Introduction:

Classifying hand written letters is both necessary and important to the study of machine learning. This assignment focus on identifying a pool of hand written letters from the MNIST dataset using the Nearest Neighbor Algorithm.

Procedures:

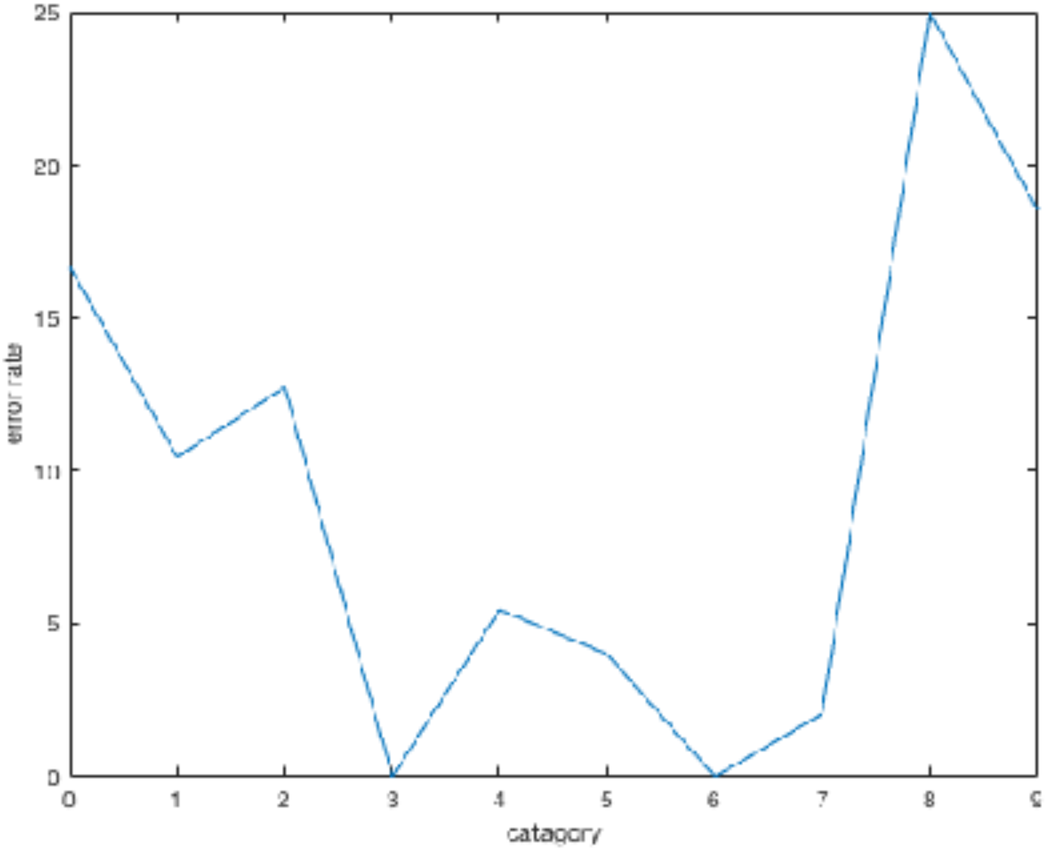
*sidenote: I initially attempted a faster method by averaging all the image with the same truth value into a "trained" model for that specific truth. Then I ran all test images to compare with these 10 "trained" images and yielded about 22% error rate in total. Though this kind of worked, the error rate is too high to be acceptable. I gave up on the attempt to optimize for computing speed and went for the 1-test-compare-all approach, comparing each test image to the entire training set to find a match.

The algorithm simply performed a euclidian distance calculation per pixel between the test and the reference. Each teat image will yield a 5000 element array with distance filled out. We then choose the minimal distance from this list of 5000 elements, then look up the truth of that match in the label_train dataset.

Results:

The algorithm performed as excepted. It made handful of mistakes, which can be characterized in the following manner:

Number 3 and 6 were classified perfectly. While we saw pretty terrible results between 8 and 9. 0 and 2 were also up in the error ally.

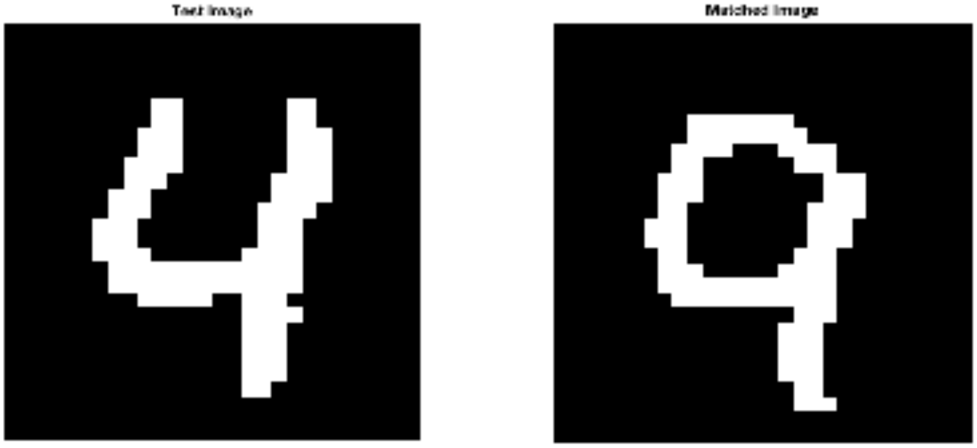


The error rate in the 0-9 order is: (%)

```
   16.6667   10.4478   12.7273         0    5.4545    4.0000         0    2.0408   25.0000   18.5185
```

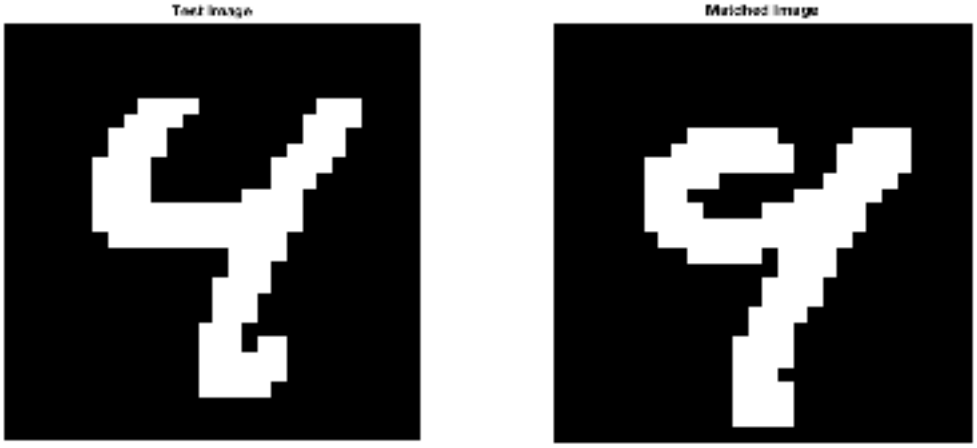
The total error rate of the classifier performed at 9.4%, getting 47 images wrong out of 500.

Below is a sample of 5 misclassified images with comments on why it failed embedded in the images.



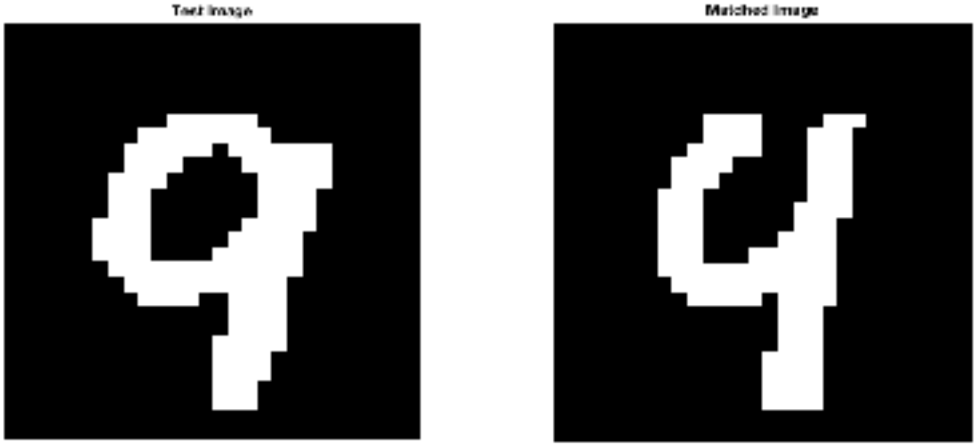
the pair matched with dist 399.507066

the 4 matches the shape of the 9 or,but the gap could be connecting for human even



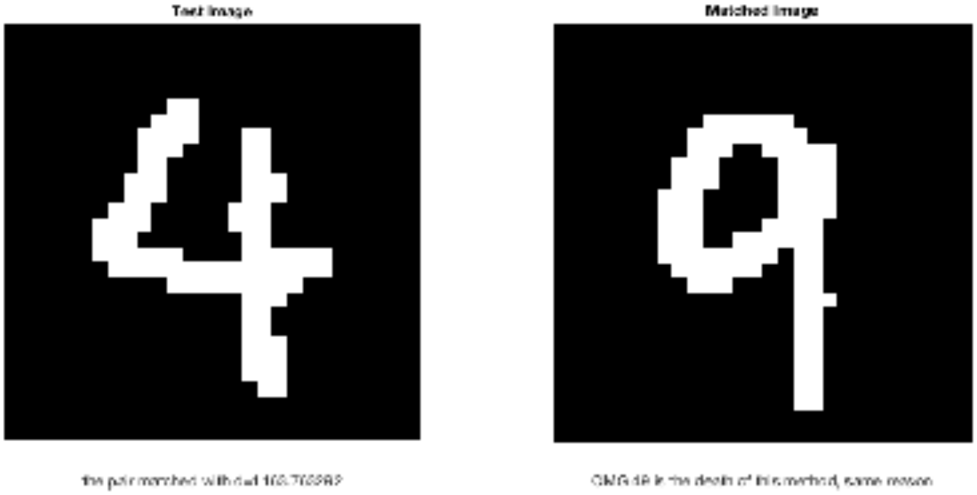
the pair matched with dist 466.441506

Another 49 mess up, same reason, the general shape is very close



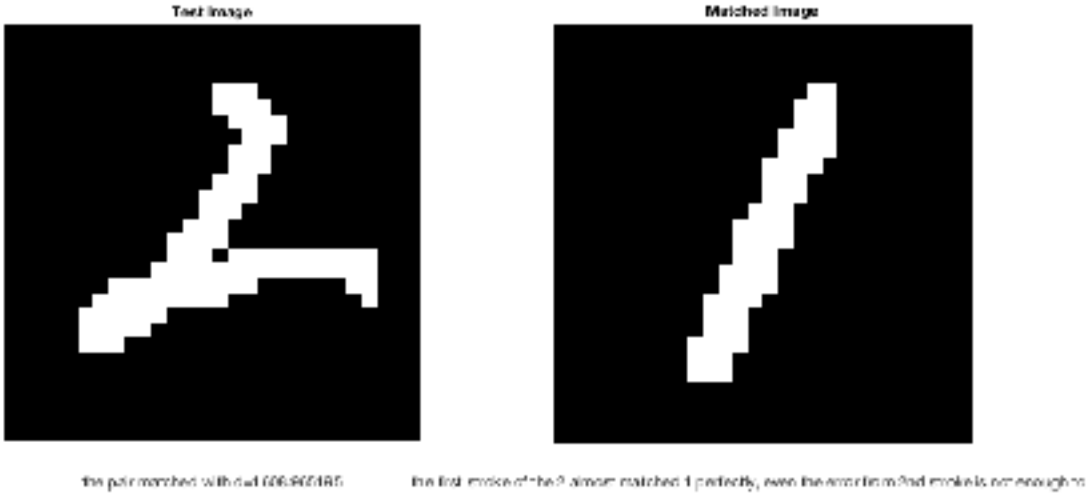
the pair matched with dist 165.333585

yet another one, this one is very bad to human even, just bad writing



the pair matched with dist 103.769982

OMG 4/9 is the death of this method, same reason



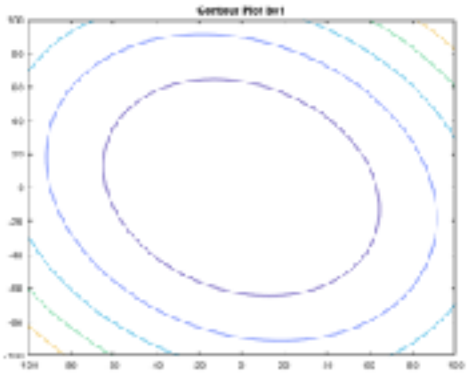
the pair matched with dist 606.665485

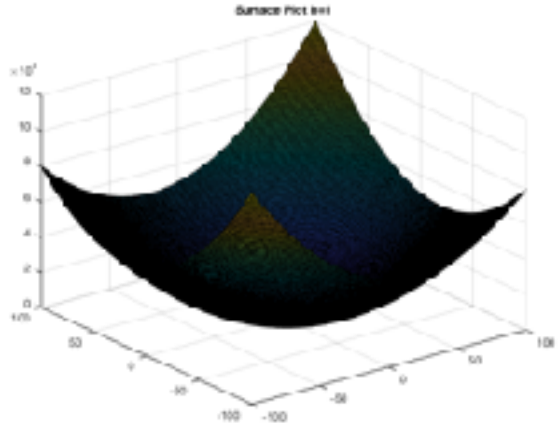
the first stroke of the 2 almost matched 1 perfectly, even the error from 2nd stroke is not enough to

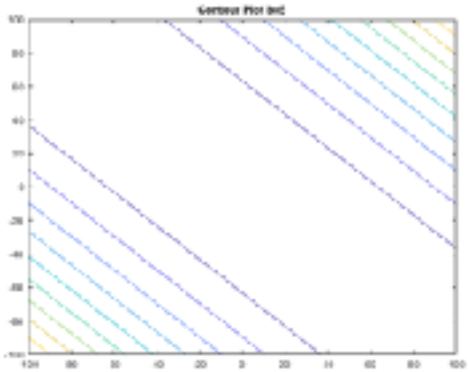
Conclusion:

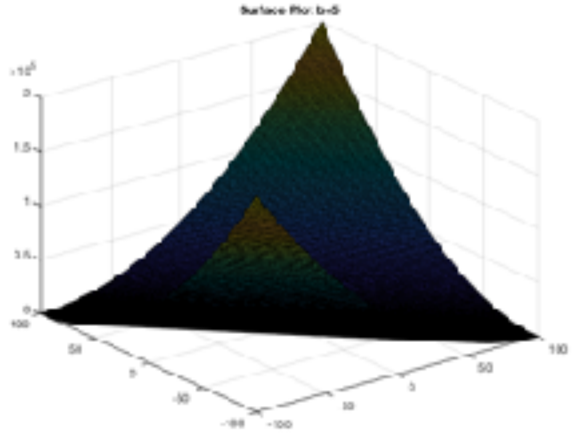
The algorithm performed as excepted. Due to the simplistic nature of this method, a large number of errors were captured. Further implementations of algorithm improvement can help the classifier perform better.

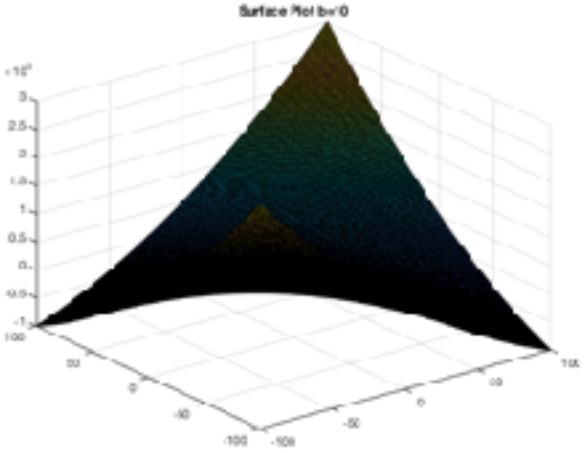
Appendix, Practice Problems:

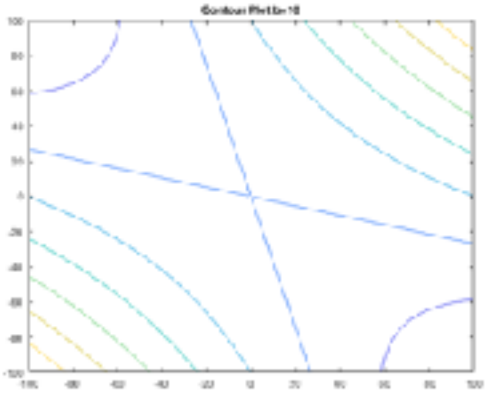


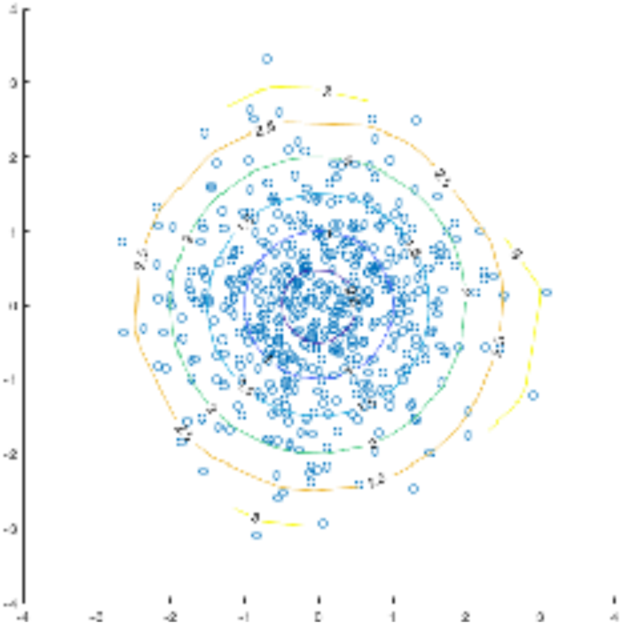


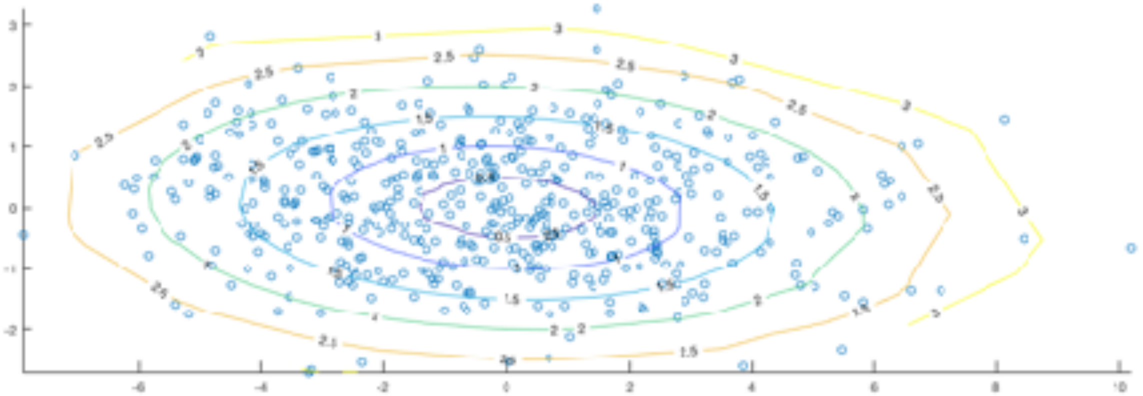


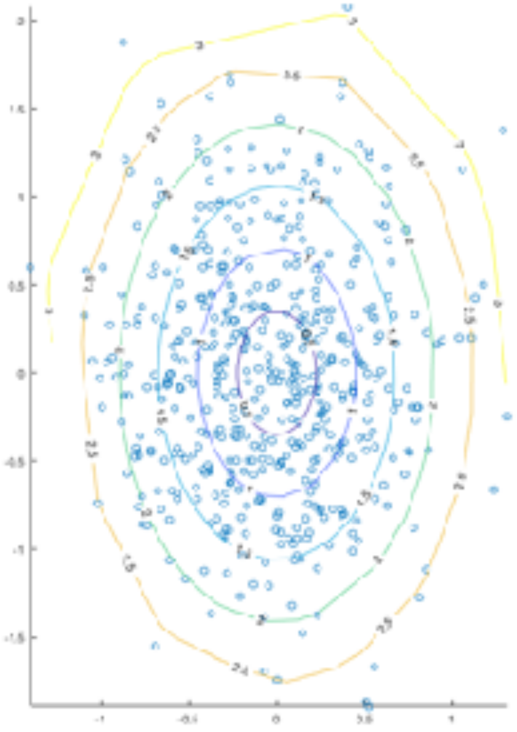







4
3
2
1
0
-1
-2
-3
-4
-4
-3
-2
-1
0
1
2
3
4
3
2.5
2
1.5
1
0.5


3
2
1
0
-1
-2
-6
-4
-2
0
2
4
6
8
10
3
2.5
2
1.5
1
0.5


2
1.5
1
0.5
0
-0.5
-1
-1.5
-1
-0.5
0
0.5
1
3
2.5
2
1.5
1
0.5