Employing supervised machine learning algorithms for fitting terrain data

Eina B. Jørgensen, Anna Lina P. Sjur and Jan-Adrian H. Kallmyr

September 19, 2019

Abstract

1 Introduction

The use of machine learning for problem solving has risen in popularity as large data sets have become available for analysis. There now exists many different methods in varying complexity for both supervised and unsupervised learning. All of these methods have advantages and drawbacks, as well as many similarities. This means we can get familiar with some of the central themes in machine learning by studying simple algorithms. In this report, we will implement three different supervised learning algorithms with increasing complexity, as well as the k-fold resampling technique.

2 Theory

2.1 Linear Regression

A linear Regression model makes a prediction of the response variable y_i by computing a weighted sum of the explanatory variables:

$$\tilde{y}_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \ldots + \beta_{p-1} x_{ip-1}$$

Here, $\tilde{y}_i$ is the prediction, $\{\beta_j\}_{j=0}^{p-1}$ are the regression parameters, while p are the number of explanatory variables.

In vectorized form, this can be written as

$$\tilde{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{\beta}$$

where $\tilde{\boldsymbol{y}} = [\tilde{y}_0, \tilde{y}_1, \ldots, \tilde{y}_{n-1}]^T$ are the predicted values, $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots . \beta_{n-1}]^T$ are the regression parameters, and $\boldsymbol{X}$ is the so called design matrix given by

$$\boldsymbol{X} = \begin{bmatrix} x_{00} & x_{01} & \ldots & x_{0p-1} \\ x_{10} & x_{11} & \ldots & x_{1p-1} \\ \vdots & \ddots & \ddots & \vdots \\ x_{n0} & \ldots & \ldots & x_{np-1} \end{bmatrix},$$

where n is the number of cases.

In this project, we are dealing with a two dimensional problem, where each row of the design matrix represents the variables of a mth order polynomial, i.e. is on the form $[\{x^i y^j : i + y \leq m\}]$

In order to compute the regression parameters, a cost function $C(\boldsymbol{\beta})$ is introduced. The $\boldsymbol{\beta}$ that are used will then be the ones that minimize the cost function. Different cost functions gives rise to different regression methods. Here we will look at Ordinary least squares, Ridge and Lasso regression.

2.2 Ordinary least squares

One form of the cost function is given as

$$\begin{aligned} C(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{n=0}^{n-1}(y_i - \tilde{y}_i)^2 \\ &= \frac{1}{n}\sum_{n=0}^{n-1}(y_i - \boldsymbol{x}_{i*}\boldsymbol{\beta})^2 \\ &= \frac{1}{n}\left((\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right) \end{aligned}$$

where $\boldsymbol{x}_i$ is the ith row of the design matrix. By setting $\frac{\partial C(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0$, one can show that the model parameters that minimizes the cost function are given by

$$\boldsymbol{\beta} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

These are the model parameters used in the Ordinary Least Squares (OLS) model.

2.3 Ridge regression

Something something restrain

In Ridge regression, the cost function takes the form

$$C(\boldsymbol{\beta}) = \frac{1}{n}\sum_{n=0}^{n-1}(y_i - \boldsymbol{x}_{i*}\boldsymbol{\beta})^2 + \lambda\sum_{i=0}^{p-1}\beta_i^2$$

Minimizing this results in model parameters on the form

$$\boldsymbol{\beta} = \left(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\lambda I}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

2.4 Lasso regression

$$C(\boldsymbol{\beta}) = \frac{1}{n}\sum_{n=0}^{n-1}(y_i - \boldsymbol{x}_{i*}\boldsymbol{\beta})^2 + \lambda\sum_{i=0}^{p-1}|\beta|$$

2.5 Score function

2.6 k-fold cross validation

There are several methods for estimating the skill of a machine learning model. One such method is the k-fold cross-validation procedure, which can be used when working with a limited data sample. The idea is to divide the data sample into k groups or folds, and then retain one of the folds to use as a test set after fitting a model to the remaining data. This is done for all folds. Algorithm 1 outlines the different steps in the procedure.

```
Shuffle the dataset randomly;
Divide the dataset into k folds;
foreach k do
    Take the kth fold out to use as test
      data set;
    Set the remaining folds as training
      data set;
    Fit a model to the training set;
    Evaluate the model on the test set;
    Retain the evaluation score and
      discard the model;
Calculate the mean of the evaluation
  scores;
```

Algorithm 1: The k-fold cross-validation algorithm.

This yields a statistical estimate for how well the model will preform on new data. The choice of k will however effect the bias and variance in the estimation of the evaluation scores. It has been shown empirically that $k = 5$ or $k = 10$ gives neither a high bias nor variance (James et al., 2013). In this project, a value of 5 was chosen for k.

2.7 The bias-variance trade-off

2.8 Franke's function and digital terrain data

In this project, the different regression methods were applied on both constructed and real data. The first was in the form of a sum of weighted exponentials, known as Franke's function:

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

In addition to the above terms, a normal distributed noise term with $\mu = 0$ and $\sigma^2 = 1$ was added.

After testing the code on the simpler Franke's function, the same regression methods were used and evaluated on real digital terrain data downloaded from `https://earthexplorer.usgs.gov/`. The data used in this project is of the Oslo fjord region.

3 Results

Figure 1

4 Discussion

5 Conclusion

References

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in R.* Springer Texts in Statistics. Springer New York, 2013. ISBN 9781461471387.