Clustering and Dimensionality Reduction

Anirudh Sharma
ID : 2015CSB1007

Aditya Ranjan
ID : 2015CSB1004

***Abstract*—We experimented with K-means clustering on MNIST dataset. We first used the inbulilt K-means function in matlab and calculated accuracy for our dataset after making clusters. Then we changed the number of clusters and noted our observations. Then we used PCA to reduce the dimensionality of the dataset. Then we again implemented K-means clustering on the dataset with reduced dimensionality and noted the results.**

Keywords*—*K-means, clusters,PCS,dimensionality reduction.

I. Question1 : K-means clustering

We used the inbuilt matlab K-means clustering function to implement K-means clustering. We passed the parameters dataset input and number of clusters and obtained the vector containing the cluster number of all the points. Then we iterated through each cluster and found the label which was most frequent. We called that label the label of that cluster. We converted the original output into the required decimal output and stored it in a vector. We made the confusion matrix and calculated the accuracy. below is the table of accuracy for different cluster numbers we obtained.

Number of clusters	accuracy
10	57
15	65
5	43

TABLE I: accuracy with number of clusters

The reason for the observation is following.

A. Increasing accuracy with increasing number of clusters

As number of cluster is increased, some clusters which were merged before mainly that for digit 1 and 7 or 6 and 9 were further divided into independent clusters which increased the accuracy.

B. decreasing accuracy with decreasing number of clusters

The reason being the as above. As the number of clusters decrease, the number 1 and 7 are classified in same clusters and hence accuracy decreases. for some iterations, clusters with 6 and 9 also got combined into a single cluster.

II. PCA

First the input was taken from the given dataset. The we created the covariance matrix. The we obtained the corrosponding eigenvalues and eigenvectors of the covariance matrix. Then the largest K eigenvalues (k is the dimension we want after reduction) were noted and eigenvectors corrosponding to to it was taken. This was transformation matrix was obtained. Using the transformation matrix, the new reduced dataset was made. The graph we obtained between reconstruction error andf number of principal components (projected dimension) is below,

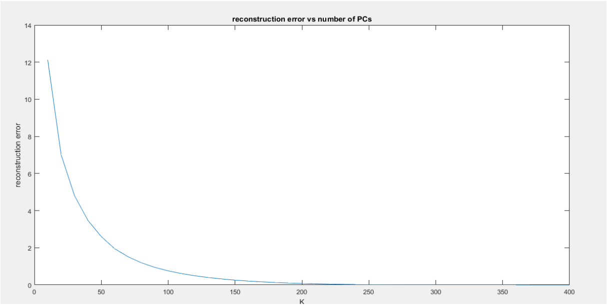

Fig. 1: reconstruction error VS principal components

III. K-means on reduced dimension

We applied K-means to the reduced matrix with 191 features and found accuracy with different cluster numbers. The table below shows the same

Number of clusters	accuracy
10	55
15	66
5	43

TABLE II: accuracy with number of clusters for reduced dataset

We can see that accuracy before and after applying PCA is almost the same. This is a good thing because we have successfully reduced the dimension of our dataset without compromising with the accuracy, and that's what is PCA for. The reasons for accuracy reduction and increase with different cluster numbers is the same as described above.

References

[1] `http://cse.iitrpr.ac.in/ckn/courses/f2017/csl603/csl603.ht`