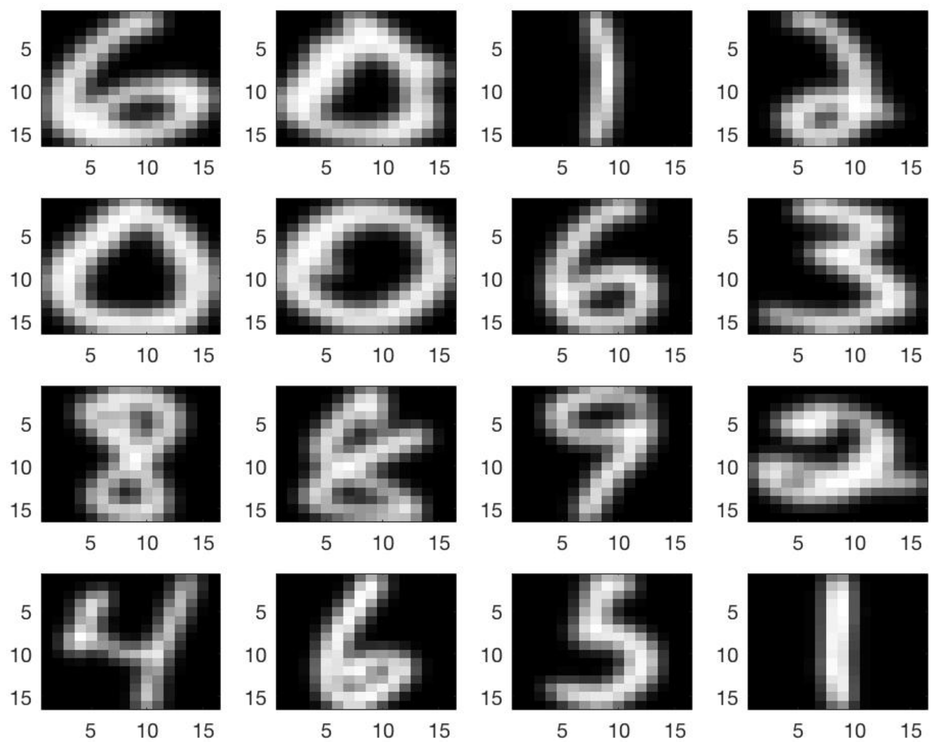
MAT 167 - Final Project

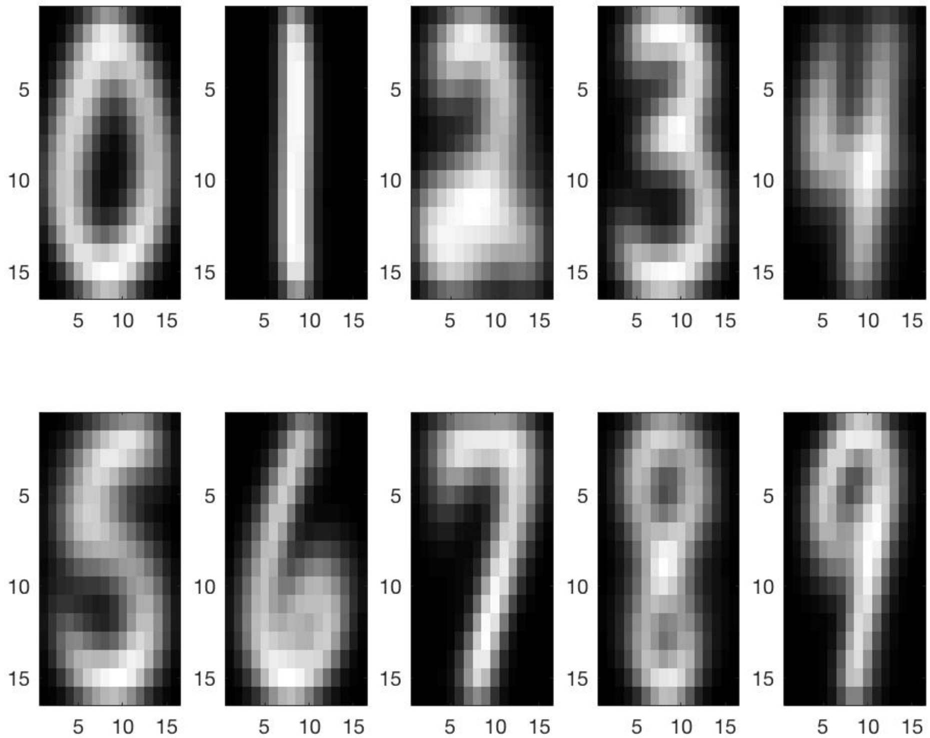
Dannie Vo - Student ID: 915004803

June 8, 2018

Step 1:
The first 16 images in train_patterns array from Figure 1 in MATLAB:



Step 2:
The 10 mean digit images from Figure 2 in MATLAB:



Step 3(c):
The confusion matrix test_confusion of size 10×10:

$$\begin{pmatrix}
656 & 1 & 3 & 4 & 10 & 19 & 73 & 2 & 17 & 1 \\
0 & 644 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\
14 & 4 & 362 & 13 & 25 & 5 & 4 & 9 & 18 & 0 \\
1 & 3 & 4 & 368 & 1 & 17 & 0 & 3 & 14 & 7 \\
3 & 16 & 6 & 0 & 363 & 1 & 8 & 1 & 5 & 40 \\
13 & 3 & 3 & 20 & 14 & 271 & 9 & 0 & 16 & 6 \\
23 & 11 & 13 & 0 & 9 & 3 & 354 & 0 & 1 & 0 \\
0 & 5 & 1 & 0 & 7 & 1 & 0 & 351 & 3 & 34 \\
9 & 19 & 5 & 12 & 6 & 6 & 0 & 1 & 253 & 20 \\
1 & 15 & 0 & 1 & 39 & 2 & 0 & 24 & 3 & 314
\end{pmatrix}$$

Step 4(d):
The confusion matrix test_svd17_confusion of size 10×10:

$$\begin{pmatrix} 772 & 2 & 1 & 3 & 1 & 1 & 2 & 1 & 3 & 0 \\ 0 & 646 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 3 & 6 & 431 & 6 & 0 & 3 & 1 & 2 & 2 & 0 \\ 1 & 1 & 4 & 401 & 0 & 7 & 0 & 0 & 4 & 0 \\ 2 & 8 & 1 & 0 & 424 & 1 & 1 & 5 & 0 & 1 \\ 2 & 0 & 0 & 5 & 2 & 335 & 7 & 1 & 1 & 2 \\ 6 & 4 & 0 & 0 & 2 & 3 & 399 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 2 & 0 & 0 & 387 & 0 & 11 \\ 2 & 9 & 1 & 5 & 1 & 1 & 0 & 0 & 309 & 3 \\ 0 & 5 & 0 & 1 & 0 & 0 & 0 & 4 & 1 & 388 \end{pmatrix}$$

Step 5:

(a)

The training set contains 4649 handwritten single digits and the testing set also contains 4649 handwritten single digits. All digits are from 0 to 9 that is used for constructing a classification algorithm. Both train_patterns and test_patterns have size 256×4649 since each digit image is stored under a 1D array with the length of 256. Each digit consists of 16 × 16 pixel image and the pixel values are ranged from -1 to 1. Both train_labels and test_labels have size 10 × 4649

(b)

We create a new figure window (2) and color it gray, then we create train_aves matrix with size 256 ×10, representing the mean digits in train_patterns matrix, then print the 10 mean digits images using subplot() function and imagesc() function. The reason why we're computing the mean digits using the training digits because it's the simplest classification algorithm.

(c)

Using the simplest algorithm: The matrix below shows the digit and its corresponding effectiveness by percentage where the digits are from 0-9:

$$\begin{pmatrix} 0 & 83.4606 \\ 1 & 99.5363 \\ 2 & 79.7357 \\ 3 & 88.0383 \\ 4 & 81.9413 \\ 5 & 76.338 \\ 6 & 85.5072 \\ 7 & 87.3134 \\ 8 & 76.435 \\ 9 & 78.6967 \end{pmatrix}$$

At digit 5, it's the most difficult to identify correctly as the percentage = 76.3380; and at digit 1, it's the easiest to identify correctly as the percentage = 99.5363.

Using the k-nearest neighbor classification algorithm: The matrix below shows the digit and its corresponding effectiveness by percentage where the digits are from 0-9:

$$\begin{pmatrix} 0 & 98.2188 \\ 1 & 99.8454 \\ 2 & 94.9339 \\ 3 & 95.933 \\ 4 & 95.7111 \\ 5 & 94.3662 \\ 6 & 96.3768 \\ 7 & 96.2687 \\ 8 & 93.3535 \\ 9 & 97.2431 \end{pmatrix}$$

At digit 8, it's the most difficult to identify correctly as the percentage = 93.3535; and at digit 1, it's the easiest to identify correctly as the percentage = 99.8454.

Based on the percentages of each digit of each algorithm: We can conclude that using the k-nearest neighbor classification algorithm will give us a more precise result based on each digit of each algorithm.

The theory behind the algorithm is to compare which algorithm is better for use by showing the chance of identifying a handwritten single digit. Moreover, using the k first k left singular vector of the SVDs of the training digits is more precise as it is the most optimized to show the accuracy since you don't want to under-optimize or over-optimize a data set by setting the rank either to low or too high.

(d)

Using the simplest algorithm: we get the percentage = 84.6634 and using the k-nearest neighbor (k-NN) classification algorithm : we get the percentage = 96.6229

This implies that the k-nearest neighbor classification algorithm is more effective as the result it gives is better than the result that uses the simplest algorithm.