ECS 130: Final Project Report

Author: Avashi Lal (916273730)

Introduction:

The objective of this project was to create a program for image processing. Given data containing different samples of handwritten digits, two programs for the centroid and PCA algorithm creates a way to identify each handwritten digit correctly.

Algorithm:

Centroid:

The centroid algorithm takes an input of a chosen test matrix and the average train digits and outputs a vector consisting of how the algorithm labeled the digit corresponding to each element. It does this by finding the closest centroid by finding the minimum distance which is equivalent to the minimum of the norm of difference of an arbitrary row from the test matrix and the average train digits. Each element of the output vector is the index value of the centroid per each row of the test matrix.

PCA:

Similar to the centroid algorithm, the input is the test matrix and average train digits, and the output is the vector of the labeled digits. However, in this algorithm, the basis length also affects the train digits. In this method, we find the minimum distance of the norm of the difference between a row of the test matrix and the single value decomposition (SVD) of the test matrix multiplied by the quantity of the product of the transposed SVD and the row of the test matrix. The output vector consists of each of these outputs based on each of the rows of the test matrix.

Matlab Scripts:

Task 1: Centroid Algorithm

Task 2: Success Rates from Centroid Algorithm

I combined task 1 and 2 into a single function. I would input the test matrix and digit that I was testing for.

```
prompt = ('What is test input?');

A = input(prompt);

n = size(A);
prompt_3 = ('What is the digit?')
dig = input(prompt_3)
for i = 1:n
   z = double(A(i,:));
  dist = zeros(10,1);
   for k=1:10
       dist(k) = norm( z - T(k,:) );
   end
   [h,I] = min(dist); %closest centroid
   B(i) = I - 1;
   N = transpose(B);
   num = numel(find(N==dig));
   sze = size(N);
   d = sze(1);
   S_Rate = (num/d)*100;
end
dist
```

Task 3: PCA Algorithm

Task 4: Success Rates from Centroid Algorithm

I also combined task 3 and 4 into a singular function. I would input the basis length, test matrix, and the digit that I tested for.

```
prompt = (['What is basis length?']);
basis_length = input(prompt);   % the number $m$ of basis vectors
   Us = zeros( 28*28, basis_length, 10);
for k=1:10
       %
   % go through each digit 0 to 9
   %
   s = strcat('train',num2str(k-1));
   A = double(eval(s));
   %
   % get first 5 singular vector
   %
   [U,~,~] = svds( A', basis_length );
   Us(:,:,k)=U;   % store the basis vectors of digit ``k-1''.
end
prompt_2 = ('What is test?')
M = input(prompt_2);
n = size(M)
prompt_3 = ('What is the digit?')
dig = input(prompt_3)
```

```
for i =1:n
    z = double(M(i,:))';  %  select a test digit
    dist = zeros(10,1);
    for k=1:10
        Uk = Us(:,:,k);
        dist(k) = norm( z - Uk*(Uk'*z) );
    end
    dist
   [h,I] = min(dist); %closest centroid
    B(i) = I - 1;
    N = transpose(B);
    num = numel(find(N==dig));
    sze = size(N);
    d = sze(1);
    S_Rate = (num/d)*100;
end
```

Tables For Success Rates:

Centroid Algorithm Success Rate:

digit	0	1	2	3	4	5	6	7	8	9
Success Rate (%)	85.41	96.21	68.90	71.72	71.54	53.92	72.86	83.27	64.67	71.72

PCA Algorithm Success Rate:

digit	0	1	2	3	4	5	6	7	8	9
m = 1	92.04%	93.74%	74.71%	83.37%	79.74%	64.91%	88.41%	81.91%	76.39%	80.23%
m = 5	97.86%	99.21%	90.21%	93.76%	89.82%	90.13%	96.24%	89.30%	90.04%	88.90%
m = 10	98.78%	99.47%	92.73%	94.26%	95.62%	91.93%	96.56%	93.39%	91.79%	93.26%
m = 20	98.78%	99.38%	93.51%	94.16%	96.84%	93.95%	97.49%	94.55%	94.35%	93.86%

Summary Statements:

The comprehensive effectiveness of each algorithm is decent. Nonetheless, the PCA algorithm is clearly more accurate. The average success rate for the centroid method is 74.93%.

Relative to the least accurate results of the PCA algorithm of basis length 1 which had the average success rate of 81.545%, the centroid algorithm is inferior to the PCA algorithm.

The centroid algorithm does not yield an accurate enough result for it to be implemented. This is because the success rate, collectively, would cause too many errors. In engineering, there should be less than 5% margin of error, so this algorithm would not fall under this criteria.

Within the PCA algorithm, it is clear that as the basis increases, so does the accuracy. For reference, the average success rate for basis length 1 is 81.545%, 5 is 84.647%, 10 is 94.779%, and 20 is 95.687%. Overall, this algorithm yields better results than the centroid algorithm, but it still accounts for quite a bit of error. Even the most successful rate with basis length 20, barely meets the 5% error margin. Hence, it is also less than perfect.

In order to find better accuracy, we could use various methods. One of the easiest would be by testing out the PCA algorithm with a greater basis length as we know that the accuracy increases this way. Also, finding other algorithms and testing those might also provide better success rates.

Acknowledgements:

Professor Zhaojun Bai Lecture (3/10/2022)

Matthew Scates OH (3/11/2022)

Discord chat

References: