Logistic Regression Project

Machine Learning Course @JCE - by Ilya Krasnov (341008159)

Project Scope and Description

Given a dataset input in form of an Excel spreadsheet, a logistic regression algorithm was implemented from scratch. The algorithm was trained on the first 300 rows of the dataset and tested on the remaining rows.

Tools and Implementation

The algorithm was implemented in Python, using Pandas for data frame manipulation as well as seaborn and matplotlib for visualisation. For convenience, the project was implemented using a jupyter notebook. The notebook as well as this report are under version control and can be found on github

Implemented Algorithm

Results

Definition of Accuracy

The accuracy was defined by the following formula:

$$\frac{CorrectPredictions}{TotalPredictions} \times 100$$

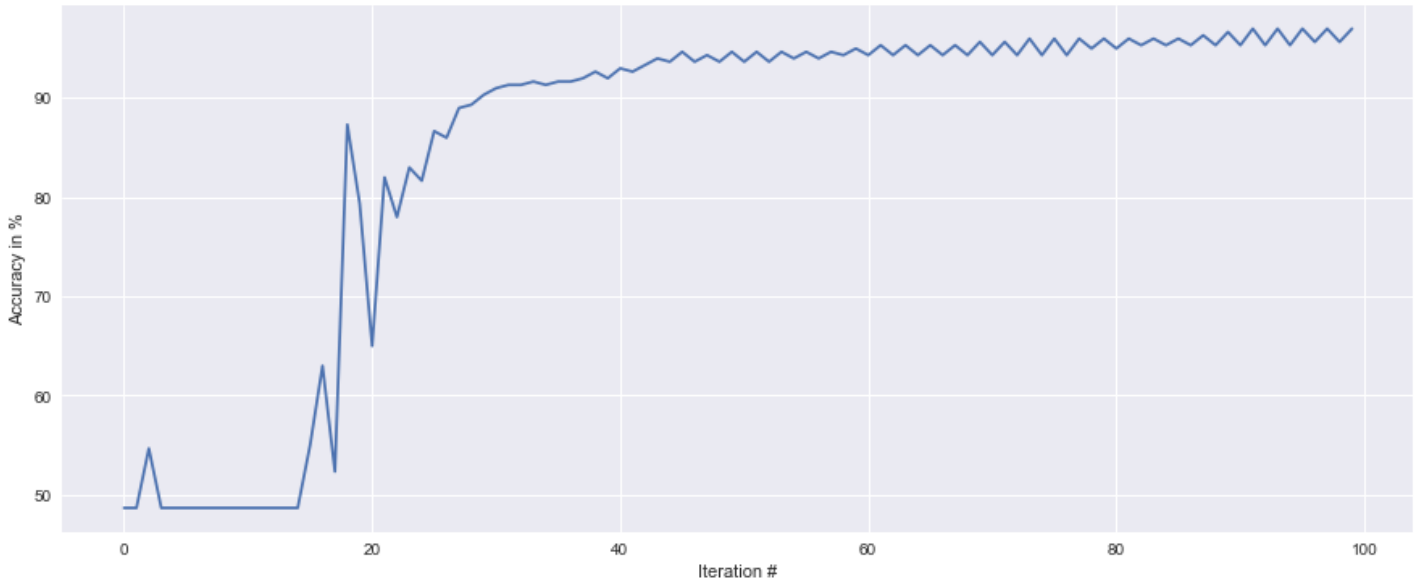
Training Accuracy

The following accuracy results were produced with 100 iterations of the gradient descent and different alpha values. For all cases the final accuracy of nearly

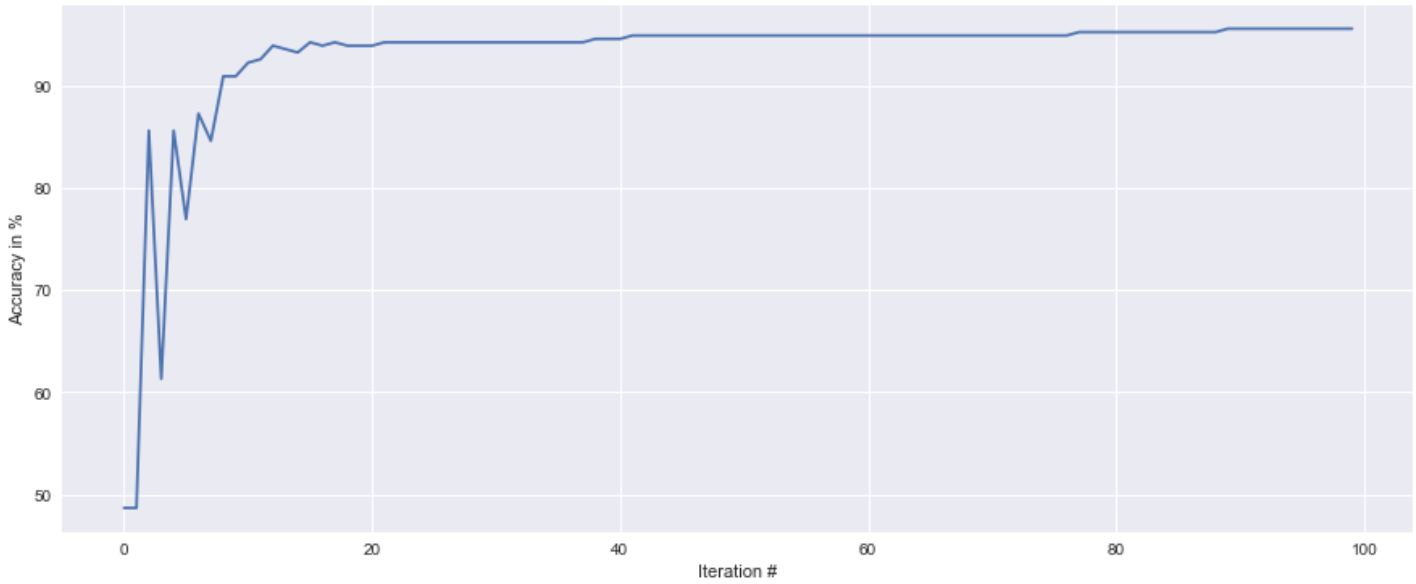
$$98\%$$

was achieved.

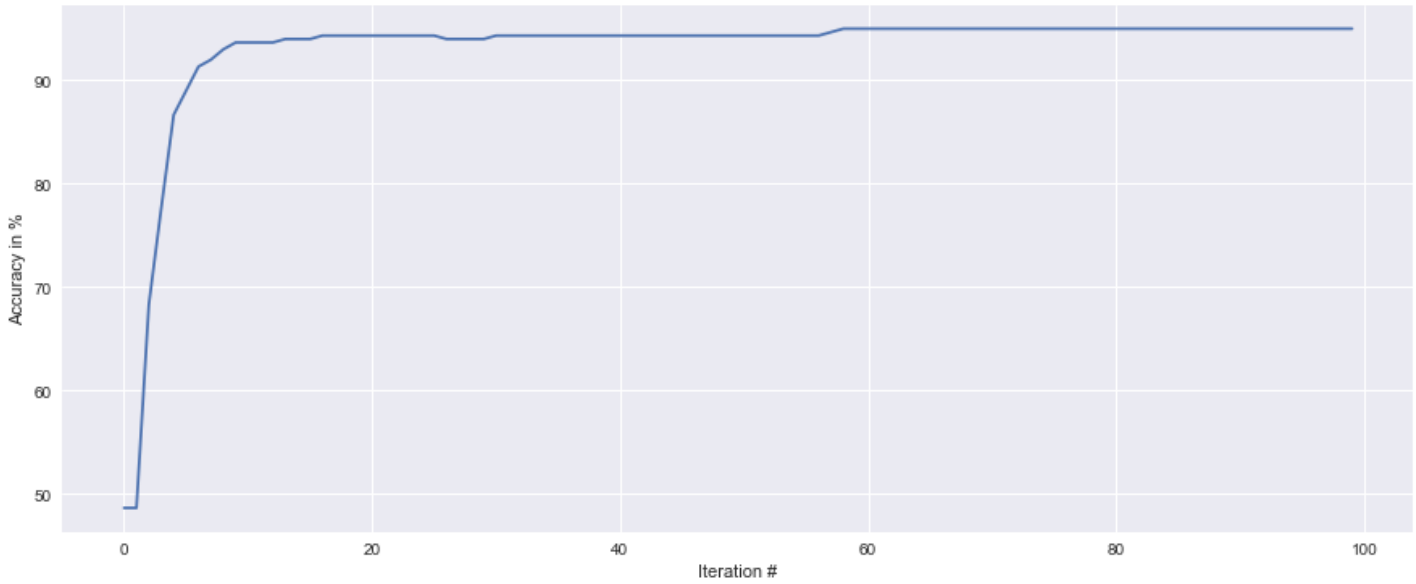
alpha = 0.2



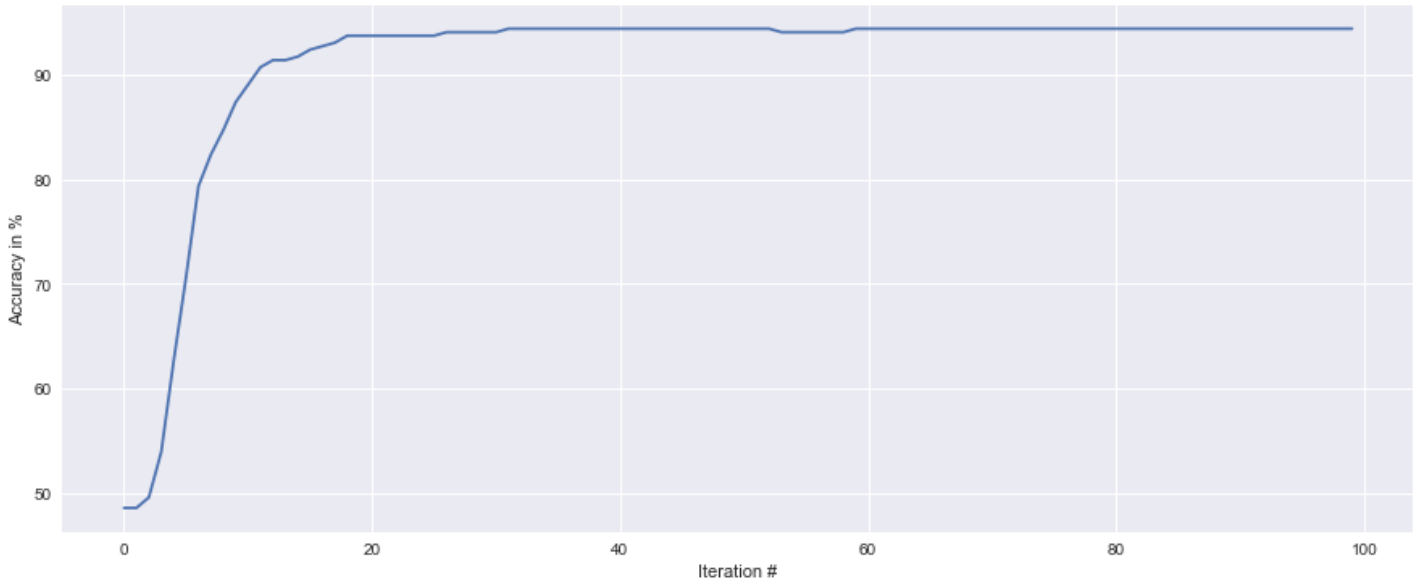
alpha = 0.03



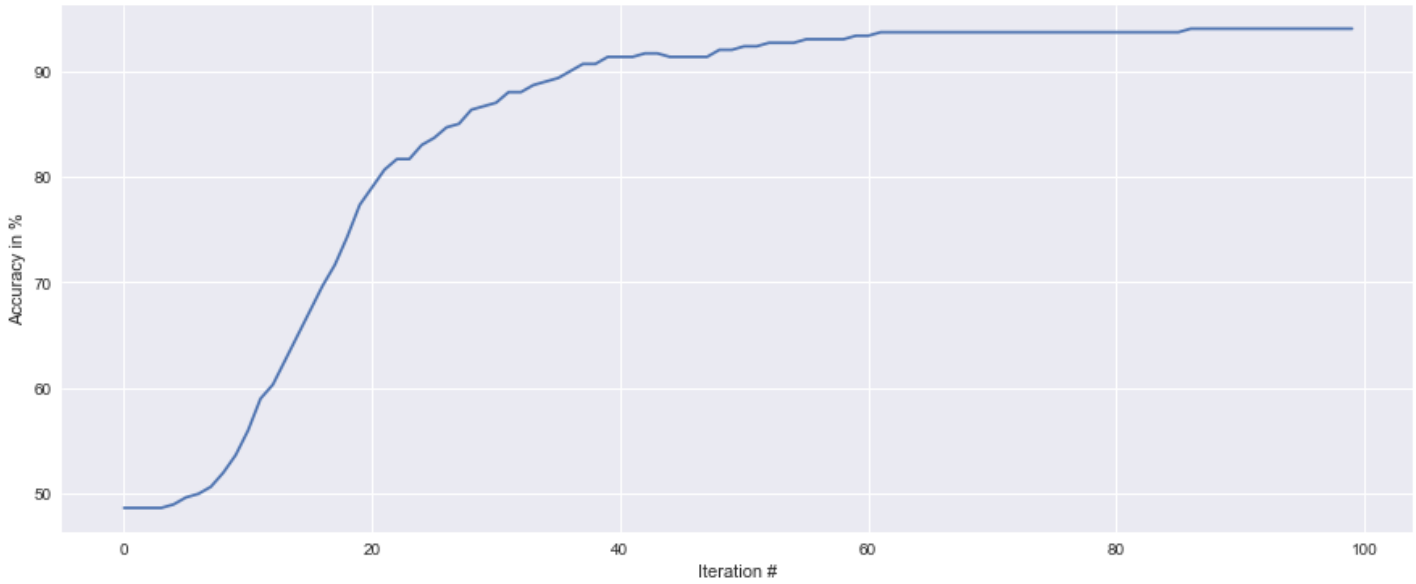
alpha = 0.02



alpha = 0.01



alpha = 0.003



Test Accuracy

The accuracy for the test data using the final θ was:

$$97.77\%$$

Conclusion

We clearly see that finding the right α values is very important. With the initial value of 0.2 we can see fluctuations in the accuarcy, although with overall trend towards 100%. Making α smaller produces much nicer results. With $\alpha = 0.02$ we see a fast convergence, almost after 10 interations. Making α too small on the other hand lets the alogirthm converge slower, thus making a higher number of iterations necessary in order to achieve high accuracy values.