University of Oslo
Department of Mathematics

Fall, 2023
Weekly Exercises: Week 39

Author: Jonas Semprini Næss

FYS-STK4155: Applied Data Analysis and Machine Learning

Outlines of the reoport

Abstract

This project aims to examine different regression methods to analyze topographic data of [insert place name], in particular the following methods: Ordinary Least Squares (OLS), Ridge regression and Lasso regression. An assessment is performed on these methods by studying their bias-variance trade-off through resampling techniques such as cross-validation and bootstrap, in addition to evaluating their mean squared error (MSE) and R^2 score. The regression methods are tested and assessed on Franke's function, a widely known function used for testing interpolation and fitting algorithms, before proceeding with fitting the topographic data. Our findings suggest [insert regression method] to be the best method for fitting terrain data of [insert place name]. It performed well under the assessment with results such as [insert results].
Keywords: (To be added when the report is more or less finished)

Introduction

In the 1800s, Sir Francis Galton did a study on sweet peas, a self-fertilizing plant, to understand how strongly the characteristics of one generation would manifest in the following generation. It all started when he noticed the sweet pea packets distributed to his friends had substantial variations. A data set was created where he looked at the size of daughter peas against the size of mother peas and illustrated the basic foundation of what statisticians still call regression [1]. This report presents a comprehensive analysis of function fitting techniques applied to a two-dimensional function known as the Franke function ([11]). The primary goal is to assess the performance of three different regression methods, namely Ordinary Least Squares (OLS) ([3]), Ridge Regression ([4]), and Lasso Regression ([4]), in terms of model accuracy, bias-variance trade-off ([5]), and generalization capabilities. The study also incorporates resampling techniques ([8]) and cross-validation ([8]) to enhance the model evaluation process. Initially, the report outlines the theoretical background of the Franke function and the mathematical formulations of OLS, Ridge, and Lasso regression. The study explores the bias-variance trade-off, to understand the trade-offs between model complexity and generalization performance. To assess these trade-offs, the report employs resampling techniques such as cross-validation. In addition to synthetic data generated from the Franke function, this report extends its analysis to real-world data, offering practical insights into the application of these regression techniques in authentic scenarios.

i References

Bibliography

[1] Stanton, J. M. (2017). *Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors*. Journal of Statistic Education, **9:3**, 1-2. DOI

[2] Jensen, M. H. (n.d.). *Week 34: Introduction to the course, Logistics and Practicalities*. Applied Data Analysis and Machine Learning. Retrieved September 22, 2023, from Jupyter book

[3] Jensen, M. H. (n.d.). *Week 35: From Ordinary Linear Regression to Ridge and Lasso Regression*. Applied Data Analysis and Machine Learning. Retrieved September 22, 2023, from Jupyter book

[4] Jensen, M. H. (n.d.). *Week 36: Statistical interpretation of Linear Regression and Resampling techniques*. Applied Data Analysis and Machine Learning. Retrieved September 22, 2023, from Jupyter book

[5] Jensen, M. H. (n.d.). *Week 37: Statistical interpretations and Resampling Methods*. Applied Data Analysis and Machine Learning. Retrieved September 22, 2023, from Jupyter book

[6] Arya, N. (2023, February 20). *The Role of Resampling Techniques in Data Science*. KDnuggets. Retrieved September 23, 2023, from [Website]

[7] Arya, N. (2021). *Resampling Techniques in Data Science*. KDnuggets. [Photograph]

[8] Jensen, M. H. (n.d.). *Week 37: Statistical Interpretations and Resampling Methods*. Applied Data Analysis and Machine Learning. Retrieved September 8, 2023, from Jupyter book

[9] Jensen, M. H. (n.d.). *Week 34: Introduction to the course, Logistics and Practicalities*. Applied Data Analysis and Machine Learning. Retrieved September 8, 2023, from Jupyter book

[10] *Mean squared error*. (2023, August 15). In Wikipedia. URL

[11] Jensen, M. H. (n.d.). *Project 1 on Machine Learning, deadline October 9 (midnight)*, 2023. Applied Data Analysis and Machine Learning. Retrieved September 7, 2023, from Jupyter Book

[12] *Central Limit Theorem*. (This page was last edited on 27 September 2023, at 23:34 (UTC).)In Wikipedia. URL