# 2024 WQ Final Programming Project Report*

Your Name

Fri Mar 15 04:26:00 PM PDT 2024

## Contents



*Revision 1.03

# 1 PROBLEM DESCRIPTION

**(20 points total)**

I want you to provide ***thoughtful*** responses to the prompts in this template.

As usual, in this section write a description of what you are doing and why it is useful or important.

You will find some helpful information in Chapter 10.01 of the first edition of the textbook.

There is also valuable information concerning this project, including background and history, here:

NS_LECTURE_21

**Do not plagiarize these lecture notes!** You may use the information in these lecture notes but **you must write this report entirely in your own words.**

Furthermore, your computer program must produce the required output for this report - including figures and confusion matrices - when the TA runs it. **(50 points - A separate from this section)**.

Even though the title of this section is "Problem Description" this is also where you **write the required** Introduction and Motivation.

**As always, keep in mind your intended audience .** Your report should be sufficiently detailed (and clear!) to give someone with an expert's understanding of (Numerical) Linear Algebra and computing, (but not necessarily the SVD) enough of an idea of what you did that they could reproduce your results.

**Make sure you address all of the following points.**

**(a)** **(05 points)** Introduction to and overview of the problem.

**(b)** **(05 points)** What are the objectives of this exercise or procedure on which you are reporting?

**(c)** **(05 points)** Why is it important?

**(d)** **(05 points)** Briefly mention where your report leads or ends up; i.e., briefly mention one or more conclusions.

**(e)** Keep this part concise and to the point.

This project involves identifying the images of the handwritten digits data from a set of images. It is part of the computer vision. It is applied in a lot of scenarios in the real life, for example, character identification. This is important because in real life, we have a lot of images containing digits and we need to use computer to identify them. In this project, we wil use Centriod Classification algorithm SVD method to identify the matrix images of digits, classify handwritten digits by rank-17 approximations. More importantly, we will compare the performance of two algorithms using matlab.

# 2 DATA SET

**(25 points total)**

Report on **Steps 01(a)-(b)** of the assignment.

(a) **Step 01(a)** **(05 points)** Briefly describe the data structure in which the images of the digits are stored.

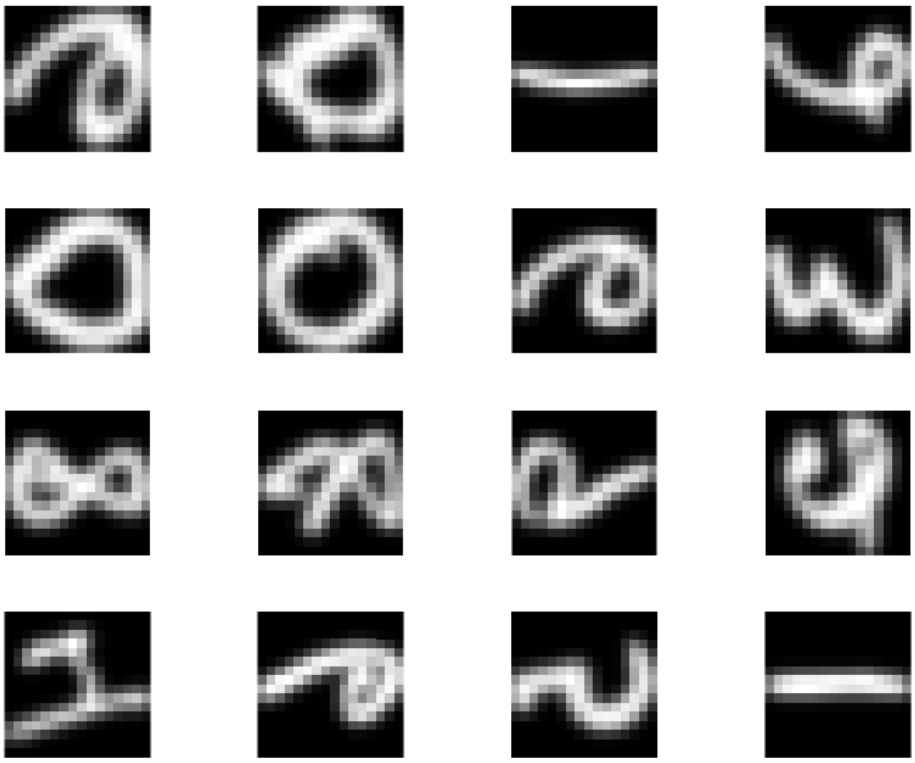
The dataset we are using is USPS.mat. We will extract the images of the digits and store them in train_patterns and test_patterns. Those variables are the matrix of 256* 4649 matrix, which means that for each digits images, it is 16*16= 256 pixels and there are 4649 images to be trained. We also have the corresponding train_labels and test_labels, which are matrix variables with normalized vectors, which means each number is from -1 to 1.

(b) **Step 01(a)** **(05 points)** Include a (brief) explanation of the difference between the training data and the test data. (This is a simple example of *machine learning*.)
For training data, it is used for the computer to study and approach the result for machine learning. For test data, it is used to evaluate the overall result of learning.

(c) **(05 points)** Where can the data be obtained by a member of the general public; i.e., someone who **does not have access** to UCD CANVAS where you got your data.

The handwritten digit data (similar to the one used here) can typically be found in public datasets like the MNIST database, link:http://www.gaussianprocess.org/gpml/data/ .

(d) **Step 01(b)** **(05 points)** Display the first 16 images in train patterns. (Include this figure here.)



(e) **(05 points)** Write something brief but sensible about this figure; e.g., "In Figure 01 we display ..." **Do not** just plop the figure down on the page without writing something about it!
From the figure, we can observe that the digits are:
9 0 1 6
0 0 9 3
8 8 9 6
4 6 5 1

# 3 THE CENTROID CLASSIFICATION ALGORITHM

**(05 points)**

**Include SEVERAL sentences here to introduce the Centroid algorithm.** For example, is this the most complicated algorithm that you can think of? The Centroid algorithm is used some center point to classify a set of data. It is an unsupervised learning algorithm for image recognition.

## 3.1 DESCRIPTION OF THE ALGORITHM

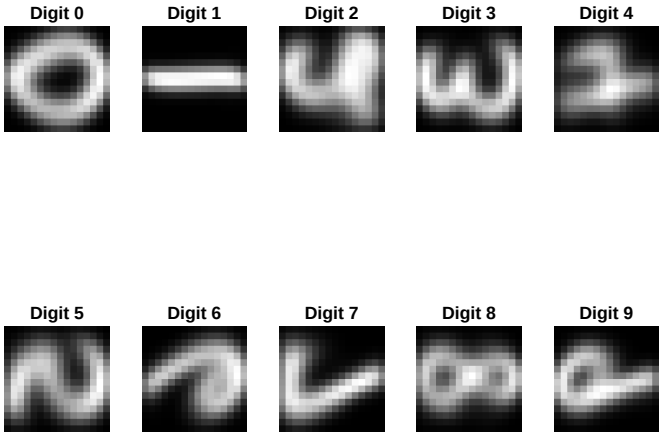
**(15 points total)**

**Report on Step 02 here.**

This section should be self explanatory. However, here are some prompts in case you want them.

1. **(05 points)** Describe the algorithm in sufficient detail that your clone could reproduce your results based only on your description.
   1, initializing centroids by selecting random data points from the dataset. 2, find the mean data for each image: the centroids of the clusters corresponding to each digit from 0 to 9.

2. **(05 points)** Display the 10 mean digit images; i.e., place the figure here.




3. **(05 points)** Write something brief but sensible about the figure; e.g., "In Figure 02 we display ..." From the figure, we observe that the images are from 0 to 9 respectively, which reaches the conclusion.

## 3.2 DESCRIPTION OF THE RESULTS

**Report on Step 03(b)–(c) here.**

(05 points) Print the confusion matrix here. **Do not just leave the confusion matrix in this subsection without any text!**

Here is confusion matrix of step 3:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 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 |

(05 points) Describe the confusion matrix. For example,

1. What is a confusion matrix?
   A confusion matrix is a result matrix describing the performance of a classification model on a set of test data for known values.
2. What do the rows, columns, diagonal entries, represent?
   rows:the actual classes
   columns:the predicted classes
   Diagonal entries: the correctly classified instances for each class.

# 4 THE SVD CLASSIFICATION ALGORITHM

**(05 points)**

**This is the beginning of Section 04! DO NOT LEAVE IT BLANK!**

Include several introductory sentences here.

Remember, the reader may not know what the SVD is. So you could start with something like the following.

The Singular Value Decomposition (SVD) is a powerful matrix factorization ...

## 4.1 DESCRIPTION OF THE ALGORITHM

Report on **Steps 04(a)–(b)** in this Section.

**IMPORTANT!** This is the most technical section of your report. Try to clearly explain the algorithm. (Don't hesitate to use equations.[1]) Your explanation should be sufficient to give someone with an expert's understanding of Numerical Linear Algebra - but not the SVD - a *basic* idea of what the SVD is and how the algorithm works.

(a) **Step 04(a) (10 points)** In your words and - if you like - equations describe Step 4(a).
In Step 04(a), we compute the rank 17 SVD for each digit, which extracts all the digit images and compute the SVD matrix for all of them. The svd formula is $X_k = U_k \sum_k V_K^T$.

(b) **Step 04(b) (10 points)** Similarly, in words and (possiby) equations describe Step 4(b).
In Step 04(b), we compute the expansion coefficients of each digit image of test dataset with the 17 singular vectors of each train digit image set,respectively. We use the formula $c_{jk} = U_k^T * x_j$.

## 4.2 DESCRIPTION OF THE RESULTS

(a) **Step 04(c) (10 points)** In words and equations(?) describe Step 4(c).

- How do you compute the error and what is the rationale for doing it that way?

We compute the error between each original test dataset for digit image and its rank 17 approximation using the left singular vectors of the corresponding matrix.

(b) **Step 04(d) (10 points)** Print the confusion matrix. Describe this confusion matrix and what it tells you about the SVD algorithm's performance on this problem.

- **(05 points)** For the correct confusion matrix and **(05 points)** for your thoughts concerning what it tells you.

The confusion matrix will tell how well the SVD perform for the digit recognition, telling the number of test cases that were classified correctly and incorrectly

# 5 ANALYSIS

**(20 points)**

**STEP 05 Analyze your results!**

**Compare and contrast the Centroid and SVD Algorithms**

(a) **(10 points** (05 points per algorithm) Summarize all of your results. How effective is each algorithm; i.e, for that particular algorithm what percentage of *each* digit is identified correctly? For example, which digit is the most difficult to identify correctly? Which digit is the easiest to identify correctly? You can

[1] An equation often communicates an idea better than a pile of words.

obtain all of this information from the confusion matrices. Include some thoughtful reasoning to support your analysis.
The percentage of both svd and centroid algorithms are approximate 90%.From each confusion matrix of both algorithms, number 8 has the least cases correctly recognized, while number 1 has the most cases of that. The reason why this happened is because image of 8 is the most complicated and the image of 1 is least complicated.

**(b)** **(10 points** (05 points per algorithm) Describe the differences in performance you observe (or - at least in principle - can measure) between each algorithm. Which of the two algorithms yields the best results when you include performance? **HOW DO YOU QUANTIFY "BETTER"?** More accurate, faster, less storage, ...? Why? How "much" better? The SVD has the better accuracy and less storage. According to the confusion matrix, the diagonal values for the confusion matrix of svd is less than that of the Centroid Classification algorithm. The reason why it happens is because SVD involves matrix decomposition and we can directly compute it(less space) and less calculation(more accuracy), while for Centroid Classification, we need to compare how well we calculate and adjust the weights.

# 6 CONCLUSIONS

**(30 points)**

**THIS IS THE SINGLE MOST IMPORTANT SECTION OF YOUR REPORT.**

**Compare and contrast the Centroid and SVD Algorithms**

It is OK - and in fact it is generally good practice - to summarize your work by pulling *one or two* sentences from each **Description** and **Results** Section / Subsection. The sentences you choose can even be identical to the ones you used to describe the algorithm(s) and / or summarize your results described in that particular section / subsection.

**BUT don't overdue the "identical" part.** It is tedious to constantly read repeated sentences! **:-)**

**Make sure you address all of the following points.**

1. **(05 points)** A brief recap of what you wrote in Section 1 "Introduction to and Overview" including the objectives of this exercise or procedure on which you are reporting and why it is important.
2. **(05 points)** Include a (brief) explanation of the data and an explanation of the difference between the training and test data from Section 2.
3. **(05 points)** Mention and briefly describe the centroid algorithm
4. **(05 points)** Mention and briefly describe the SVD algorithm
5. **(05 points)** Summarize your analysis from Section 5.
6. **(05 points)** Write a two to four sentence long OVERALL CONCLUSION. This project compares the performance of two algorithms, centroid classification algorithm and SVD algorithm by using the example

of image classification and write matlab code for both of them. We will first split the data into traning and testing data, which is commonly used in machine learning. Training data is used to train the computer to recognize the digits and test data is used to test how well the trainings are. For Centroid algorithm, we find the center point and cluster the data to classify the digit images. The SVD decompse the matrix into parts to identify. According to our results, 8 is the hardest digit to identify while 1 is the easiest. SVD has a better performance because it decompose the matrix(using less space), and less calculation(more accuracy). This project really let us learn computer vision and we can feel the power of mathematics.

# 7 COMPUTER PROGRAM

When the TA runs your computer program it must produce the required output for this report - including all figures and confusion matrices.

**(10 points)** per figure or matrix.