University of Oslo

FYS-STK3155

Project 1

Eirik eirik@ifi.uio.no
Felix felix@ifi.uio.no
Brage Wiseth bragewi@ifi.uio.no

October 2, 2023

https://github.com/bragewiseth/MachineLearningProjects

Contents

Abstract

In this paper, we delve into the realm of machine learning model optimization and evaluation. Our study encompasses various regression techniques, including Ordinary Least Squares (OLS), Ridge, and Lasso regression, to analyze their effectiveness in handling simple and more complex datasets. Additionally, we employ bootstrap resampling and cross-validation methodologies to rigorously assess model performance and enhance generalization. A significant portion of our investigation is dedicated to understanding the delicate balance between bias and variance. We explore how regularization methods like Ridge and Lasso impact bias-variance trade-offs, offering insights into the stability and predictive power of these models. Furthermore, we provide empirical evidence of the benefits of cross-validation and bootstrap techniques in mitigating overfitting and improving model robustness. We found that { ..results.. }. Additionally we verify and compare our findings with well established theory and libraris such as SKLearn.

Keywords: Linear Regression, Scaling, Bias & Variance

1. Introduction

Machine learning has emerged as a powerful tool in data analysis, providing the ability to uncover complex patterns and relationships in diverse datasets. But, at its core, machine learning is all about finding functions that capture the underlying structure of the data. The use of machine learning algorithms to approximate functions is the essence of this paper.

Our motivation for this research lies in the exploration of machine learning techniques to approximate the terrain on our planet, which can perhaps be described by such a function. Earth's terrain exhibits peaks and troughs, hills and valleys, much like some polynomial functions. Fortunately, we can employ standard linear regression techniques to approximate polynomials, but the terrain presents its own set of challenges. Firstly, the terrain's true underlying function may not be a polynomial at all, and its complexity may vary significantly from one location to another. Secondly, our landscape is teeming with small, intricate details. Some regions are characterized by flat and smooth surfaces, while others are marked by rough and uneven terrain. Focusing too much on these minute details can lead to model overfitting, making it crucial to strike a careful balance between model complexity and generalization. In this context, regularization and resampling techniques, including Ridge and Lasso regression with bootstrap and cross validation, have proven indispensable. By introducing regularization and resampling, we aim to find the sweet spot between bias and Variance. And getting the best predictions we can with our assumptions.
To embark on this exploration, we will begin with a simpler case: "Franke's function." which mimics our real terrain data. This function serves as a foundational starting point, allowing us to assess our model's performance in a controlled environment before venturing into the complexity of real-world terrain data. Through this gradual progression, we provide ourselves wih a framework that can be applied to more complex and varied real-world terrain datasets.

Data: We begin by introducing the dataset used for our analysis, highlighting data collection and preprocessing procedures. Understanding the characteristics of the terrain data is fundamental to our modeling endeavor.

Methods and Scaling: Next, we delve into the methodology, encompassing the implementation of polynomial regression models and the application of regularization techniques such as Ridge and Lasso. Additionally, we will discuss the importance of proper scaling for model stability and convergence.
Bias-Variance Trade-off: A significant portion of our study will revolve around the critical concept of bias and variance. We'll explore how regularization methods influence this trade-off and delve into the fine balance between model complexity and generalization.
Results: In this section, we will present the outcomes of our experiments, showcasing the performance of different models and regularization techniques. Through empirical evidence, we aim to provide insights into the effectiveness of our approach.
Conclusion: Finally, we will summarize the key findings and their implications for terrain modeling with machine learning. Our conclusion will underscore the importance of regularization in achieving accurate representations of complex terrains and provide a perspective on future research directions.

Appendix

appendix

analytical

adding these two we get

$$\frac{\partial \left[\frac{1}{n}||\mathbf{y}-\mathbf{X}\beta||_2^2+\lambda||\beta||_2^2\right]}{\partial\beta}=0=\frac{2}{n}\left(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I}\right)\beta-\frac{2}{n}\mathbf{X}^T\mathbf{y} \implies \beta=\left(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

The minimization of ridge addition alone is

$$\min_{\beta\in\mathbb{R}^p}\frac{1}{n}\lambda||\beta||_2^2=\frac{2}{n}\lambda\beta$$

We know that

$$\min_{\beta\in\mathbb{R}^p}\frac{1}{n}||\mathbf{y}-\mathbf{X}\beta||_2^2=\frac{2}{n}\mathbf{X}^T\mathbf{X}\beta-\frac{2}{n}\mathbf{X}^T\mathbf{y}$$

$$\min_{\beta\in\mathbb{R}^p}\frac{1}{n}\sum_{i=0}^{n-1}(y_i-\tilde{y}_i)^2=\frac{1}{n}||\mathbf{y}-\mathbf{X}\beta||_2^2,$$

$$\min_{\beta\in\mathbb{R}^p}\frac{1}{n}||\mathbf{y}-\mathbf{X}\beta||_2^2+\lambda||\beta||_2^2$$

$$\hat{\beta}_{\text{OLS}}=\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y},$$

$$\hat{\beta}_{\text{Ridge}}=\left(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y},$$

SVD

Use the singular value decomposition of an $n\times p$ matrix $\mathbf{X}$ (our design matrix)

$$\mathbf{X}=\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T,$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices of dimensions $n\times n$ and $p\times p$, respectively, and $\mathbf{\Sigma}$ is an $n\times p$ matrix which contains the ingular values only. This material was discussed during the lectures of week 35.

Show that you can write the OLS solutions in terms of the eigenvectors (the columns) of the orthogonal matrix $\mathbf{U}$ as

$$\tilde{\mathbf{y}}_{\text{OLS}}=\mathbf{X}\beta=\sum_{j=0}^{p-1}\mathbf{u}_j\mathbf{u}_j^T\mathbf{y}.$$

$$\mathbf{X}^T\mathbf{X}=\mathbf{V}\mathbf{\Sigma}^T\mathbf{U}^T\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T=\mathbf{V}\mathbf{\Sigma}^T\mathbf{\Sigma}\mathbf{V}^T=\mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^T$$

$$\tilde{\mathbf{y}}_{\text{OLS}} = \mathbf{X}\beta = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\left(\mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^T\right)^{-1}\mathbf{V}\mathbf{\Sigma}^T\mathbf{U}^T\mathbf{y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{V}^{T^{-1}}\mathbf{\Sigma}^{2^{-1}}\mathbf{V}^{-1}\mathbf{V}\mathbf{\Sigma}^T\mathbf{U}^T\mathbf{y}$$

using the orthogonality of $\mathbf{V}$ and $\mathbf{U}$ we get. Multiplying with $\mathbf{\Sigma}$ removes columns from $\mathbf{U}$ with eigenvalues equal to zero.

$$\tilde{\mathbf{y}}_{\text{OLS}} = \mathbf{U}\mathbf{U}^T\mathbf{y} = \sum_{j=0}^{p-1}\mathbf{u}_j\mathbf{u}_j^T\mathbf{y}$$

$$\mathbf{y} = f(\mathbf{x}) + \varepsilon$$

$$\tilde{\mathbf{y}} = \mathbf{X}\beta.$$

$$\mathbb{E}(y_i) = \sum_j x_{ij}\beta_j = \mathbf{X}_{i,*}\,\beta$$

$$Var(y_i) = \sigma^2$$

$$Var(\hat{\beta}) = \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}.$$

Math Behind the Confidence Interval section

We can assume that y follows some function f with some noise ϵ

$$\mathbf{y} = f(\mathbf{x}) + \epsilon$$

$$\mathbb{E}[\mathbf{y}] = \mathbb{E}[f(\mathbf{x}) + \epsilon] = \mathbb{E}[f(\mathbf{x})] + \mathbb{E}[\epsilon_i]$$

The expected value of ϵ_i is 0, $f(x)$ is a non-stochastic variable and is approimated by $\mathbf{X}\beta$

$$\mathbb{E}[y_i] = \mathbf{X}_{\mathbf{i},*}\beta$$

The variance is defined as

$$\begin{aligned}
Var(y_i) &= \mathbb{E}\left[(y_i - \mathbb{E}[y_i])^2\right] = \mathbb{E}\left[y_i^2 - 2y_i\mathbb{E}[y_i] + \mathbb{E}[y_i]^2\right] = \mathbb{E}[y_i^2] - \mathbb{E}[y_i]^2 \\
Var(y_i) &= \mathbb{E}\left[\left(\mathbf{X}_{i,*}\beta\right)^2 + 2\epsilon\mathbf{X}_{i,*}\beta + \epsilon^2\right] - (\mathbf{X}_{i,*}\beta)^2 \\
Var(y_i) &= (\mathbf{X}_{i,*}\beta)^2 + 2\mathbb{E}[\epsilon]\mathbf{X}_{i,*}\beta + \mathbb{E}[\epsilon^2] - (\mathbf{X}_{i,*}\beta)^2 \\
Var(y_i) &= \mathbb{E}[\epsilon^2] = \sigma^2
\end{aligned}$$

for the expected value of β we can insert the definition of β from earlier

$$\mathbb{E}[\hat{\beta}] = \mathbb{E}[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{E}[\mathbf{Y}] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\beta = \beta$$

$$
\begin{aligned}
Var(\hat{\beta}) &= \mathbb{E}\{[\beta - \mathbb{E}(\beta)][\beta - \mathbb{E}(\beta)]^T\} \\
&= \mathbb{E}\{[(\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\mathbf{y} - \beta]\,[(\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\mathbf{y} - \beta]^T\} \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\,\mathbb{E}\{\mathbf{y}\,\mathbf{y}^T\}\,\mathbf{X}\,(\mathbf{X}^T\mathbf{X})^{-1} - \beta\,\beta^T \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\,\{\mathbf{X}\,\beta\,\beta^T\,\mathbf{X}^T + \sigma^2\}\,\mathbf{X}\,(\mathbf{X}^T\mathbf{X})^{-1} - \beta\,\beta^T \\
&= \beta\,\beta^T + \sigma^2\,(\mathbf{X}^T\mathbf{X})^{-1} - \beta\,\beta^T \quad = \quad \sigma^2\,(\mathbf{X}^T\mathbf{X})^{-1},
\end{aligned}
$$

ridge

$$
\mathbb{E}[\hat{\boldsymbol{\beta}}^{\text{Ridge}}] = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}(\mathbf{X}^\top\mathbf{X})\boldsymbol{\beta}.
$$

$$
Var[\hat{\boldsymbol{\beta}}^{\text{Ridge}}] = \sigma^2[\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}]^{-1}\mathbf{X}^T\mathbf{X}\{[\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}]^{-1}\}^T,
$$

$$
\mathbb{E}[\hat{\boldsymbol{\beta}}^{\text{Ridge}}] = \mathbb{E}\left[\left(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_{pp}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}\right] = \left(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_{pp}\right)^{-1}\boldsymbol{X}^T\mathbb{E}[\mathbf{Y}] = \left(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_{pp}\right)^{-1}\left(\mathbf{X}^T\mathbf{X}\right)\boldsymbol{\beta}
$$

$$
\begin{aligned}
Var(\hat{\boldsymbol{\beta}}^{\text{Ridge}}) &= \mathbb{E}\{[\boldsymbol{\beta} - \mathbb{E}(\boldsymbol{\beta})][\boldsymbol{\beta} - \mathbb{E}(\boldsymbol{\beta})]^T\} \\
&= \mathbb{E}\left\{\left[\left(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_{pp}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y} - \left(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_{pp}\right)^{-1}\left(\mathbf{X}^T\mathbf{X}\right)\boldsymbol{\beta}\right]\left[\left(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_{pp}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y} - \left(\boldsymbol{X}^T\right.\right.\right. \\
&= \mathbb{E}\left\{\left(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_{pp}\right)^{-1}\boldsymbol{X}^T\,\mathbf{y}\,\mathbf{y}^T\,\boldsymbol{X}\,\left(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_{pp}\right)^{-1} - \left(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_{pp}\right)^{-1}\left(\mathbf{X}^T\mathbf{X}\right)\boldsymbol{\beta}\,\boldsymbol{\beta}^T\left(\mathbf{X}^T\mathbf{X}\right.\right. \\
&= \left(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_{pp}\right)^{-1}\boldsymbol{X}^T\,\mathbb{E}\{\mathbf{y}\,\mathbf{y}^T\}\,\mathbf{X}\,\left(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_{pp}\right)^{-1} - \left(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_{pp}\right)^{-1}\left(\mathbf{X}^T\mathbf{X}\right)\boldsymbol{\beta}\,\boldsymbol{\beta}^T\left(\mathbf{X}^T\mathbf{X}\right. \\
&= \left(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_{pp}\right)^{-1}\boldsymbol{X}^T\,\left(\mathbf{X}\,\boldsymbol{\beta}\,\boldsymbol{\beta}^T\,\mathbf{X}^T + \sigma^2\mathbf{I}\right)\,\mathbf{X}\,\left(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_{pp}\right)^{-1} - \left(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_{pp}\right)^{-1}\left(\mathbf{X}^T\mathbf{X}\right)
\end{aligned}
$$

Bibliography

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1 edition, 2007.

[2] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. Book in preparation for MIT Press.

[3] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[4] Wessel N. van Wieringen. Lecture notes on ridge regression, 2023.

[5] Brage Wiseth, Eirik ?, and Felix ? MachineLearningProjects. `https://github.com/bragewiseth/MachineLearningProjects`, September 2023.