The Prediction on Specific Datasets Using Linear Regression with Regularization and Linear Regression with Gradient Descent

Yichen Ma*[1,a],Xinyue Li[2,b],Botao Cai[3,c]

2543667553@qq.com[a], lxynflsyh2022@outlook.com[b], 3268305420@qq.com[c]

Macau University of Science and Technology, Macau, China[1]
Nanjing Foreign Language School Yuhua International Academy, Nanjing, China[2]
Nanjing Foreign Language School Xianlin Campus, Nanjing, China[3]

Abstract—This article mainly discusses the methods of predicting specific datasets using linear regression with regularization and linear regression with gradient descent. Firstly, the basic principles and mathematical models of these two methods are introduced in detail. Then, through case studies, the authors demonstrate how to apply these two methods for prediction in real-world problems. Finally, the article discusses the advantages and disadvantages of these two methods and how to choose the appropriate method in practical applications.

Keywords: Linear Regression, Gradient Descent, Norm, Mean Squared Error.

1 INTRODUCTION

The problem at hand is to analyze and predict the values of different suburbs or towns in the city of Boston using a given dataset. This dataset encompasses a variety of attributes, including urban population, city size, river water quality, species diversity, and urban consumption levels. The primary objective is to build models that can accurately predict the median value of a home while gaining insights into the relationship between various attributes and home values.

There are several important reasons why addressing this issue is crucial. Firstly, predicting home values can greatly benefit buyers, sellers, and investors in understanding the real estate market and making informed decisions regarding buying, selling, and investing. This allows them to make strategic decisions based on the projected performance of the real estate market.

Additionally, governments and urban planners can utilize these predictive models to develop better housing policies and promote sustainable urban development. By understanding the relationship between various attributes and home values, policymakers can make informed decisions to create more livable and affordable communities.

Furthermore, investors can leverage these predictive models to assess potential investment returns and risks in different regions. This enables them to make informed investment decisions based on the projected performance of the real estate market in specific areas.

To address this problem, we will first consider the methodology, which involves utilizing Simple Linear Regression and Multiple Linear Regression to analyze the dataset. Additionally,

PMBDA 2023, December 15-17, Nanjing, People's Republic of China

DOI 10.4108/eai.15-12-2023.2345298

other methods of linear regression will also be explored. We will then substitute and analyze the experimental dataset, while also setting up the experiment to ensure accurate and reliable results.

2 METHODOLOGY

2.1Linear Regression

Regression is a statistical technique used to describe the dependence of a response variable on a set of explanatory variables. There are two reasons for using regression model. One is to identify how changes in an explanatory variable affects the response. This can be useful for identifying patterns and trends in data that may not be immediately apparent. Another is for the purpose of prediction. When used for it, the model provides an estimate of the predicted values of the response as a function of the explanatory variables. Moreover, regression has many practical applications in fields such as economics, finance, and social science.[1]

Linear Regression is a popular approach in machine learning. It is useful to allow us to model the relationship between input variables and the output variable, which is a general way for data fitting.

Nowadays, there are many studies on optimization of Linear Regression. In this report, we apply the normal equation and gradient descent on Linear Regression for the prediction of Boston Housing Price, which aims to study the performance of these several models.[1]

2.1.1Simple Linear Regression

Simple Linear Regression has only one input variable. It is often used in various areas such as finance, mathematics, clinical medicine, and others.[2]

Suppose there are n observation values in a dataset, each with two features x and y. Our goal is to find a linear model that best fits these observation values.

Assume the linear regression function is: $\hat{y} = \beta_0 + \beta_1 x$, and we need to fit $\beta = [\beta_0, \beta_1]$ by minimizing the sum of squared errors (SSE, according to the ordinary least squares (OLS) method)

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 x_i)]^2 \quad (1)$$

We assume that the loss function $J(\beta_0, \beta_1)$ equals to Equation (1).

And we partially derive β_0 and β_1 on $J(\beta_0, \beta_1)$ and let the derivatives equal to 0, then we get:

$$\beta_0 = \bar{y} - \beta_1 \bar{x} \quad (2)$$

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad (3)$$

where β_1 is the slope of the line, β_0 is the intercept, n is the number of observations (x, y), x and y are the observation values. [3]

2.1.2Multiple Linear Regression

Multiple Linear Regression is defined as a model of the association between several inputs and an output, which is often used,[4] for example, in salary and many factors such as education level,

race, age, work experience, gender etc. Assume that there is a hyperplane $\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$, where $\hat{y}$ is the target variable, $[\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_p}]$ are the features, $[\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \ldots, \boldsymbol{\beta_p}]$ are the coefficients and β_0 is the parameter.

A more complex method of Ordinary Least Squares (OLS) than above is to find the optimal solution for the matrix $[\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \ldots, \boldsymbol{\beta_p}]$:

$$\mathbf{A} = (\mathbf{X'}^T \mathbf{X'})^{-1} X'^T \mathbf{Y} \tag{4}$$

where the size of matrix $\mathbf{X}$ is m rows by n columns, which means the number of observations is m, and the number of independent variables is n. $\mathbf{X'}^T$ means the transpose of matrix $\mathbf{X'}$, $(\mathbf{X'}^T \mathbf{X'})^{-1}$ denotes the inverse of the product of $\mathbf{X'}$ transpose and $\mathbf{X'}$. The result $\mathbf{A}$ represents the optimal solution, where $[\beta_1, \beta_2, ..., \beta_p]$ are the first n elements of the parameter matrix $\mathbf{A}$ and β_0 is the last element of matrix $\mathbf{A}$.[4]

Linear regression models can be solved using standard least squares methods, making them computationally efficient and easy to implement. But it is also sensitive to outliers, and outliers can have a large impact on the model's fit, leading to unreliable results.

2.2Linear Regression with Regularization

To avoid the over-fitting situation, a regularization term can be added to the Equation (1) to avoid the parameters of the regression to be extremely large. Ridge Regression and Lasso Regression is a variation of linear regression that aims to reduce the complexity of the model by adjusting the loss function. This adjustment involves introducing a penalty parameter that corresponds to the squared magnitude of the coefficients.

2.2.1Ridge Regression

Firstly, we define the loss function:

$$L = SSE + \lambda Reg = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda |\beta| \tag{5}$$

where λ is the regularization parameter or penalty parameter, $|\beta|$ is the norm of the vector $\beta = [\beta_0, \beta_1, \beta_2, \ldots, \beta_p]^T$.

In Ridge Regression, the L2 norm is applied on the Equation (5): $\left|\left|\beta\right|\right|_2 = \sqrt{\sum_{i=0}^{p} {\beta_i}^2}$.[5]

The regularization term in ridge regression penalizes each regression coefficient and drives them towards smaller values. By shrinking the size of the regression coefficients, the regularization term reduces the variance of parameter estimates and helps mitigate the impact of multicollinearity so that the smooth parameter estimates can be obtained. Its parameter estimates typically do not become exactly zero, making it relatively weak in terms of feature selection.

2.2.2Lasso Regression

The loss function is the same as the Equation (5). But what the difference in Lasso Regression is that the L1 norm is applied: $\left|\left|\beta\right|\right|_1 = \sum_{i=0}^{p} \beta_i$.

The L1 norm is the sum of the absolute values of regression coefficients. It has the effect of the sparsity to the regression coefficients. When L1 norm is used as a regularization term, LASSO regression can shrink some regression coefficients to zero, effectively achieving feature

selection by automatically identifying the most important features. It tends to produce sparse parameter estimates, meaning it sets some of the coefficients of unimportant features to zero. Therefore, LASSO regression is very useful for feature selection, as it can identify the most relevant features and eliminate redundant and noisy features.[6]

So compared to the general linear regression, the difference is the regularization term $\lambda|\beta|$. The parameter λ controls the trade-off between the data fitting term (SSE) and the regularization term. A larger λ increases the penalty on the model complexity, encouraging the model to select fewer features or simpler parameter estimates, thereby reducing the risk of over-fitting. A smaller value of λ leans towards better fitting the training data but may lead to over-fitting and high variance. Therefore, by adjusting the value of λ, one can balance the trade-off between fitting and regularization to achieve better model performance.[7]

To obtain the optimal vector β , we can also use the same formula in 2.1.2, which is the normal equation. It obtains the closed-form solution for the optimal parameters by solving the equation where the derivative of the loss function is zero. This method is suitable for small datasets but may not be computationally efficient for large datasets.

There is another method to find the optimal vector β: Gradient Descent. It iteratively updates the β vector to gradually minimize the loss function until it reaches the optimal solution. The details are as follows.

2.3Linear Regression with Gradient Descent

Gradient descent is an iterative optimization method used to locate the optimal value (maximum or minimum) of a specific objective function. This technique has gained widespread recognition in machine learning projects for refining model parameters and minimizing lost functions.

The main goal of gradient descent is to find the best parameters of the model that give it the highest accuracy on both the training and test datasets. In gradient descent, the gradient refers to a vector that represents the direction of the fastest growth of the objective function at a certain point. By moving in the opposite direction of the gradient, the algorithm gradually descends towards lower function values and eventually reaches the minimum of the function. Gradient descent is a fundamental technique used in various machine learning algorithms to optimize model performance.

During model training, the model calculates a lost function, such as root mean square error, by comparing the predicted value ($pred$) with the true value (y). The purpose of our model is to minimize this lost function.

To minimize the lost function, the model needs to determine the optimal values of θ_1 and θ_2. Initially, the model randomly chooses the values of θ_1 and θ_2, and then iteratively updates these values to minimize the lost function until the minimum value is reached. Once the model reaches the minimum lost function, it will have the best values for θ_1 and θ_2.

Using these updated values of θ_1 and θ_2 in the hypothesis equation of the linear regression model, our model can predict the output value y. The gradient descent algorithm for linear regression is:

$$\theta_j = \theta_j - \frac{\alpha}{m}\sum_{i=1}^{m}[(h_\theta(x_i) - y)x_i] \quad (6)$$

where θ_j is the weights of the hypothesis, $h_\theta(x_i)$ is the first i input the y value, i is the feature index $(0, 1, 2, \ldots\ldots, n)$ and α is the learning rate.[8]

Gradient Descent is applicable to different lost functions and has the capability to handle regression problems that involve non-linear relationships. And it can efficiently handle large by updating the parameters sequentially for each training example.

3 EXPERIMENTS

3.1 Dataset

The provided passage pertains to a database containing information about various suburbs and towns in the Boston area. This dataset was compiled from the Boston Standard Metropolitan Statistical Area (SMSA) in the year 1970. Each entry or record within this database serves as a description for a specific Boston suburb or town, encompassing a range of attributes that characterize these locales (Table 1).

TABLE 1. Attributes of the dataset

CRIM	This attribute denotes the per capita crime rate associated with each town.
ZN	This attribute represents the proportion of residential land zoned for large lots, typically measuring over 25,000 square feet, within the town or suburb.
INDUS	This attribute represents the proportion of non-retail business acres per town or suburb.
CHAS	This attribute represents a binary value that indicates whether a given town or suburb is bounded by the Charles River or not.
NOX	This attribute represents the concentration of nitric oxides (NOx) in parts per 10 million (ppm) in the air of a given town or suburb.
RM	This attribute represents the average number of rooms per dwelling in a given town or suburb.
AGE	This attribute represents the proportion of owner-occupied units built before 1940 in a given town or suburb. It provides information about the age distribution of the housing units within the area.
DIS	This attribute stands for "weighted distances to five Boston employment centers".
RAD	This attribute stands for "index of accessibility to radial highways". It represents a measure of how easily a given town or suburb's location can access radial highways, which are major roads leading outward from the city center in a hub-and-spoke pattern.
TAX	This attribute represents the full-value property tax rate for each town or suburb. It indicates the annual property tax amount levied on property owners based on the assessed value of their properties.
PTRATIO	This attribute represents the pupil-teacher ratio in schools within each town or suburb. It's a measure of the number of students in relation to the number of teachers.
B - 1000(Bk - 0.63)^2	The expression B−1000(Bk−0.63)^2 appears to be a component of a formula or equation, but without further context or information about the variables and their meanings, it's challenging to provide a specific detailed explanation.
LSTAT - %	lower status of the population
MEDV	This attribute represents the median value of owner-occupied homes in thousands of dollars for each town or suburb.

3.2Experimental Settings and Metrics

The environment of experiments is as follows:

The experimental equipment is a MacBook Pro (15-inch, 2016) with a 2.6 GHz quad-core Intel Core i7 processor and 16GB of 2133 MHz LPDDR3 memory.

The operation system is macOS Monterey 12.6.7 and the programming tool is PyCharm 2023.1.1 (Community Edition). Python version is Python 3.11.3, and the dependency library modules are matplotlib 3.7.2, pandas 2.0.3 and scikit-learn 1.3.0.

3.3Experimentation

There are four models to predict the Boston Housing Price: Linear Regression, Ridge Regression, Lasso Regression and Linear Regression with Gradient Descent. We figured all of them and compared their performance and scores (Table 2).

TABLE 2. Python packages

import matplotlib.pyplot, pandas, numpy
from sklearn.linear_model import LinearRegression, Ridge, Lasso, SGDRegressor
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.metrics import mean_squared_error
from sklearn.preprocessing import StandardScaler

We first define four functions to implement four models. After downloading the Boston housing price dataset into the project, we use the pandas in python to read the dataset. Then, we split it into features and target variables (x, y) and divide them into training and testing sets. The data is standardized using the StandardScaler package. The estimators used are LinearRegression, Ridge, Lasso, and SGDRegressor from the sklearn.linear_model package. Finally, we train the models on the training set, obtain the prediction values, compare them with the actual values, and calculate the mean squared error (MSE).

There is something creative of how to choose the regularization term α in Ridge Regression and Lasso Regression. We use GridSearchCV in python to try to select the optimal α.

In Ridge Regression, through experiments with specific values of 0.01, 0.1, 1.0, 2.0, 5.0, and 10.0, we found that the optimal regularization term is 2.0. So, we set the range of alpha from 0.1 to 3.1 with an interval of 0.1. The final best alpha value is determined to be 2.9000000000000004.

In Lasso Regression, using the same specific values for testing, we found that the optimal alpha is 0.01. We speculate that for Lasso Regression, the smaller the alpha, the better the regression effect. So, we kept trying until alpha reached 1^{-14}, but the program threw an error indicating overfitting. Therefore, we concluded that the best regularization term for Lasso Regression, under successful fitting, is 1^{-13}.

Based on the obtained optimal regularization terms, we calculated the mean squared error (MSE) for both models.[7]

3.4Experimental Result

The following four diagrams respectively represent Linear Regression, Ridge Regression, Lasso Regression, and Linear Regression with Gradient Descent, along with their corresponding Mean Squared Errors (MSE):

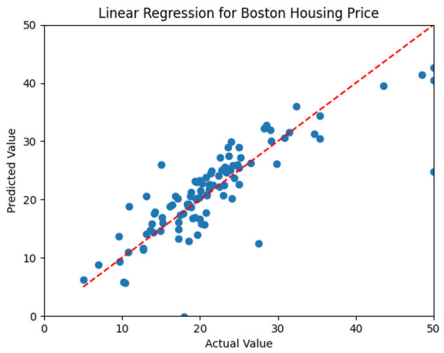

Figure 1 Linear Regression

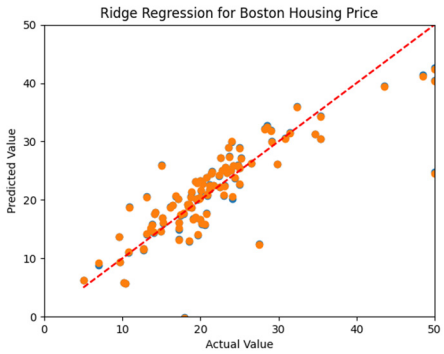

Figure 2 Ridge Regression

Figure 3 Lasso Regression

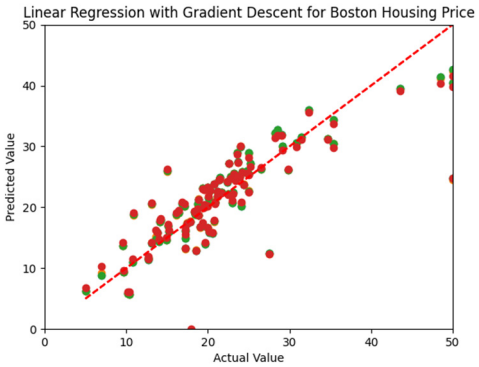

Figure 4 Linear Regression with Gradient Descent

Through these figures (figure 1-4), it can be observed that if we do not consider individual outliers, the predicted results of these models are quite close. For Ridge regression and Lasso regression, we adjusted the value of the alpha parameter in a timely manner. Without adjustment, the differences are more pronounced.

TABLE 3. The values of MSE

Models	MSE
Linear Regression	24.291119474973513
Ridge Regression	24.35406348400126
Lasso Regression	24.29111947497387
Gradient Descent	24.620096310125522

According to the experimental results, in the Boston housing price dataset, the MSE of Lasso Regression and Linear Regression is minimum, while that of Linear Regression with Gradient Descent is maximum. Therefore, on smaller datasets, the predictive performance of Linear Regression and LR with Regularization is better (Table 3).

4 CONCLUSION

Linear Regression is the fundamental linear regression algorithm used for predicting continuous target variables. Ridge Regression and Lasso Regression are improvements to Linear Regression, introducing regularization terms to reduce the risk of overfitting. Ridge Regression uses the regularization term with the L2 norm, which is suitable for situations where multiple correlations exist. Lasso Regression uses the regularization term with the L1 norm, which is suitable for scenarios involving feature selection and sparse modeling. Linear Regression with Gradient Descent solves the linear regression model parameters using gradient descent algorithms, which is suitable for large-scale datasets and high-dimensional feature scenarios.[9]

Choosing an appropriate regression algorithm depends on the characteristics of the dataset and the requirements of the problem. If there is a high degree of correlation between features, consider using Ridge Regression; if feature selection or sparse modeling is required, consider

using Lasso Regression; if the dataset is large or the feature dimension is high, consider using Linear Regression with Gradient Descent.[10]

ACKNOWLEDGEMENT. All the authors contributed equally to this work and should be considered as co-first author.

REFERENCE

[1] Fitzmaurice. (2016). Regression. Diagnostic Histopathology (Oxford, England: 2008), 22(7), 271–278. https://doi.org/10.1016/j.mpdhp.2016.06.004

[2] Marill. (2004). Advanced Statistics: Linear Regression, Part I: Simple Linear Regression. Academic Emergency Medicine, 11(1), 87–93. https://doi.org/10.1197/j.aem.2003.09.005

[3] Altman, & Krzywinski, M. (2015). Points of Significance: Simple linear regression. Nature Methods, 12(11), 999–1000. https://doi.org/10.1038/nmeth.3627

[4] Nimon, & Oswald, F. L. (2013). Understanding the Results of Multiple Linear Regression: Beyond Standardized Regression Coefficients. Organizational Research Methods, 16(4), 650–674. https://doi.org/10.1177/1094428113493929

[5] Saleh, Arashi, M., & Kibria, B. M. G. (2019). Introduction to Ridge Regression. In Theory of Ridge Regression Estimation with Applications (Vol. 285, pp. 1–13). John Wiley & Sons, Incorporated. https://doi.org/10.1002/9781118644478.ch1

[6] Komatsu, Yamashita, Y., & Ninomiya, Y. (2019). AIC for the group Lasso in generalized linear models. Japanese Journal of Statistics and Data Science, 2(2), 545–558. https://doi.org/10.1007/s42081-019-00052-0

[7] Doebler, Doebler, A., Buczak, P., & Groll, A. (2023). Interactions of Scores Derived From Two Groups of Variables: Alternating Lasso Regularization Avoids Overfitting and Finds Interpretable Scores. Psychological Methods, 28(2), 422–437. https://doi.org/10.1037/met0000461

[8] Ighalo, Adeniyi, A. G., & Marques, G. (2020). Application of linear regression algorithm and stochastic gradient descent in a machine-learning environment for predicting biomass higher heating value. Biofuels, Bioproducts and Biorefining, 14(6), 1286–1295. https://doi.org/10.1002/bbb.2140

[9] Massaron, & Boschetti, A. (2016). Regression analysis with Python: learn the art of regression analysis with Python / Luca Massaron, Alberto Boschetti. (1st edition). Packt Publishing.

[10] Chen, & Messner, M. C. (2023). Training material models using gradient descent algorithms. International Journal of Plasticity, 165, 103605–. https://doi.org/10.1016/j.ijplas.2023.103605