Math 425   Spring 2021   Project 2   due: 5PM on Fri April 23

Project Directions

- Include a report on every group member's contribution.
- Submit the group's well commented code used for the project with instructions on how to compile and run.
- Make an **8** to **15** minute video presentation of your results.

The project consists of one problem

Problem 1

Classification of Handwritten Digits

On iLearn you will find the following data files of handwritten digits[1]:

- `handwriting_training_set.txt`: 4000 training examples of handwritten digits. Each training example is a 20 pixel by 20 pixel grayscale image of a digit reshaped into a 400-dimensional vector. Each pixel is represented by a floating point number that indicates the grayscale intensity at that location. Thus the set is a 4000 by 400 matrix.
- `handwriting_training_set_labels.txt`: This data set contains the labels of the corresponding digits in the training set. The digits "1" to "9" are labeled as they are. However, because MATLAB has no zero index, the digit zero is represented as the value ten, i.e. "0" is labeled as "10."
- `handwriting_test_set.txt`: 1000 test set of handwritten digits with the same format as the training set. Thus this set is a 1000 by 400 matrix.
- `handwriting_test_set_labels.txt`: The labels for the test set.

A. Construct an algorithm for classification of handwritten digits. Use the training set and compute the SVD of each class/digit matrix. Note that in the training set, the first 400 are examples of the digit 0, the next 400 are examples of the digit 1, etc.. Identify the unknown test digits by using the singular value decomposition of the each digit matrix. Do the classification using 5, 10, 15 and 20 singular vectors as a basis.

SPECIFIC TASKS:

i. Give a table or graph of the percentage of correctly classified digits as a function of the number of basis vectors.

ii. Check if all digits are equally easy or difficult to classify. Also look at some of the difficult ones, and see that in may cases they are very badly written.

[1] This is a subset of the MNIST handwritten digit dataset (http://yann.lecun.com/exdb/mnist)

iii. Check the singular values of the different classes. Is there evidence to support using different number of basis for different digits?

B. Implement the following **two-stage** algorithm:

In the first stage compare the unknown digit only to the first singular vector in each class. If for one class/digit the residual is significantly smaller than for the others, classify as that digit. Otherwise perform the algorithm above. Is it possible to get as good a result for this version? How frequently is the second stage necessary?