Group 24
Xiaoyi GONG 54947670
Shuyi TANG 54952471

Project Proposal

1. Introduction

MNIST database is a large dataset of handwritten digits that is commonly used for training various image processing systems. The purpose of this project is to predict the correct label of handwritten digit which from a subset of the MNIST digits dataset.

2. Project Description

Feature: In this project, we use image's pixel grayscale values as features. Each feature vector is a image with 28×28 dimensions. Each dimension contains grayscale values from 0 to 255.

Dataset: We will use provided dataset which contains the following files:

- digits4000_digits_vec.txt – a 784×4000 matrix, where each column is vectorized image.
- digits4000_digits_labels.txt – a 1×4000 matrix with the corresponding label from zero through nine.
- digits4000_trainset.txt - a 2×2000 matrix, where each row is set of indices to be used for training the classifier.
- digits4000_testset.txt - a 2×2000 matrix, where each row is the corresponding set of indices to be used for training the classifier.

Methodology: In this project, we have three steps as following:

- **Pre-process phase**
  To get a lower dimensional space, we will use PCA to pre-process data. PCA method will find directions that contain the largest subset of data points relative to all the data points present.

- **Training phase**
  In this step, we plan to use SVM model and KNN model. We will use SVM model to create a vector classifier, then pass our trainset and labels to the classifier's fit method, which trains our model. Since SVM can only do binary classification, we will use one-against-one method for multi-class learning problems. For KNN, we just need to import trainset and corresponding labels.

- **Validation and prediction phase**
  After training the classifier, apply the SVM classifier and check accuracy on data that wasn't used in training. KNN needs to take each test point and find the closest sample to it in our trainset. Then, get the K smallest distances and their corresponding label values. Lastly, check the accuracy.

Evaluation: we will evaluate the outcome on the categorization accuracy of our predictions. (number correct predictions / total number we predict). Comparing accuracy value between different methods, choose algorithm with the greatest accuracy to be the most suitable algorithm.