Akanksha Bansal	Assignment 1
Machine Learning	Feb 7, 2015

Basis Of the Assignment The problem statement requires us to build a model which after sufficient training is able to classify the digits with significant accuracy. We have been provided with images of the digits which can be used directly for training without having to bother with feature extraction. Otherwise, the images of the digits would have needed to translated and scaled so that each digit is contained within a box of a fixed size. This greatly reduces the variability within each digit class, because the location and scale of all the digits are almost the same, which makes it much easier for the training to take place.

As part of the experiment we train the model to recognize the Eigen Vectors which can be used to reconstruct the digits and effectively determine the label of the image. The 28*28 pixels of the image are treated as distinct features of the image and are then used to determine the Principal Components of the images. The division of the data into these Eigen vectors allows us to project the vectors in the space which maximizes the differences between various digits images. Thus maximizing the variance between the classes. Another advantage of this dimension reduction is that it increases the efficiency of the model when making use of the classification model (Clustering) like Knn.

Why Dimensionality reduction Dimension reduction is usually performed for data sets with high dimensions so as to avoid the effect of curse of dimensionality. It is performed prior to applying any kind of clustering algorithm. For this assignment, we perform dimension reduction by making use of Eigen vectors. We only make use of the top k Eigen vectors as these should be able to capture the most of the variance with in the training. We then project the training and the test data into the Vector space determined by the Eigen Vectors. Then a clustering algorithm is applied to the data set in order to determine the label of the test dataset.

Then in order to perform clustering I make use of K-Nearest neighbors algorithm to chose the label for the test data set. Another way could be making use of K-mean Clustering. This process is also called low-dimensional embedding.

In KNN we assign the test datapoint the label which occurs most frequently in the nearest k neighbors of the test data point.

Sample images which indicate how the data sets look like when projected on the Eigen Space.

Figures represent the Train set images and their Eigen Digits once Reconstructed followed by the Test Set 1 images and their Eigen Digits once Reconstructed followed by the Test Set 2 images and their Eigen Digits once Reconstructed.

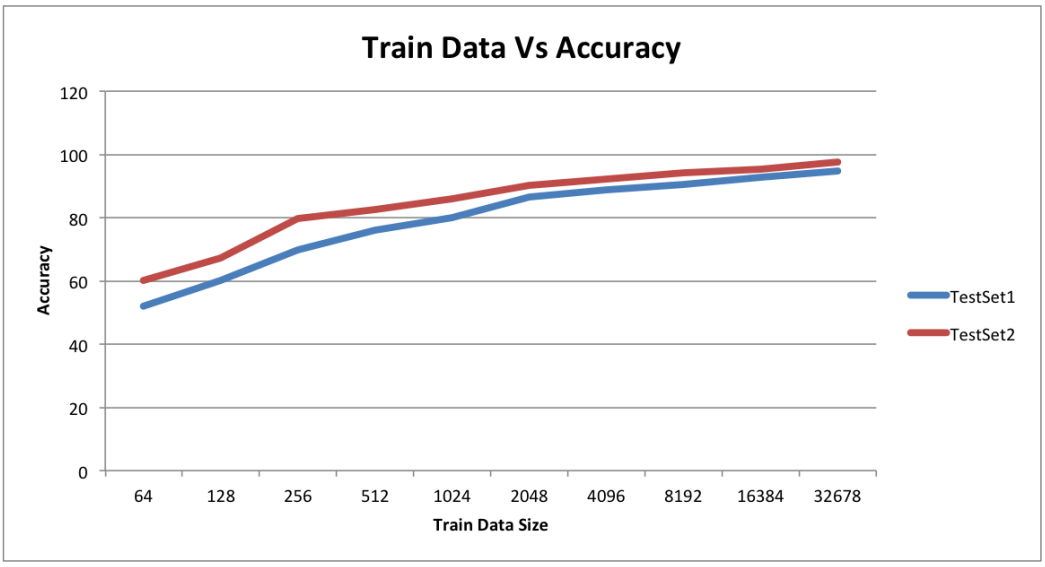
Experiments We have been given with two types of Test data. The first half of 5000 images are of better quality(TestSet1) and the rest 5000 images are of the poorer image quality((TestSet2). So I

have performed all the tests on both set of Test images. Surprisingly the model worked better for poorer images than for better quality images. This could be because of the fact that poorer quality images have higher variance and is thus making clustering more effective.

In order to understand the behavior of the model under various variables I performed the following set of Tests :

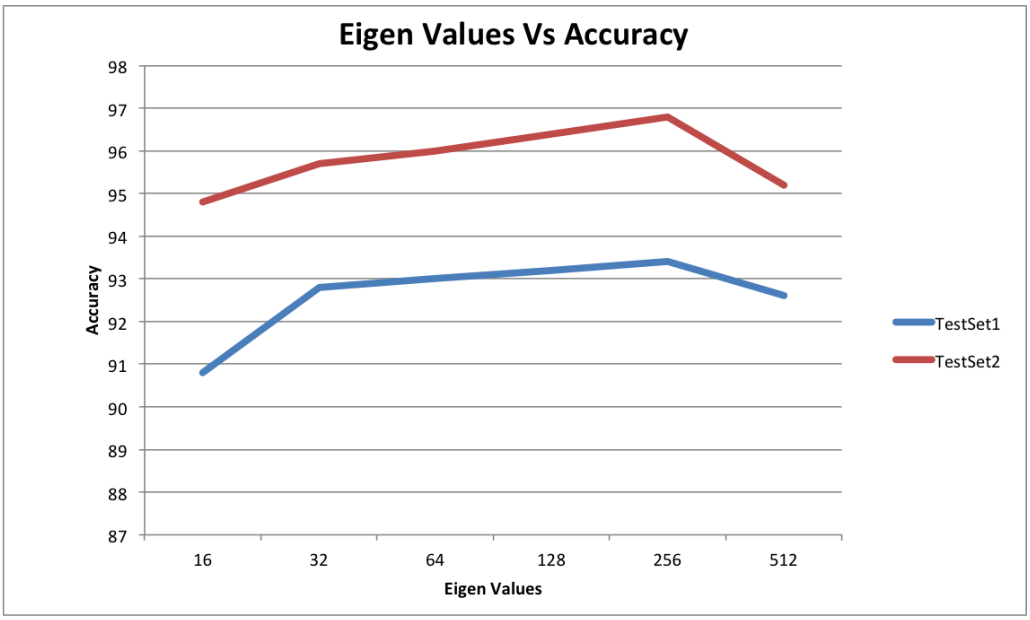
- Vary Size of the training data.

  I increased the training data starting from $2^6 = 64$ training images to 2^{15} for the various tests. As we would expect, with increase in the training images set, the accuracy of the model increased from 52% to 97%. As can be seen in the image below, with increase in the training data we see a increase in the accuracy. This increase is not linear. As can be seen, when training with low amount of data, doubling the training data size has a 10% increase in the accuracy, but not at higher levels of training data size.



- Varying no of top Eigen vectors chosen

  In order to understand the behavior of the model when we change the no of Eigen vectors used for projecting the data from the 2^4 to 2^8. We would expect that with increasing the no of Eigen vectors we would get an increase in the accuracy. However, after certain point, increasing the no of Eigen vectors tends to reduce the accuracy. Thus making use of top Eigen Vectors is more effective and efficient.



- Vary No of nearest neighbors used for KNN classification

  Varying the no of nearest neighbors chosen to determine the test data label doesn't have any kind of significant change on the accuracy of the model. The best results are observed with keeping the nearest neighbors to a minimum with regards to both performance and accuracy achieved. When the training data is large, i.e. we have more than 1000 training images and about 64 top Eigen vectors being chosen for the training model, I didn't know see any kind of difference in the accuracy when varying the no of top nearest neighbors chosen for the Clustering. However, when the training data is small, example making use of only 50-200 training images and about 16 top Eigen vectors, I did see a small change in the accuracy, however the change was less than 2% variation in the values.