Machine Learning Task Report

September 3, 2019

The two algorithms logistic regression and K Nearest Neighbours have been implemented for a handwritten digit recognition task using MNIST dataset. The training data has been divided into training and development sets, with respectively 54000 and 6000 data points. Test set consists of 10000 samples.

1 Logistic Regression

In order to increase computational speed the solver has been set to "lbfgs" and the number of iterations has been increased to 150 for better convergence. An accuracy of 92.24% is reached in a relatively short computation time. Other performance measures are also calculated using the `precision_recall_fscore_support` method in sklearn (all 0.922). The confusion matrice is built using the `confusion_matrix` method.

n(iterations)	50	100	150	200	250
acc.	91.03%	91.23%	91.27%	91.00%	90.92%

Table 1: Accuracy of logistic regression models with n iterations

$$N = \begin{bmatrix} 957 & 0 & 2 & 3 & 0 & 5 & 6 & 4 & 2 & 1 \\ 0 & 1117 & 4 & 2 & 0 & 1 & 3 & 1 & 6 & 1 \\ 5 & 13 & 915 & 19 & 9 & 4 & 12 & 10 & 43 & 2 \\ 5 & 0 & 15 & 920 & 3 & 24 & 3 & 10 & 21 & 9 \\ 1 & 1 & 8 & 1 & 915 & 0 & 8 & 6 & 13 & 29 \\ 12 & 3 & 3 & 36 & 12 & 770 & 13 & 5 & 31 & 7 \\ 9 & 2 & 8 & 2 & 6 & 19 & 908 & 2 & 2 & 0 \\ 1 & 6 & 20 & 10 & 5 & 1 & 0 & 946 & 6 & 33 \\ 8 & 11 & 7 & 26 & 7 & 26 & 11 & 11 & 854 & 13 \\ 9 & 6 & 0 & 9 & 22 & 5 & 0 & 25 & 11 & 922 \end{bmatrix}$$

Table 2: confusion matrix of the logistic regression model with 150 iterations

2 KNN

Models have been trained using different numbers of neighbours from 1 to 20 with a step length of 2. The accuracy of each of these models on the development set is listed below. As it can be seen, the model with three neighbours has gained the highest accuracy. This model has been used to to recognise the digits in the test set, which resulted in an accuracy of 96.97%. Precision, recall and F-score are 0.9697 and the confusion matrix is shown below.

K	1	3	5	7	9	11	13	15	17	19
acc.	96.63%	96.69%	96.47%	96.23%	96.03%	95.97%	95.92%	95.70%	95.68%	95.50%

Table 3: Accuracy of KNN models with K neighbours

$$M = \begin{bmatrix} 974 & 1 & 1 & 0 & 0 & 0 & 2 & 1 & 1 & 0 \\ 0 & 1133 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 11 & 10 & 992 & 3 & 0 & 0 & 0 & 13 & 3 & 0 \\ 0 & 1 & 3 & 977 & 1 & 13 & 1 & 7 & 2 & 5 \\ 2 & 7 & 0 & 0 & 949 & 0 & 4 & 2 & 0 & 18 \\ 6 & 1 & 0 & 10 & 2 & 861 & 4 & 1 & 3 & 4 \\ 5 & 3 & 0 & 0 & 4 & 3 & 943 & 0 & 0 & 0 \\ 0 & 23 & 4 & 0 & 3 & 0 & 0 & 989 & 0 & 9 \\ 8 & 2 & 5 & 15 & 7 & 11 & 3 & 4 & 915 & 4 \\ 6 & 5 & 2 & 7 & 10 & 2 & 1 & 10 & 2 & 964 \end{bmatrix}$$

Table 4: confusion matrix of the KNN model

3 conclusion

Comparing the two algorithms, logistic regression with limited memory BFGS optimisation algorithm works much faster than KNN but KNN yields higher accuracy on this dataset. Both algorithms have difficulty recognising number 7, since in hand-written digits it could look like 1 or 2. Numbers 8 and 9 also cause problems for both algorithms according to the confusion matrices.