Project 1

FYS-STK4155

Aaby I., Tandberg V., Ludvigsen E. *

[September 2024]

Department of physics, University of Oslo, P.O. Box 1048 Blindern, N-0316 Oslo, Norway

Abstract

This article seeks to expand our understanding of the use of linear regression models for fitting of both theoretical and real world data. In order to do so, we implement three linear regression models (Ordinary Least Squares, Ridge, and Lasso), and study their performance and statistical characteristics via resampling methods such as bootstrap and cross-validation.

We implement a pipeline following the theory presented in this report, and verify it both by comparing components of our implementation against references provided by the scikit-learn library for Python, and also by applying the pipeline to synthetic data generated by the analytical Franke function. This verifcation shows that the implementation provides reasonable results for the Franke function. Thus, the code developed is proven capable of fitting linear models to 2-dimensional data.

Finally, we apply the pipeline to real terrain data from Norway. The best-performing model suggested by resampling analysis was Lasso with a λ-value of 0.01. After fitting the model to the training set, this model achieved an R^2-score of 0.831 and an MSE of 0.0127 on the test set. NOTE: The last sentence is just a placeholder.

I Introduction

Accurate modeling and representation of terrain data play a critical role in various fields, including geospatial analysis, environmental studies, and urban planning. Terrain surfaces, which are inherently complex and irregular, present significant challenges for mathematical and computational modeling. A variety of techniques have been developed to fit and predict these surfaces based on scattered data points. As discussed in [5, 2], linear regression is a simple but powerful method that can provide a machine learning foundation for fitting and predicting all kinds of data, including such heightmaps.

*isakaab@student.matnat.uio.no, vebjort@uio.no, ericlu@uio.no

In this article, we aim to explore the application of linear regression methods for fitting terrain data, assessing their effectiveness in capturing the underlying patterns of elevation variations.

The goal of this article is to assess the feasibility and limitations of using linear regression in terrain modeling and highlight possible improvements or alternatives for future research. Our focus will include both the implementation of standard linear regression and its variations, such as Ridge and Lasso regression. In addition, we also want to implement resampling methods such as the bootstrap method and cross validation, to enhance the generalization capabilities of our models. It is also our intention to evaluate obstacles such as overfitting and matrix inversion problems.

To do so, we seek to introduce a machine learning pipeline and the theoretical framework upon which this pipeline is built. The concept is explored in existing machine learning literature, such as [5], and [2]. In the end, we seek to construct a pipeline containing the following stages:

1. split data into validation, training, and test sets,
2. perform k-fold cross-validation and find the optimal model structure and corresponding hyperparameters,
3. train the best model on the training set,
4. extract statistical properties such as bias and variance via bootstrap resampling,
5. evaluate the performance of the model on the test set.

Numerical methods and programming are supported by [1] and [4]. We seek to use the Franke function in order to verify the implementation of our model and pipeline. By first establishing the robustness of our linear regression approach on the Franke function, we can then extend our analysis to more complex and noisy datasets derived from actual terrain measurements.

First, we shall present the analytical methods and mathematical background for our regression modelling in Section 2. Then, the code implementation and verification of the pipeline will be introduced in the second part of Section 2. The last part of Section 2 will present how our pipeline generates a model from actual terrain data. The model fitting results are then presented in Section 3, before we cover how we verified that our implementation was correct, as well as the challenges we faced when testing various components with the Franke function. We will discuss the performance of the model and how it performs on the test set. At the end of this article, we will discuss how the model may be used, state its limitations, and finally, suggest how the model may be improved through further research. [1]

[1] The first draft for the introduction was generated via ChatGPT [3]

II References

[1] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.

[2] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer.

[3] OpenAI. Chatgpt conversation for intro draft (gpt-4 model). https://chatgpt.com/share/66f40fc0-b81c-800b-9f04-f5c45de00edd, 2024. Accessed: 2024-09-25.

[4] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[5] Sebastian Raschka and Vahid Mirjalili. *Python Machine Learning*. Packt Publishing Ltd., Livery Place 35 Livery Street Birmingham B3 2PB, UK, second edition, September 2017.