Data Mining Project 2

Darshan Lal : 1001667684
Harsha Boppana: 1001675386

Task A

For Task A to carryout classification on the handwritten data, we used various classification methods like

Linear Regression:

In this method with the help of data handler the classes A, B, C, D, E have been selected. To classify these classes the set of first 30 images are used as training images and the remaining 9 images are used as testing images. In this way the classifier has been trained in 30 X 5 = 150 images. For the training part the classifier is tested by a different training data of 9 X 5 = 45 images. Utilizing the code which has been provided for linear regression the achieved accuracy is 88.23%

K Nearest Neighbors:

The classes A, B, C, D, E have been selected with the help of data handler. To classify these classes the set of first 30 images are used as training images and the remaining 9 images are used as testing images. In this way the classifier has been trained in 30 X 5 = 150 images. For the training part the classifier is tested by a different training data of 9 X 5 = 45 images. Euclidean distance has been used for KNN to calculate the distance between the training data points and testing data points. The calculated distances are then stored in a data structure dictionary in python which has key which is equal to the distance between each test and train data point, and the value was the label or the class which it belongs to.
To achieve a higher accuracy the K is taken as 5. Nearest neighbors are selected using the election function. The label with the maximum occurrence in K is noted and the test data point is classified as that label. The accuracy in the handwritten data was 93.75%

Nearest Centroid:

In this method with the help of data handler the classes A, B, C, D, E have been selected. To classify these classes the set of first 30 images are used as training images and the remaining 9 images are used as testing images. In this way the classifier has been trained in 30 X 5 = 150 images. For the training part the classifier is tested by a different training data of 9 X 5 = 45 images. The function centroid is used to find the centroid for each of the training data point which belongs to the same class. We have centroids for

each of the classes A, B, C, D, E. For the nearest centroid the distance between the centroids and testing data points the Euclidean distance is calculated. The accuracy in the handwritten data was 91.50%

Support Vector Machine:

The classes A, B, C, D, E have been selected with the help of data handler. To classify these classes the set of first 30 images are used as training images and the remaining 9 images are used as testing images. In this way the classifier has been trained in 30 X 5 = 150 images. For the training part the classifier is tested by a different training data of 9 X 5 = 45 images. Sci-Kit learn library has been used to implement SVM. Class svc is imported, trainX and trainY data was fitted in this function kernel used was linear. The function predict is used to classify the test data. The accuracy in the handwritten data was 95.21%

Task B Cross Validation

In task B we carried out cross validation on ATNTFaceimages400 using the four classifiers. The data handler has been used to split the data for cross validation. 5-fold cross validation has been done. At every fold we chose different dataset. The 320 images data has been used as training and the 80 images has been used for testing for each iteration since it was 5-fold cross validation. For the first fold first 2 images from each class was used as testing data and other 320 images were used as training data. The next two images from each class was used as testing data and other 320 images as training data for the second fold. This process is done similarly for each process and for each classifier we get 5 different accuracies and the average of those accuracies is calculated.

Linear Regression Accuracies:

Fold	**Accuracy(%)**
Fold1	96.25
Fold2	91.25
Fold3	96.25
Fold4	85.00
Fold5	87.50
Average Accuracy	91.25

Support Vector Machine Accuracies:

Fold	**Accuracy(%)**
Fold1	100.00
Fold2	97.25
Fold3	100.00
Fold4	96.50
Fold5	96.50
Average Accuracy	98.05

Nearest Centroid Classifier Accuracies

Fold	Accuracy(%)
Fold1	93.25
Fold2	97.25
Fold3	97.75
Fold4	89.5
Fold5	90.5
Average Accuracy	93.65

Since KNN has a huge time complexity we have confined cross validation to first 5 classes.

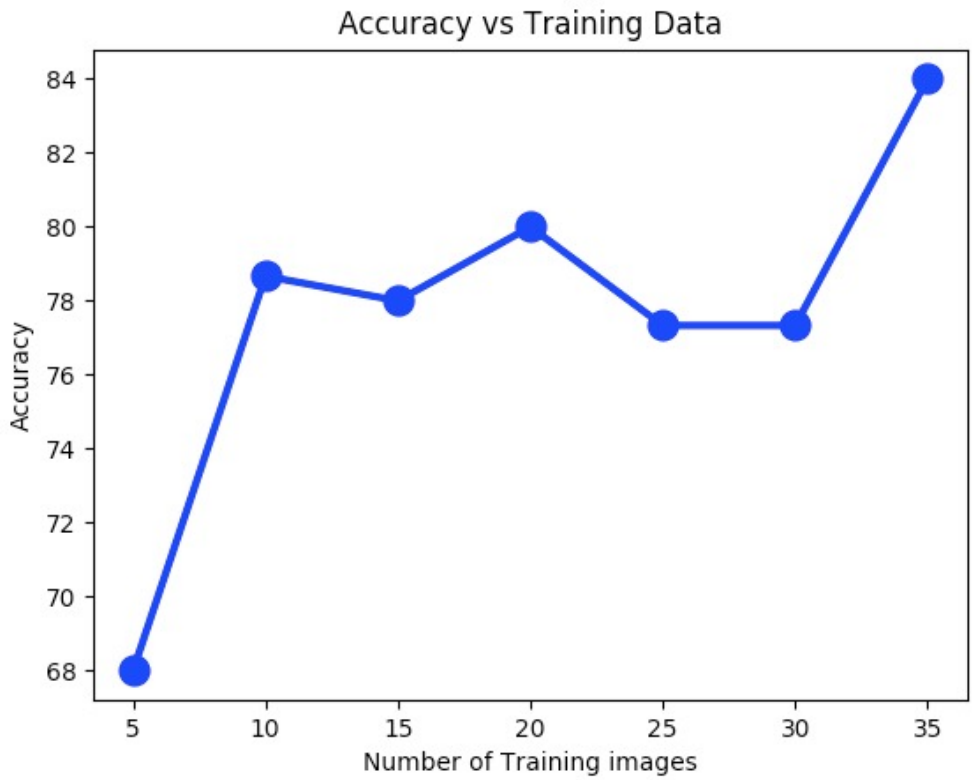
Fold knn	**Accuracy(%)**
Fold1	90.00
Fold2	90.00
Fold3	90.00
Fold4	100.00
Fold5	100.00
Average Accuracy	94.00

Task C and D

Task C and D was carried out on handwritten data and the classifier used was centroid classifier. For task C the 10 classes we choose were A, B, C, D, E, F, G, H, I, J at each iteration we increased the number of training data.

The Accuracies for the same are as follows.

Splits		Accuracies(%)
Train		
Test		
5	34	68.20
10	29	78.68
15	24	77.97
20	19	80.11
25	14	77.20
30	9	77.40
35	4	84.16

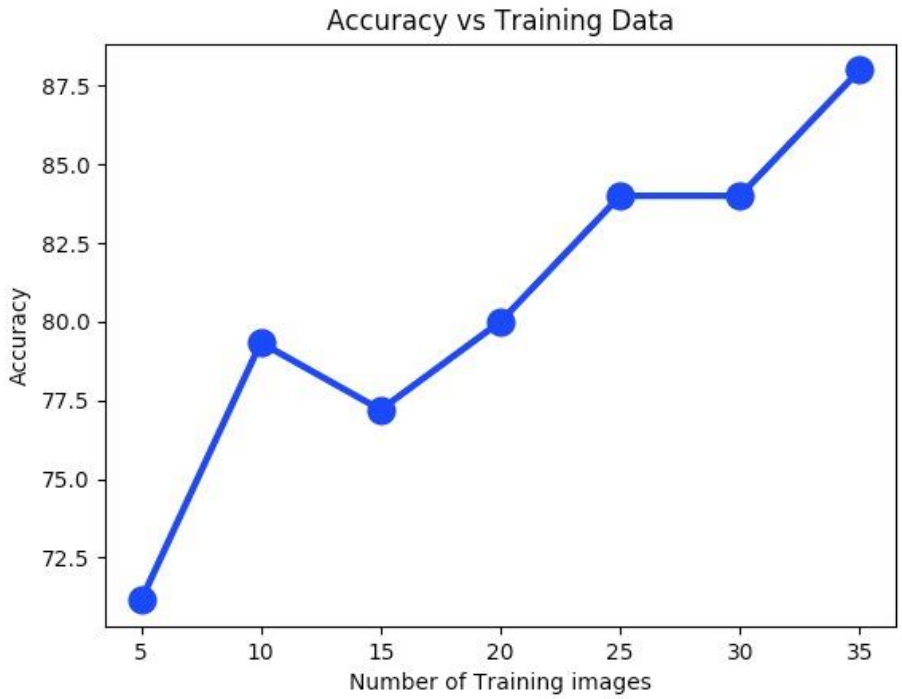


From the graph we can see that if the number of train data is increased the accuracy gets better. The highest accuracy is when the training data 35 and testing data is 4 **from each class.**However But there is an increase when we take 20 data instances for training and 19 for testing.

For **Task D** the 10 classes we choose were L,M,N,O,P,Q,R,S,T,U at each iteration we increased the number of training data

The accuracies for the dame are as follows

Splits		
Train		**Accuracies(%)**
Test		
5	34	71.11
10	29	79.29
15	24	77.18
20	19	79.95
25	14	83.92
30	9	83.95
35	4	87.95



From the graph we can see that if the number of train data is increased the accuracy gets better.The highest accuracy is when the training data 35 and testing data is 4 **from each class.**

References

- https://github.com/koreaditya/Face-Image-and-Handwritten-Letter-Image-Recognition-
- https://stackoverflow.com/questions/1401712/how-can-the-euclidean-distance-be-calculated-with-numpy