Neural networks 1

March 2018

Task 1

	0	1	2	3	4	5	6	7	8	9
0	0.000000	14.449608	9.334556	9.143734	10.769844	7.519296	8.154443	11.864555	9.907902	11.488875
1	14.449608	0.000000	10.125323	11.733233	10.173786	11.118800	10.614700	10.743154	10.086777	9.932094
2	9.334556	10.125323	0.000000	8.178285	7.932541	7.906796	7.331808	8.872531	7.077516	8.887748
3	9.143734	11.733233	8.178285	0.000000	9.087608	6.118750	9.302065	8.922401	7.020425	8.354350
4	10.769844	10.173786	7.932541	9.087608	0.000000	8.001517	8.782233	7.583012	7.380909	6.010408
5	7.519296	11.118800	7.906796	6.118750	8.001517	0.000000	6.698692	9.211954	6.967386	8.258538
6	8.154443	10.614700	7.331808	9.302065	8.782233	6.698692	0.000000	10.888237	8.587222	10.440004
7	11.864555	10.743154	8.872531	8.922401	7.583012	9.211954	10.888237	0.000000	8.467785	5.426474
8	9.907902	10.086777	7.077516	7.020425	7.380909	6.967386	8.587222	8.467785	0.000000	6.401166
9	11.488875	9.932094	8.887748	8.354350	6.010408	8.258538	10.440004	5.426474	6.401166	0.000000

Figure 1: Distances between centers

Centers with the smallest distances are the pairs (7,9), (4,9), (3,5), (5,6), (8,9), (5,6) and (5,8) in the order, with the Euclidean distances circled in the table. The radius values of the clouds are ['15.89','9.48', '14.17', '14.74', '14.53', '14.45', '14.03', '14.91', '13.71', '16.14'] from Cloud0 to Cloud9. As the radius of the clouds are longer than their center distances (except for Cloud2), which means that the center of another cloud will be within the radius of the most clouds, we can already expect that the classification accuracy will not be 100% for the training data.

Task 2

-Euclidean train
The overall prediction accuracy was 0.864. Normalized confusion matrix and its visualization is shown below.

```
Normalized confusion matrix
[[ 0.85   0.     0.     0.     0.006  0.013  0.113  0.     0.019  0.   ]
 [ 0.     1.     0.     0.     0.     0.     0.     0.     0.     0.   ]
 [ 0.015  0.     0.827  0.045  0.045  0.005  0.015  0.02   0.03   0.   ]
 [ 0.     0.     0.015  0.916  0.008  0.023  0.     0.008  0.023  0.008]
 [ 0.     0.066  0.008  0.     0.779  0.     0.025  0.     0.     0.123]
 [ 0.034  0.     0.023  0.034  0.045  0.761  0.034  0.011  0.023  0.034]
 [ 0.066  0.026  0.033  0.     0.013  0.     0.854  0.     0.007  0.   ]
 [ 0.     0.024  0.     0.     0.012  0.012  0.     0.843  0.006  0.102]
 [ 0.007  0.014  0.007  0.069  0.014  0.021  0.007  0.     0.84   0.021]
 [ 0.     0.023  0.     0.008  0.076  0.     0.     0.045  0.     0.848]]
```

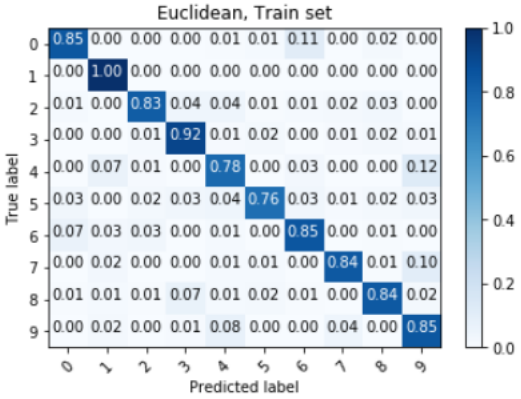


Figure 2: Euclidean Training set

-Euclidean test

The overall prediction accuracy was 0.804. Normalized confusion matrix and its visualization is shown below.

```
Normalized confusion matrix
[[ 0.795  0.     0.013  0.009  0.018  0.009  0.103  0.004  0.045  0.004]
 [ 0.     0.992  0.     0.     0.     0.     0.008  0.     0.     0.   ]
 [ 0.02   0.     0.683  0.059  0.079  0.01   0.     0.02   0.129  0.   ]
 [ 0.038  0.     0.038  0.772  0.013  0.101  0.     0.     0.013  0.025]
 [ 0.012  0.035  0.035  0.     0.802  0.     0.012  0.012  0.     0.093]
 [ 0.055  0.     0.     0.109  0.055  0.691  0.018  0.     0.     0.073]
 [ 0.078  0.     0.022  0.     0.022  0.011  0.867  0.     0.     0.   ]
 [ 0.     0.031  0.016  0.     0.078  0.     0.     0.781  0.     0.094]
 [ 0.033  0.022  0.     0.065  0.033  0.033  0.     0.     0.793  0.022]
 [ 0.     0.057  0.     0.     0.091  0.     0.     0.057  0.023  0.773]]
```

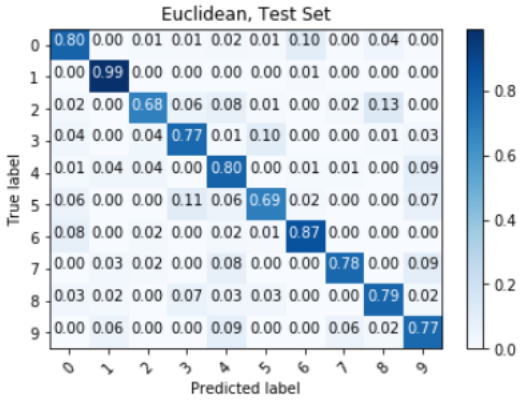


Figure 3: Euclidean Test set

The overall prediction accuracy is 6% higher in the train set. The misclassified pairs are similar in both train and test set- in train set, the pairs (0,6), (4,9) and (7,9) had the classification error around 10%. For the pair (0,6), the distance between the cloud centers were not as close, but the misclassification rate was high. For test set, the most confused pairs were (0,6), (2,8), (3,5), (4,9) and (7,9). The pairs (2,8) and (3,5) had relatively high misclassification rate(10%) only in the test set, whereas for the train set the rate was around mere 3%. This difference might root from the small train set size for the clouds 5 and 8, each with sample size 88 and 144, which are relatively small compared to the sample size of the other clouds. The characteristics of Cloud5 and Cloud8 in the train set might differ from the test set.

As the Euclidean distance measure was used so far to calculate the distance from the center, 3 other distance measures were tried as well- cosine measure and Manhattan measure were used on the test set.

```
Normalized confusion matrix
[[ 0.795  0.     0.009  0.009  0.022  0.009  0.103  0.004  0.045  0.004]
 [ 0.     1.     0.     0.     0.     0.     0.     0.     0.     0.   ]
 [ 0.02   0.02   0.653  0.059  0.089  0.01   0.     0.02   0.129  0.   ]
 [ 0.038  0.     0.025  0.759  0.013  0.114  0.     0.013  0.013  0.025]
 [ 0.012  0.07   0.023  0.     0.802  0.     0.     0.012  0.     0.081]
 [ 0.055  0.     0.     0.145  0.055  0.655  0.018  0.     0.     0.073]
 [ 0.056  0.011  0.011  0.     0.022  0.     0.9    0.     0.     0.   ]
 [ 0.     0.062  0.     0.     0.047  0.     0.     0.797  0.     0.094]
 [ 0.022  0.033  0.     0.087  0.033  0.022  0.     0.011  0.772  0.022]
 [ 0.     0.08   0.     0.     0.091  0.     0.     0.057  0.023  0.75 ]]
```

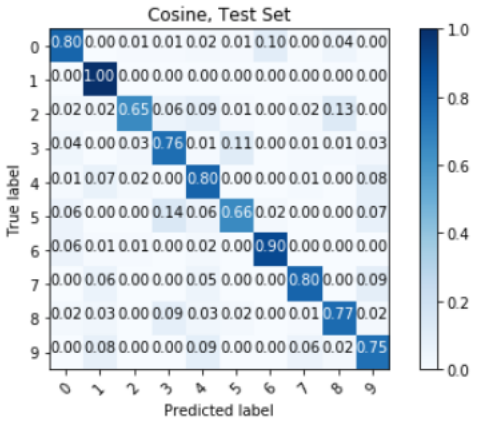


Figure 4: Cosine distance, Test set

```
Normalized confusion matrix
[[ 0.772  0.018  0.     0.013  0.013  0.004  0.121  0.022  0.013  0.022]
 [ 0.     1.     0.     0.     0.     0.     0.     0.     0.     0.   ]
 [ 0.02   0.198  0.465  0.079  0.05   0.     0.02   0.04   0.129  0.   ]
 [ 0.038  0.076  0.     0.785  0.     0.025  0.     0.025  0.013  0.038]
 [ 0.012  0.163  0.     0.     0.535  0.     0.     0.023  0.     0.267]
 [ 0.036  0.055  0.     0.2    0.073  0.4    0.091  0.036  0.036  0.073]
 [ 0.044  0.122  0.     0.     0.     0.     0.833  0.     0.     0.   ]
 [ 0.     0.094  0.     0.     0.016  0.     0.     0.812  0.     0.078]
 [ 0.011  0.174  0.     0.109  0.     0.011  0.     0.022  0.62   0.054]
 [ 0.     0.148  0.     0.     0.034  0.     0.     0.057  0.011  0.75 ]]
```

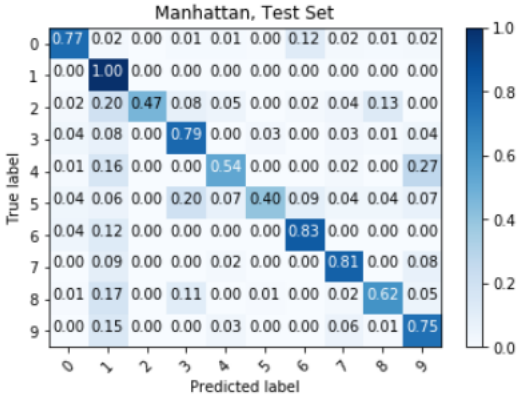


Figure 5: Manhattan distance, Test set

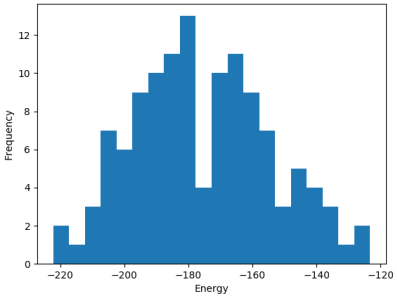
The overall accuracy were 0.799 for Cosine measure and 0.72 for Manhattan distance, out of which the Euclidean measure remained the best.

Task 3

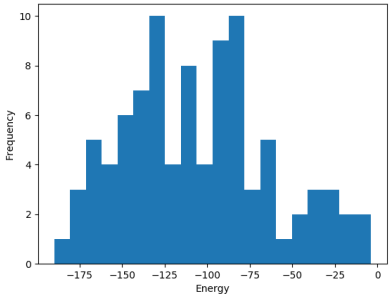
Task three entails developing a feature for discerning between two different digits to our choice. We developed two different features for the digits.

- **Energy:** All pixels in the image are added up into a single value. This value represents the total energy value of the image.
- **XY ratio:** First the image is iterated in a row major fashion. We check if the rows contain any positive values. If they do we remember the first and last occurrence. We then calculate the delta of these row numbers. The image is transposed and the process is repeated. With these two delta‘s we calculate the X:Y ratio.

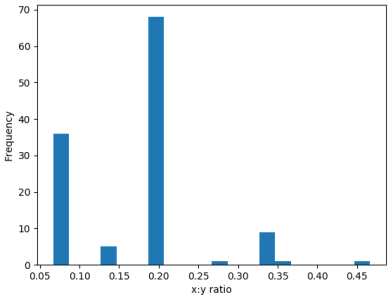
The histograms of both features can be seen in Figure 6. For both features twenty bins were used.



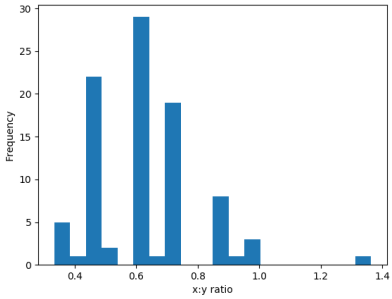
(a) Energy histogram of digit '1'



(b) Energy histogram of digit '8'



(c) Ratio histogram of digit '1'



(d) Ratio histogram of digit '8'

Figure 6: Histogram comparison

The digit 1 occurs 252 times in the training set and 121 times in the test set. The digit 8 occurs 144 times in the training set and 92 times in the test set.

$$P(C) = (121/(121 + 252)) = 0.3244$$

$$P(X|C) = 0.992063492063$$

Here P(X—C) is calculated from the program by checking with what accuracy the classifier predicts a 1 from a dataset of just 1's.

$$P(X) = 0.664141414141$$

P(X) is derived from the amount of times the classifier predicts a digit is a 1 on the training set. We can now calculate the following.

$$P(C|X) = \frac{0.992063492063 * 0.3244}{0.664141414141} = 0.4846$$

Task 4

For the sake of efficiency this part of the assignment was implemented in C++. As activation function of the nodes the standard Sigmoid function was used and the learning rate was set at 0.1. The network only contains 2 layers, The input layer and the output layer. The input layer has 256 + 1 nodes and the output layer 10 nodes. The network is fully connected. After 1000 epochs the network achieves 80 to 90 % accuracy. Further training does not increase the accuracy on the training set by much and possibly increases the hazard of overtraining. In Figure 7 the results of the neural network can be seen. All results are averaged over 10 runs.

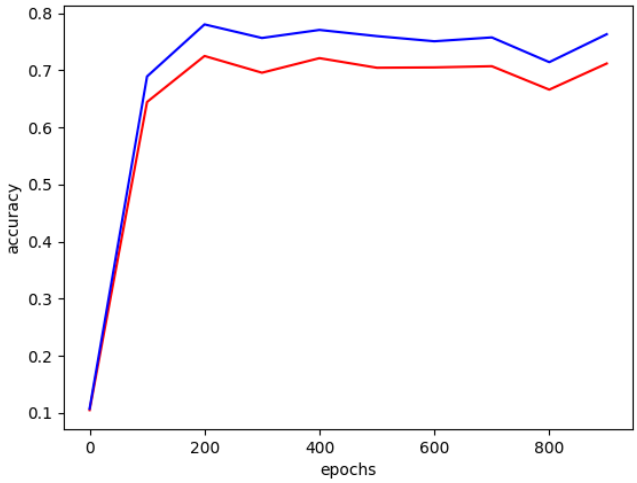


Figure 7: Training set accuracy(red) and Test set accuracy(blue)

An observation on the results is that the test set scores higher than the training set. A possible explanation for this is that the test set contains less outliers compared to the training set. However if the training set and test set are switched the same result can be observed. Another observation is that the network seems to converge after roughly 150 to 200 epochs.

Task 5

Choice of the parameters

1 random weights initiator
uniform[-1,1], uniform[-4,4], normal(0,1), normal(0,4). The two distributions uniform and normal are one of the common choices for the inital weights initiator. As the final updated weights had the values between

(-9,9) in the first trial, uniform (-4,4) and normal (0,4) were tried as 4 is about the middle value of 0 and 9.

2 learning rate(eta)
Values 0.1 and 0.01 are tried. eta = 0.001 was experimented for one time but the convergence rate was too slow so was discarded from the experiment.

3 activation functions
sigmoid function, linear rectifier and hyperbolic tangent functions were tried.

4 Others: iteration numbers of gradient descent algorithm
n = 50000 was chosen as the number of weights updates in the gradient descent algorithm. This number could be experimented as well, but for the sake of the simulation number we did not but n=50000 was a good number to be experimented as for the different parameters differences were shown