Machine Learning Assignment 2 – MNIST Data

Introduction

This report explores the MNIST dataset, and different methods to correctly predict the correct digit from 0-9. This problem is useful in a multitude of ways, such as scanning handwritten documents (cheques, mailing addresses, tax reports etc). Investigated characteristics will include basic classification techniques along with machine learning techniques, and an analysis of the accuracy of these methods, and the limitations of the dataset itself.

Part 1 – Data Analysis

The level of data analysis and manipulation needed for this dataset is not as vast as other sets, due to the fact it has been refined to be a useable set without removing anomalies. The data set is simply split between the first 60,000 for training, and last 10,000 for testing. Next, the data is normalised, as the values for each pixel are in a range of 0-255, it is normalised to a range of 0-1. This is done to help with the convergence of the later neural networks, as the process is more stable when using a smaller numerical range.

Part 2 – Methodology

The explored methods can be split into two categories: Simple machine learning algorithms, and CNNs (Convolutional Neural Networks). Typically, CNNs will offer better results, but require much higher computational power than standard machine learning approaches. As for models, ones used in this report are:

Logistic Regression, chosen for its computational simplicity and ease to implement as a base model, albeit potentially struggling with its precision compared to other, more complex models.

Linear Discriminant Analysis (LDA), chosen for its ability with smaller datasets, which albeit not applicable to this MNIST dataset, felt to be a good addition to explore regardless.

Support Vector Machine (SVM), chosen for being incredibly flexible and typically very accurate, however a noted drawback being its very high computational intensity. This model was run with a random 90% of the data removed, and using a linear kernel, to allow it to run in a reasonable time.

Decision Tree, chosen for also being very simple, but does not assume linearity in the same way Logistic Regression and LDA do. However, it is prone to overfitting, which is a common problem with MNIST and one that will be explored more later.

LeNet, a very early CNN, chosen here for its simplicity and the fact it was designed primary for digit recognition, which should make it very effective for the MNIST dataset, albeit a very outdated CNN.

A Sequential CNN, in this case using Keras, is very simple and flexible to use, and is more useable for a wider range of problems then LeNet, but may be overfit for a task like MNIST.

For evaluation of these models, a combination of factors was used. First, for each model a confusion matrix was created. This is a graph of predicted digit on the X axis, and true digit on the Y axis, in a heatmap form. This shows not only the accuracy of the model visually, but also its false positive rates and which specific digits it struggles with most. Each model was also evaluated for its precision, recall and F1 score for each digit, and an overall score for the whole model. Precision score represents true positives, being the True Positive rate, divided by the True + False positive rate. A higher precision score means the model produces more relevant results. Recall score represents sensitivity, being the True Positive rate, divided by the True Positive + False Negative rate. A higher recall score means the model misses less positive, but this can come with more false positives, hence why we combine both precision and recall into a third metric, F1 Score. F1 score is the harmonic mean of precision and recall, aiming to balance both metrics, and will be the main numeric metric used for comparing models.

A third method of evaluation was considered, ROC curves, however (see fig1), these graphs are less applicable for this task due to the high accuracy of the models. Typically, an ROC curve shows true postive rate against false positive rate, with a more accurate model being closer to the top left. As visible in figure 1, these models are too accurate for useful visualtion using ROC curves. The reasons behind this will be explored more in the results and analaysis section of this report.

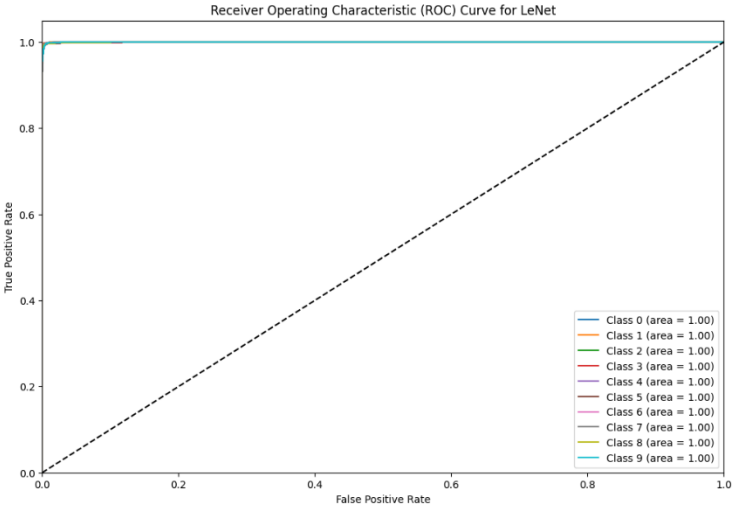


Figure 1, a ROC curve for LeNet.

Other models were considered (AlexNet and ResNet), most other models require datasets of a different format (larger sizes than 28x28 for example), so they were excluded from this report. They also weren't deemed necessary once the accuracy of the two used CNNs was established,

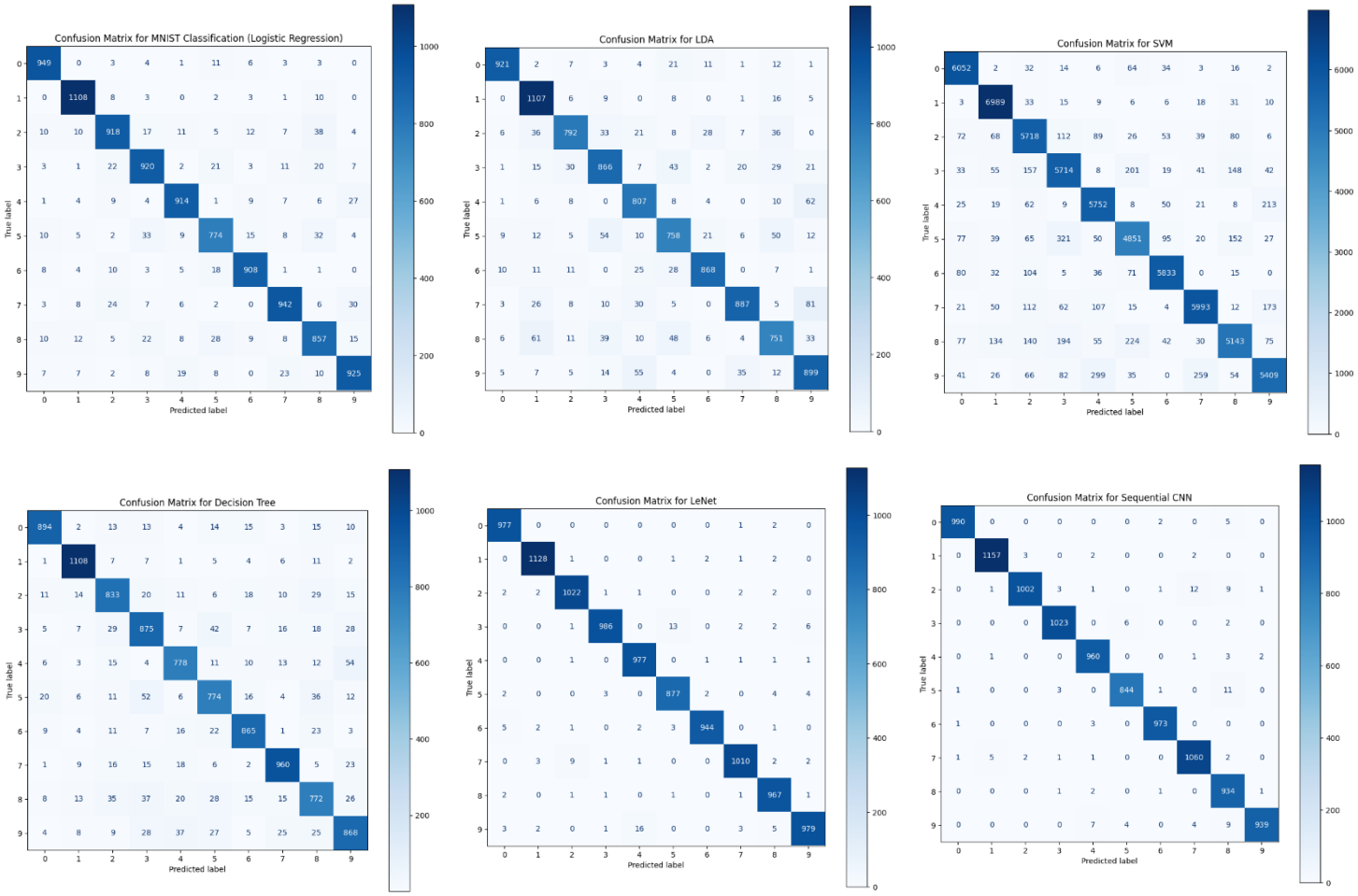
Part 3 – Results, Analysis and Discussion

Comparing each model, we get the following results and graphs:

-	Precision Score	Recall Score	F1 Score	Run Time (Seconds)
Logistic	0.92	0.92	0.92	22.5
LDA	0.87	0.86	0.86	11.8
SVM	0.91	0.91	0.91	$54.0_{(1)}$
Decision Tree	0.87	0.87	0.87	19.1
LeNet	0.99	0.99	0.99	23.7
Sequential	0.99	0.99	0.99	69.7

A thing to note with these results, the SVM run time (1) is 54.0 seconds, however this is for only 10% of the dataset. It would take considerably longer to run for 10x the data set, but would also be likely more accurate (SVM does only use a subset, so it tends to be more accurate than others with a reduced sample size, so most likely would not improve to an accuracy worth the increased time with a larger sample size).

The confusion matrix heatmaps for these 6 models are as follows:



From left to right, top to bottom. Confusion matrix heatmaps for Logistic, LDA, SVM, Decision Tree, LeNet, Sequential.

These graphs concur the numerical statistics of the table before in a visual manner. It is clear the best model for a combination of time and accuracy is LeNet. This makes sense given LeNet was designed with MNIST dataset in mind, but also consolidates the idea the more complex models will be more accurate. This is shown as the sequential CNN is very similar in accuracy, but takes 3x as long to run.

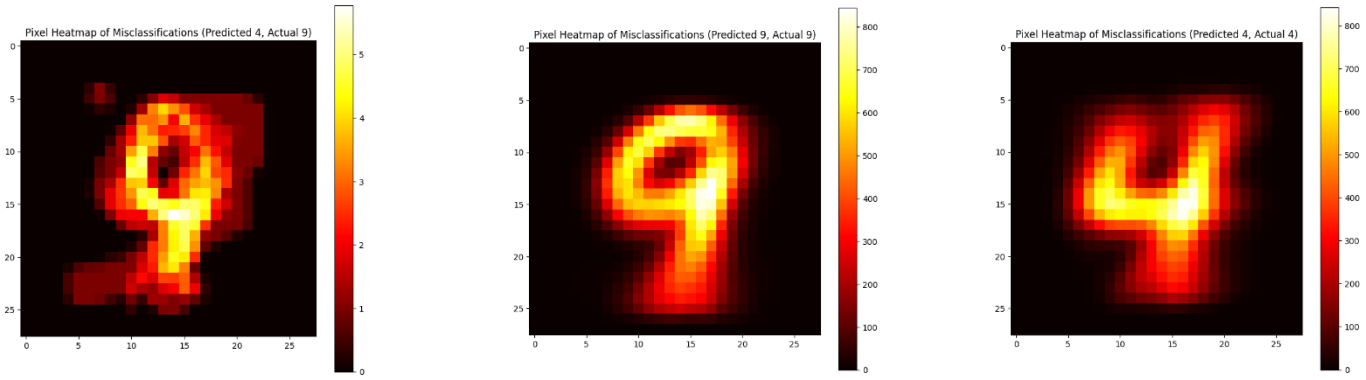
LDA, Decision Tree and Logistic Regression all run quicker than the more complex models, at the cost of accuracy, with the best of these being Logistic Regression, providing a 0.92 F1 score. The only model which appears to be without benefit is SVM, due to its long run time and poor accuracy in comparison to other models. This may be for a multitude of reasons, most likely the high dimensionality of the MNIST data set, having 784 data points (28x28), which massively increases the computational requirements for SVM, reducing how large of a sample can be used.

Looking at both CNN models, but mainly LeNet (quicker with same accuracy), it becomes clear one of two things is happening. Either the dataset itself is too easy, or the model is overtrained to this dataset. It is most likely a combination of both. The MNIST dataset is regarded as too simple for modern computational power and modern CNN models, with it being possible to return the correct number out of a pair of digits using a single pixel[1] . It is also plausible the models are overtrained, and could be examined against another set of 28x28 digits to see the models have been overtrained on this data set. Further tests could be done using completely different MNIST sets, such as Fashion-MNIST or Kuzushiji-MNIST, both of which are considered more difficult for modern machine learning algorithms.

[1] Research and code for the 1 pixel comparison found via https://lucaf.eu/2021/04/05/mnist-pairwise-one-pixel.html

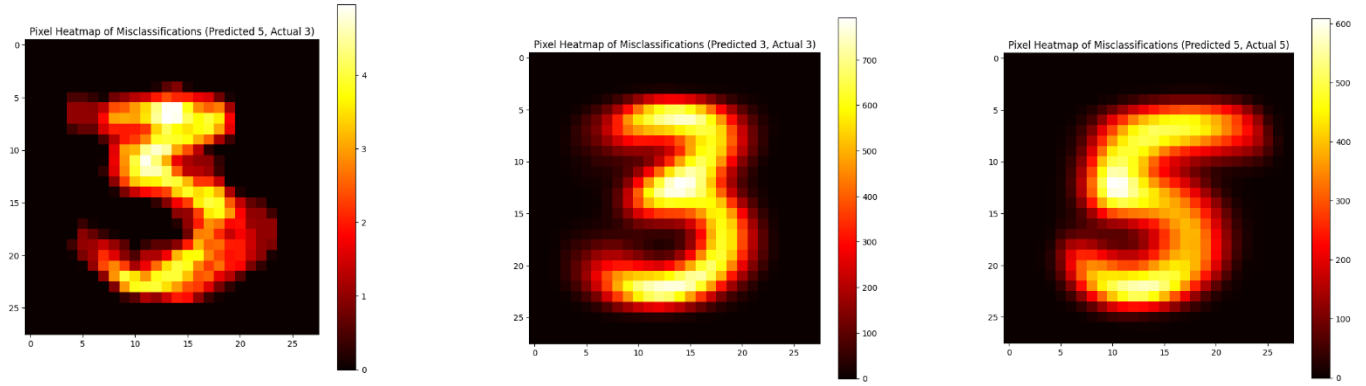
This high accuracy returns to the point made earlier, about how ROC curves simply do not show the differences between the most accurate models well. Looking at specific number pairs, the pairs the LeNet model performs worst on are:

Predicting 4, actual answer 9 (but not the other way round).



From left to right, heatmap of predicted 4 with real digit 9, predicted 9 with real digit 9, predicted 4 with real digit 4.

Predicting 5, actual answer 3 (but not the other way round).



From left to right, heatmap of predicted 5 with real digit 3, predicted 3 with real digit 3, predicted 5 with real digit 5.

These were the only pairs of numbers in which the LeNet module failed more than 10 times (16 and 13 respectively), with the model having an abundance of pairs it never got incorrect (7 against 5, 1 against 0, to name a few).

Looking at these erroneous pairs (4 instead of 9, 5 instead of 3) on different models, we see no real crossover. Sequential CNN struggles by most predicting 7 instead of 2 and 8 instead of 5 being its only pairs with more than 10 errors. This suggests that these errors are less systematic and more random deviation we can expect when dealing with an incredibly large testing set. For example, training using a slightly different 60,000 may return a completely different pair of two digits which they struggle most with. It appears the only concrete trend across all models is the ability to very accurately predict 0s and 1s, with all models doing well on these two digits. This makes sense given the uniqueness of the shape of the digits 0 and 1, and their lack of complex shape, compared to other digits.

Part 4 – Conclusion

To conclude, this report shows that LeNet is the best tested model for predicting the MNIST dataset, but this does not guarantee its sustained success on a different data set, as it may potentially be overtrained, or just performing well due to the simplicity of MNIST. However, it does clearly show that CNN models are considerably more accurate than normal, simple machine learning algorithms, with both of those tested outperforming the simple models considerably. Considerations going forward could be to test more CNN models, or on a more complex dataset to determine which CNN model is truly stronger for digit prediction.

APPENDIX: STATEMENT OF AI USAGE

I have used AI in this report to help understand and debug code issues. Some more complex programming was written initially with GPT but modified, checked and thoroughly understood by me before implementing into the report. All text in this report is written solely by me without the use of AI. AI has simply been used to enhance my understanding of the topics in this report.