Classification of Handwritten Digits

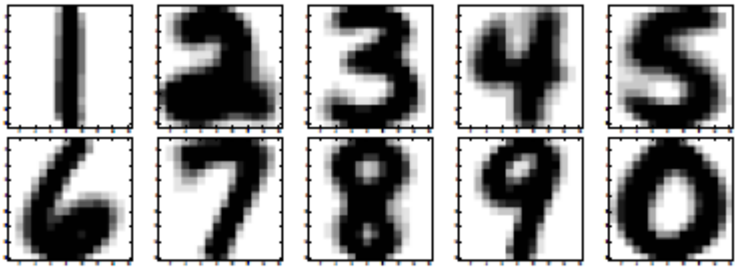
Problem Statement

Classification by computer of handwritten digits is a standard problem in pattern recognition. The typical application is automatic reading of zip codes on envelopes. In this assignment you'll address the following problem: Given a set of manually classified digits (the training set), classify a set of unknown digits (the test set) using SVD method.

Data Set

We will be using the US postal Service database that contains 1707 training and 2007 test digits (uploaded to Canvas). Each image is a grayscale 16x16 image that is converted to a 256x1 column vector by stacking all the columns of each image matrix above each other.

- The training images are stored in trainInput.csv. (256x1707).
- The correct digit corresponding to each column of trainInput is stored in trainOutput.csv. (1x1707).
- The test images are stored in testInput.csv. (256x2007).
- The correct digit corresponding to each column of testInput is stored in testOutput.csv. (1x2007).



Methodology

- Form a matrix A for each digit, such that each row in A represents an image of that digit. (You will have 10 A's).
- Determine the singular value decomposition for each A. (Right singular vectors V_i are an orthogonal basis in the image space of that digit. We will refer to the right singular vectors as "singular images.") You should get 10 sets of singular images, one for each digit.
- Express test images as a linear combination of the first k=20 singular images of each digit. (This is a least square problem of the form Ax=b).
- Compute the distance between test images and their least square approximations.
- Classify each test image to be the digit corresponding to the smallest residual.
- Calculate the overall correct classification rate, as well as correct classification rate for each digit in a confusion matrix.