Assignment 1 FYS-2021

Markus Leander Wilhelmsen

September 8, 2024

1 Introduction

This assignment taught us how to extract and filter out data in python to which can be used in a machine learning algorithm. Github link `https://github.com/markuslw/FYS-2021-1`

2 Design & Implementation

To retrieve data to be used in machine learning, we load a `CSV` file, filter out the relevant columns, create a new column with values to represent the filtered columns, and create a count of the different values. From here we create a test and train set using two new columns and the value label.
Before we push the data through the logistic regression classification method, we define the logistic function as sigmoid's. This function's characteristic is the S-shape curve, or the sigmoid curve. We also define the loss function as cross entropy to minimize the probability of incorrect predictions. We also define an accuracy function to measure the percentage of correctly predicted instances.
Finally we define the logistic discrimination classifier with stochastic gradient descent. For each epoch we run through the samples, getting the prediction value, updating the gradients and the weights. Once all samples have been iterated, we use the loss function and append the result to a list to be visualized.

3 Performance

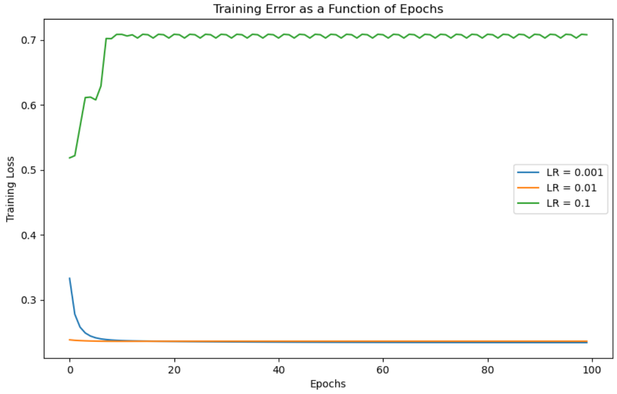
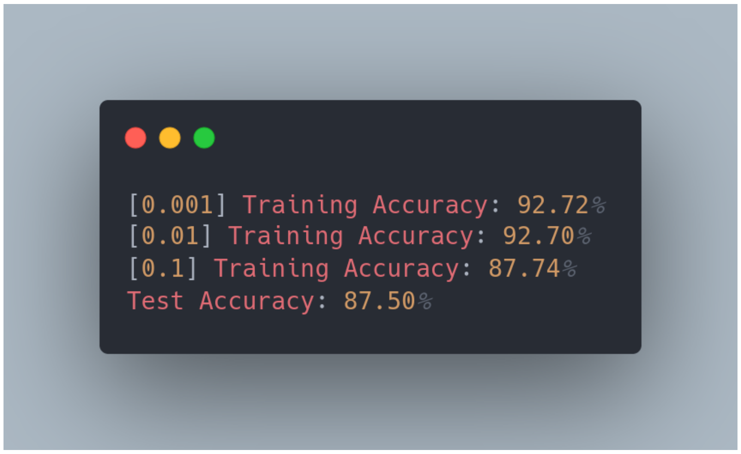


Figure 1: Learning rates visualized 0.001, 0.01, 0.1.

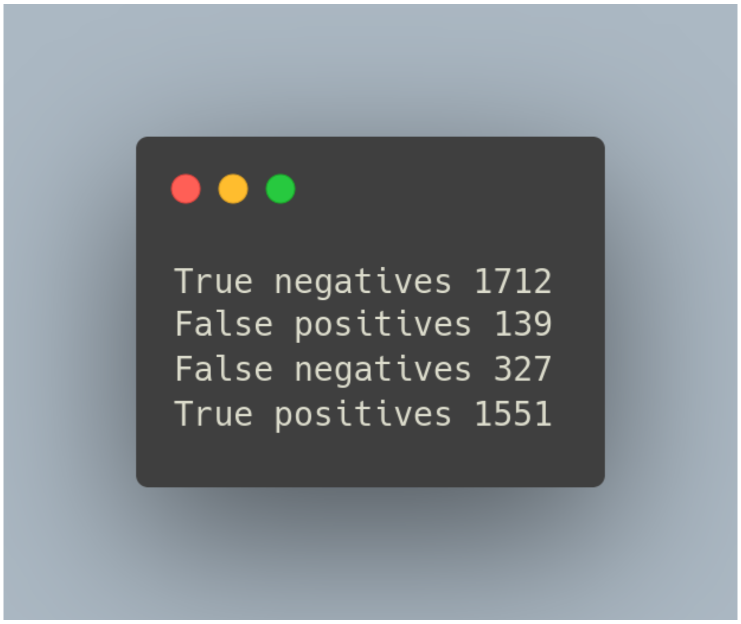
Figure 1 shows how with a high learning rate such as `0.1`, the loss skyrockets, fluctuates and fails to converge. The smallest rate of `0.001` shows a typical result from such a small rate, with weights barely being updated the loss curve will appear near linear. The `0.01` rate shows that the model learns from the data and converges to minimized loss.



```
[0.001] Training Accuracy: 92.72%
[0.01] Training Accuracy: 92.70%
[0.1] Training Accuracy: 87.74%
Test Accuracy: 87.50%
```

Figure 2: Train and test accuracy.

Figure 2 shows the training accuracy decreasing with each step, and the final test accuracy at 87.5%. This reflects that with each larger step, it tries to speed the convergence, but overshoots. The fact that the test accuracy is lower than the training accuracy simply means it cant generalize the data as well as on the training set.



```
True negatives 1712
False positives 139
False negatives 327
True positives 1551
```

Figure 3: Confusion matrix.

The confusion matrix in figure 3 provides a detailed breakdown of the accuracy rather than just a percentage. It shows us the counts of the actual classifications versus the predicted classifications. With the true values being higher than its counterparts, which of course is good.

4 References

- Lecture slide [Linear & logistic regression]
- Lecture slide [Optimization and gradient descent]
- `https://www.geeksforgeeks.org/understanding-logistic-regression/`
- `https://ml-cheatsheet.readthedocs.io/en/latest/gradient_descent.html`
- `https://www2.imm.dtu.dk/pubdb/edoc/imm3274.pdf`
- `https://chandhana520.medium.com/implementing-sgd-stochastic-gradient-descent-for-linea`