FYS-STK4155 - Project1

Magnus Kristoffersen, Markus Bjørklund

September 2023

Abstract

We investigated the Mean Squared Error (MSE) and R^2 score for Ordinary Least Squares (OLS), Ridge and Lasso regression, and found that the error decreased as a function of complexity, here represented by polynomial order. According to our results, Ordinary Least squares fit the data best. As expected, OLS was closely followed by Ridge regression for small values λ, as this Ridge regression approaches OLS in the limit $\lambda \to 0$. We implemented the Bootstrap and Cross-Validation resampling methods, which we found to increase the robustness of the MSE measure.

1 Introduction

The evaluation of different models with respect to computational cost and accuracy is a recurring theme within machine learning. In this project, we aim to investigate different types of linear regression models, with varying degrees of complexity, and compare them. A motivating question for such an inquiry can be "what is the relationship between computational cost and prediction accuracy, and will more complex models always perform better?". Generally, one can expect a higher variance and a lower bias when increasing model complexity [1]. The relationship between model complexity and error measures can be summarized with the concepts of under and overfitting, where you either tune your model too tightly to the training examples, making it worse at generalization (overfit), or your model is too simple to capture all the features in the data (underfit).

We will investigate three different models: Ordinary Least Squares (OLS), Ridge regression and Lasso regression. We will also look into bootstrapping

and cross-validation methods for increasing the stability of our models. Section 2 will contain the underlying theory, and a description of our framework. We present our results in section section 3, which are further discussed in section 4. Finally, we provide our concluding remarks in section 5.

2 Method

2.1 Theory

2.1.1 The Franke Function

We will begin by validating our methods with the Franke function. This function has been widely used as a benchmarking tool for fitting and interpolation algorithms [2]. The Franke function reads as follows,

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

2.1.2 Analytical expressions for statistical measures

Our original assumption is that $\mathbf{y}$ is given by

$$\mathbf{y} = f(x) + \epsilon, \tag{2}$$

where $f(x)$ is a continuous function and ϵ is Gaussian distributed. Now, we approximate this function $f(x)$ by $f(x) \approx \tilde{\mathbf{y}} = \mathbf{X}\beta$, such that

$$\mathbf{y} \approx \mathbf{X}\beta + \epsilon, \tag{3}$$

or

$$y_i \approx \sum_j X_{ij}\beta_j + \epsilon_i = \mathbf{X}_i \cdot \beta + \epsilon_i. \tag{4}$$

Now, if we take the expectation value of this, we get

$$\begin{aligned}\mathbb{E}\left[y_i\right] &\approx \mathbb{E}\left[\mathbf{X}_i \cdot \beta\right] + \mathbb{E}\left[\epsilon_i\right] \\ &= \mathbf{X}_i \cdot \beta, \end{aligned} \tag{5}$$

because $\mathbf{X}$ and β are assumed to be non-stochastic variables (enforced by design), and $\mathbb{E}\left[\epsilon\right] = 0$ by the properties of a normal distribution with zero mean.

For the variance, we have that

$$\begin{aligned} Var\left[y_i\right] &= \mathbb{E}\left[\left(y_i - \mathbb{E}\left[y_i\right]\right)^2\right] \\ &= \mathbb{E}\left[y_i^2 - 2y_i\mathbb{E}\left[y_i\right] + \mathbb{E}\left[y_i\right]^2\right] \\ &= \mathbb{E}\left[\left(\mathbf{X}_i \cdot \beta + \epsilon_i\right)^2 - 2\left(\mathbf{X}_i \cdot \beta + \epsilon_i\right)\mathbf{X}_i \cdot \beta + \left(\mathbf{X}_i \cdot \beta\right)^2\right] \\ &= \mathbb{E}\left[\left(\mathbf{X}_i \cdot \beta\right)^2 + 2\mathbf{X}_i \cdot \beta\epsilon + \epsilon^2 - 2\left(\mathbf{X}_i \cdot \beta\right)^2 - 2\mathbf{X}_i \cdot \beta\epsilon + \left(\mathbf{X}_i \cdot \beta\right)^2\right] \\ &= \mathbb{E}\left[\epsilon^2\right] \\ &= \sigma^2 \end{aligned} \tag{6}$$

For the expectation value of $\hat{\beta}$, we note that since $\mathbf{X}$ is non-stochastic, the inverse $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$ and product with $\mathbf{X}^T$ is also non-stochastic, so

$$\begin{aligned}\mathbb{E}\left[\hat{\beta}\right] &= \mathbb{E}\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}\right] \\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbb{E}\left[\mathbf{y}\right] \\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\left(\mathbb{E}\left[\mathbf{X}\beta\right] + \mathbb{E}\left[\epsilon\right]\right) \\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\beta \\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\tilde{\mathbf{y}} \\ &= \beta \end{aligned} \tag{7}$$

Finally, the variance of $\hat{\beta}$, given by

$$\begin{aligned} Var\left[\hat{\beta}\right] &= \mathbb{E}\left[\left(\hat{\beta} - \mathbb{E}\left[\hat{\beta}\right]\right)^2\right] \\ &= \mathbb{E}\left[\hat{\beta}^2 - 2\hat{\beta}\beta + \beta^2\right] \end{aligned} \tag{8}$$

Now let's do some intermediate, simplifying calculations:

$$\begin{aligned}\hat{\beta} &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\left(\mathbf{X}\beta+\epsilon\right) \\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\tilde{\mathbf{y}} + \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\epsilon \\ &= \beta + \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\epsilon \end{aligned} \tag{9}$$

$$\hat{\beta}^2 = \beta^2 + 2\beta\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\epsilon + \left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\epsilon\right]^2 \tag{10}$$

$$-2\hat{\beta}\beta = -2\beta^2 - 2\beta\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\epsilon \tag{11}$$

Putting it all together we get

$$\begin{aligned}Var\left[\hat{\beta}\right] &= \mathbb{E}\left[\beta^2 + 2\beta\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\epsilon + \left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\epsilon\right]^2 - 2\beta^2 - 2\beta\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\epsilon + \beta^2\right] \\ &= \mathbb{E}\left[\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\epsilon\right)^2\right] \\ &= \mathbb{E}\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\epsilon\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\epsilon\right)^T\right] \\ &= \mathbb{E}\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\epsilon\epsilon^T\mathbf{X}^T\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right)^T\right] \\ &= \mathbb{E}\left[\epsilon^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\left(\left(\mathbf{X}^T\mathbf{X}\right)^T\right)^{-1}\right] \\ &= \mathbb{E}\left[\epsilon^2\right]\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} \\ &= \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\end{aligned} \tag{12}$$

assuming that all compounds of $\mathbf{X}$ are non-stochastic.

2.1.3 The Bias-variance trade-off

$$C\left(\mathbf{X},\beta\right) = \frac{1}{n}\sum_{i=0}^{n-1}\left(y_i - \tilde{y}_i\right)^2 = \mathbb{E}\left[\left(\mathbf{y}-\tilde{\mathbf{y}}\right)^2\right] \tag{13}$$

Now, writing out the terms,

$$\begin{aligned}\mathbb{E}\left[\left(\mathbf{y}-\tilde{\mathbf{y}}\right)^2\right] &= \mathbb{E}\left[\mathbf{y}^2 - 2\mathbf{y}\tilde{\mathbf{y}} + \tilde{\mathbf{y}}^2\right] \\ &= \mathbb{E}\left[\mathbf{y}^2\right] - 2\mathbb{E}\left[\mathbf{y}\tilde{\mathbf{y}}\right] + \mathbb{E}\left[\tilde{\mathbf{y}}^2\right] \end{aligned} \tag{14}$$

Term by term, and simplifying notation with $f = f(\mathbf{x})$, we have that

$$\begin{aligned}\mathbb{E}\left[\mathbf{y}^2\right] &= \mathbb{E}\left[(f+\epsilon)^2\right] \\ &= \mathbb{E}\left[f^2 + 2f\epsilon + \epsilon^2\right] \\ &= \mathbb{E}\left[f^2\right] + 2\mathbb{E}\left[f\epsilon\right] + \mathbb{E}\left[\epsilon^2\right] \\ &= f^2 + f\underbrace{\mathbb{E}\left[\epsilon\right]}_{=0} + \sigma^2 \\ &= f^2 + \sigma^2 \end{aligned} \tag{15}$$

$$\begin{aligned}\mathbb{E}\left[\mathbf{y}\tilde{\mathbf{y}}\right] &= \mathbb{E}\left[(f+\epsilon)\,\tilde{\mathbf{y}}\right] \\ &= \mathbb{E}\left[f\tilde{\mathbf{y}}\right] + \mathbb{E}\left[\epsilon\tilde{\mathbf{y}}\right] \\ &= f\mathbb{E}\left[\tilde{\mathbf{y}}\right] + \mathbb{E}\left[\tilde{\mathbf{y}}\right]\underbrace{\mathbb{E}\left[\epsilon\right]}_{=0} \\ &= f\mathbb{E}\left[\tilde{\mathbf{y}}\right] \end{aligned} \tag{16}$$

$$\mathbb{E}\left[\tilde{\mathbf{y}}^2\right] = \mathrm{Var}\left[\tilde{\mathbf{y}}\right] + \mathbb{E}\left[\tilde{\mathbf{y}}\right]^2 \tag{17}$$

where the last term is simply from the definition of variance, $\mathrm{Var}\left[\tilde{\mathbf{y}}\right] = \mathbb{E}\left[\tilde{\mathbf{y}}^2\right] - \mathbb{E}\left[\tilde{\mathbf{y}}\right]^2$. Putting everything together, we get

$$\begin{aligned}\mathbb{E}\left[\left(\mathbf{y}-\tilde{\mathbf{y}}\right)^2\right] &= f^2 + \sigma^2 - 2f\mathbb{E}\left[\tilde{\mathbf{y}}\right] + \mathrm{Var}\left[\tilde{\mathbf{y}}\right] + \mathbb{E}\left[\tilde{\mathbf{y}}\right]^2 \\ &= \left(f^2 - 2f\mathbb{E}\left[\tilde{\mathbf{y}}\right] + \mathbb{E}\left[\tilde{\mathbf{y}}\right]^2\right) + \mathrm{Var}\left[\tilde{\mathbf{y}}\right] + \sigma^2 \end{aligned} \tag{18}$$

Looking at all the terms inside the paranthesis, we can see that since f is non-stochastic we have that

$$\left(f^2 - 2f\mathbb{E}\left[\tilde{\mathbf{y}}\right] + \mathbb{E}\left[\tilde{\mathbf{y}}\right]^2\right) = \mathbb{E}\left[\left(f^2 - 2f\mathbb{E}\left[\tilde{\mathbf{y}}\right] + \mathbb{E}\left[\tilde{\mathbf{y}}\right]^2\right)\right] = \mathbb{E}\left[\left(f - \mathbb{E}\left[\tilde{\mathbf{y}}\right]\right)^2\right] \tag{19}$$

and we are left with

$$\mathbb{E}\left[(\mathbf{y}-\tilde{\mathbf{y}})^2\right] = \mathbb{E}\left[(f-\mathbb{E}\left[\tilde{\mathbf{y}}\right])^2\right] + \mathrm{Var}\left[\tilde{\mathbf{y}}\right] + \sigma^2 \tag{20}$$

3 Results

4 Discussion

5 Conclusion

References

[1] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer Verlag, Berlin, 2009.

[2] Morten Hjort-Jensen. *Project 1 on Machine Learning*. Sept. 2023. URL: https://github.com/CompPhysics/MachineLearning/blob/master/doc/Projects/2023/Project1/pdf/Project1.pdf.