Analysis of Regression and Resampling Methods

Håkon Olav Torvik, Vetle Vikenes and Sigurd Sørlie Rustad

University of Oslo
Norway
September 30, 2021

CONTENTS

I. INTRODUCTION

Regression analysis is a statistical method for fitting a function to data. It is useful for building mathematical models to explain observations. There are several regression methods to achieve this, all with their strengths and weaknesses. We will in this paper study three different methods; ordinary least squares, Ridge and Lasso regression. All the code, results and instructions on running the code can be found in our GitHub repository here[1]

Larger datasets contain more information, giving more accurate models. However, real-world datasets usually have a fixed size, and getting more is practically impossible. For smaller datasets it is then useful to have tools mitigating the effects of little data. Resampling methods does this by running the same data through the regression, without over-fitting the model to the sample data. In addition to the regression methods, we will also study the effect of bootstrapping and cross-validating the data.

In order to study this, we need data to analyze. We are going to use two data sets to study the different methods. One data set we will generate ourselves using the Franke function, given by equation (1). To make the data more realistic we also add normally distributed noise. The other data is real terrain data from here[2].

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

[1] https://github.com/sigurdru/FYS-STK4155/tree/main/project1

[2] https://github.com/CompPhysics/MachineLearning/tree/master/doc/Projects/2021/Project1/DataFiles

II.   ORDINARY LEAST SQUARE

III.   BIAS-VARIANCE TRADE-OFF AND BOOTSTRAPPING

IV.   CROSS-VALIDATION

V.   RIDGE REGRESSION

VI.   LASSO REGRESSION

VII.   ANALYSIS OF REAL DATA

VIII.   TESTING

---

[1] Ref.