UNIVERSITY OF SOUTH FLORIDA

Project 4

Pattern Recognition

By:

Mohammed Alghamdi

Saurabh Hinduja

In this project, we were given 4 datasets with ten classes each, the classes are the letters a,c,e,m,n,o,r,s,x,z and also with ten samples for each class. We were expected use clustering to classify the samples. We classified the samples using MATLAB function “kmeans”. The syntax for kmeans is

IDA = kmeans(X,K)

Where

IDA – Gives us the a *n x 1* matrix, where n is the number of samples, with the cluster number of each sample

X – Matrix of the samples

K – Number of clusters in which the samples are to be clustered

For this project we took the 8 moments, instead of the bitmap, for clustering. X was a 100 x 8 matrix for all the four datasets.

Result:

For dataset A we conducted the experiment 3 times. Clustering the dataset into 9 , 10 and 11 clusters .

classified	1	2	3	4	5	6	7	8	9	error
a	0	10	0	0	0	0	0	0	0	0
c	0	0	0	0	0	10	0	0	0	0
e	0	0	0	0	0	10	0	0	0	10
m	9	0	0	0	0	0	0	1	0	1
n	0	0	10	0	0	0	0	0	0	0
o	0	0	0	0	0	0	9	0	1	1
r	0	0	0	0	10	0	0	0	0	0
s	0	0	0	0	0	0	0	0	10	0
x	0	0	10	0	0	0	0	0	0	0
z	0	0	0	10	0	0	0	0	0	0

Table 1: Dataset A clustered in 9 clusters

Table 1, shows the table for dataset A clustered into 9 clusters. In this we notice that the samples of letter *c* and *e* are clustered into one clusters. This is due to the fact letters *c* and *e* have similar moments and their clusters may be close to each other, giving an impression of one cluster when the dataset is divided into 9 clusters.

classified	1	2	3	4	5	6	7	8	9	10	error
a	0	0	10	0	0	0	0	0	0	0	0
c	0	10	0	0	0	0	0	0	0	0	0
e	0	0	0	0	10	0	0	0	0	0	0
m	0	0	0	0	0	0	0	0	9	1	1
n	0	0	0	10	0	0	0	0	0	0	0
o	0	0	0	0	0	3	0	7	0	0	3
r	6	0	0	0	0	0	4	0	0	0	4
s	0	0	0	0	0	10	0	0	0	0	0
x	0	0	0	10	0	0	0	0	0	0	0
z	0	0	8	0	0	2	0	0	0	0	2

Table 2: Dataset A clustered in 10 clusters

Table 2, is the table when dataset A is divided into 10 clusters, we notice that the samples are mostly clustered correctly. But there is a tradeoff, when we increased the number of clusters the errors in the other classes increased.

Classified	1	2	3	4	5	6	7	8	9	10	11	error
a	0	0	0	7	3	0	0	0	0	0	0	3
c	0	0	10	0	0	0	0	0	0	0	0	0
e	0	0	10	0	0	0	0	0	0	0	0	10
m	9	0	0	0	0	0	0	0	0	0	1	1
n	0	2	0	0	0	0	7	0	0	1	0	3
o	0	0	0	0	0	1	0	9	0	0	0	9
r	0	0	10	0	0	0	0	0	0	0	0	0
s	0	0	0	0	0	0	0	10	0	0	0	0
x	0	1	0	0	0	0	0	0	0	9	0	1
z	0	0	0	0	0	5	0	0	5	0	0	5

Table 3: Dataset A clustered in 11 clusters

Table 3, is the table for dataset A when it is clustered into 11 clusters. We notice that as the number of clusters increase the errors also increase. More number of samples are misclassified.

Table 4, 5 and 6 are the table for dataset B, C and D, respectively, clustered into 10 clusters.

classified	1	2	3	4	5	6	7	8	9	10	error
a	0	0	0	0	0	0	0	10	0	0	0
c	0	0	0	0	0	0	0	0	0	10	0
e	0	0	10	0	0	0	0	0	0	0	0
m	10	0	0	0	0	0	0	0	0	0	0
n	0	0	0	0	0	0	4	0	6	0	4
o	0	6	0	0	4	0	0	0	0	0	4
r	0	0	0	10	0	0	0	0	0	0	0
s	0	0	0	0	10	0	0	0	0	0	0
x	0	2	0	0	0	7	0	1	0	0	3
z	0	10	0	0	0	0	0	0	0	0	0

Table 4: Dataset B clustered in 10 clusters

classified	1	2	3	4	5	6	7	8	9	10	error
a	5	5	0	0	0	0	0	0	0	0	5
c	0	0	0	0	0	0	0	0	10	0	0
e	0	0	0	0	0	0	0	0	10	0	10
m	0	0	0	0	10	0	0	0	0	0	0
n	0	0	0	10	0	0	0	0	0	0	0
o	0	0	0	0	0	10	0	0	0	0	0
r	0	0	0	0	0	0	10	0	0	0	0
s	0	0	0	0	0	0	0	10	0	0	0
x	0	0	0	10	0	0	0	0	0	0	0
z	0	0	3	0	0	0	0	0	0	7	3

Table 5: Dataset C clustered in 10 clusters

classified	1	2	3	4	5	6	7	8	9	10	error
a	0	0	10	0	0	0	0	0	0	0	0
c	10	0	0	0	0	0	0	0	0	0	0
e	0	0	0	0	0	0	0	10	0	0	0
m	0	0	0	7	3	0	0	0	0	0	3
n	0	0	0	0	0	0	0	0	0	10	0
o	0	3	0	0	0	0	0	0	7	0	3
r	0	0	0	0	0	0	10	0	0	0	0
s	0	0	0	0	0	0	0	0	10	0	0
x	0	10	0	0	0	0	0	0	0	0	0
z	0	0	0	0	0	8	0	0	2	0	2

Table 6: Dataset D clustered in 10 clusters

Conclusion:

By this project we can conclude that when we cluster the dataset into the correct number of clusters we get the least errors. As we increase or decrease the number of clusters, the error increases. Therefore, when we do not know the number of clusters we can do a number of iterations with changing the number of clusters with each iteration until we get the least error.

Appendix:

We used a website to know how to use the Matlab which is http://www.mathworks.com/