# Report (uri-ml-hw-2-f20)

Baheem Ferrell

The MNIST database of handwritten digits contains a number of grayscale images with fixed size (28 x 28 pixels).

Each image is labeled with digits ranging from 0 to 9.

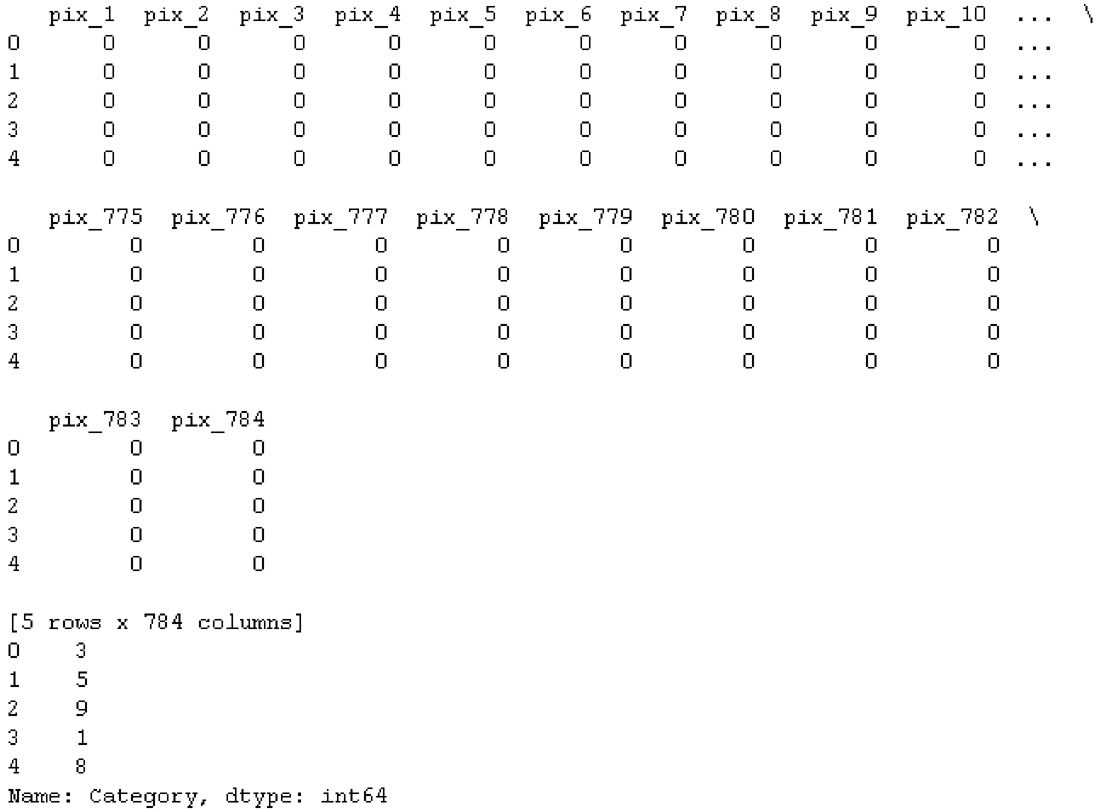
The top-performing models for this dataset are the convolutional neural networks that achieve a classification accuracy of 99% or above with error rate on test dataset as low as 0.26%.

Here, the task is to classify a given image into one of 10 classes (0-9).

In this assignment, it is expected to use a subset of the MNIST dataset to train and validate the best performing model among the classification methods covered in class.

The training dataset contains 40,000 rows of 28x28 pixel data described by *pix_1*, *pix2*, …, *pix_784* fields. It also contains ground truth labels for each of the training samples ranging from 0 to 9 stored in *Category* field.

The following figure shows some of the input data from the first 5 rows in the train dataset.



```
   pix_1  pix_2  pix_3  pix_4  pix_5  pix_6  pix_7  pix_8  pix_9  pix_10  ...  \
0      0      0      0      0      0      0      0      0      0       0  ...
1      0      0      0      0      0      0      0      0      0       0  ...
2      0      0      0      0      0      0      0      0      0       0  ...
3      0      0      0      0      0      0      0      0      0       0  ...
4      0      0      0      0      0      0      0      0      0       0  ...

   pix_775  pix_776  pix_777  pix_778  pix_779  pix_780  pix_781  pix_782  \
0        0        0        0        0        0        0        0        0
1        0        0        0        0        0        0        0        0
2        0        0        0        0        0        0        0        0
3        0        0        0        0        0        0        0        0
4        0        0        0        0        0        0        0        0

   pix_783  pix_784
0        0        0
1        0        0
2        0        0
3        0        0
4        0        0

[5 rows x 784 columns]
0    3
1    5
2    9
3    1
4    8
Name: Category, dtype: int64
```

Figure 1. First 5 rows of input data from training dataset

The same data is plotted in the following figure.

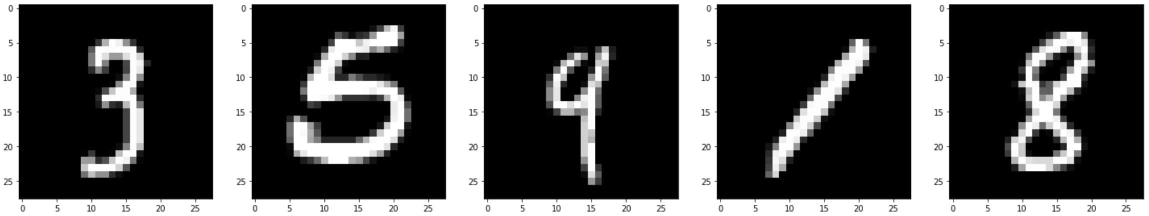


Figure 2. First 5 images of the training dataset

The test dataset contains 10,000 rows of 28x28 pixel data, the labels of which are expected to be predicted using the best performing classification model.

Decision tree and k-nearest neighbor classifiers are trained and compared in this assignment.

Decision trees are one of the simplest models for classification. They are easy to understand and can be combined with other decision techniques. But they often become unstable with a small change in the data and perform worse than several other models.

In this assignment decision trees are trained and validated on the raw training dataset (no preprocessing performed) using various parameters and the best performing model is selected. In order to find the best performing model, a grid search is performed on the training dataset with maximum tree depth ranging from 10 to 20. (The maximum tree depth is increased from 1 until there is no significant improvement on the model accuracy.) In order to avoid overfitting, 5-cross validation is used on the training dataset, where 80% of the data is used for training and the remaining 20% of the data is used for validating the trained model. The best performing model (maximum tree depth = 15) has an accuracy of 0.997 on the training data. Upon submission of the predicted result, it is found that the accuracy of the above model is 0.865 on the test data significantly lower than that on the training data. It was concluded that pixel-by-pixel comparison (all 784 columns are used as features) may not be efficient for recognizing handwritten data.

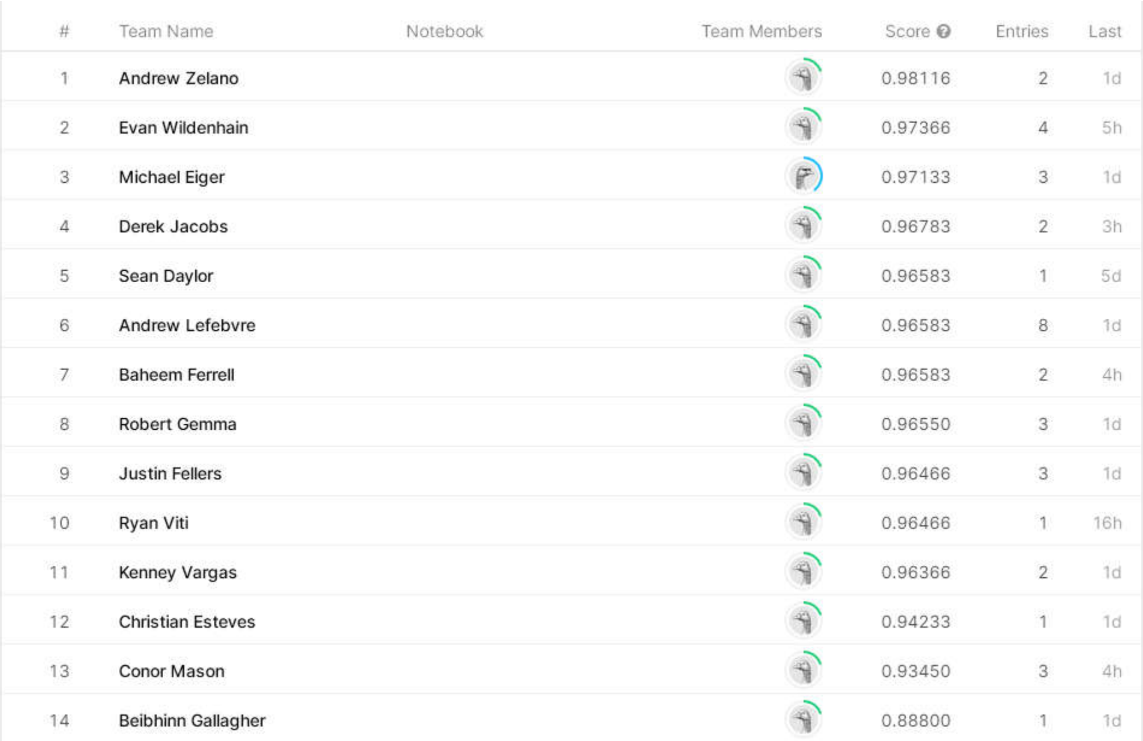
Next, k-nearest neighbor classifier is trained and validate on the raw training dataset (no preprocessing performed) using various parameters and the best performing model is selected. In order to find the best performing model, a grid search is performed on the training dataset with k value ranging from 2 to 11. In order to avoid overfitting, 5-cross validation is used on the training dataset. The best performing model (k = 3) has an accuracy of 0.997 on the training data. Upon submission of the predicted result, it is found that the accuracy of the above model is 0.966.

The following table summarizes the training and test results.

| Model | Decision Tree | k-nn |
|---|---|---|
| Parameter | Tree depth = 15 | k = 3 |
| Training Accuracy | 0.997 | 0.983 |
| Test Accuracy | 0.865 | 0.966 |

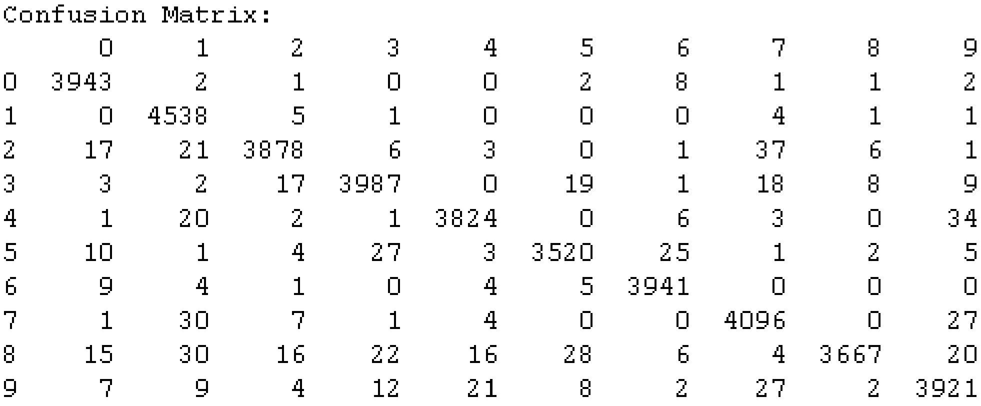
Table 1. Training and test results for decision tree and k-nn classifiers.

The following figure shows the final leaderboard result.



| # | Team Name | Notebook | Team Members | Score | Entries | Last |
|---|---|---|---|---|---|---|
| 1 | Andrew Zelano | | | 0.98116 | 2 | 1d |
| 2 | Evan Wildenhain | | | 0.97366 | 4 | 5h |
| 3 | Michael Eiger | | | 0.97133 | 3 | 1d |
| 4 | Derek Jacobs | | | 0.96783 | 2 | 3h |
| 5 | Sean Daylor | | | 0.96583 | 1 | 5d |
| 6 | Andrew Lefebvre | | | 0.96583 | 8 | 1d |
| 7 | Baheem Ferrell | | | 0.96583 | 2 | 4h |
| 8 | Robert Gemma | | | 0.96550 | 3 | 1d |
| 9 | Justin Fellers | | | 0.96466 | 3 | 1d |
| 10 | Ryan Viti | | | 0.96466 | 1 | 16h |
| 11 | Kenney Vargas | | | 0.96366 | 2 | 1d |
| 12 | Christian Esteves | | | 0.94233 | 1 | 1d |
| 13 | Conor Mason | | | 0.93450 | 3 | 4h |
| 14 | Beibhinn Gallagher | | | 0.88800 | 1 | 1d |

Figure 3. Leaderboard at the final submission



```
Confusion Matrix:
      0     1     2     3     4     5     6     7     8     9
0  3943     2     1     0     0     2     8     1     1     2
1     0  4538     5     1     0     0     0     4     1     1
2    17    21  3878     6     3     0     1    37     6     1
3     3     2    17  3987     0    19     1    18     8     9
4     1    20     2     1  3824     0     6     3     0    34
5    10     1     4    27     3  3520    25     1     2     5
6     9     4     1     0     4     5  3941     0     0     0
7     1    30     7     1     4     0     0  4096     0    27
8    15    30    16    22    16    28     6     4  3667    20
9     7     9     4    12    21     8     2    27     2  3921
```

Figure 4. Confusion matrix for the final model

The final submission result is from the k-nn classifier obtained from grid search with 5-cross validation on the training dataset. The number of neighbors for this model is 3.

Figure 4 illustrates the confusion matrix for the final model. The most frequent classification error occurs between 2 and 7.

This may be due to the fact that handwriting style vary depending on the writer and that similarity between digits may look the same based on the handwriting style of the writer.

On the other hand, the test result indicates that the trained model generalizes well to unseen data.