Bias-variance tradeoff in simple linear models

Anna Aasen, Carl Martin Fevang, Håkon Kvernmoen

September 20, 2022

Abstract

This is the abstract.

1 Introduction

In this project, we will introduce some of the most important general properties of machine learning models in the context of linear regression. Specifically the mean squared error (MSE), the bias and the variance of data predictions are essential in evaluating the quality of a statistical model. We look at the phenomena of underfitting and overfitting, and explore them in the context of bias-variance trade-off. The three models we will use explicitly are Ordinary Linear Regression (OLS), Ridge regression and Least-Absolute-Shrinkage-and-Selection-Operation (LASSO) regression. These will first be tested on the 2D Franke function, and later on real geographical terrain data.
The aforementioned models will be explored using resampling techniques, which will be employed to extract the relevant statistical quantities in an accurate way. We will use the standard methods of Bootstrapping and Cross Validation and compare the results we get between them across the three linear models.
All the models will be applied by fitting a two-dimensional polynomial expansion to the observed data.

Figure 1: Some figure caption

2 Theory

2.1 Statistical introduction to linear regression

A way to motivate the process of statistical learning is that we want to develop a process by which we use observed data to inform our belief in a future outcome. Such a process can be developed from the ideas of Bayes' theorem, which we can use to inform us what confidence we should have in the parameters of a given model, given data points we have observed.
Let $\boldsymbol{y}$ denote a vector of a series of measured values y_i at points X, where X is a matrix where the rows x_i correspond to the input values for the measurement y_i. Further let $\boldsymbol{\theta}$ denote the parameters of our given model – then Bayes' theorem tells us that our *posteriori* confidence in the parameters $\boldsymbol{\theta}$ given the available data follows

$$P(\boldsymbol{\theta}|X,\boldsymbol{y}) = \frac{P(\boldsymbol{y}|X,\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(X,\boldsymbol{y})}. \tag{1}$$

Here $P(X,\boldsymbol{y})$ plays the role of a normalisation, and can safely be ignored – either because we look for values of $\boldsymbol{\theta}$ that maximise the probability, or because we assume the probabilities $P(\boldsymbol{y}|X,\boldsymbol{\theta})$ and $P(\boldsymbol{\theta})$ which we then can properly normalise. Now we have a framework through which we can use assumptions about the distributions $P(\boldsymbol{y}|X,\boldsymbol{\theta})$ and $P(\boldsymbol{\theta})$ to give us a probability distribution for the parameters $\boldsymbol{\theta}$ that we can maximise.

2.1.1 Ordinary least squares

Now we can define the ordinary least squares method by assuming that the data y that we want to make a fit to has a linear noise term that is normally distributed with mean nought. That is, the data is generated as $y(x) = f(x) + \epsilon$, where $f(x)$ is some analytic, non-stochastic function of the data input x, and $\epsilon \sim N(0,\sigma^2)$. Further, if we make a model fitting $y(x)$ by a $\tilde{y}(x)$, we should expect the error at every data point (y_i, x_i) to be be normally distributed with the same variance, i.e., independently and identically distributed (i.i.d.). This can be summarised

$$P(\boldsymbol{y}|X,\boldsymbol{\theta}) = \prod_{i=0}^{n-1} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\tilde{y}_i)^2}{2\sigma^2}}. \tag{2}$$

Furthermore, we assume the parameters $\boldsymbol{\theta}$ to be distributed evenly, and as such, $P(\boldsymbol{\theta})$ just contributed to an overall normalisation. This means that maximising $P(\boldsymbol{y}|X,\boldsymbol{\beta})$ amounts to the same as maximising $P(\boldsymbol{\theta}|X,\boldsymbol{y})$, and can be done analytically.
Knowing that the Gaussian distribution has a single extremum that is a maximum, we can find the values for $\boldsymbol{\theta}$ maximising the probability as

$$\begin{aligned} \frac{\mathrm{d}P(\boldsymbol{y}|X,\boldsymbol{\theta})}{\mathrm{d}\theta_k} &= 0 \\ \Rightarrow \prod_{i=0}^{n-1} \frac{1}{\sqrt{2\pi\sigma^2}} \frac{1}{\sigma^2} X_{ik}(y_i - X_{ij}\theta_j) e^{-\frac{(y_i - X_{ij}\theta_j)^2}{2\sigma^2}} &= 0, \end{aligned} \tag{3}$$

which as we can see, is maximisable point by point. This means that the optimal parameters $\hat{\boldsymbol{\theta}}_{\text{OLS}}$ that maximise $P(\boldsymbol{y}|X,\boldsymbol{\theta})$ can be found as

$$\boxed{\hat{\boldsymbol{\theta}}_{\text{OLS}} = \left(X^T X\right)^{-1} X^T \boldsymbol{y}.} \tag{4}$$

2.2 Ridge and LASSO regression

So far, we have assumed that the parameters $\boldsymbol{\theta}$ are uniformly distributed, and as such have no *bias* towards any particular value. In practice, this means that the OLS model will contort to fit itself to all values in the data. As such, a motivation for adding a bias could be for the sake of stability in the fit. This is done in ridge regression, where one assumes that the parameters are normally distributed, such that parameter values far from the mean are thought less likely to occurr. This in practice means that the model is less willing to contort to outliers in the dataset, trading it for a bias towards certain parameter values. Assuming the probability distrbution for $\boldsymbol{\theta}$

$$P(\boldsymbol{\theta}) = \prod_{i=0}^{p-1} \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{\theta_i^2}{2\tau^2}}, \tag{5}$$

with a derivative with respect to $\boldsymbol{\theta}$

$$\frac{\mathrm{d}P(\boldsymbol{\theta})}{\mathrm{d}\theta_k} = -\prod_{i=0}^{p-1} \frac{1}{\sqrt{2\pi\tau^2}} \frac{1}{\tau^2}\theta_k e^{-\frac{\theta_i^2}{2\tau^2}} = -\frac{1}{\tau^2}\theta_k P(\boldsymbol{\theta}), \tag{6}$$

the condition for maximising $P(\boldsymbol{\theta}|X,\boldsymbol{y})$ becomes

$$\begin{aligned} \frac{\mathrm{d}P(\boldsymbol{y}|X,\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}} P(\boldsymbol{\theta}) + P(\boldsymbol{y}|X,\boldsymbol{\theta}) \frac{\mathrm{d}P(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}} &= 0 \\ \left(\frac{1}{\sigma^2} X^T(\boldsymbol{y} - X\boldsymbol{\theta}) - \frac{1}{\tau^2}\boldsymbol{\theta}\right) P(\boldsymbol{y}|X,\boldsymbol{\theta}) P(\boldsymbol{\theta}) &= 0. \end{aligned} \tag{7}$$

Rewriting with a parameter $\lambda = \sigma^2/\tau^2$, the optimal parameters $\hat{\boldsymbol{\theta}}_{\text{ridge}}$ are given by

$$\boxed{\hat{\boldsymbol{\theta}}_{\text{ridge}} = \left(X^T X + \lambda\,\mathbb{I}\right)^{-1} X^T \boldsymbol{y}.} \tag{8}$$

This has the added bonus of ensuring that $X^T X + \lambda\,\mathbb{I}$ is always invertible, as $\det(\lambda\,\mathbb{I}) \neq 0$ for $\lambda > 0$.
LASSO regression builds on the same ideas as ridge regression, by assumes rather a Laplace distribution $L(0,\tau)$ for the parameters $\boldsymbol{\theta}$. This gives

$$P(\boldsymbol{\theta}) = \prod_{i=0}^{p-1} \frac{1}{2\tau} e^{-\frac{|\theta_i|}{\tau}}, \tag{9}$$

which as we can see is not differentiable at $\theta_i = 0$. This means that the optimisation of this the probability $P(\boldsymbol{\theta}|X,\boldsymbol{y})$ is not necessarily possible analytically without putting bounds on the data $X, \boldsymbol{y}$, and even numerical algorithms like gradient descent will encounter run into trouble. Furthermore, it is not optimisable point by point, unless bounds like $X^T X$ being diagonal are enforced, which makes LASSO generally less used. Nevertheless, we can coax a cost function out of the probability that is minimalisable. Letting the cost function be written $C_{\text{lasso}}(\boldsymbol{\theta}) = -A\log\left(P(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{y})\right) + B$ we can choose the constants A, B such that

$$\boxed{C_{\text{lasso}}(\boldsymbol{\theta}) = \|\boldsymbol{y} - X\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1,} \tag{10}$$

where $\lambda = \frac{2\sigma^2}{\tau}$ and $\|\boldsymbol{v}\|_p = \left(\sum_i |v_i|^p\right)^{1/p}$ denotes the L_p norm of the vector $\boldsymbol{v}$.

A Theoretical proof of concept of the OLS model

$$\mathbb{E}(y_i) = \mathbb{E}\left[X_{ij}\boldsymbol{\beta}_j + \epsilon_i\right] = X_{ij}\boldsymbol{\beta}_j + \mathbb{E}(\epsilon) = X_{ij}\boldsymbol{\beta}_j \tag{11}$$

$$\begin{aligned} \mathrm{Var}(y_i) &= \mathbb{E}\left[\left(y_i - \mathbb{E}\left(y_i\right)\right)^2\right] = \mathbb{E}\left[\left(X_{ij}\boldsymbol{\beta}_j + \epsilon_i - X_{ij}\boldsymbol{\beta}_j\right)^2\right] \\ &= \mathbb{E}\left(\epsilon_i^2\right) = \mathrm{Var}\left(\epsilon_i\right) = \sigma^2 \end{aligned} \tag{12}$$

$$\begin{aligned} \mathbb{E}(\hat{\boldsymbol{\beta}}) &= \mathbb{E}\left((X^T X)^{-1} X^T \boldsymbol{y}\right) = \mathbb{E}\left((X^T X)^{-1} X^T \left[X\boldsymbol{\beta} + \boldsymbol{\epsilon}\right]\right) \\ &= \mathbb{E}\left(\boldsymbol{\beta}\right) + \left(X^T X\right)^{-1} X^T \mathbb{E}\left(\boldsymbol{\epsilon}\right) = \boldsymbol{\beta} \end{aligned} \tag{13}$$

$$\begin{aligned} \mathrm{Var}(\hat{\boldsymbol{\beta}}) &= \mathbb{E}(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^T) - \mathbb{E}(\hat{\boldsymbol{\beta}})\,\mathbb{E}(\hat{\boldsymbol{\beta}}^T) \\ &= \mathbb{E}((X^T X)^{-1} X^T \boldsymbol{y}\boldsymbol{y}^T X (X^T X)^{-1}) - \boldsymbol{\beta}\boldsymbol{\beta}^{\boldsymbol{T}} \\ &= (X^T X)^{-1} X^T \left[X\boldsymbol{\beta}\boldsymbol{\beta}^T X^T + \sigma^2\,\mathbb{I}\right] X (X^T X)^{-1} - \boldsymbol{\beta}\boldsymbol{\beta}^T \\ &= \sigma^2 (X^T X)^{-1}, \end{aligned} \tag{14}$$

where we have used that $\mathbb{E}(\boldsymbol{y}\boldsymbol{y}^T) = \mathbb{E}\left((X\boldsymbol{\beta} + \boldsymbol{\epsilon})(X\boldsymbol{\beta} + \boldsymbol{\epsilon})^T\right) = X\boldsymbol{\beta}\boldsymbol{\beta}^T X^T + \sigma^2\,\mathbb{I}$.