LASSO: A Feature Selection Technique In Predictive Modeling For Machine Learning

Muthukrishnan R
Dept. of Statistics
Bharathiar University
Coimbatore, India
muthukrishnan1970@gmail.com

Rohini R
Dept. of Statistics
Bharathiar University
Coimbatore, India
rohinirajan92@gmail.com

***Abstract*—Feature selection is one of the techniques in machine learning for selecting a subset of relevant features namely variables for the construction of models. The feature selection technique aims at removing the redundant or irrelevant features or features which are strongly correlated in the data without much loss of information. It is broadly used for making the model much easier to interpret and increase generalization by reducing the variance. Regression analysis plays a vital role in statistical modeling and in turn for performing machine learning tasks. The traditional procedures such as Ordinary Least Squares (OLS) regression, Stepwise regression and partial least squares regression are very sensitive to random errors. Many alternatives have been established in the literature during the past few decades such as Ridge regression and LASSO and its variants. This paper explores the features of the popular regression methods, OLS regression, ridge regression and the LASSO regression. The performance of these procedures has been studied in terms of model fitting and prediction accuracy using real data and simulated environment with the help of R package.**

Keywords—LASSO; Ridge regression; OLS; R software

I. INTRODUCTION

Statistics is a field that faces the challenges and problems from the areas of science and industry. These challenges and problems have exploded due to the invention and rapid growth of computers and information technologies. As the amount of data increased predominantly, extracting the patterns and trends, and understanding the data (i.e., learning from data) is required. This led to a new field known as "machine learning". The amount of data is increasing each and every second and making sense of this enormous amount of data is a challenging task for data-driven industry people. The data is composed of a number of features. It mainly focuses on supervised and unsupervised learning. Supervised learning postulates the machine to learn from the data when a target variable is specified. Thus, the machine's task is reduced to only divine some pattern to get the target variable from the input data. Unsupervised learning is just the contrary of supervised learning. In which, the target value or label is not mentioned for the data. It also visualizes the data in two or three dimensions.

Regression is one of the most popular and essential tasks of machine learning, which falls in the category of supervised learning. Regression analysis deals with two main problems namely, parameter estimation and variable selection. Parameter estimation is commonly carried out using OLS regression estimation method. Variable selection is yet another crucial part in regression analysis. A model with more number of regressors can possibly reduce modeling biases. But, it can lead to less accurate predictions and affect the efficiency of the estimation procedure. Hence, selecting the most important variables is vital.

To overcome these problems, number of methods have been proposed and established for performing feature selection, that is, choosing only relevant variables from a model. Variable selection can be carried out using various approaches namely, subset selection, shrinkage and dimension reduction. This paper explores only the shrinkage approach restricting to the methods Ridge and LASSO, which are the types of regularized linear regression variants. Section 2 presents the theory behind Ridge and LASSO type regressions and their limitations. The performance of these regression techniques has been studied with the usual OLS approach in real and simulating environment is discussed in the section 3 and the last section provides summary of the results.

II. FEATURE SELECTION PROCEDURES

Shrinkage methods minimize the residual sum of squares of the model using OLS and also reduce the intricacy of the model such as the number or absolute size of the sum of all coefficients in the model. This method is an alternative to the subset selection and dimension reduction methods, a model including all p predictors can be fitted using a technique that regularizes the estimated coefficients comparative to the least squares. Thus this method, also recognized as regularization, can significantly reduce the variance and can also perform variable selection. The two important methods that shrink the estimated regression coefficients approaching to zero are ridge regression and LASSO. Ridge regression minimizes the squared sum of the coefficients (L2 regularization) and LASSO minimizes the absolute sum of the coefficients (L1 regularization). In case, if there is collinearity in the input values, these methods can perform effectively while OLS would overfit the data.

978-1-5090-3770-4/16/$31.00©2016 IEEE

A. Ridge regression

Ridge regression, originally proposed by Hoerl and Kennard [1970(a), (b)] is a method for analyzing the data which are affected by multicollinearity. If multicollinearity is present in the data, least squares estimates are unbiased, but the variances are large hence they are far away from the true value. Ridge regression is identical to least squares, unless the ridge coefficients are estimated by minimizing a slightly different quantity. It is hoped that the net effect will give estimates that are more reliable. The ridge coefficients minimize a penalized residual sum of squares,

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

where $\lambda \geq 0$ is a tuning parameter that controls the amount of shrinkage. The coefficients are shrunk towards zero.

The ridge problem can also be written as

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

$$\text{subject to } \sum_{j=1}^{p}\beta_j^2 \leq t,$$

where there exists one-to-one correspondence between the parameters λ and t.

Multiple linear regression coefficient estimates depend on the independence of terms of the model. If there exists correlation in the terms and approximate linear dependence in the design matrix X, the matrix $(X^TX)^{-1}$ attains singularity. Thus, the least squares estimate

$$\hat{\beta} = (X^TX)^{-1}X^Ty$$

gets affected by random errors in the determined response y, giving rise to a large variance. This is known as multicollinearity.

Ridge regression handles the problem by estimating the coefficients using

$$\hat{\beta} = (X^TX + hI)^{-1}X^Ty$$

where h is the parameter of ridge regression and I is the identity matrix. Small positive values of h amend the conditioning of the problem and lessen the variance of the coefficient estimates. Hoerl and Kennard (1970) have suggested that using the ridge trace, an appropriate value of h can be determined. The ridge estimates gives an equation that predicts future observations better than least squares. While biased, the decreased variance of the ridge estimates frequently result in a minimum mean square error compared to the estimates of least squares method.

The assumptions of ridge regression are same as the least squares regression except the normality is not assumed. It reduces the variability by shrinking the coefficients, resulting in more prediction accuracy at the cost of usually only a small increase of bias. It shrinks the coefficients nearly to zero but not exactly to zero, thus not good for feature selection. When the number of predictors is large, ridge regression will not provide an easily interpretable sparse model.

B. LASSO

The least absolute shrinkage and selection operator (LASSO) was put forwarded by Tibshirani (1996) for parameter estimation and also variable (model) selection simultaneously in regression analysis. The LASSO is a particular case of the penalized least squares regression with L1-penalty function.

The LASSO estimate can be defined by

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

which can also be written as

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

$$\text{subject to } \sum_{j=1}^{p}|\beta_j| \leq t,$$

LASSO transforms each and every coefficient by a constant component λ, truncating at zero. Hence it is a forward-looking variable selection method for regression. It decreases the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. LASSO was originally defined in the context of least squares, but it can also be extended to a wide variety of models.

LASSO improves both prediction accuracy and model interpretability by combining the good qualities of ridge regression and subset selection. If there is high correlation in the group of predictors, LASSO chooses only one among them and shrinks the others to zero. It reduces the variability of the estimates by shrinking the some of the coefficients exactly to zero producing easily interpretable models.

III. Experimental Study

This section presents the summary of the results which are carried out to study the performance of ridge and lasso along with OLS, based on real data and simulation.

A. Real Data

The data used for experiment is the Diabetes data which was originally used by Efron et al. (2003). There are ten baseline variables such as age, sex, body mass index (BMI), blood pressure (BP) and six different blood serum measurements. There are about 422 observations of diabetes patients. The response variable y represents a measurement of growth of disease one year after baseline. The regression analysis was executed for the data. The estimated coefficients and the variables selected under the methods LASSO, Ridge and OLS are summarized in the table I.

B. Simulation

The performance of LASSO and Ridge regression along with OLS has been studied under simulation. The data are simulated from the model

$$y = \beta^{T} x + \sigma\varepsilon, \quad (3)$$

where, β = (3, 1.5, 0 , 0, 2, 0, 0, 0)^{T} and ε is distributed normally with mean 0 and variance 1. The correlation between x_i and x_j is $\rho^{|i-j|}$ with ρ = 0.5 and σ = 3. The computed mean squared error (MSE), median mean squared error and average number of zero coefficients under various methods are summarized in table II. Consider different numbers of simulation to obtain best Mean Squared Error.

C. Tables

The tables I and II of the experimental study under the real data and simulation are presented in this section.

TABLE I. ESTIMATED COEFFICIENTS AND THE SELECTED VARIABLES

Predictor	Coefficients		
	OLS	***Ridge***	***LASSO***
Intercept	-339.453	-179.947	-226.739
AGE	0.020	0.082	.
SEX	-22.443	-11.603	.
BMI	6.095	4.350	5.796
BP	1.103	0.842	0.596
S1	-1.078	-0.018	.
S2	0.742	-0.094	.
S3	0.360	-0.611	-0.283
S4	5.215	4.138	.
S5	69.963	30.430	39.784
S6	0.150	0.418	.
Variables selected	**All**	**All**	**BMI,BP,S3,S5**

It is observed from the table I, the methods OLS and ridge regression fitted by using all the variables which includes the variables that are not significantly contributed. But LASSO fits the most significantly contributed variables BMI, BP, S3 and S5 only. It is noted that LASSO ignores the variables by making the coefficients exactly to zero. This shows that LASSO performs better than the other two methods by selecting the features.

From the table II, it can be observed that the method LASSO has the minimum median MSE, which shows that the method LASSO predicts the model with much accuracy than the other methods. The LASSO exactly shrinks the non-significantly contributed variables to zero while OLS and Ridge regression methods fail to estimate the exactly zero coefficients.

TABLE II. MSE, MEDIAN MSE AND AVERAGE NUMBER OF ZERO COEFFICIENTS

No. of Simulation	***OLS***			***Ridge***			***LASSO***		
	MSE	***Median MSE***	***Avg. no. of zero coefs***	***MSE***	***Median MSE***	***Avg. no. of zero coefs***	***MSE***	***Median MSE***	***Avg. no. of zero coefs***
100	9.114	8.056	0.0	4.808	4.967	0.0	4.646	4.598	0.75
500	5.930	8.505	0.0	7.418	4.556	0.0	4.595	4.246	0.25
1000	9.114	8.509	0.0	6.130	4.539	0.0	1.021	4.334	0.63
5000	5.935	8.347	0.0	7.974	4.499	0.0	2.124	4.292	0.50
10000	10.972	8.265	0.0	2.160	4.454	0.0	2.370	4.219	0.63
50000	7.512	8.217	0.0	5.670	4.469	0.0	3.071	4.188	0.63

IV. SUMMARY AND CONCLUSION

There are many conventional methods for feature selection, which is a significant technique in machine learning. In which regression analysis plays an important role. This paper is mainly focused on the regression methods namely LASSO, Ridge regression and OLS. The performance of these methods has been carried out under real and simulated data with the help of R software. The experimental results show that the LASSO works better than the other methods by shrinking the coefficients exactly to zero. Hence, the LASSO method can be applied as an alternative to the conventional feature selection methods. It would be beneficial to the research communities especially, those who are working in machine learning tasks.

References

[1] A.E. Hoerl, and R.W. Kennard, "Ridge regression: Applications to nonorthogonal problems," Technometrics, vol. 12, 1970a, pp. 69-82.

[2] A.E. Hoerl, and R.W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," Technometrics, vol. 12, 1970b, pp. 55-67.

[3] D.C. Montgomery, E.A. Peck, and V.G. Geoffrey, Introduction to Linear Regression Analysis, 5th ed., Wiley: New Jersey, 2012.

[4] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning with Applications in R, Springer: New York, 2015.

[5] M. Atashbar and M.H. Kahaei, "Direction-Of-Arrival Estimation Using AMLSS Method," IEEE Latin America Transactions, vol. 10, No. 5, 2012, pp.2053-2058.

[6] R. Muthukrishnan, and Radha Myilsamy, "M-Estimators in Regression Models," Journal of Mathematics Research, vol. 2, No. 4, 2010, pp.23-27.

[7] R. Tibshirani, "Regression shrinkage and selection via the lasso," Journal of the Royal Statistical Society, Series B, vol. 58(1), 1996, pp. 267-288.

[8] R. Tibshirani, "Regression shrinkage and selection via the lasso: a retrospective," Journal of the Royal Statistical Society, Series B, vol. 73(3), 2011, pp. 273-282.

[9] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference and Prediction, 2nd ed., Springer: Canada, 2009.