CS 6375

Project Report

Hand Written Digit Recognition

Number of free late days used: \_\_\_\_\_\_\_\_0\_\_\_\_\_\_\_\_\_\_\_\_   
Note: You are allowed a **total** of 4 free late days for the **entire semester**. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

Paril Doshi - PSD170000

Vivek Shah - VXS173830

Debanjana Gupta – DXD170016

Dataset Description:

The project is on Digit Recognition and the dataset selected is from Kaggle Competition.

The dataset is of MNIST type.

The dataset contains 60000 images of hand written digits. Out of which 40000 images are used for Training and other 20000 for Test.

There are following number of training samples available for each digit:

0 = 4132 1 = 4684 2 = 4177 3 = 4351

4 = 4072 5 = 3795 6 = 4137 7 = 4401

8 = 4063 9 = 4188

The training dataset has 785 columns. The first column is for labels and the other 784 columns are pixel values for a 28 x 28 image.

The test dataset has 784 columns of pixel value for a 28 x 28 image.

The pixel value ranges from 0(Black) to 255(White) and middle values are for shades of grey colour

Techniques we plan to use:

The algorithms that we plan to implement are support vector machine, neural net, k-nearest neighbours and compare their results.

Then we plan to choose the best algorithm among them and try to improve the results by tuning the parameters.

For our initial run we have tried SVM.

Experimental Methodology

The dataset is already flattened(In 1-d array). And then the individual pixels are divided by 255(The range of pixel value is 0(Black) to 255(White)) to standardize the data.

We also plan to standardize the label data into binary form for CNN.

For our initial run the training is split into test and train test using cross\_validation package of sklearn.

*X\_train, X\_test, y\_train, y\_test = cross\_validation.train\_test\_split(X/255.,y,test\_size=0.1,random\_state=0)*

Coding Language / Techniques:

We are using python as a coding language for algorithm. We are using sevral machine language library like Panda, Sklearn, Matplolib, TensorFlow, Numpy.

We are using spyder and google colab for coding and github for code integration.

Experimental Results:

Initially we divided the training data into two sets one for training and other for testing and then used Linear SVC from sklearn and got 91% accuracy.

As the training data was split for training and testing the accuracy count got low.

And further by applying the Convoluted Neural Network the accuracy can be increased further.