Programming Exercise 1

Regression Task

Choose a regression dataset and apply linear regression on a random subset of the training set of increasing size. You should select training sets that include more and more data points.

1. Plot the approximation error (square loss) on the training set as a function of the number of samples N, i.e., data points in the training set.

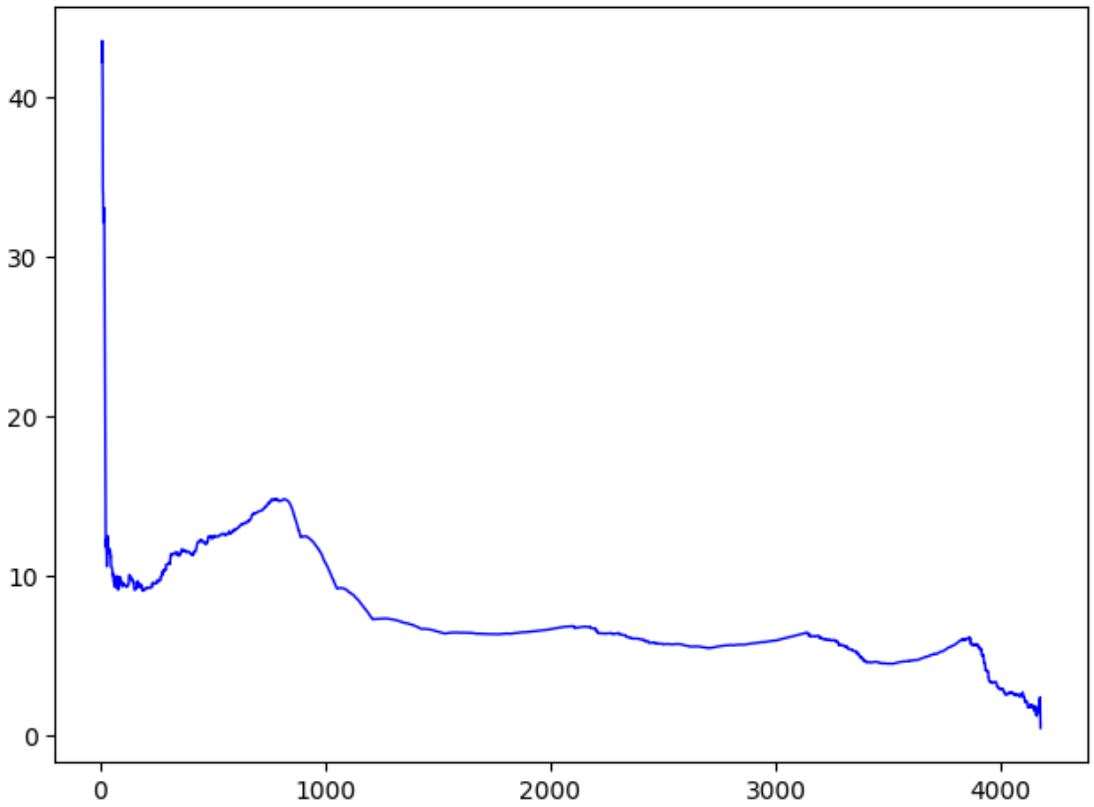


Figure 1: Approximation error (square loss) on the training set as a function of the number of samples N for the *abalone.txt* example

2. Plot the cpu-time as a function of N.

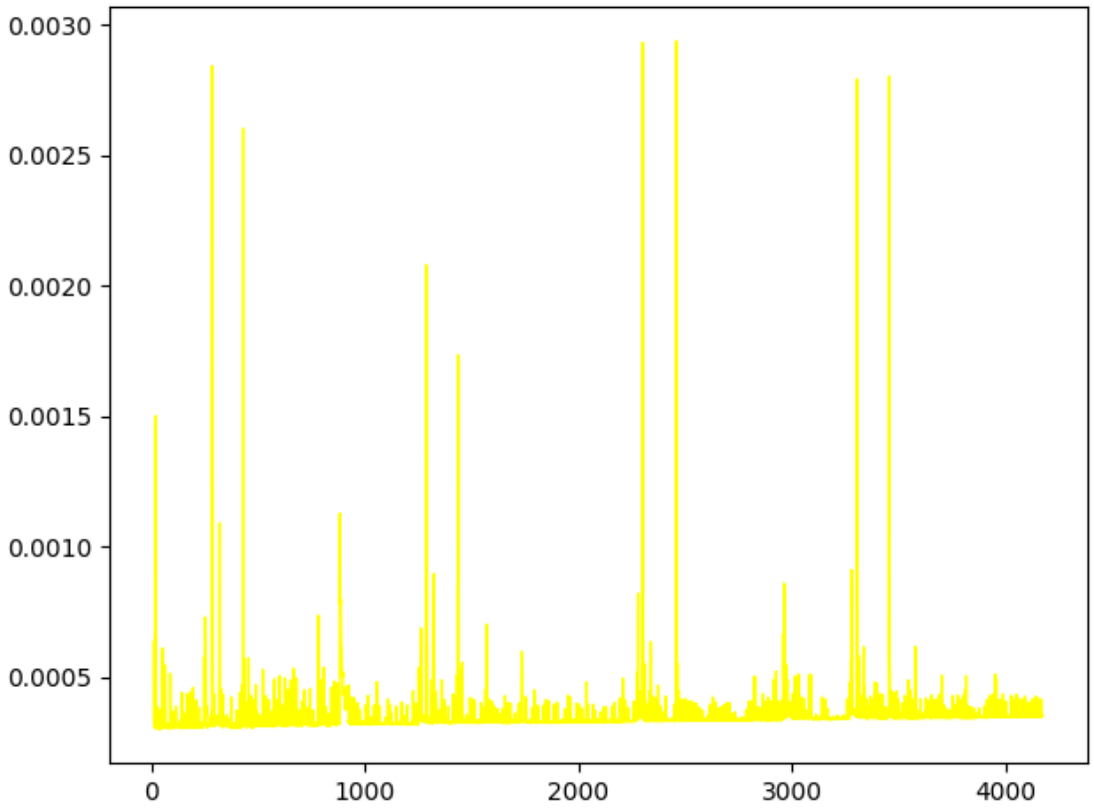
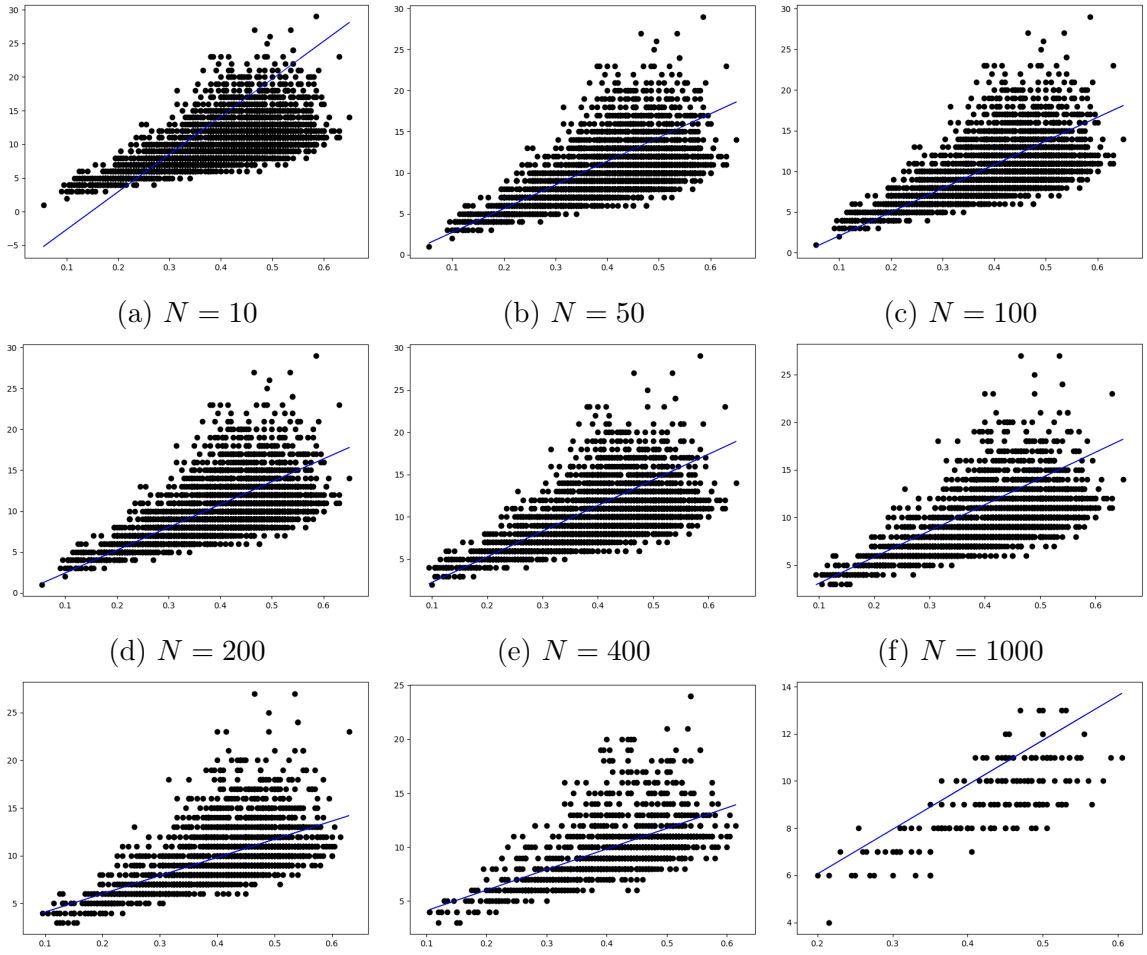


Figure 2: CPU-time as a function of the number of samples N for the *abalone.txt* example

3. Explain in detail the behaviour of both curves: what is the trend you observe? does it stabilize? why?

   The approximation error (square loss) decreases when N increases and stabilizes, because when N increases we are actually increasing the training dataset, and it allows a better fit of the model. The CPU-time does not stabilize.

4. Explore how the learned weights change as a function of N. For this, you can make separate stem plots for several values of N. Can you find an interpretation for the learned weights?



(a) $N = 10$ (b) $N = 50$ (c) $N = 100$

(d) $N = 200$ (e) $N = 400$ (f) $N = 1000$

(g) $N = 2000$ (h) $N = 3000$ (i) $N = 4000$

Figure 3: Separate stem plots for several values of N, for the *abalone.txt* example. Black points are the test set and the blue line is the prediction of the linear regression model, fit with the training set of size N

Classification Task

Choose a classification dataset and apply logistic regression. Repeat the previous four steps using as error the mean accuracy.

1. Plot the mean accuracy on the training set as a function of the number of samples N, i.e., data points in the training set.

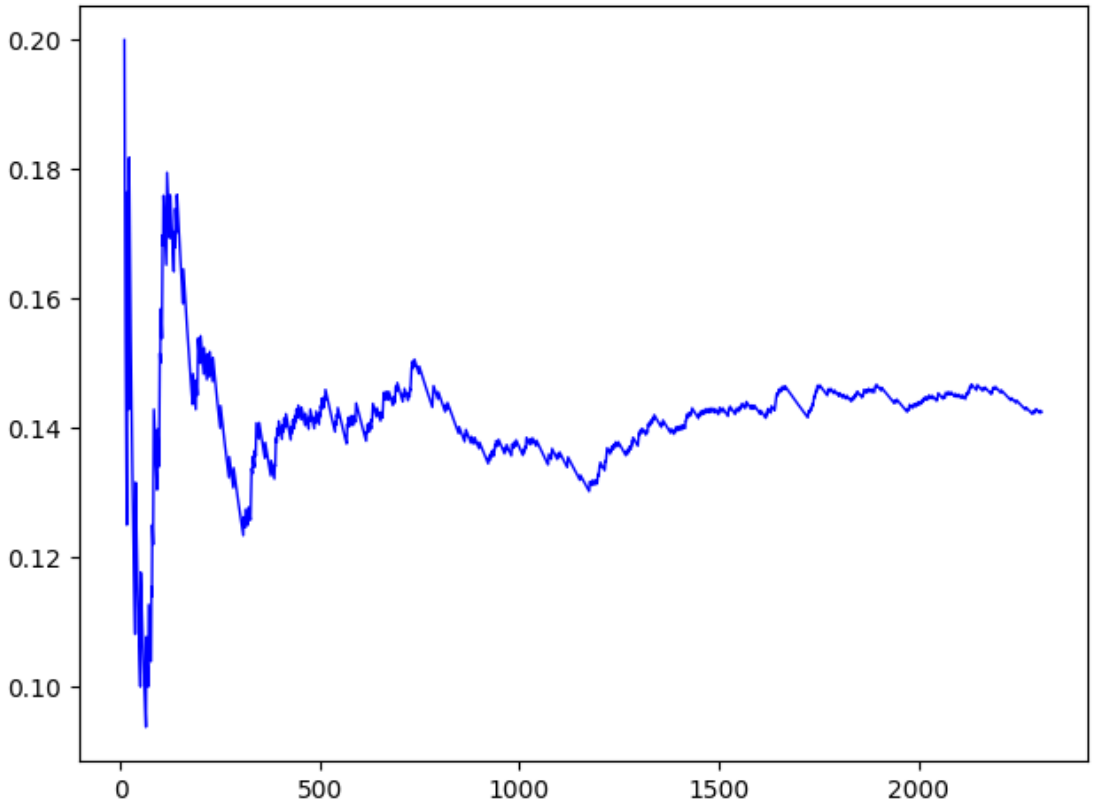


Figure 4: Mean accuracy on the training set as a function of the number of samples N, for the *segment.scale.txt* example

2. Plot the cpu-time as a function of N.

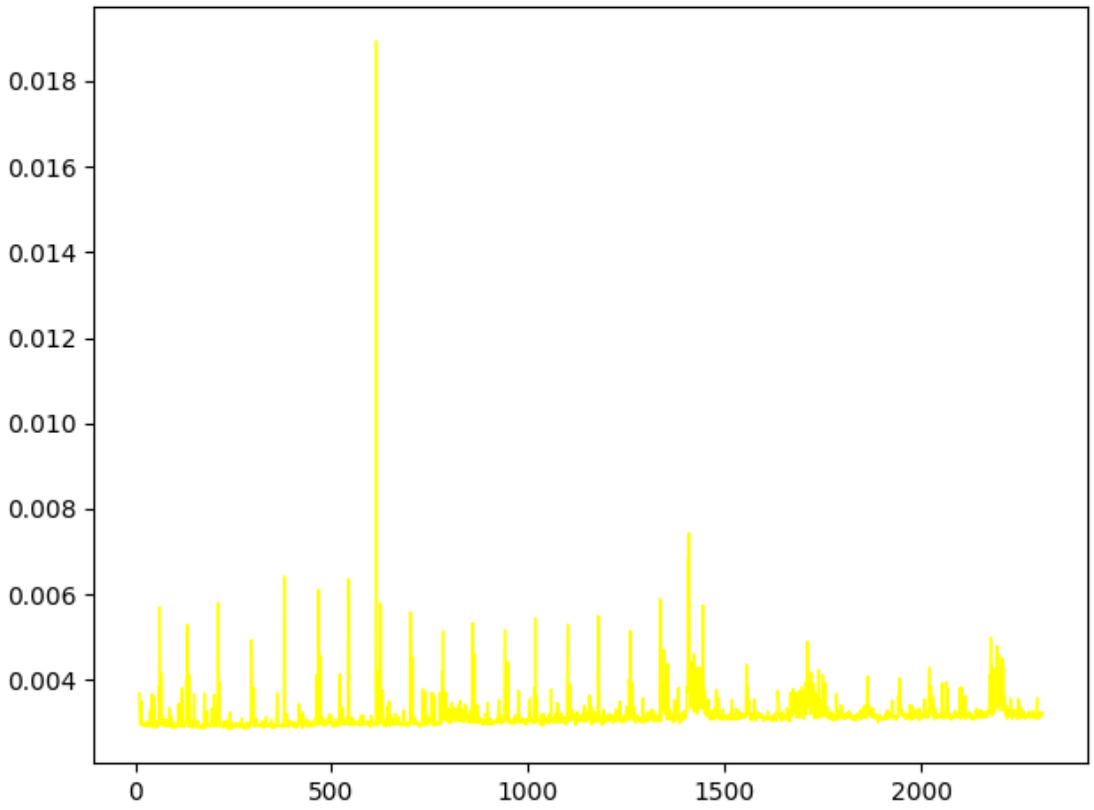
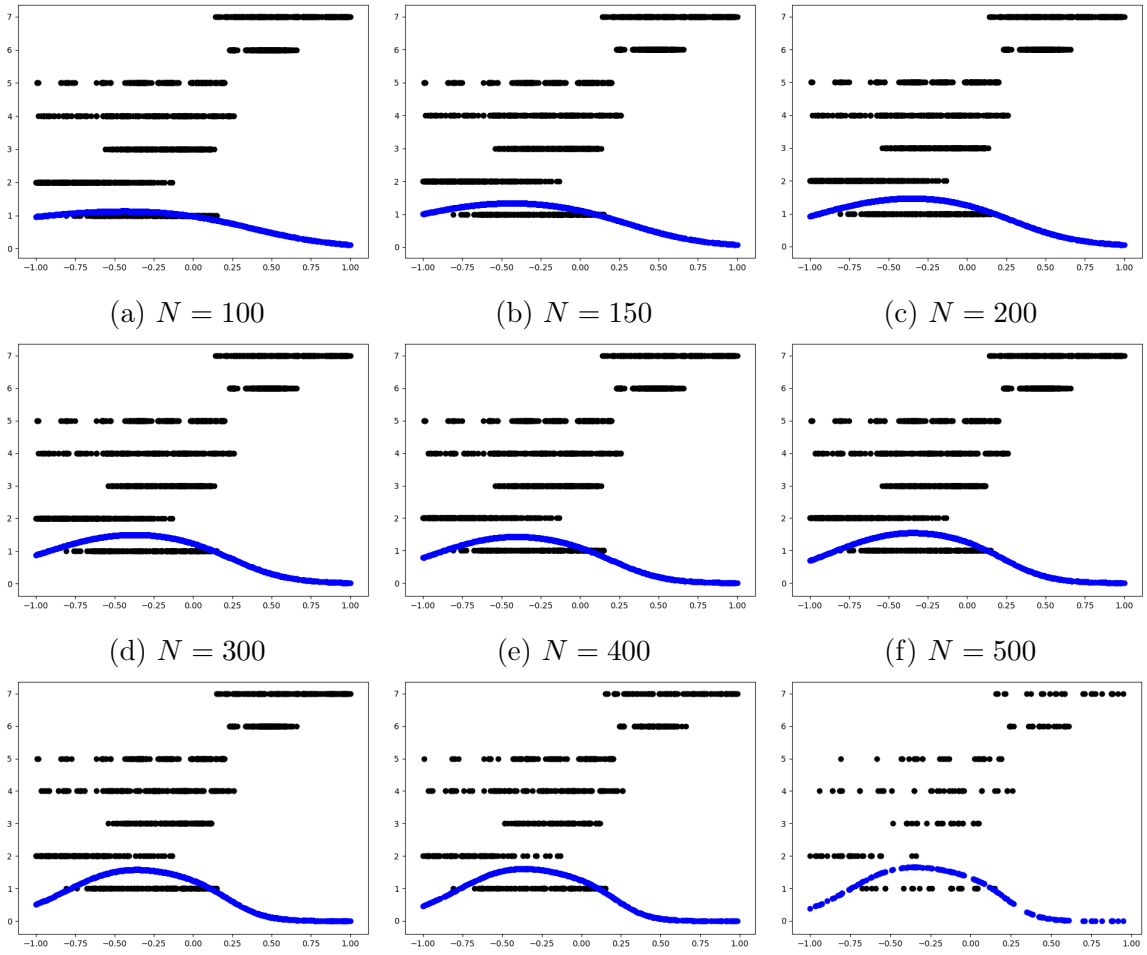


Figure 5: CPU-time as a function of the number of samples N, for the *segment.scale.txt* example

3. Explore how the learned weights change as a function of N. For this, you can make separate stem plots for several values of N. Can you find an interpretation for the learned weights?



(a) $N = 100$ (b) $N = 150$ (c) $N = 200$

(d) $N = 300$ (e) $N = 400$ (f) $N = 500$

(g) $N = 1000$ (h) $N = 2000$ (i) $N = 2100$

Figure 6: Blue points are the probability of which the data belong to class 2, deduced with the logistic regression model, fit with the training set of size N. The figures are separate stem plots for several values of N, for the *segment.scale.txt* example