STUDY OF LASSO AND RIDGE REGRESSION USING ADMM

E.Annapoorna
dept of computer science
Amrita Vishwa Vidyapeetham
kollam, India
amenu4aie20123@am.students.amrita.edu

Rohit Narayanan.M
dept of computer science
Amrita Vishwa Vidyapeetham
kollam, India
amenu4aie20160@am.students.amrita.edu

Sreepriya.S
dept of computer science
Amrita Vishwa Vidyapeetham
kollam, India
amenu4aie20166@am.students.amrita.edu

Sreya.V.Sujil
dept of computer science
Amrita Vishwa Vidyapeetham
kollam, India
amenu4aie20167@am.students.amrita.edu

Sudhin.S
dept of computer science
Amrita Vishwa Vidyapeetham
kollam, India
amenu4aie20168@am.students.amrita.edu

***Abstract*—Many recent statistics and machine learning challenges may be stated in the perspective of convex optimization. The ability to solve problems with a large number of features is becoming increasingly vital as recent datasets increase in complexity. As a result, the alternating direction method for multiplier is particularly suited to distributed convex optimization, especially large-scale problems in statistics, machine learning, and related fields. This project include the comparison of ordinary lasso and ridge regression with ADMM implementation of both regression techniques. It also determines whether implementation is better in terms of performance, time, and the impact of optimization on the outcome.**

I. INTRODUCTION

Regression is a widely used data analytical technique for forecasting, modelling time series, and prediction. It is a supervised learning strategy which approximates the connection between the goal and the independent variables, used in analysis of information trends, helps to expect actual/continuous values.

For larger and complex datasets, the usual regression models could cost very high and some are not efficient. This issue is resolved by optimising of the models using different optimization techniques. The ADMM algorithm is a simple but powerful algorithm that is ideally suited to distributed convex optimization.

Tibshirani (1996) proposed the least absolute shrinkage and selection operator (LASSO) for simultaneous parameter estimation and variable selection in regression analysis. The penalised least squares regression with L1-penalty function is known as the LASSO.

Ridge regression, originally proposed by Hoerl and Kennard [1970], is a method for analyzing data affected by multicollinearity. The ridge regression penalty is the amount of deviation brought into the version. The normally linear or polynomial regression will fail if the independent variables have an excessive degree of collinearity, so ridge can be used to solve this problem. This is an L2 regularisation technique that reduces the number of versions in a database.

II. ADMM

It is currently standard in all practical fields to approach problems through data analysis, particularly through the use of statistical and machine learning methods on large datasets. Despite the fact that these problems exist in a variety of application domains, they share some common characteristics. First, the datasets are extremely large, containing hundreds of millions or billions of training examples; second, the data is frequently very high-dimensional, as it is now possible to measure and store extremely detailed information about each example; and third, the data is frequently stored or even collected in a distributed manner, due to the large scale of many applications. As a result, developing algorithms that are both rich enough to capture the complexity of modern data and scalable enough to analyse massive datasets in a parallelized or totally decentralised manner has become critical.

Convex optimization can be used to solve a variety of such problems. Given the optimization community's focus on decomposition methods and decentralized algorithms, parallel optimization algorithms are a suitable choice for handling large-scale statistical problems. This method also has the advantage of allowing a single algorithm to solve many problems.

ADMM algorithm takes the form of a decomposition-coordination technique, in which small local subproblem solutions are coordinated to solve a large global problem. Many more algorithms are found to be comparable to or substantially related to ADMM, such as Douglas-Rachford splitting from numerical analysis, Spingarn's method of partial inverses, Dykstra's alternating projections method, Bregman iterative algorithms for L1 problems in signal processing, proximal methods, and many others.

ADMM was developed before large-scale distributed computing systems and enormous optimization problems were readily available. The algorithm would also be a good fit for more difficult problems, such as graphical models. The algorithm can be used in a number of cases, including engineering design, multi-period portfolio optimization, time series analysis, network flow, and scheduling, in addition to statistical learning challenges. [10]

III. REGRESSION

Regression analysis is a statistical data collection technique which evaluates the relation between a response variable and one or more explanatory variables, used to predict the value of response variables. The response variables are used to predict or explain the model. They are also known as dependent or outcome variables whereas explanatory variables are used to predict the value of response variables. They are also referred by independent variables, predictors, covariates, regressors, factors and carriers. The relation between response variable and explanatory variable:

$$y = b_0 + b_1 x_1 + \ldots + b_n x_n + \epsilon \tag{1}$$

where y is the response variable, x is the explanatory variable, n is the number of explanatory variables, ϵ is the error mismatch in appropriation and $b_0, b_1, \ldots b_n$ is the unknown coefficients that are determined from the data. [2]

There are over 15 types of regression among them Linear Regression is the most ordinary form of regression. Conceptional purposes of regression analysis are to discern causal relationships between the independent and dependent variables, having a great deal in common with machine learning and used for prediction and forecasting.

A. REGULARIZATION

Regularization technique is used to avoid overfitting of data. The lasso regularization has a greater impact than the ridge regularization. To implement regularization, penalty term is added to the best fit derived from the trained dataset. Lasso regression (L1 regularization) shrinks the coefficient towards zero while ridge regression (L2 regularization) does not shrinks the coefficient close to zero. [1]

B. PENALTY TERM LAMBDA (λ)

The amount of the penalty can be bettered using a constant called lambda (λ). The penalty term (lambda) regularises the coefficients so that the optimization function is penalised if the coefficients take the high values. Lambda's value can range from 0 to infinity. Lambda controls the amount of shrinkage, hence when lambda closes to infinite it will greatly impact the shrinkage penalty. Penalty term does not affect when lambda tends to 0 where penalty increases with the value of penalty grows to infinity. [4]

C. CROSS-VALIDATION

Cross validation is the procedure in regression analysis to examine the predictive validity of the regression equation. Here the penalty term lambda can be determined by Cross-Validation method. K fold cross validation used for determining the best value for lambda. Cross validation splits data into different categories of train and test to get better accuracy. When performing K-fold cross-validation, the data is first partitioned into k parts of (approximately) equal size, called folds. Next, a sequence of models is trained. The first model is trained using the first fold as the test set, and the remaining folds are used as the training set then the accuracy is evaluated on fold 1. This process is repeated using the remaining folds as test sets. In the end, accuracy of the k values is calculated and collected. Cross-validation can help avoiding overfitting to a validation set easier.

D. MULTICOLLINEARITY

Multicollinearity is a statistical condition which occurs when the two or more independent variables are significantly corelated with one another, and the dependent variables are not. It should be excluded from the dataset as it causes issues while determining the most influential variable.

IV. DATASET

This project involves analyzation and comparison of lasso and ridge regression on both regular and ADMM methods on two high-dimensional datasets: Student performance and heart disease dataset. This paper provides regression models for datasets using the Python programming language.

A. STUDENT PERFORMANCE DATASET

The first is a set of information on student performance. The impact of parents' education on students' grades in three different subjects is included in this dataset, which aids in improving students' test performance. The students' performance dataset contains 1000 rows and 8 columns which includes Gender, race/ethnicity, parental level of education, lunch, test preparation course, math score, reading score and writing score.

B. HEART DISEASE DATASET

The second dataset used in this study is a heart disease dataset. This dataset's purpose is to uncover any accurate heart health markers or forecast cardiovascular events. The heart disease dataset includes 303 rows and 14 columns containing age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise induced angina, old peak, the slope of the peak exercise, number of major vessels, Thal and target.

C. PREPROCESSING

Real-world data lacks attribute values or it contains errors which make the data noisy. Preprocessing data overcomes the above problem. Data preprocessing is a data mining approach that involves converting raw data into an understandable format. Label encoding () function in data preprocessing converts the labels or categorical values into numeric format so that it will be machine-readable. The program reads the preprocessed datasets in the form of csv file, splits it into testing and training datasets, in the ratio 3:1(75% : 25%). Finally, regression model is created for the training dataset. Drop () function in python eliminates the attributes which the model wants to predict.

V. RIDGE

Ridge regression is a more robust form of linear regression that introduces a small amount of bias in order to obtain better long-term predictions. If multi-collinearity is present, the estimates of at least squares are unbiased, but the variations are large and therefore far from the true value. The ridge regression is identical to the least-squares, except that the ridge coefficient is estimated by minimizing a slightly different amount. It is hoped that the net effect will provide a more reliable estimate. The ridge coefficients minimize a penalized residual sum of squares, [5]

$$\widehat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{P} X_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{2}$$

where y_i is the value of the response variable in the i^{th} trial, β_0 is the intercept coefficient, β_j is the regression coefficient for j = 1,...,k, X_{ij} is the j^{th} component of X_i which X_i is a known constant namely the value of the predictor variable in the i^{th} trial. $\lambda \geq 0$ is a tuning parameter determines the amount of shrinkage. The coefficients are shrunk towards zero. The ridge problem can also be written as , [6]

$$\widehat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{P} X_{ij}\beta_j \right)^2 \tag{3}$$

subject to

$$\sum_{j=1}^{P} \beta_j^2 \leq t \tag{4}$$

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

Ridge regression follows the same assumptions as least squares regression, with the exception that normality is not assumed. It decreases variability by shrinking coefficients, leading to better prediction accuracy at the cost of a modest increase in bias in most cases. It reduces the coefficients to nearly zero but not precisely zero, making it unsuitable for feature selection. Ridge regression will not produce a readily interpretable sparse model when the number of predictors is big. [5]

VI. RIDGE USING ADMM

A. PARALLEL IMPLEMENTATION

Ridge's objective function for executing in distributed/parallel mode by dividing over all samples,

$$\min_{x^{(i)},z} \frac{1}{2} \sum_{i=1}^{N} ||A^{(i)}x^{(i)} - b^{(i)}||_2^2 + \lambda ||z||_2^2 \tag{5}$$

$$\text{st } x^{(i)} - z = 0 \tag{6}$$

for i=1....N,

The modified update rule for each variable is as follows:
1. Minimization of x_i

$$x_{t+1}^{(i)} = \underset{x^{(i)}}{\operatorname{argmin}} \frac{1}{2}||A^{(i)}x^{(i)} - b^{(i)}||_2^2 + \frac{\rho}{2}||x^{(i)} - z_t + \frac{1}{\rho}v_t^{(i)}||_2^2 \tag{7}$$

which has a closed form solution

$$x_{t+1}^{(i)} = (A^{(i)T}A^{(i)} + \rho I)^{-1}(A^{(i)T}b^{(i)} + \rho z_t - v_t^{(i)}) \tag{8}$$

for i=1....N

2. Minimization of z

$$z_{(t+1)} = \underset{z}{\operatorname{argmin}}\ \lambda||z||_2^2 + \frac{\rho}{2}||\overline{x}_{t+1} - z + \frac{1}{\rho}\overline{v}_t||_2^2 \tag{9}$$

$$z_{t+1} = \frac{\rho\overline{x}_{(t+1)} + \overline{v}_t}{2\lambda + \rho} \tag{10}$$

3. Update of dual variable v_t

$$v_{t+1}^{(i)} = v_t^{(i)} + \rho(x_{t+1}^{(i)} - z_{t+1}) \tag{11}$$

B. SERIES IMPLEMENTATION

The objective function by renaming one of the two variables and introduce an equality constraint to bring in the ADMM form. The modified Ridge objective is

$$\min_{x,z} \frac{1}{2}||Ax - b||_2^2 + \lambda||z||_2^2 \tag{12}$$

$$\text{st } x = z \tag{13}$$

The modified update rule for each variable is as follows:
1. Minimization of x

$$x_{t+1} = \underset{x}{\operatorname{argmin}} \frac{1}{2}||Ax - b||_2^2 + \frac{\rho}{2}||x - z_t + \frac{1}{\rho}v_t||_2^2 \tag{14}$$

which has a closed form solution

$$x_{t+1} = (A^TA^+\rho I)^{-1}(A^Tb + \rho z_t - v_t) \tag{15}$$

2. Minimization of z

$$z_{(t+1)} = \underset{z}{\operatorname{argmin}}\ \lambda||z||_2^2 + \frac{\rho}{2}||x_{t+1} - z + \frac{1}{\rho}v_t||_2^2 \tag{16}$$

$$z_{t+1} = \frac{\rho\ x_{t+1} + v_t}{2\lambda + \rho} \tag{17}$$

3. Update of dual variable v_t

$$v_{t+1} = v_t + \rho(x_{t+1} - z_{t+1}) \tag{18}$$

VII. LASSO

Lasso is the abbreviated version of the least absolute shrinkage and selection operator. Lasso can be useful for calculating regression coefficients and selecting variables.The model will be developed utilizing feature selection, which implies that only a set of characteristics from the dataset will be chosen, which prevents model overfitting.The penalised least squares regression with L1-penalty function is known as the LASSO. [6]
The LASSO estimate can be defined by,

$$\widehat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{P} X_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{19}$$

which can also be written as

$$\widehat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{P} X_{ij}\beta_j \right)^2 \tag{20}$$

subject to

$$\sum_{j=1}^{P} \beta_j^2 \leq t \tag{21}$$

Lasso can be useful for calculating regression coefficients and selecting variables. The Ridge regression is similar to this. If the set of predictors has a high degree of correlation, LASSO selects one and reduces the others to zero. It lowers the variability of the estimates by decreasing some of the coefficients to zero, resulting in models that are simple to understand. [5]

VIII. LASSO USING ADMM

A. PARALLEL IMPLEMENTATION

LASSO'S objective function for executing in distributed/parallel mode by dividing over all samples,

$$\min_{x^{(i)},z} \frac{1}{2} \sum_{i=1}^{N} ||A^{(i)}x^{(i)} - b^{(i)}||_2^2 + ||z||_1 \tag{22}$$

$$\text{st } x^{(i)} - z = 0 \tag{23}$$

for i=1....N,

The modified update rule for each variable is as follows:
1. Minimization of x_i

$$x_{t+1}^{(i)} = \underset{x^{(i)}}{\operatorname{argmin}} \frac{1}{2}||A^{(i)}x^{(i)} - b^{(i)}||_2^2 + \frac{\rho}{2}||x^{(i)} - z_t + \frac{1}{\rho}v_t^{(i)}||_2^2 \tag{24}$$

which has a closed form solution

$$x_{t+1}^{(i)} = (A^{(i)T}A^{(i)} + \rho I)^{-1}(A^{(i)T}b^{(i)} + \rho z_t - v_t^{(i)}) \tag{25}$$

for i=1....N,

2. Minimization of z

$$z_{t+1} = \underset{z}{\operatorname{argmin}}\ \lambda||z||_1 + \frac{\rho}{2}||\overline{x}_{t+1} - z + \frac{1}{\rho}\overline{v}_t||_2^2 \tag{26}$$

$$z_{t+1} = \overline{x}_{t+1} + \frac{\overline{v}_t}{\rho} - \frac{\lambda}{\rho}sign(z) \tag{27}$$

3. Update of dual variable v_t

$$v_{t+1}^{(i)} = v_t^{(i)} + \rho(x_{t+1}^{(i)} - z_{t+1}) \tag{28}$$

B. SERIES IMPLEMENTATION

The objective function by renaming one of the two variables and introduce an equality constraint to bring in the ADMM form. The modified LASSO objective is:

$$\min_{x,z} \frac{1}{2}||Ax - b||_2^2 + ||z||_1 \tag{29}$$

$$\text{st } x = z \tag{30}$$

The update rule for each of the two variables is as follows:
1. Minimization of x

$$x_{t+1} = \underset{x}{\operatorname{argmin}} \frac{1}{2}||Ax - b||_2^2 + \frac{\rho}{2}||x - z_t + \frac{1}{\rho}v_t||_2^2 \tag{31}$$

which has a closed form solution

$$x_{t+1} = (A^TA + \rho I)^{-1}(A^Tb + \rho z_t - v_t) \tag{32}$$

2. Minimization of z

$$z_{t+1} = \underset{z}{\operatorname{argmin}}\ \lambda||z||_1 + \frac{\rho}{2}||x_{t+1} - z + \frac{1}{\rho}v_t||_2^2 \tag{33}$$

$$z_{t+1} = x_{t+1} + \frac{v_t}{\rho} - \frac{\lambda}{\rho}sign(z) \tag{34}$$

3. Update of dual variable v_t

$$v_{t+1} = v_t + \rho(x_{t+1} - z_{t+1}) \tag{35}$$

IX. RESULT

The models was tested using two data sets as classification and regression problems. The heart disease data set is used for checking the performance of the models on classification problems because such problems predicts binary values (0 and 1) and the student performance estimates continuous values, allowing the model's performance in regression problems to be checked.

The processing time was estimated for both ordinary regression models and that by using ADMM, by which the models were evaluated and the results are tabulated.

TABLE I
TIME CONSUMPTION OF REGRESSION MODELS

Regression Model	Student Performance Dataset	Heart Disease Dataset
lasso	0.025	0.027
$lasso_{admm^s}$	0.0354	0.1252
$lasso_{admm^p}$	0.6487	0.1878
ridge	0.0001	0.0002
$ridge_{admm^s}$	0.001	0.001
$ridge_{admm^p}$	0.1869	0.2104

Where $admm^s$ denotes admm implementation in series, and $admm^p$ represents admm implementation in parallel.

Table I shows the time taken for each of the model to train using the given data. As the values indicate, the time taken by standard models is greater than those optimized with admm. We can see that admm significantly reduces the time required for lasso and ridge.

As the heart disease dataset was used to test the model as a classification problem, the performance of the model was assessed using accuracy, and as the student performance data set was used to test the model as a regression problem, the performance of the model was assesed using R2 Score.

TABLE II
PERFORMANCE OF REGRESSION MODELS

Regression Model	**Student Performance Dataset (R2 score)**	**Heart Disease Dataset (Accuracy)**
lasso	0.9018	0.6721
$lasso_{admm^s}$	0.9359	0.8852
$lasso_{admm^p}$	0.8367	0.4262
ridge	0.9292	0.8852
$ridge_{admm^s}$	0.9380	0.8852
$ridge_{admm^p}$	0.8351	0.4262

Where $admm^s$ denotes admm implementation in series, and $admm^p$ represents admm implementation in parallel.

Table II shows the accuracy and r2 score of the regression models. It is evident that as we optimise the model with admm, the accuracy of the model improves. It can also be seen that when admm is applied, the r2 score is higher.

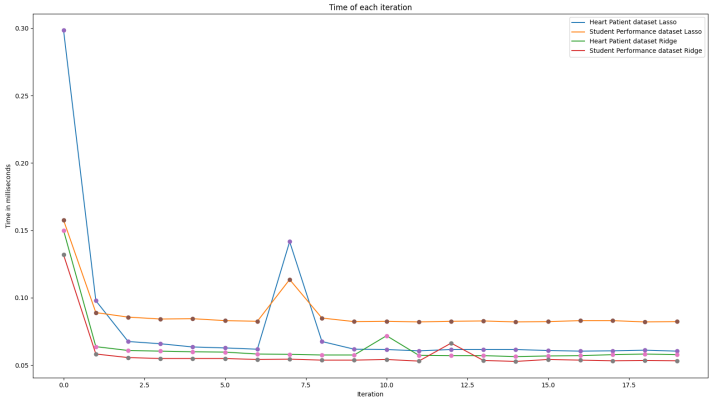

Fig. 1. Time took for iterations on different runs

The Figure 1 shows the variation in time for each iteration for the optimized regression models. The graph depicts that the time for the first iteration is higher for all the models.

X. CONCLUSION

In this paper, we present the comparison of ordinary lasso and ridge regression with ADMM implementation of both regression techniques. Simulation results on two typical datasets demonstrate that the time taken by the regression models optimised with ADMM is lower than the time taken by the standard regression models. Since two datasets were used to test the model as classification and regression problems, the accuracy and r2 score of the optimised regression models is higher than the standard models. In modern distributed frameworks, the ADMM algorithm can be easily parallelized and implemented. ADMM has been recognised as a versatile strategy for handling large-scale machine learning and signal processing problems quickly, as demonstrated by the simulation results.

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.

REFERENCES

[1] Zaikarina, Hilda, Anik Djuraidah, and Aji Hamim Wigena. "Lasso and ridge quantile regression using cross validation to estimate extreme rainfall." Global Journal of Pure and Applied Mathematics 12.3 (2016): 3305-3314.

[2] Freund, Rudolf J., William J. Wilson, and Ping Sa. Regression analysis. Elsevier, 2006.

[3] Chatterjee, Samprit, and Ali S. Hadi. Regression analysis by example. John Wiley Sons, 2013.

[4] García-Nieto, Paulino José, Esperanza García-Gonzalo, and José Pablo Paredes-Sánchez. "Prediction of the critical temperature of a superconductor by using the WOA/MARS, Ridge, Lasso and Elastic-net machine learning techniques." Neural Computing and Applications 33.24 (2021): 17131-17145.

[5] Muthukrishnan, R., and R. Rohini. "LASSO: A feature selection technique in predictive modeling for machine learning." 2016 IEEE international conference on advances in computer applications (ICACA). IEEE, 2016.

[6] Xin, Seng Jia. "Modelling house price by using ridge regression and lasso regression." (2018).

[7] Refaeilzadeh, P., Lei Tang, and Huan Liu. "Cross Validation, Encyclopedia of Database Systems (EDBS)." Arizona State University, Springer 6 (2009).

[8] Daoud, Jamal I. "Multicollinearity and regression analysis." Journal of Physics: Conference Series. Vol. 949. No. 1. IOP Publishing, 2017.

[9] Alin, Aylin. "Multicollinearity." Wiley Interdisciplinary Reviews: Computational Statistics 2.3 (2010): 370-374.

[10] Boyd, Stephen, Neal Parikh, and Eric Chu. Distributed optimization and statistical learning via the alternating direction method of multipliers. Now Publishers Inc, 2011.

[11] Thevaraja, Mayooran, and Azizur Rahman. "Assessing robustness of regularized regression models with applications." International Conference on Management Science and Engineering Management. Springer, Cham, 2019.