Project 1

Knut Magnus Aasrud
Philip Karim Niane

September 16, 2020

Abstract

Abstract

1 Introduction

The goal of this project is to explore a number of different methods which is quiet popular within the field of machine learning. The methods that will be explored are the ordinary least square method (OLS), ridge regression and LASSO regression which will be used to find unknown parameters when performing interpolation and fitting tasks. The methods will be tested on a two-dimensional function called Franke's function.
After the methods are settled and ready to go, a few different resampling techniques will be combined with the methods to evaluate the effectiveness. The resampling techniques to be used are the bootstrap method and the cross-validation method. The Bias-Variance trade off will also be studied in the article.
To finish of the project, the different methods will be tested on digital terrain data which descends from real life landscape.

2 Theory

2.1 Franke's function

Franke's function is a function which is often used to test different regression and interpolation methods. The function has two Gaussian peaks of differing heights, and a smaller dip. It's expression is

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

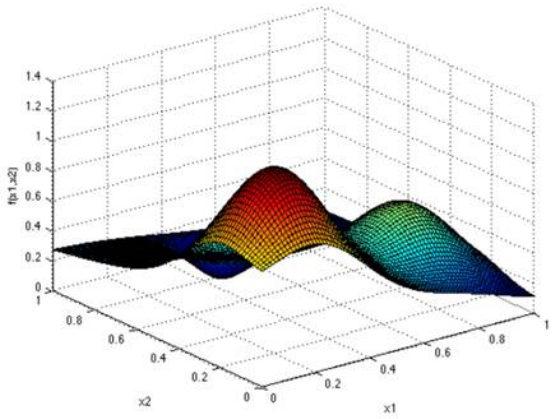

Figure 1: Shape of Franke's function which will be used as a interpolating goal. Note to self: Might be better to plot the graph instead of using a figure from the web. Remember to add the source for the figure by using for example bib. https://www.sfu.ca/ ssurjano/franke2d.html

and we define it for the interval $x, y \in [0, 1]$. An illustration of Franke's function can be seen in figure 1

2.2 Linear regression

The goal of a linear regression model is to find the coefficients $\hat{\beta}$ best suited for predicting new data via this expression:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta},$$

where $\mathbf{X}$ is the design matrix constructed from the input data, and $\hat{\mathbf{y}}$ is the resulting prediction. To achieve this, we define a "cost function" of sorts, that evaluates each coefficients ability to predict the initial training dataset. We then find the $\hat{\beta}$ that minimizes this cost function. More formally put, we want to find

$$\hat{\beta} = \arg\min_{\beta} C(\beta), \tag{1}$$

where $C(\beta)$ is the cost function.

2.2.1 Ordinary least squares regression (L_0)

Ordinary least squares (OLS) regression uses the residual sum of squares (RSS) function as the cost function. Given N datapoints and the predicted output $\mathbf{y}$,

it reads

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2. \tag{2}$$

As found in appendix 6.2, a cost function like (2) gives the following matrix equation for $\hat{\beta}$

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

OLS regression has very low variance, but high bias - as a consequence of the bias-variance tradeoff. This makes OLS regression an "accurate" predictor of its own training data, but susceptible to overfitting, which the following two models are better suited to handle.

2.2.2 Lasso regression (L_1)

Lasso regression expands upon the above cost function by adding a term that penalizes the size of each coefficient. This is done by a factor of λ, as shown here:

$$\hat{\beta} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{4}$$

The first sum is from the RSS function, while the second sum is the imposed penalty. The Lasso decreases bias from the OLS model, by decreasing the size of the coefficients, and thus making the variables more equally weighted.

Lasso regression has no closed form expression for $\hat{\beta}$, which means it must be calculated programatically.

2.2.3 Ridge regression (L_2)

Ridge regression has a penalty corresponding to the coefficient's squared sizes, further decreasing the bias from The Lasso, but obviously also increases variance. It defines $\hat{\beta}$ like this

$$\hat{\beta}^{\mathrm{ridge}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\},$$

which - by the same proceedure as shown in appenix 6.2, has this closed form expression

$$\hat{\beta}^{\mathrm{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}. \tag{5}$$

3 Results

4 Discussion

Discussion

5 Conclusion

Conclusion

6 Appendix

6.1 Source code

All the source code is located in this GitHub repository.

6.2 L_0 regression on matrix form

The cost function we use for OLS regression is the residual sum of squares (RSS) function:

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 .$$

Changing into matrix notation, we get

$$\mathrm{RSS}(\beta) = (\mathbf{y} - \hat{\mathbf{y}})^T (\mathbf{y} - \hat{\mathbf{y}}) = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta),$$

which we can differentiate with respect to β to find the minimum.

$$\frac{\partial \mathrm{RSS}}{\partial \beta} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta).$$

Assuming full column rank for $\mathbf{X}$, $(\mathbf{X}^T\mathbf{X})$ is thus positive definite (and importantly, invertible). Setting the first derivative to 0, we get

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0$$

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