

Project 1 FYS-STK4155

Mia Synnøve Frivik, Andrea Myrvang, Max Jan Willem Schuringa, Janita Ovidie Sandtrøen Willumsen
(Dated: September 27, 2023)

The aim of this report is to see how different regression method affects the data it is applied to. More concretely, we will look at the three different methods ordinary least squares (OLS), Ridge and LASSO. We will also apply bias variance trade of as well as cross validation on the data sets used to evaluate our models. What we found was that ... regression with parameter ... best fitted the topological data analysis...

I. INTRODUCTION

Machine learning is a powerful tool usfull in many fields of reasarche. One illustration of its utility is its application to terrain data analysis. Through the creation of terrain models from real data of a specific geographic area, one can effectively anticipate high-risk avalanche zones[4], potentially leading to life-saving interventions. This methods can extends to addressing concerns related to floods, which has become a hot topic this past month following the storm Hans. It can also help in aiding with spatial planning challenges. which is usefull in big citys all over the world. It is fair to say machine learning possesses immense potential to contribute to the solutions of complex and relevant challenges in our modern society, encompassing climate-related issues, urban planning, and life-saving endeavors.

In this report er are going to study three different regression methods, ordinary least squares (OLS), Ridge and LASSO and see how these method compare to eachother when applied to different data sets. First we are going to use the Franke function to make dummy data to validate if our models works. When plotted in the interval $[0, 1]$ this function looks like a mountain and a valley, which is a perfect starting point when we later want to apply these methods on real digital terrain data taken from <https://earthexplorer.usgs.gov/>. To more accurately simulate the realism of practical machine learning scenarios, we will impose limitations on our datasets. Additionally, we will employ techniques such as bootstrapping and cross-validation to expand our dataset size and assess their impact on model validation.

II. THEORY

A. Regression methods

While numerous regression methods exist, in this particular context, we will narrow our focus to three fundamental techniques, Ordinary Least Squares, Ridge and Lasso Regression."

1. Ordinary least squares (OLS)

Ordinary Least Squares (OLS) stands as a fundamental technique in regression analysis, where the aim is to create a model that minimize the diffrence from the observed data and the predicted model.

A linear regression model has the following form :

$$f(\mathbf{X}) = \beta_0 + \sum_{i=1}^p X_i \beta_i [2] \quad (1)$$

Where β is the coefficients and \mathbf{X} the design matrix. It is important to note that thise method assume that either the function is linear or approxematly linear.

The cost function for OLS is given by:

$$C(\beta) = \frac{1}{n} = \{(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)\} \quad (2)$$

From this we obtain that the expression for the optimal β is given by:

$$\beta = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (3)$$

2. Ridge

$$\beta = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y} \quad (4)$$

Calculations for bothe the optimal β and variance can be found in the appendix

3. LASSO

B. MSE

C. Resampling techniques

The main restriction in machine learning is the amount of data points available to create the model out of. It may be the case where one have done a costfull and time consuming experiments and are left with a small number of data. It is therfull extremely useful to have methods where one can reuse the data multiple times thereby creating a relatively large dataset from the small number of

datapoints. In this report we are going to use two different methods, the first is called bootstrap and the second one is cross validation.

1. Bootstrap

The bootstrap method is a resampling procedure that uses data from one sample to generate a sampling distribution by repeatedly taking random samples from the known sample, with replacement[3] This means that if we have a data set D with n data points. The elements in this data set can be represented in the following way:

$$D = d_1, d_2, d_3, d_4, \dots, d_n \quad (5)$$

Then by applying the bootstrap method on this data set one possible output D^* can be:

$$D^* = d_3, d_n, d_4, d_4, \dots, d_2 \quad (6)$$

From this example we see that one observation can appear multiple times in the new dataset. We can take this method a step further and create "new" data points by extracting multiple data points from the dataset and take the mean of all these values:

$$d_{new} = \frac{1}{k}(d_1, \dots, d_k) \quad (7)$$

This gives us a method of producing lots of "new" dataset from limited data points to train our model with. For each of these data-sets the mean and standard deviation can be calculated to evaluate the model statistically. [1]

One huge advantage of using the bootstrap method is that the data can be split into test and train before shuffling the data, this means that the test data can be kept entirely separate from the creation of the model. When we then test the model it will be on a dataset that has nothing to do with creating the model and will therefore show how good the model represents real data. **Write something about large numbers law, also disadvantages and advantage**

2. Cross validation

Cross validation is another method of creating "new" datasets from the original data. This method works by splitting the data in k -folds

D. Bias-variance trade-off

The Bias-variance trade-off is a measurement on how accurately a model fits the real data.

III. METHOD

In the first part of this project a function called Franke function was used as the data analysed. The Franke function is given by the following equation:

$$\begin{aligned} f(x, y) = & \frac{3}{4} \exp\left(-\frac{(9x-2)^2}{4} - \frac{(9y-2)^2}{4}\right) \\ & + \frac{3}{4} \exp\left(-\frac{(9x+1)^2}{49} - \frac{(9y+1)}{10}\right) \\ & + \frac{1}{2} \exp\left(-\frac{(9x-7)^2}{4} - \frac{(9y-3)^2}{4}\right) \\ & - \frac{1}{5} \exp(-(9x-4)^2 - (9y-7)^2) \end{aligned}$$

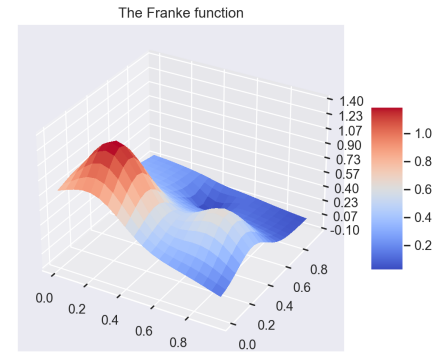


Figure 1. A plot of the Franke function

This function was fitted with the OLS method, where a polynomial with degree 5 was used to create the design matrix. Since the design matrix in this case was noninvertible, singular value decomposition was used to create the β -values needed to create a model of the dataset. The mean square error and the R^2 score were calculated for both the testing and training datasets.

Next Ridge regression was used on the Franke function, to see if this method had a better fit than what was obtained with OLS. Different values for λ were used to obtain the best fit as possible.

IV. RESULTS

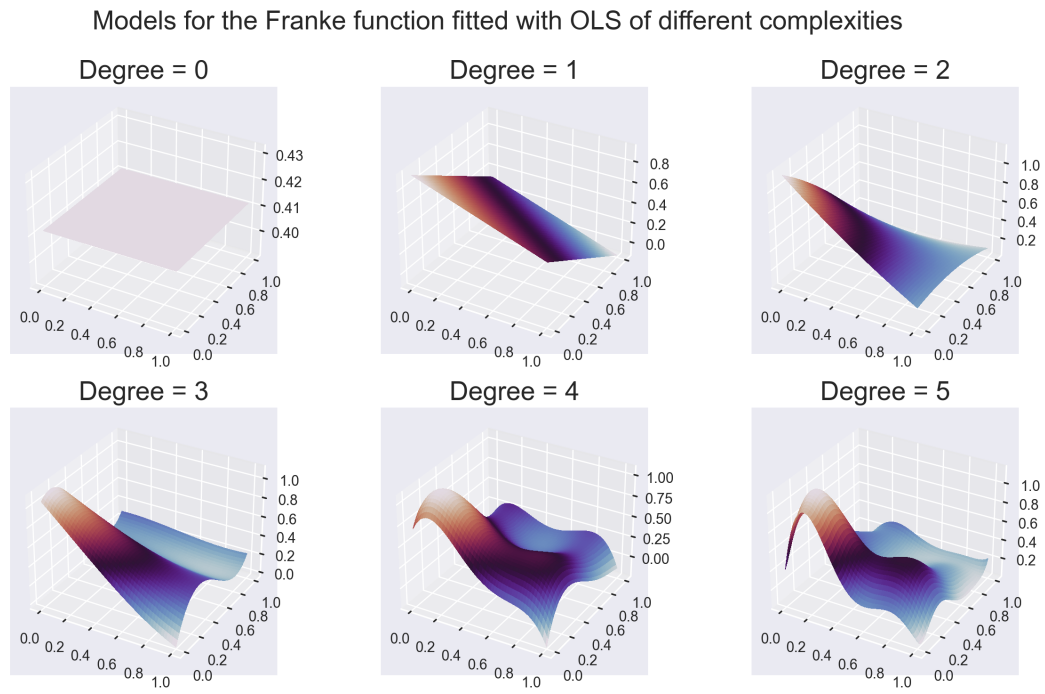


Figure 2. A plot showing how model with different complexities fit the franke function when OLS recession has been used.

V. DISCUSSION

VI. CONCLUSION

REFERENCES

-
- [1] Jason Brownlee. A Gentle Introduction to the Bootstrap Method. <https://machinelearningmastery.com/a-gentle-introduction-to-the-bootstrap-method/>, 2019.
 - [2] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edition, 2009.
 - [3] Eberly College of Science. Bootstrapping Methods. <https://online.stat.psu.edu/stat500/lesson/11/11.2/11.2.1>.
 - [4] Hong Wen, Xiyong Wu, Xin Liao, Dong Wang, Kaiyang Huang, and Bernd Wünnemann. Application of machine learning methods for snow avalanche susceptibility mapping in the parlung tsangpo catchment, southeastern qinghai-tibet plateau. *Cold Regions Science and Technology*, 198:103535, 2022.

Appendix A: Mean values and variances calculations

The main regression method used in this report is the ordinary least squares method. This appendix shows the calculations for some of the equations used to produce the results shown in this report.

We have assumed that our data can be described by the continuous function $f(\mathbf{x})$, and an error term $\epsilon \sim N(0, \sigma^2)$. If we approximate the function with the solution derived from a model $\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta}$ the data can be described with $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. The expectation value

$$\begin{aligned}\mathbb{E}(\mathbf{y}) &= \mathbb{E}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\ &= \mathbb{E}(\mathbf{X}\boldsymbol{\beta}) + \mathbb{E}(\boldsymbol{\epsilon}) && \text{where the expected value } \boldsymbol{\epsilon} = 0 \\ \mathbb{E}(y_i) &= \sum_{j=0}^{P-1} X_{i,j}\beta_j && \text{for the each element} \\ &= X_{i,*}\beta_i && \text{where } * \text{ replace the sum over index } i\end{aligned}$$

The variance for the element y_i can be found by

$$\begin{aligned}\mathbb{V}(y_i) &= \mathbb{E}[(y_i - \mathbb{E}(y_i))^2] \\ &= \mathbb{E}(y_i^2) - (\mathbb{E}(y_i))^2 \\ &= \mathbb{E}((X_{i,*}\beta_i + \epsilon_i)^2) - (X_{i,*}\beta_i)^2 \\ &= \mathbb{E}((X_{i,*}\beta_i)^2 + 2\epsilon_i X_{i,*}\beta_i + \epsilon_i^2) - (X_{i,*}\beta_i)^2 \\ &= \mathbb{E}((X_{i,*}\beta_i)^2) + \mathbb{E}(2\epsilon_i X_{i,*}\beta_i) + \mathbb{E}(\epsilon_i^2) - (X_{i,*}\beta_i)^2 \\ &= (X_{i,*}\beta_i)^2 + \mathbb{E}(\epsilon_i^2) - (X_{i,*}\beta_i)^2 \\ &= \mathbb{E}(\epsilon_i^2) = \sigma^2\end{aligned}$$

The expression for the optimal parameter

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

We find the expected value of $\hat{\boldsymbol{\beta}}$

$$\begin{aligned}\mathbb{E}(\hat{\boldsymbol{\beta}}) &= \mathbb{E}((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}) \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbb{E}(\mathbf{y}) && \text{using that } \mathbf{X} \text{ is a non-stochastic variable} \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \boldsymbol{\beta} && \text{using } \mathbb{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta}\end{aligned}$$

we can find the variance by

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

Appendix B: Bias-variance trade-off