Analysis of Lasso and Ridge Regression on A High Dimensional Dataset

E. Annapoorna
AM.EN. U4AIE20123
B. Tech Computer Science, AI
Amrita Vishwa Vidyapeetham
Kollam, India
amenu4aie20123@am.students.amrita.edu

Rohit Narayanan.M
AM.EN. U4AIE20160
B. Tech Computer Science, AI
Amrita Vishwa Vidyapeetham
Kollam, India
amenu4aie20160@am.students.amrita.edu

Sreepriya.S
AM.EN. U4AIE20166
B. Tech Computer Science, AI
Amrita Vishwa Vidyapeetham
Kollam, India
amenu4aie20166@am.students.amrita.edu

Sreya.V. Sujil
AM.EN. U4AIE20167
B. Tech Computer Science, AI
Amrita Vishwa Vidyapeetham
Kollam, India
amenu4aie20167@am.students.amrita.edu

Sudhin. S
AM.EN. U4AIE20168
B. Tech Computer Science, AI
Amrita Vishwa Vidyapeetham
Kollam, India
amenu4aie20168@am.students.amrita.edu

***Abstract* – Modern IT relies on big data and analytics. Components that serve as abstractions of real-world objects are used to create a data model. Entities and relationships are the most basic components of a data model. Additional information and complexity are added to the data model as model progresses, such as characteristics, domains, constraints, keys, cardinality, requirements, and connections. Regression analysis is a statistical approach for predictive modelling that is widely utilized in a variety of scientific areas, including engineering, physics and chemistry, economics, management, and life and biological sciences. In data science, satisfying assumptions like collinearity between variables should be a key concern. Advanced techniques such as Lasso and Ridge regression methods are designed to overcome such problem. In this study we are comparing Lasso regression and Ridge regression. The Kaggle's titanic dataset is used to compare these two regression techniques. All the required calculations and graphical displays are performed using the MATLAB software**.

Keywords – Regression, Lasso, Ridge, MSE, MAE

I. INTRODUCTION

Regression analysis is a statistical method that uses one or more independent variables to model the relationship between the dependent variable (target) and the independent variable (predictor). More specifically, regression analysis helps us understand how the value of the dependent variable changes in response to the independent variable when other independent variables remain fixed.

Regression is a supervised learning technique that helps find the correlation between variables and allows us to predict continuous output variables based on one or more predictor variables. It is mainly used to forecast, model time series, and determine the causal relationship between variables. Regression analysis is a statistical method used in machine learning and data science. Here are some other reasons to use regression analysis:
Regression estimates the relationship between the target and the independent variables.
Used to search for data trends.
Helps to predict actual / continuous values.
By regression, we can confidently determine the most important factors, the least important factors, and how each factor affects other factors.
The Ridge regression is one of the more robust versions of linear regression, introducing a small amount of bias so that we can get better long-term predictions. The amount of deviation added to the model is called the ridge regression penalty. If there is a high degree of collinearity between the independent variables, the generally linear or polynomial regression will fail, so ridge regression can be used to solve such problems. Ridge Regression is a regularization technique used to reduce model complexity. Also called L2 regularization.
Lasso is another regularization technique to reduce model complexity. It is like ridge regression; the difference is that the penalty term only contains the absolute weight instead of the square of the weight. Since it takes the absolute value, it can reduce the slope to 0, while ridge regression can only reduce it to close to 0. It is also called L1 regularization.

II. REGRESSION

Regression analysis is a collection of statistical procedures for evaluating the relationships between a dependent variable ('outcome' or 'response' variable) and one or more independent variables ('predictors', 'covariates', 'explanatory variables' or 'features').

- **Dependent Variable:** The dependent variable is the main factor in regression analysis that we wish to predict or understand.
- **Independent Variable:** Independent variables are the elements that influence the dependent variables or are used to predict the values of the dependent variables.

Regression analysis is primarily used for two conceptually different purposes.

- Regression analysis is commonly used for prediction and forecasting, and it shares a lot of ground with machine learning.
- To infer causal links between the independent and dependent variables, regression analysis can be utilized.

There are over 15 types of regression:

- Linear Regression
- Polynomial Regression
- Logistic Regression
- Quantile Regression
- Ridge Regression
- Lasso Regression
- Elastic Net Regression
- Principal Components Regression
- Partial Least Squares Regression
- Support Vector Regression
- Ordinal Regression
- Poisson Regression
- Negative Binomial Regression
- Quasi Poisson Regression
- Cox Regression
- Tobit Regression

The most ordinary form of regression analysis is linear regression. It's a method in which the dependent variable is always the same. The dependent variable's relationship with the independent factors is linear.

A. REGULARIZATION

Regularization is used to avoid overfitting of the data, especially when the trained and tested data are quite different. Regularization is implemented by adding a "penalty" term to the best fit derived from the trained data, to achieve a lesser variance with the tested data and restricts the influence of predictor(independent) variables over the outcome(dependent) variable by compressing their coefficients.

B. PENALTY TERM LAMBDA (λ)

The amount of the penalty can be fine-tuned using a constant called lambda (λ). Selecting a good value for λ is very important. Lambda determines the severity of the penalty. When λ=0, the penalty term has no effect. Lasso regression will produce the sum of magnitude of the coefficient whereas ridge regression will produce the classical least square coefficients. However, as λ increases to infinite, the impact of the shrinkage penalty grows as lambda controls the amount of shrinkage.

C. CROSS VALIDATION METHOD

The penalty term lambda can be determined by Cross-Validation method. Small λ value can lead to overfitting (Overfitting means that our algorithm works well on the training set but is unable to perform better on the test sets) when the model would tend to describe the noise in the data. On the other side, large λ values would lead to underfitting (When our algorithm works so poorly that it is unable to fit even training set well then it is said to underfitting) when the procedure cannot capture the underlying relationship.

D. MULTICOLLINEARITY

Multicollinearity is such a condition which occurs when the independent variables are highly correlated with each other than other variables. It should not be present in the dataset, because it creates problem while ranking the most affecting variable.

III. DATASET

For the dataset of this project, we have chosen the Titanic dataset. The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships. One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew.

Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class. In this challenge, we need to complete the analysis of what sorts of people were likely to survive. We want to apply the tools of machine learning to predict which passengers survived the tragedy.

Applying and analyzing of lasso and ridge regression on a high dimensional dataset based on existing regression principles. The program reads the csv file, splits it into testing and training datasets, in the ratio 3:1(75%: 25%). Then, we create regression models using the functions in MATLAB for the training dataset. After that we supply the testing dataset to the model for predicting the outcome and score the model using evaluation metrices MSE and MAE. This model predicts what sort of people are most likely to survive by finding the relationship between different parameters like age, sex, ticket class in the dataset.

IV. METHOD

We typically define the design matrix X as a matrix having n rows and p columns, representing the p variables for n instances. The problem can then be formulated as finding a "good" value for the length p coefficient vector β.

A. Lasso Regression

Lasso Regression performs regularization along with feature selection. It forbids the absolute size of the regression coefficient. As a result, the coefficient value approaches to zero. As a result, feature selection is used in Lasso Regression to develop the model, which allows select a collection of features from the dataset. Only the relevant characteristics are used in Lasso Regression, and the others are set to zero. This helps in preventing the overfitting in the model.

The Lasso is also formulated with respect to the center matrix, X. Moreover, the L1-penalty is solely applied to the slope coefficients, and thus the intercept, β0, is excluded from the penalty term. Hence the Lasso can be expressed as a constrained minimization problem,

$$\beta_{lasso} = \sum_{i=1}^{n} (Y_i - X_i^T \beta)^2$$

It can again be re-formulated using the Lagrangian for the L1- penalty, as follows

$$\beta_{lasso} = \left\{ \sum_{i=1}^{n} (Y_i - X_i^T \beta)^2 + \lambda \sum_{i=1}^{p} |\beta_j| \right\}$$

B. Ridge Regression

This is another form of regression in machine learning that is typically employed when there is a high correlation between the independent variables. This is because the least square estimations produce unbiased results in the case of multi collinear data. But, in case the collinearity is very high, there can be some bias value. As a result, a bias matrix is introduced in the Ridge Regression equation. This is a powerful regression method where the model is less susceptible to overfitting.

It places a constraint on the sum of squares of the coefficient's weights through a constraint on the p coefficients. It can be formulated as,

$$\beta_{ridge} = \sum_{i=1}^{n} (Y_i - X_i^T \beta)^2$$

It can be formulated using Lagrange multiplier as,

$$\beta_{ridge} = \left\{ \sum_{i=1}^{n} (Y_i - X_i^T \beta)^2 + \lambda \sum_{i=1}^{p} {\beta_j}^2 \right\}$$

C. Cross Validation

To perform cross validation, the initial titanic dataset is divided into K blocks of equal length. One of the blocks is assigned the role of the test set while the remaining K-1 blocks together constitute the train set. In practice the number of blocks K is usually selected to be 5 or 10. Then we choose a grid of values λ= λs and calculate the regression coefficients for each λs value. Given these regression coefficients, then compute the residual sum of square.

$$RSS_{\lambda_s k} = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{k-1} \hat{\beta}(k, \lambda_s) x_{ij} \right)^2$$

where K=1,2,3... K is the index of the block selected as the test set.

V. EVALUATION METRICS

The evaluation metrics used to evaluate the model performance are

A. Mean Absolute Error (MAE)

The absolute difference between actual and predicted values is calculated using MAE metric. MAE is a measure that estimates the average magnitude of errors in a set of predictions without considering their direction. It is the average of the absolute differences between prediction and actual observation over the test sample, where all individual deviations are given equal weight.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}|$$

B. Mean Squared Error (MSE)

Mean Squared Error (MSE) or Mean Squared Deviation (MSD) of an estimator calculates the average of squares of difference between the estimated values and the actual value. It measures the variance of the residuals.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$$

In comparison to a non-differentiable function like MAE, MSE is a differentiable function that makes mathematical procedures easier. MAE is more robust to data that contains outliers.

VI. RESULT

A. Confusion Matrices

The confusion matrix is a matrix that is used to evaluate the classification models' performance for a given set of testing data. Only if the true values for test data are known can it be determined. We can determine the model's various parameters, such as accuracy, precision, and so on, using the confusion matrix.

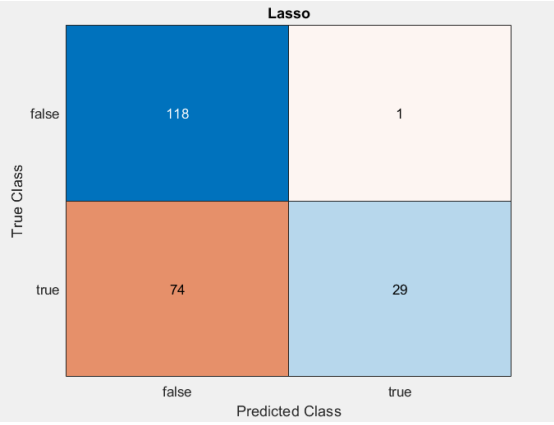

Figure 1: Confusion matrix output for Lasso.

Form this figure 1,
True negative value –118
False positive value – 1
False negative – 74
True positive – 29

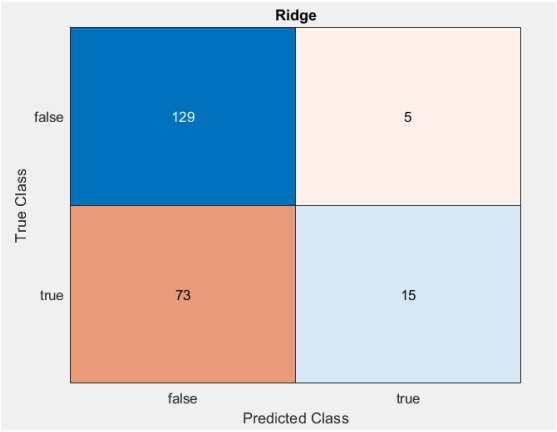

Figure 2: Confusion matrix output for ridge.

Form this figure 2,
True negative value –129
False positive value – 5
False negative – 73
True positive – 15

B. Efficiency

Regression	MSE	MAE	accuracy
Lasso	0.263	0.265	74.4%
Ridge	0.287	0.31	66.1%

The above table contains the average data collected from multiple runs of the programs, which shows their efficiency.
In this data set, we can infer that Lasso regression works more efficiently.

C. Comparing coefficients

The coefficient values generated by the regression models are listed in the table below.

coefficient	Lasso	Ridge
β_0	-5.454e-04	-0.0677
β_1	0	-0.0425

β_2	0	-0.0190
β_3	0	0.0242
β_4	0.1141	0.0662
β_5	0	0.0108
β_6	-0.1083	-0.0663
β_7	0.4369	0.1305
β_8	-8.470e-14	-0.1305
β_9	0	-0.0533
β_{10}	0	-0.0259
β_{11}	0	-0.0897

VII. CONCLUSION

The Ridge regression and the LASSO tend to "shrink" to zero coefficient estimates in the sense that they reduce the norm of the estimate vector as λ increases. The Ridge regression does not produce zero estimates even for large values of λ. The Lasso performs variable selection i.e., some of the coefficient estimates become exactly equal to zero, which makes the regression model easier to interpret. The lower the MAE and MSE, the more accurate the regression model. According to our observations, The MSE and MAE of the Lasso model is comparatively lower than that of Ridge model, and also the accuracy of prediction is more for Lasso, when compared to ridge making it the best fit.

ACKNOWLEDGMENT

We would like to express our sincere thanks to Gopakumar sir and Mithun sir, for guiding us to do our work and report. We would also like to express our sincere gratitude to our beloved college Amrita Vishwa Vidyapeetham for providing and presenting us an opportunity to work on this report.

REFERENCE

[1] https://www.javatpoint.com/regression-analysis-in-machine-learning

[2] *Regression analysis - Wikipedia*, https://en.wikipedia.org/wiki/Regression_analysis.

[3] *15 Types of Regression in Data Science*, https://www.listendata.com/2018/03/regression-analysis.html.

[4] *Expression Prediction - GitHub Pages*, https://seandavi.github.io/ITR/expression-prediction.html.

[5] *Learn Coding Neural Network in C#: Solve Titanic Survival ...*, https://www.tech-quantum.com/learn-coding-neural-network-in-c-solve-titanic-survival-problem/

[6] *6 Types of Regression Models in Machine Learning You Should Know About* (July 27, 2020). https://www.upgrad.com/blog/types-of-regression-models-in-machine-learning/

[7] *Know the Best Evaluation Metrics for Your Regression Model!* (May 19, 2021). https://www.analyticsvidhya.com/blog/2021/05/know-the-best-evaluation-metrics-for-your-regression-model/

[8] Mayooran Thevaraja, Azizur Rahman, Mathew Gabirial (April 2019). *Recent Developments in Data Science: Comparing Linear, Ridge and Lasso Regressions Techniques Using Wine Data.* https://www.researchgate.net/publication/324870033_Recent_Developments_in_Data_Science_Comparing_Linear_Ridge_and_Lasso_Regressions_Techniques_Using_Wine_Data