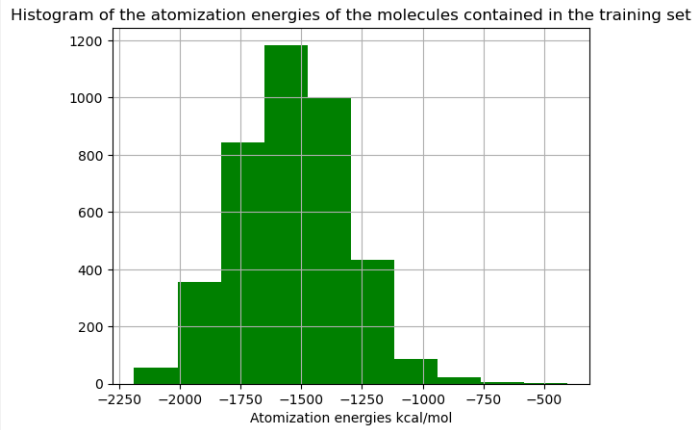
Machine Learning for Physicists - Third Assignment

Guilherme Simplício

December 30, 2022

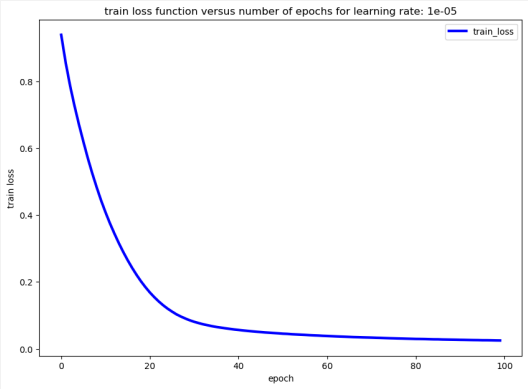
1 Plots & Answers

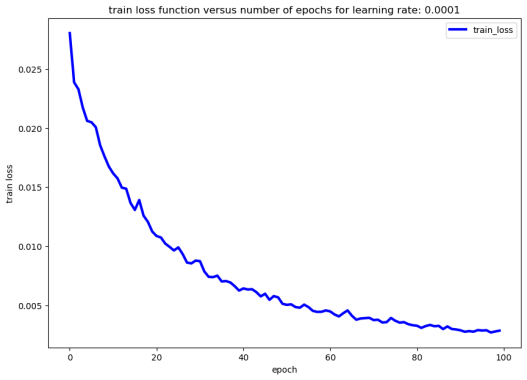
Question 2 Histogram of the atomization energies of the molecules contained in the training set.

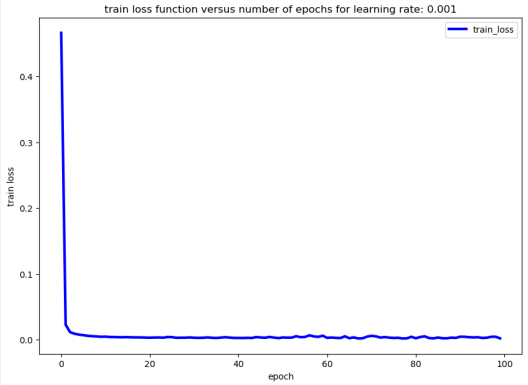


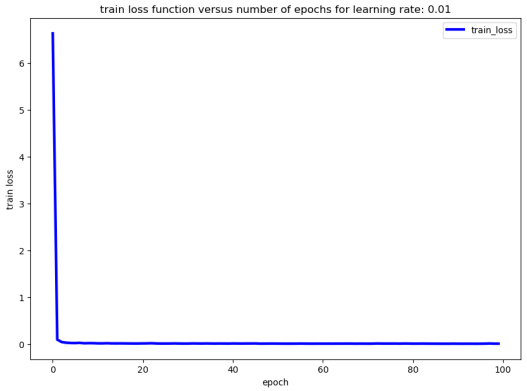
Question 5 First, plot and analyze the behavior of the train and validation losses during training . Finally, evaluate the performance of this model on the test set, for a giving learning rate, contained in the set $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$.

First of all, we did a quick inspection how the learning rate influenced the training data, for a smaller number of epochs (= 100), in order to have a more informed choice on the learning rate.

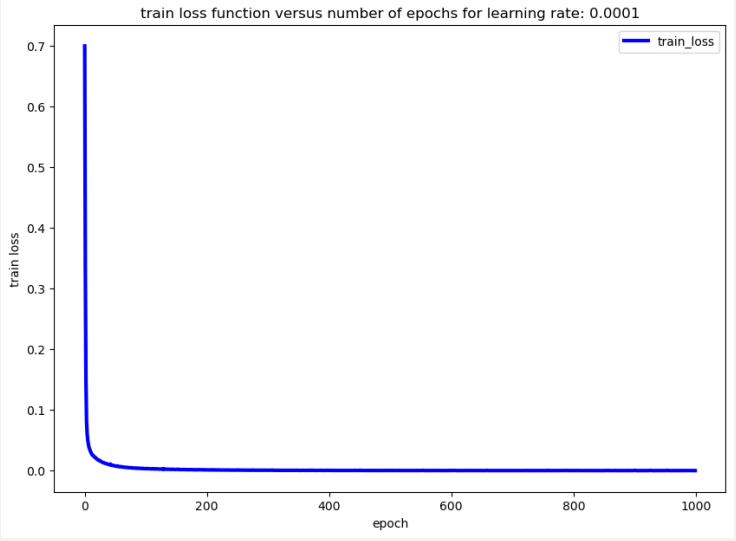


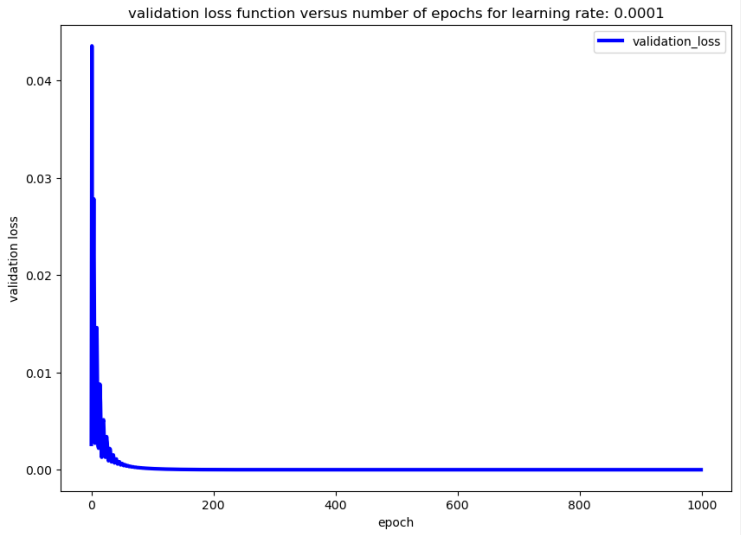






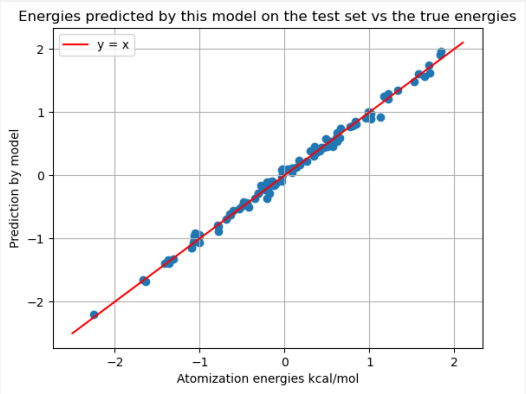
A quick inspection can show as that as expected, a learning rate of 10^{-4} would give us the lower, more steady train loss. Proceeding with that value, we can *plot and analyze the behavior of the train and validation losses during training.*





One can easily observe a fast convergence to a very low loss value, as it was expected! Afterwards, we calculated the test loss (mean square error) = 0.00414, with just the test dataset. This result is fairly satisfying.

Question 6 Plot the energies predicted by this model on the test set vs the true energies in kcal/mol units.



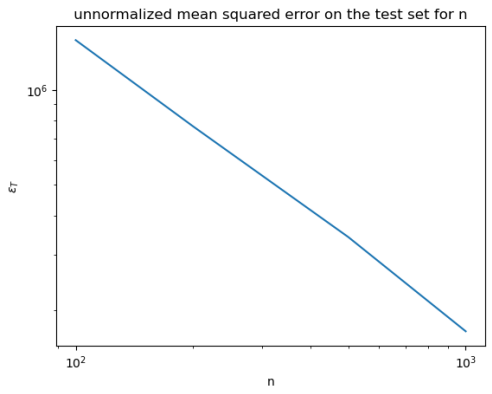
Here we can see that the energies predicted by this model on the test set vs the true energies have a almost perfect one to one correlation (y = x).

Compute the unnormalized root mean square error on the test set.

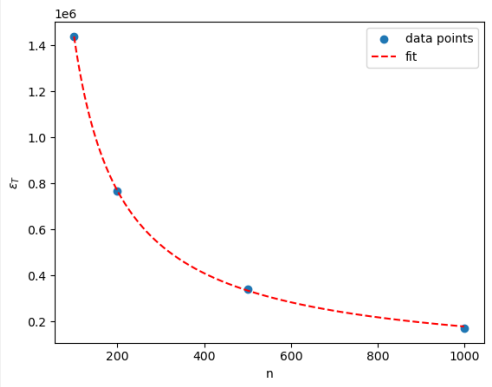
Using the sklearn library, the unnormalized RMSE is 15.46.

Question 7 Plot the learning curve , i.e., unnormalized mean squared error on the test set ε_T as a function of the size of the training set n on a log-log scale.

Note that throughout this exercise, a new data loader was created with the unnormalized data.



Identify what function fits well ε_T (n).



The Function fits fairly well the equation $\varepsilon_T = 92785015.35705 \cdot n^{-0.90487}$.

Estimate the test error that this model would achieve with n = 4000 training points, i.e., compute ε_T(n = 4000).

The test error that this model would achieve with n = 4000 training points is 168605941279.45822

2 Acknowledgments

I would like to thank Albert Riber, Arianna Alonso Bizzi,Francisco Simões and Tomás Feith(not in the course), for having very useful debates with me regarding this assignment and the course, in general.