5	6	7	8	9
0	495	0	3	0	1	0	0	1	0	0
1	82	307	30	6	9	46	4	9	4	3
2	29	7	393	5	11	5	3	34	12	1

3	20	2	1	371	5	13	12	0	76	0
4	7	0	217	18	202	12	0	11	27	6
5	24	0	5	0	11	454	0	0	0	6
6	34	1	0	21	3	0	344	0	97	0
7	34	3	143	13	11	4	2	258	32	0
8	8	1	10	184	2	4	120	1	168	2
9	1	0	32	0	47	35	1	1	0	383

From above table most of diagonal entries have greater value than other entry, which

Represents most of data goes to corresponding label cluster

K = 5

For 5 runs of K-means average accuracy is 43.104 %

Corresponding Confusion matrix is given below.

Actual Label

	0	1	2	3	4	5	6	7	8	9
0	494	0	4	0	0	0	2	0	0	0
1	99	0	48	0	0	337	10	0	0	6
2	55	0	412	0	0	8	23	0	0	2
3	32	0	0	0	0	49	419	0	0	0
4	168	0	248	0	0	18	59	0	0	7

5	45	0	10	0	0	437	1	0	0	7
6	61	0	1	0	0	2	432	0	0	4
7	218	0	219	0	0	15	48	0	0	0
8	60	0	11	0	0	5	422	0	0	2
9	1	0	32	0	0	29	2	0	0	436

From above table most of diagonal entries have greater value than other entry, which

Represents most of data goes to corresponding label cluster

Reference for question2. PCA

http://www.holehouse.org/mlclass/14_Dimensionality_Reduction.html