Approximation of Topographical Datasets Using Linear Regression Project 1 – FYS-STK4155

Anders Thorstad Bø

Department of Mathematics (Fluid Mechanics), University of Oslo, Oslo, Norway[*]

This project uses three different linear regression analysis methods to investigate their performance against 1D- and 2D-datasets, using n^{th}-order polynomial functions for prediction models. The methods are Ordinary Least Squares (OLS), Ridge and Lasso. It also displays how to make better predictions and get better estimates of hyperparameters in the model by using resampling techniques such as bootstrapping and K-fold – cross-validation. For testing the different methods, I use the first Franke-function as a topographical analog. As a final showcase, I make use of the method to try to approximate the shapes from topographical data from two mountainous areas in Norway to see how the models handles real-world data.

INTRODUCTION

When using machine learning techniques to predict the outcome of datasets, a common issue that arises is overfitting and underfitting, often related to the complexity of the prediction model(Hastie et al., 2009). These effects will decide whether or not the resulting model are able to accurately make predictions on test-, benchmarking-, and new data after training. One important factor is to strike a balance between the performance of the model on the training data sets and how generally applicable it is to similar dataset. That way the finished model will be able to give a satisfactory representation of the new datasets of the same type, as it will be general enough to capture different effects in the new dataset that necessarily were not present in the training data (Goodfellow et al., 2016).

The aim of this project is to show how to use the linear regression methods *Ordinary Least Squares* (OLS), *Ridge*, and *Lasso* to make prediction models for 2D – datasets. Together with these methods, the project uses resampling techniques like *bootstrapping* and *K-fold – cross-validation* in order to tuning the hyperparameters in Ridge and Lasso-regression, as well as investigating the *Bias-Variance – trade-off* to indicate a suitable polynomial degree for the model (Hastie et al., 2009; Raschka et al., 2022).

For testing the methods and models, I am using a 1D exponential-function, a similar 2D exponential-function, and the first Franke-function (Franke, 1979) as data functions. This allows for seeing how the regression measures like mean square error and R^2-score evolve with more elaborate data functions. In order to measure the effect from the resampling algorithms, I will also show comparisons between the resampled models and the ones obtained from a straight-forward regression analysis.

As final test of the methods and algorithms, I am using real topographical dataset from two different places in Norway. The first dataset is from a small mountain range in Etnedal, Oppland, and the second dataset is from the mountains in the western part of Jotunheimen, near Årdal. Both these datasets has exhibit substantial fluctuations between areas in the dataset, and the expectation is that they will challenge the methods.

The article will first go through relevant theoretical concepts like the regression methods, resampling techniques, and give a general overview on how to interpret the measures and model performance.Next, it gives a breakdown of the most important algorithms and method implementations, and a description on how I use them to generate the results.

Finally, the article presents the results from the benchmarking, and important discussion points that comes up in these results. The results and discussion on the model performance on the topographical datasets follows after that.

The very last parts of the article gives some final remarks and conclusions from the work, as well as some future , followed by a brief appendix explaining where and how to access the full program files, and how to use the different program files.

[*] andetb@uio.no; https://github.com/andersthorstadboe/project-1-anders-boe

REFERENCES

Richard Franke. A critical comparison of some methods for interpolation of scattered data. *Calhoun: The NPS Institutional Archive*, 1979.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

T. Hastie, R. Tibshirani, and J.H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. Springer, 2009. ISBN 9780387848846. URL `https://books.google.no/books?id=eBSgoAEACAAJ`.

Sebastian Raschka, Yuxi (Hayden) Liu, and Vahid Mirjalili. *Machine Learning with PyTorch and Scikit-Learn*. Packt Publishing, 2022.