DISP '19, Oxford, United Kingdom
ISBN: 978-1-912532-09-4

Recent Developments in Data Science: Comparing Linear, Ridge and Lasso Regressions Techniques Using Wine Data

Mayooran Thevaraja[1]
Lecturer in Engineering Mathematics,
Department of Interdisciplinary studies,
Room 38, Faculty of Engineering, University of Jaffna,
Ariviyal Nagar, Killinochchi, Sri Lanka.
mayooran@eng.jfn.ac.lk

Azizur Rahman[2]
Senior Lecturer in Mathematics and Statistics
Charles Sturt University
Room 310 | B001, Boorooma Street
Wagga Wagga, NSW 2678, Australia
azrahman@csu.edu.au

Mathew Gabirial[3]
Department of Mathematics and Statistics,
Minnesota State University, Mankato, USA.
mathew.gabriel@mnsu.edu

ABSTRACT

Big data is the reality of the 21st century. However, big data modeling and prediction require advanced level analytics which encompasses both the computing-intensive and statistics-oriented analysis tools in data science. Regression analysis is the statistical method for predictive modeling, and it is one of the most commonly used methods in many scientific fields such as engineering, the physical and chemical sciences, economics, management, life and biological sciences, and the social sciences, sociology, geology, etc. Satisfying the assumptions such as collinearity between variables ought to be a significant issue in data science. Advanced level tools such as Lasso and Ridge regression methods are designed to overcome such problem. In this study we discussed about comparing linear regression with the Ridge and Lasso. The Vinho Verde white wine test data from the Minho (northwest) region of Portugal is used to analyze advantages of each of the three regression analysis methods. All the required calculations and graphical displays are performed using the R software for statistical computing.

Keywords

Standard regression, Ridge, Lasso, White wine data, Validation of estimates.

1. INTRODUCTION

In data science, the analyst might be interested in the estimated values for the objects in the training set, predicting responses for new objects. However, identifying which input variables are most important in making good predictions, is challenging due to complex relations between these variables (Rahman, 2008). For example, in wine industry physicochemical laboratory tests are routinely used to characterize quality of wine such as its density, alcohol or pH values. Also sensory tests are required which mainly depend on experts people within the industry. Classification of wine is a challenging task though. The links between the physicochemical and sensory analysis are fairly complex and require advance level modelling to understand them better (Legin et al. 2003).

A big dataset can preserve useful facts and figures which can be utilized to effective decision making process (Rahman, 2017). For instance, data manning techniques including the linear regression focus on extracting advance level knowledge by modelling raw data with continuous variables (Turban et al. 2007; Rahman and Harding, 2016). The performance of any data mining tools also very much depends on a right set variables and robust model selection (Das et al. 2018; Chowdhury et al. 2018; Rahman et al. 2018). Typically, a simple model such as linear regression may fail in mapping the underlying concept due to collinearity and too complex ones such as generalized regression tend to over fit the data due to lack of functional link and collinearity as well (Hastie et al. 2001; Guyon and Elisseeff 2003). Often the traditional regression model can show higher correlation coefficients, but can suffer from overfitting and its lower prediction accuracy (Aleixandre-Tudo et al. 2015).

One of the ways of avoiding overfitting is using cross validation that helps in estimating the error over test data set, and in deciding what parameters work best for your model. For examples, regularized regression techniques such as Lasso and Ridge methods are designed to minimize the consequences of collinearity between variables by avoiding overfitting and also increasing model interpretability with an increased level of accuracy. Regularization process imposes an upper threshold on the values taken by the coefficients that constrains or shrinks the coefficient estimates towards zero especially for the variables which influence the overfitting.

DOI: http://dx.doi.org/10.17501......................................