Machine Learning

MNIST Dataset

Matthew Ahern-Hailu

EC Utbildning

Data Scientist Program

2025-03

Abstract

This report looks at how two machine learning models (logistic regression and random forest) classify hand written digits from the MNIST dataset. The work uses a machine learning workflow, whereby I did preprocessing, model training, validation and final analysis on some test data. Based on how these models performed, the random forest model was stronger than the logistic regression model and was used for the finakl model. My report also includes confusion matrixes and a classification table to visualise the findings of my work and to help with analysis.

Innehållsförteckning

1 Inledning

This report will explain, discuss and analyse how two Machine Learning models were used on the MNIST dataset to see which model was more effective in determining handwritten digits. The models used were a Logistic Regression and Random Forrest model. We will look at how each model performed on a set a of validation data, and discuss which model was determined to be more effective and why. This was all done using Python.

1.1 Underrubrik

Text är skriven på formatet Calibri med textstorlek 11.

2 Teori

The following are some of the core theories that came up during my work.

2.1 Logistic Regression

Generally logistic regression is an algorithm which predicts output of a categorical dependent variable, and gives an output in a yes or no, 1 or 0, true or false in a probabilistic value. That is to say where it predicts. Since out dataset has 10 different classes (0-9), we needed to use Multinominal logistic regression. (understanding-logistic-regression, n.d.)

The equation for Logistic Regression:

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(w \cdot x + b)}}$$

2.1.1 Multinomial logistic regression

Whereas binominal regression can only classify between two outputs, Multinomial considers all classes simultaneously during optimization (understanding-logistic-regression, n.d.). It calculates the probability of each class using a the softmax function, so that all probabilties add up to 1. After that the model will assign any input to the class with the probility that was highest. (plot logistic multinominal, n.d.)

$$P(y = k \mid x) = \frac{e^{w_k \cdot x + b_k}}{\sum_{j=1}^{K} e^{w_j \cdot x + b_j}}$$

2.2 Random Forrest

Random Forest is a learning method that makes decision trees to classify input, it combines these decsion trees to make predictions accurately. For example, it would as a series of questions of the input, if we are guessing who in my family did this report it would first ask, is the person <= 5 years old? If yes we move to the next question, boy or girl etc until it reaches a prediction from the possible classes. (random forest regression in python, n.d.)

3 Metod

Firstly, I loaded in the MNIST dataset into python, then processed the data into a 1d array, converted the lables into integers and used StandardScaler to normalise the pixel values.

Then I split the data set into Training (60%), Validation (15%) and test (15%), with a plan of using the training data to learn the models, validiation to be able to compare which is best and test to analyse the final model i chose. The models I used were a logistic regression model and random forest. I chose first the logistic regression because its simple, easy to interpret its results and serves as a good base model, I chose the random forest model because its more powerful and its nature suits itself better to working with more complex data such as the MNIST dataset.

After using my models on the validation set I made two confusion matrixes to compare the results, after testing the final chosen model I made a table from the classification report and a final confusion matrix to evaluate the models effectiveness.

I also included a grid with some random samples of digits that were incorrectly predicted so we could actually have a little look where things were wrong.

4 Resultat och Diskussion

After running both models, it was immediately apparent that the random forest model was more accurate in predicting digits than the logistic regression model. I got an accuracy percentage of Logistic Regression: 0.9151 and Random Forest: 0.9668 on the validation data. Thats fairly simple, but we can look at a confusion matrix to better see which is best and where they both have had issues:

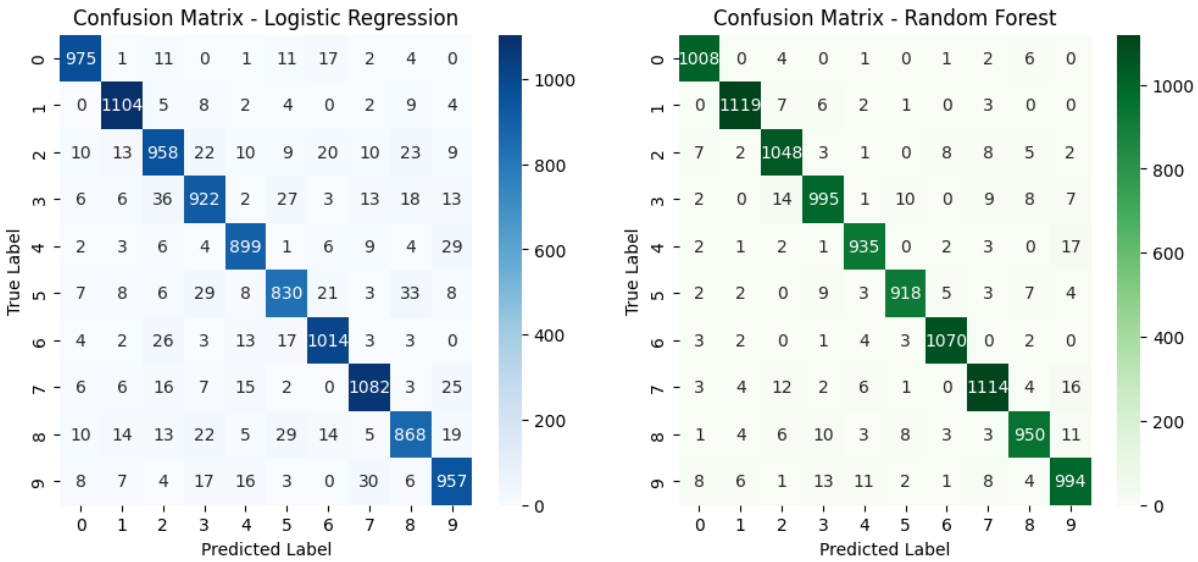

Figure 1: Confusion Matrix which shows how both models scored on the validation set. Left in Blue we have the Logistic Regression model, and on right in Green we have the Random Forest.

As we can see the Logistic Regression model was less accurate and its errors were more random, there doesnt seem to be a pattern we can identify in which numbers it had issues with and why, whereas with the random forest model, we get a better view of its problems. We know it was more accurate and we can see there are a few little patterns, like getting 3's mixed up for 8's and 9's for 4's.

I quickly chose to continue with the random forest model and we can see its results in the confusion matrix of figure 2, as we can see, much similar in pattern to the validiation set.

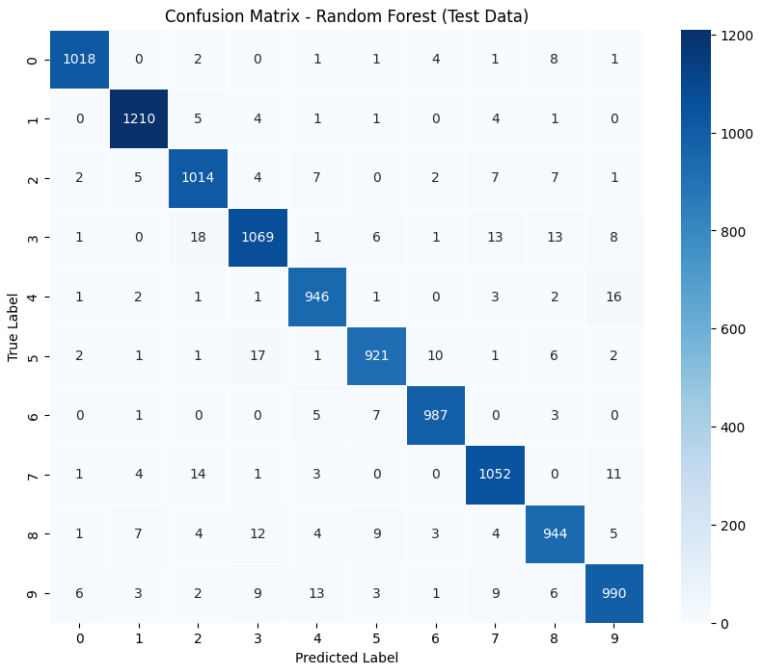

Figure 2: A confusion matrix showing how the Random Forest model worked on the Test data. The higher the intensity of blue, the higher the number of predictions.

Classification Report				
Class (Number)	**Precision**	**Recall**	**F1-Score**	**Support**
0	0.99	0.98	0.98	1036
1	0.98	0.99	0.98	1226
2	0.96	0.97	0.96	1049
3	0.96	0.95	0.95	1130
4	0.96	0.97	0.97	973
5	0.97	0.96	0.96	962
6	0.98	0.98	0.98	1003
7	0.96	0.97	0.97	1086
8	0.95	0.95	0.95	993
9	0.96	0.95	0.95	1042
Accuracy			0.97	10500
Macro Avg	0.97	0.97	0.97	10500
Weighted Avg	0.97	0.97	0.97	10500

Table 1: Classification report which shows how accurate the model was in predictions, how often it caught each digit and how many times each digit appeared (support).

Just to check, we can also look a the classification report in table 1. The model did pretty well over all the classes 0-9, with all scores between 95% and 99%. It was best at predicting 0, 1 and 6 and as we can confirm from looking at figure 2, it did slightly worse with 3, 8 and 9.

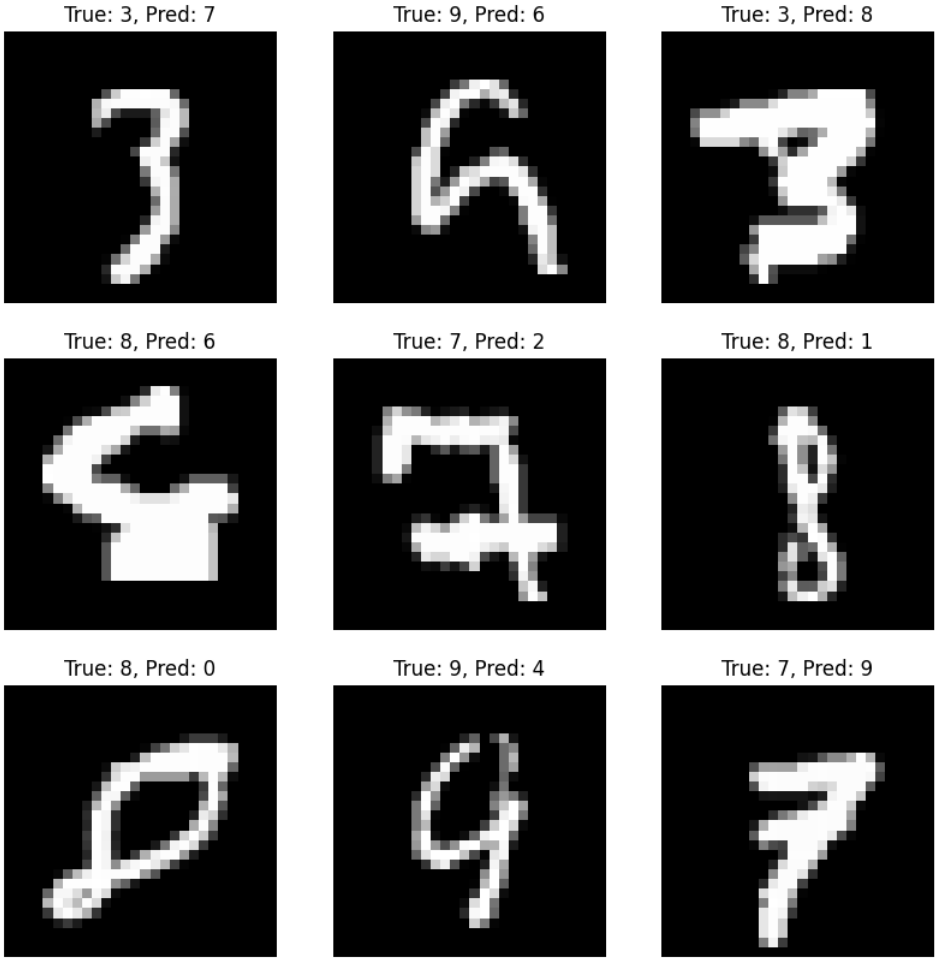

Figure 3: A grid of randomly selected digits from the test data which were incorrectly classificted. Above each image, True is the actual digit and Pred is what the model thought it was.

We can see in the above figure 3, and random selection of digits my model predicted incorrectly, as we see here we have a couple of examples where the model predicted 4 when it was a 9, or 8 when it was 3. They do look a little similar sometimes. I was curious about this mistake, and asked my 5 year old twins to look at figure 3, as well as a few others and tell me what they thought the number was. They both did a little better than the machine, but also had trouble with 9s and 4s. Perhaps theres an issue with how some of the digits are drawn as opposed to the model?

5 Slutsatser

This report used two different machine learning models, to predict digits from the MNIST dataset. The models chosen were logistic regression and random forest. The latter performed better during validation and was chosen to be used on the test data for further evaluation, it got 97% accuracy on the test data and showed strong results across all digit classes.

For further research I would use more models, and perhaps skip the logistic regression. As mentioned in my analysis, when looking at some of the digits that were predicted wrong with my kids, I would like to research whether errors made are based on the drawing of the digits or the method of the model itself.

6 Teoretiska frågor

1. Kalle delar upp sin data i ”Träning”, ”Validering” och ”Test”, vad används respektive del för?

Its splitting up the data set for your models, the first, Träning is used to teach the model, then the validering is used for tuning the model and the test is the final set we use to analyse the model.

2. Julia delar upp sin data i träning och test. På träningsdatan så tränar hon tre modeller; ”Linjär Regression”, ”Lasso regression” och en ”Random Forest modell”. Hur skall hon välja vilken av de tre modellerna hon skall fortsätta använda när hon inte skapat ett explicit ”validerings dataset”?

She could do some cross validation based on the data from the training set and use that to see which model is most useful.

3. Vad är ”regressionsproblem? Kan du ge några exempel på modeller som används och potentiella tillämpningsområden?

Thats when we need to predict a numerical/continous value from data we have, so we could use when analysis bowling in Cricket, if we wanted to predict how much spin or seam movement we'd get on the ball. Some different models are linear regression or lasso regression.

4. Hur kan du tolka RMSE och vad används det till: $\boldsymbol{RMSE = \sqrt{\sum_i (y_i - \hat{y}_i)^2}}$

Root mean square error is a formula we use to measure how accurate our models predict values, vs the actual values. Part of the formula shows this: ” yi-ŷi” where yi is the actual value,minus ŷi is the predicted, the formula takes the sum of all these squared values, and the average and find the root. For example, if we are trying to predicy runs a batsman will score in Cricket and we calculate a RMSE of 12, we could say our model predicts the amount of runs within about 12 runs of the actual value.

5. Vad är ”klassificieringsproblem? Kan du ge några exempel på modeller som används och potentiella tillämpningsområden? Vad är en ”Confusion Matrix”?

This is when we're workling with predictions for categorical variables, like in my assignment, we want to predict discrete categories, so as I used Linear Regression or Random forrest are examples of models, and we could use to predict image data. A confusion matrix is a table which shows the perfomance of a models predictions by plotting what the model predicted by what the actual values were. You can these in my report figures 1 & 2.

6. Vad är K-means modellen för något? Ge ett exempel på vad det kan tillämpas på.

Thats an algorithm for unsupervised mnachine learning which groups data into clusters. It could be used when analysing customers buying behaviour.

7. Förklara (gärna med ett exempel): Ordinal encoding, one-hot encoding, dummy variable encoding. Se mappen ”l8” på GitHub om du behöver repetition.

Ordinal encoding gives numerical values into ordered categories, like 1=poor, 2=average, 3=rich. Onehot encoding makes binary columns for categories, so you could have 3 categories say short, medium, long and then wiothin 1 and 0 within, which tells us whether or not that column is or isnt in the category. Dummy variabel encoding is similar to one hot, but it drops categories by inference, so if we had the same three, short, medium and tall, and we have a 0 in both short and medium, we infer that the category must be tall.

8. Göran påstår att datan antingen är ”ordinal” eller ”nominal”. Julia säger att detta måste tolkas. Hon ger ett exempel med att färger såsom {röd, grön, blå} generellt sett inte har någon inbördes ordning (nominal) men om du har en röd skjorta så är du vackrast på festen (ordinal) – vem har rätt?

Julia, the context is important because we can assign value to nominal data, like she said, if red becomes nice then its better than others, then its ordinal.

9. Kolla följande video om Streamlit: https://www.youtube.com/watch?v=ggDa RzPP7A&list=PLgzaMbMPEHEx9Als3F3sKKXexWnyEKH45&index=12 Och besvara följande fråga: - Vad är Streamlit för något och vad kan det användas till?

Its a python library for building and sharing data appplications. You can use it to make your reports or dashboards to share them with others. Its a bit like Jupyter but better if you wanted to share with someone who isnt necessarily used to python code, looks prettier too.

7 Självutvärdering

1. Utmaningar du haft under arbetet samt hur du hanterat dem.

   I feel like ML is such a huge field that I often felt like I was barely able to swim. When looking for help it was so easy to follow the wrong track because I'm not experienced. I'm not sure I could say I overcame this challenge, because this was definetly the most challenging so far, but i just tried to keep my report simple and stick to what i feel comfortable explaining.

2. Vilket betyg du anser att du skall ha och varför.

   G, It's been a hard course, both in terms of material and balancing my own family life, so I just want to make sure I've done enough to pass.

3. Något du vill lyfta fram till Antonio?

   No, just looking forward to working with R again, a language I've used a lot in the past.

Källförteckning

https://www.geeksforgeeks.org/understanding-logistic-regression/

https://scikit-learn.org/stable/auto_examples/linear_model/plot_logistic_multinomial.html

https://www.geeksforgeeks.org/random-forest-regression-in-python/