Machine Learning Project 1 Report
Arpit Saxena: 1001553587

Task A

Linear Regression:
For Task A we used linear regression to carry out classification on the handwritten data. Using Data handler, we selected classes A, B, C, D, E. to carry out classification, out of these classes we used the first 30 images as training images and other 9 images as the testing Images. Hence, we trained our classifier in 30 X 5 = 150 images and tested our classifier in different training data which had 9 X 5 = 45 images. Using the linear regression code provided by the professor we got the accuracy to be 88.8889 %.

K Nearest Neighbors:

For Task A we also used KNN to carry out classification on the handwritten data. Using Data handler, we selected classes A, B, C, D, E. to carry out classification, out of these classes we used the first 30 images as training images and other 9 images as the testing Images. Hence, we trained our classifier in 30 X 5 = 150 images and tested our classifier in different training data which had 9 X 5 = 45 images. For KNN we used the Euclidean distance to calculate the distance between the training data points and testing data points. We stored the distances in a data structure dictionary in python where the key was the distance between each test and train data point and the value was the label or the class which it belonged to. The value of K was taken to be 5 for better accuracy. Election function was used to select the nearest neighbors. The test data point was classified to the label which had maximum occurrence in K. The accuracy in the handwritten data was 93.33%.

Nearest Centroid:

For Task A we also used Nearest Centroid Method to carry out classification on the handwritten data. Using Data handler, we selected classes A, B, C, D, E. to carry out classification, out of these classes we used the first 30 images as training images and other 9 images as the testing Images. Hence, we trained our classifier in 30 X 5 = 150 images and tested our classifier in different training data which had 9 X 5 = 45 images. The function Centroid calculated the centroid for each training data point which belonged to same class, hence here we had 5 centroids for each class A, B, C, D, E. For nearest centroid we used the Euclidean distance to calculate the distance between the centroids and testing data points. The accuracy in the handwritten data was 91.111%.

Support Vector Machine:

For Task A we also used SVM to carry out classification on the handwritten data. Using Data handler, we selected classes A, B, C, D, E. to carry out classification, out of these classes we used the first 30 images as training images and other 9 images as the testing Images. Hence, we

trained our classifier in 30 X 5 = 150 images and tested our classifier in different training data which had 9 X 5 = 45 images. We used Sci-Kit learn library to implement SVM. Class svc was imported, trainX and trainY data was fitted in this function kernel used was linear. Classification was carried out using the function predict on the test data. The accuracy on this classifier was 95.5556%.

Task B Cross Validation

In task B we carried out cross validation on ATNTFaceimages400 using the four classifiers. The data was split using the data handler. Using the data handler, we split the data for cross validation. We carried out 5-fold cross validation. Where at every fold we choose different dataset. Since the cross validation was 5-fold the 320 images data was used as training and 80 images was used for testing at each iteration. For the first fold first 2 images from each class was used as testing data and other 320 images were used as training data. For second fold the next two images from each class was used as testing data and other 320 images as training data. The similar process was carried for each process and we get 5 different accuracies for each classifier and we calculate the average from those accuracies.

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	98.75
Fold3	100.00
Fold4	96.25
Fold5	96.25
Average Accuracy	98.25

Nearest Centroid Classifier Accuracies

Fold	Accuracy(%)
Fold1	93.75
Fold2	97.5

Fold3	97.50
Fold4	90.00
Fold5	90.00
Average Accuracy	93.75

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.

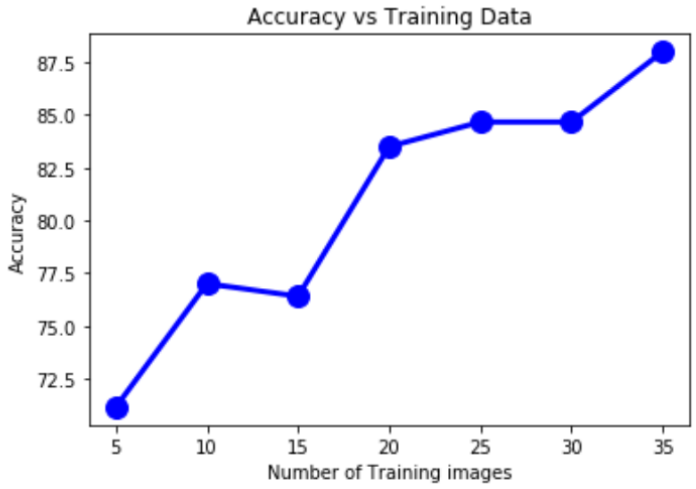
Splits		Accuracies(%)
Train	Test	
5	34	68.00
10	29	78.66
15	24	78.00
20	19	80.50
25	14	78.00
30	9	82.00
35	4	82.00

The trend seen in this graph is that if we increase the number of train data we get a better accuracy. The highest accuracy is when the training data 35 and testing data is 4 from each class. However there is an increase when we select 20 data instances for training and 19 for testing.

For Task D the 10 classes we choose were Q, R, S, T, U, V, W, X, Y, Z at each iteration we increased the number of training data

The accuracies for the dame are as follows

Splits		Accuracies(%)
Train	Test	
5	34	71.15
10	29	77.00
15	24	76.40
20	19	83.50
25	14	84.66
30	9	87.00
35	4	88.00



The trend seen in this graph is that if we increase the number of train data we get a better accuracy. The highest accuracy is when the training data 35 and testing data is 4 from each class.

Reference:

- https://github.com/koreaditya/Face-Image-and-Handwritten-Letter-Image-Recognition-
- https://github.com/ageitgey/face_recognition
- https://github.com/akshaybahadur21/Facial-Recognition-using-Facenet