Supervised Learning Algorithms Comparison with MNIST-Data

Introduction

The aim of this project is to compare the power of different machine-learning algorithms, when applied on the MNIST-Data. The MNIST database (Modified National Institute of Standards and Technology database) is a large database of handwritten digits that is commonly used for training various image processing systems. More information about MNIST can be found here:

https://en.wikipedia.org/wiki/MNIST_database

For the algorithms Scikit Learn was used, as it offers all the needed tools necessary for this analysis.

Data acquisition and cleaning

Data source

The data source is – of course – the MNIST-dataset. Here a version already prepared for data analysis was used.

Cleaning

No cleaning was necessary, as the version of the dataset is already ready to use. The data is provided in form of four files: two files containing image data and lables for training, two files containing image data and labels for testing.

For better efficiency all image data was divided by 255, to transform the numbers between 0 and 255 to numbers between 0 and 1, as the algorithms are optimized for number range.

Further the image data had to be reshaped. Every image was in the format of an 28 x 28-array. The algorithms don‘t understand arrays, they need flat data. So the images for flattened to a size of 28 x 28 = 784 when computing the models.

Algorithms discussion

The following algorithms were applied on the data: Logistic Regression, KNN, Decision tree, Naive Bayes and SVM (with and without kernel).

First a sample of only 1000 images were used, to get a first impression if everything works and to decide which hyperparameters to use for the full dataset, as the full dataset could come along with long computing time. After this test run, the full dataset was used, printing out the time of computing and the score.

The following results were obtained:

Logistic Regression

Test run:

with solver-hyperparameter „sag“: score 0.848
with solver-hyperparameter „saga“: score 0.847
with solver-hyperparameter„newton-cg“: score 0.843

The results were very similar, for the full-dataset-run the „sag“-solver-hyperparameter was chosen.

The result: score 0.92, computing time 4.51 minutes

KNN

Full-dataset-run: score 0.9665, computing time 9.12 minutes

Decision tree

Full-dataset-run: score 0.8878, computing time 13.6 seconds

Naive Bayes

Full-dataset-run: score 0.5558, computing time 5.38 seconds

SVM

Without kernel, full-dataset-run: score 0.9404, computing time 7.04 minutes

With kernel, a pipeline was built to test different hyperparameters (on the sample dataset). With these parameters set, the results of the full-dataset-run were: score 0.9833, computing time 6.01 minutes.

Conclusion

The SVM-model with kernel delivered the best score, but took quite long (6 minutes). If time is not an issue, this is the best algorithm.

If computing time matters, the decision tree seems to be a good option. The score drops to a still reasonable 0.8878, but it took only 13.6 seconds to get the score.

The analysis shows that there are significant differences between the algorithms. As the results may differ when using other data sources and/or asking other questions, it is always worth testing the different algorithms to find the best one.