Volume 9, Issue 09, September 2024 ISSN (online): 2456-0006

International Journal of Science Technology Management and Research

Available online at: www.ijstmr.com

GRADIENT DESCENT A BETTER APPROACH FOR PERFORM OPTIMIZATION FOR REGRESSION MODELING

Rahul Singh Kanasiya
M E 4th semester
Jawaharlal Institute of Technology Borawan
Khargone (M.P.) 451228

Mr. Kamlesh Patidar
Assistant Professor CSE department
Jawaharlal Institute of Technology Borawan
Khargone (M.P.) 451228

Abstract*: Outliers are unusual values in dataset, and they can distort statistical analyses and violate their assumptions. Analysts will confront outliers and be forced to make decisions about what to do with them. Given the problems they can cause, might think that it's best to remove them from data. Removing outliers is legitimate only for specific reasons. Outliers can be very informative about the subject-area and data collection process. It's essential to understand how outliers occur and whether they might happen again as a normal part of the process or study area. Unfortunately, resisting the temptation to remove outliers inappropriately can be difficult. Outliers increase the variability in data, which decreases statistical power. Consequently, excluding outliers can cause results to become statistically significant. One cannot recognize outliers while collecting data; we won't know what values are outliers until we begin analyzing the data. Many statistical tests are sensitive to outliers and therefore, the ability to detect them is an important part of data analytics. The interpretability of an outlier model is very important, and decisions seeking to tackle an outlier need some context or rationale. Outliers sometimes can be helpful indicators. We have used R language to implement the Hierarchical clustering approach. We found that the global cluster is generate by all three basic approaches are same but the number of object in clusters are in each are different. By the experimental analysis we found that all three methods has different approach but by the Dendrogram we easy find the global

Keywords: Outlier, Single, Average, Complete, Clusters, Global

I. INTRODUCTION

The gradient descent algorithm is an optimization algorithm used to minimize an objective function, commonly called the cost or loss function in machine learning. This cost function is a mathematical function used to calculate loss or error which is the difference between the model's predicted value and the actual value in the real world for a given input data set. While training a model, the ask is to find a high-performance model (model with a minimal loss). To find the best model, we need to minimize the cost function. The algorithm used to minimize the cost function is called the gradient descent algorithm[9,10].

The term "gradient descent" is derived from two key concepts: "gradient" and "descent."

1. Gradient: In mathematics and physics, the gradient of a function at any point is a multi-dimensional vector that points in the direction of the steepest increase of that function. The magnitude of the gradient vector at a point corresponds to the rate of greatest increase in the function's value. In the context of a cost function in machine learning, the gradient indicates the direction in which the cost function rises most rapidly.
2. Descent: The word "descent" refers to the method's objective of moving downwards to find the minima of the function. Since the gradient points toward the greatest increase, the opposite direction, in which the algorithm moves, will lead to the steepest decrease. By iteratively moving in the direction opposite to the gradient (hence "descending"), the algorithm finds the point where the cost or loss function attains its minimum value.

Thus, "gradient descent" is aptly named as it describes the process of descending (minimizing) a function by moving against (opposite to) its gradient. The following plot can be used to understand the gradient descent algorithm.

III. GRADIENT DESCENT ALGORITHMS

Several other gradient descent algorithms are prevalent. Below is a detailed definition of each of the following: [11,12]

1. Momentum: This algorithm introduces a momentum term that allows the optimization process to continue in the same direction as the previous iteration. This helps the algorithm to converge faster and smoother than standard gradient descent.
2. Nesterov Accelerated Gradient: This algorithm is an extension of momentum that improves the convergence rate by computing the gradient ahead of the current position. This can lead to a more accurate direction of descent and faster convergence.
3. Adagrad: This algorithm adapts the learning rate to the parameters by scaling the learning rate based on the historical sum of the squared gradients. This means the learning rate decreases for parameters that have received significant updates.
4. Adadelta: This algorithm is similar to Adagrad, but instead of accumulating all past squared gradients, it limits the window of accumulated past gradients. This helps reduce the aggressive learning rate decay and adapt quickly to changing gradients.

5. RMSprop: This algorithm also adapts the learning rate based on the past gradients but uses an exponentially weighted moving average to limit the window of accumulated past gradients. This makes the algorithm more stable than Adagrad for non-convex problems.
6. Adam: This algorithm combines the ideas of momentum and RMSprop. It estimates adaptive learning rates for each parameter and stores the exponentially weighted moving averages of the gradients and squared gradients. This algorithm is known for its speed and robustness in optimizing complex cost functions.

TYPES OF REGRESSION TECHNIQUES

There are various kinds of regression techniques available to make predictions. These techniques are mostly driven by three metrics number of independent variables, type of dependent variables and shape of regression line[12,13,14,15]

A. Linear Regression

It is one of the most widely known modeling techniques. Linear regression is used while learning predictive modeling. In this technique, the dependent variable is continuous, independent variable(s) can be continuous or discrete, and nature of regression line is linear.Linear Regression establishes a relationship between dependent variable (Y) and one or more independent variables (X) using a best fit straight line (also known as regression line). It is represented by an equation Y=a+b*X + e**,** where a is intercept, b is slope of the line and e is error term. This equation can be used to predict the value of target variable based on given predictor variable(s).

B. Logistic Regression

Logistic regression is used to find the probability of event=Success and event=Failure. Logistic regression is used when the dependent variable is binary (0/ 1, True/ False, Yes/ No) in nature. Logistic regression is widely used for classification problems. Logistic regression doesn't require linear relationship between dependent and independent variables. It can handle various types of relationships because it applies a non-linear log transformation to the predicted odds ratio. To avoid over fitting and under fitting, we should include all significant variables. A good approach to ensure this practice is to use a step wise method to estimate the logistic regression It requires large sample sizes because maximum likelihood estimates are less powerful at low sample sizes than ordinary least square

C. Polynomial Regression

A regression equation is a polynomial regression equation if the power of independent variable is more than 1. The equation below represents a polynomial equation. While there might be a temptation to fit a higher degree polynomial to get lower error, this can result in over-fitting. Always plot the relationships to see the fit and focus on making sure that the curve fits the nature of the problem. Here is an example of how plotting can help. Especially look out for curve towards the ends and see whether those shapes and trends make sense. Higher polynomials can end up producing wierd results on extrapolation.

D. Stepwise Regression

This form of regression is used when we deal with multiple independent variables. In this technique, the selection of independent variables is done with the help of an automatic process, which involves no human intervention. This feat is achieved by observing statistical values like R-square, t-stats and AIC metric to discern significant variables. Stepwise regression basically fits the regression model by adding/dropping covariates one at a time based on a specified criterion. Some of the most commonly used Stepwise regression methods are listed below. Standard stepwise regression does two things. It adds and removes predictors as needed for each step Forward selection starts with most significant predictor in the model and adds variable for each step. Backward elimination starts with all predictors in the model and removes the least significant variable for each step. The aim of this modeling technique is to maximize the prediction power with minimum number of predictor variables. It is one of the method to handle higher dimensionality of data set.

E. Ridge Regression

Ridge Regression is a technique used when the data suffers from multicollinearity (independent variables are highly correlated). In multicollinearity, even though the least squares estimate (OLS) are unbiased; their variances are large which deviates the observed value far from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors. y=a+ b*x. The assumptions of this regression are same as least squared regression except normality is not to be assumed. Ridge regression shrinks the value of coefficients but doesn't reaches zero, which suggests no feature selection feature. This is a regularization method and uses l2 regularization.

F. Lasso Regression

Similar to Ridge Regression, Lasso (Least Absolute Shrinkage and Selection Operator) also penalizes the absolute size of the regression coefficients. In addition, it is capable of reducing the variability and improving the accuracy of linear regression models. Look at the equation below: Lasso regression differs from ridge regression in a way that it uses absolute values in the penalty function, instead of squares. This leads to penalizing (or equivalently constraining the sum of the absolute values of the estimates) values which causes some of the parameter estimates to turn out exactly zero. Larger the penalty applied, further the estimates get shrunk towards absolute zero. This results to variable selection out of given n variables. The assumptions of lasso regression is same as least squared regression except normality is not to be assumed. Lasso Regression shrinks coefficients to zero (exactly zero), which certainly helps in feature selection. Lasso is a regularization method and uses l1 regularization. If group of predictors are highly correlated, lasso picks only one of them and shrinks the others to zero

III. LITERATURE SURVEY

In 2015 Supichaya Sunthornjittanon "Linear Regression Analysis on Net Income of an Agrochemical Company in Thailand." . They analyze the ABC Company's data and verify whether the regression analysis methods and models would work effectively in the ABC Company based in Bangkok, Thailand. After the data are collected, models are created to examine the contribution of each of the company's financial factors to the net income of the company. The final model is selected using Stepwise Regression Methods. After the significant category is found deeper analysis is conducted of the fungi category. Each individual fungicide product line is then regressed on the net income. After running these additional models, it was discovered that the Net Income from Fungicide remained the most significant model. Time is also taken into account to see if it plays some role in the net income, but after the analysis, it was found that time is not significant in this case[1].

In 2016 Sandhya Jain , Sunny Chourse proposed "Regression Analysis – Its Formulation and Execution In Dentistry". Prediction and estimation is the mainstay in the treatment planning in dentistry. The aim of this article is to provide a simple yet holistic approach to the understanding of the concepts of Regression Analysis along with its use and misuse, advantages and disadvantages pertaining to the art and science of dentistry In the formulation and execution of an dental treatment plan, the variables involved in the decision making are often poorly characterized and incompletely validated. For these reasons we have to rely on the mean values or go for a wild guess. The regression analysis is a statistical technique that deals with the analysis of relationship between such variables. They help us to predict the duration, course and outcome of the treatment parameters. It is a complex process of predicting/estimating the magnitude of some unknown characteristics which might be involved in the growth and treatment of a given patient [2].

In 2017 Radek Silhavy and Petr Silhavy " Analysis and selection of a regression model for the Use Case Points method using a stepwise approach". They investigates the significance of use case points (UCP) variables and the influence of the complexity of multiple linear regression models on software size estimation and accuracy. Stepwise multiple linear regression models and residual analysis were used to analyze the impact of model complexity.. The best performing model (Model D) contains an intercept, linear terms, and squared terms. The results of several evaluation measures show that this model's estimation ability is better than that of the other models tested. Model D also performs better when compared to the UCP model, whose Sum of Squared Error was 268,620 points on Dataset 1 and 87,055 on Dataset 2. Model D achieved a greater than 90% reduction in the Sum of Squared Errors compared to the Use Case Points method on Dataset 1 and a greater than 91% reduction on Dataset 2[3].

In 2017 N J Gogtay, S P Deshpande and UM Thatte proposed "Principles of Regression Analysis". Regression analysis is a statistical tool that helps evaluate relationships between a dependent variable and one or more independent or predictor variables. More specifically, it helps us understand how the dependent variable changes with changes in the independent variable and thus finds its application in forecasting and predicting. The technique must however be used with clear understanding of the assumptions in each type of regression analysis, their limitations and the potential error that can occur when models. are applied to a larger population. They apply this equation to the population for making a prediction, and able to predict either the systolic blood pressure perfectly. Hence, they need to taken into account an "error" or "deviation" that is likely to occur when this equation is used[4].

In 2018 Ira Sharma and Sampurna Kakchapati proposed "Linear Regression Model to Identify the Factors Associated with Carbon Stock in Chure Forest of Nepal". Their aims to assess the factors associated with carbon stock in Chure forest of Nepal. The linear regression showed a good fit of the model (adjusted $R2$ = 83.75%) with the results that the stem volume (sv), diameter at breast height (dbh), and the number of trees per plot showed statistically significant (p value ≤ 0.05) positive association with carbon stock. The highest carbon stock was associated with sv more than 199m3/ha, average dbh more than 43.3 cm/plot, and number of trees more than 20/plot, whereas the altitude, geographical location, and ownership had no statistical associations at all. The results can be of use to the government for enhancing carbon stock in Chure that supports both natural resource conservation and United Nations-Reducing Emission from Deforestation and Forest Degradation program to mitigate carbon emission issues[5].

In 2019 Anjali Pant and R.S. Rajput proposed "Linear Regression Analysis Using R for Research and Development" .The future forecasting opportunities and risks estimation are the most prominent prerequisite for a successful business. Regression analysis can go far beyond forecasting. The linear regression analysis technique is a statistical method that allows examining the linear relationship between two or more quantitative variables of interest. The chapter covers the fundamentals of linear regression, regression model equation, the test of significance, coefficient of determination, and residual with residual analysis. R is a potent statistical computation tool, all the computation of chapter conducted by using R. They explain R computations for the regression model with the help of two examples. Regression model also visualized with the help of some plots that are created with the help of R[6].

In 2019 Hazlina Darman, Sarah Musa, and Rajasegeran Ramasamy, proposed "Predicting Students' Final Grade in Mathematics Module using Multiple Linear Regression" Learning analytics is the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs. They proposed, multiple linear regression model is developed to predict the students' score in Final Exam using their assessments' score. a regression model has been developed with the aid of Statistical Package for Social Sciences (SPSS) analysis tool. The graphical representations and tables are presented to illustrate the models The findings from this study has achieved the objective of developing a model that can predict the students' performance in final exam. The analysis has shown that the students who perform well in Test 1 and Test 2 have better chances of getting good scores in final exam, and vice versa[7].

In 2019 Gaurav Pandeya, Poonam Chaudharya and Rajan Guptab, proposed " SEIR and Regression Model based COVID-19 outbreak predictions in India". They proposed a study to outbreak of this disease has been analyzed for India till 30th March 2020 and predictions have been made for the number of cases for the next 2 weeks. SEIR model and Regression model have been used for predictions based on the data collected from John Hopkins University repository in the time period of 30th January 2020 to 30th March 2020. The performance of the models was evaluated using RMSLE and achieved 1.52 for SEIR model and 1.75 for the regression model. The RMSLE error rate between SEIR model and Regression model was found to be 2.01. Also, the value of R0 which is the spread of the disease was calculated to be 2.02. Expected cases may rise between 5000-6000 in the next two weeks of time. This study will help the Government and doctors in preparing their plans for the next two weeks. Based on the predictions for short-ter In this study, two machine learning models SEIR and Regression were used to analyse and predict the change in spread of COVID-19 disease[8].

In 2019 K. K. Baseer, Vikram Neerugatti and Akella Amarendra Babu “Analysing various Regression Models for Data Processing”. Regression Analysis (RA) is utilized for prediction and determination, where its utilization has generous cover with the field of Artificial Intelligence. RA is a measurable procedure’s for assessing the relationship among variables (one dependent and one or more independent). Its helps us to predict and that is why it is also called as predictive analysis model. They used vehicle data like velocity with which traffic move’s, gradient, actual velocity to predict the velocity profile of the vehicle. They analyzed various regression models like linear regression, multivariate linear regression and nonlinear regression. The outcome of this work is to write a function for every model that everyone can reuse that without using pre-defined functions in languages and plotting the given data to best fit for analyzing. They discussed various regression models and their uses. Regression analysis helps us to predict things accurate; we can develop this using exponential, logistic type of regression. They used matlab software and wrote a code for that without using pre-defined functions like fitln(), fitlm()[9].

In 2020 Khushbu Kumari, Suniti Yadav “Linear Regression Analysis Study”. Linear regression is a statistical procedure for calculating the value of a dependent variable from an independent variable. Linear regression measures the association between two variables. It is a modeling technique where a dependent variable is predicted based on one or more independent variables. Linear regression analysis is the most widely used of all statistical techniques. They explain the basic concepts and explain how we can do linear regression calculations in SPSS and excel. The techniques for testing the relationship between two variables are correlation and linear regression. Correlation quantifies the strength of the linear relationship between a pair of variables, whereas regression expresses the relationship in the form of an equation. They used simple examples and SPSS and excel to illustrate linear regression analysis and encourage the readers to analyze their data by these techniques[10].

In 2020 Samit Ghosal , Sumit Sengupta and Milan Majumder proposed “Linear Regression Analysis to predict the number of deaths in India due to SARS-CoV-2 at 6 weeks from day 0 (100 cases - March 14th 2020)” They analysis tracing a trend related to death counts expected at the 5th and 6th week of the COVID-19 in India.Material and methods:. Using auto-regression technique and using week 5 death count as input the linear regression model predicted week 6 death count in India to be 467, while keeping at the back of our mind the risk of over-estimation by most of the risk-based models. Projected death rate (n) is 211 and467 at the end of the 5th and 6th week from now, respectively. According to our analysis, if situation continue in present state; projected death rate (n) is 211 and 467 at the end of the 5th and 6th week from now, respectively. Keeping these projected mortality data in mind, current measured for containment of COVID-19 must be strengthened or supplemented [11].

IV. Proposed Approach

The steps for performing gradient descent are as follows:
Step 1: Select a learning rate
Step 2: Select initial parameter values as the starting point
Step 3: Update all parameters from the gradient of the training data set, i.e. compute
Step 4: Repeat Step 3 until a local minimum is reached
Under batch gradient descent, the gradient, is calculated at every step against a full data set. When the training data is large, computation may be slow or require large amounts of computer memory.

Experimental Analysis

Regression line the first updated value of m= 3.2787 and b=0.5324

TABLE 6.1

Updated value of m= 3.2787 and b=0.5324 with error

S. No.	Value of m	Value of c	Error
1	3.2787	0.5324	-61.9312

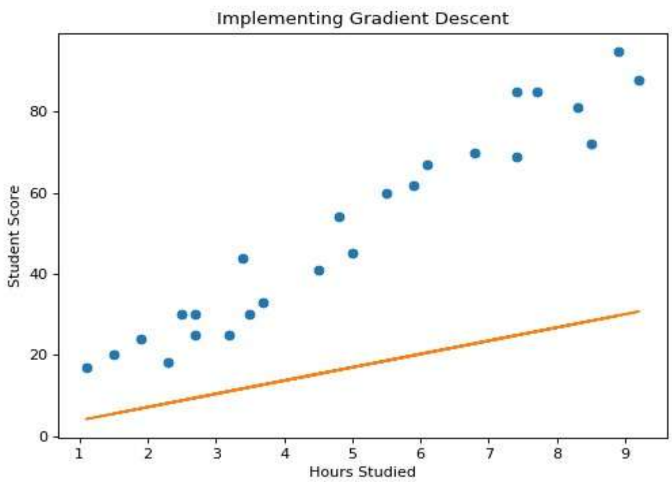

Fig 1 Updated value of m= 3.2787 and b=0.5324 with error -61.9312

CONCLUSION

Regression analysis is a form of predictive modeling technique which investigates the relationship between a dependent or target and independent variable or predictor. This technique is used for forecasting, time series modeling and finding the causal effect relationship between the variables. Regression analysis is an important tool for modeling and analyzing data. Regression analysis is the study of two variables in an attempt to find a relationship, or correlation A regression line is a straight line that attempts to predict the relationship between two points, also known as a trend line or line of best fit.

REFERENCE

1. Supichaya Sunthornjittanon "Linear Regression Analysis on Net Income of an Agrochemical Company in Thailand" Portland State University PDXScholar University Honors Theses University Honors College
2. Sandhya Jain , Sunny Chourse "Regression Analysis – Its Formulation and Execution In Dentistry ". Journal of Applied Dental and Medical Sciences NLM ID: 101671413 ISSN:2454-2288 Volume 2 Issue 1 January - March 2016
3. Radek Silhavy , Petr Silhavy, Zdenka Prokopova " Analysis and selection of a regression model for the Use Case Points method using a stepwise approach" The Journal of Systems and Software 125 (2017) 1–14 Contents lists available at Science Direct The Journal of Systems and Software journal homepage: www.elsevier.com/locate/jss.
4. NJ Gogtay, SP Deshpande, UM Thatte "Principles of Regression Analysis" Journal of The Association of Physicians of India Vol. 65 April 2017
5. Ira Sharma and Sampurna Kakchapati "Linear Regression Model to Identify the Factors Associated with Carbon Stock in Chure Forest of Nepal" Hindawi Scientific a Volume 2018, Article ID 1383482, 8 pages https://doi.org/10.1155/2018/1383482.
6. Anjali Pant R.S. Rajput "Linear Regression Analysis Using R for Research and Development" See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/336981868
7. Hazlina Darman, Sarah Musa Predicting Students' Final Grade in Mathematics Module using Multiple Linear Regression International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7 Issue-5S, January 2019.
8. Gaurav Pandeya, Poonam Chaudharya "SEIR and Regression Model based COVID-19 outbreak predictions in India" Department of CSE & IT, The NorthCap University, India DeenDayalUpadhyaya College, University of Delhi, India Defence Research & Development Organization, India a{Email: gaurav16csu120@ncuindia.edu}.
9. K. K. Baseer, Vikram Neerugatti, Analysing various Regression Models for Data Processing International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-8 June, 2019.
10. Khushbu Kumari, Suniti Yadav Linear Regression Analysis Study [Downloaded free from http://www.j-pcs.org on Friday, July 17, 2020, IP: 157.34.76.130] Journal of the Practice of Cardiovascular Sciences | Published by Wolters Kluwer – Medknow.
11. Samit Ghosal , Sumit Sengupta "Linear Regression Analysis to predict the number of deaths in India" due to SARS-CoV-2 at 6 weeks from day 0 (100 cases - March 14th 2020) Contents lists available at Science Direct Diabetes & Metabolic Syndrome: Clinical Research & Reviews journal homepage: www.elsevier.com/locate/dsx
12. Shen Rong, Zhang Bao-wen " The research of regression model in machine learning field MATEC Web of Conferences 176, 01033 (2018) https://doi.org/10.1051/matecconf/201817601033IFID 2018.
13. Syarifah Diana Permaia, Heruna Tantyb Linear regression model using bayesian approach for energy performance of residential building 2018 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0/) 3rd International Conference on Computer Science and Computational Intelligence 2018.
14. Marno Verbeek "Using linear regression to establish empirical relationships Using linear regression to establish empirical relationships". IZA World of Labor 2017: 336 doi: 10.15185/izawol.336 | Marno Verbeek © | February 2017 | wol.iza.org
15. Katarina Valaskova , Tomas Kliestik, Lucia Svabova and Peter Adamko Financial Risk Measurement and Prediction Modelling for Sustainable Development of Business Entities Using Regression Analysis Faculty of Operation and Economics of Transport and Communications, Received: 20 April 2018; Accepted: 20 June 2018; Published: 23 June 2018.