Research on linear regression algorithm

Kecheng Qu [*]

Shandong Xiehe University, 250100, JiNan, Shandong, China

Abstract. Linear regression is one of the most widely used predictive models in statistics and machine learning. This paper aims to comprehensively discuss the theoretical basis, mathematical principle and application of linear regression algorithm in various fields. Firstly, this paper introduces the research background and significance of linear regression, and summarizes its important role in modern data analysis. Then, the paper elaborates the basic theory of linear regression, including its definition, assumptions, parameter estimation methods and model diagnosis and selection. In addition, different types of linear regression are classified and discussed, such as simple linear regression, multiple linear regression and logistic regression, and the specific application scenarios of each type are analyzed.

Keywords: Linear regression, Machine learning, Model optimization.

1 Introduction

1.1 Research background

Linear regression[1], as a classical algorithm in statistics, is widely used in various data analysis tasks because of its simplicity and explanatory power. Since Gauss and least squares were proposed in the 19th century, linear regression has been an important tool used by researchers and practitioners to understand the relationships between variables. With the improvement of computing power and the advent of the era of big data, linear regression presents a unique advantage when dealing with complex data.

1.2 Research significance

Although linear regression is a time-honored algorithm, it is still central to the field of data science today. Exploring the theory and practice of linear regression can help deepen our understanding of data modeling and facilitate the development of more efficient algorithms. In addition, as new questions and new data emerge, traditional linear regression models need to be continuously improved and optimized to adapt to changing needs.

[*] Corresponding author: 287486820@qq.com

1.3 Research status at home and abroad

The theoretical research of linear regression began in the 19th century, but in recent decades, with the development of computer technology, the research focus in this field has gradually shifted to the computational efficiency, scalability and application of algorithms on high-dimensional data. Many international scholars have proposed a variety of improved linear Regression algorithms, such as Ridge Regression[2] and Lasso Regression, which perform well in dealing with multicollinearity and overfitting problems. Domestic researchers have also proposed a variety of innovative linear regression models for specific problems in the field, and have achieved remarkable results in practical applications.

1.4 Paper structure arrangement

This thesis is divided into six chapters. Following the introduction of this chapter, the second chapter will introduce the basic theory of linear regression, including definitions, mathematical principles, types of models and their application scenarios. The third chapter focuses on the implementation of linear regression algorithm. Chapter 4 discusses the advanced techniques of model optimization. Chapter five summarizes the full text and puts forward the future research direction.

2 The basic theory of linear regression

2.1 Definition of linear regression

Linear regression is a statistical method used to establish a linear relationship between the independent variable X and the dependent variable Y. The goal is to find an optimal linear function, that is, to determine a set of coefficients (weights) so that the function can predict the value of the dependent variable as accurately as possible. Formally, the linear regression model can be expressed as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon \tag{1}$$

where β0 is the intercept, β1 to βn is the regression coefficient, X1 to Xn is the independent variable, and ε is the error term.

2.2 Assumptions of linear regression

In order to ensure the validity of the linear regression model and the properties of the best linear unbiased estimator (BLUE), the following assumptions need to be met:

- Linear relationship: There is a linear relationship between independent variables and dependent variables.
- Independence: The sample observations are independent of each other.
- Homoscedasticity: The error term has the same variance (σ²) for all independent variable levels.
- Normality: The error terms are normally distributed.
- No multicollinearity: there is no complete or highly linear relationship between the independent variables.

2.3 Parameter estimation

In linear regression analysis, parameter estimation refers to the process of determining the regression coefficient in the model. The most commonly used method is least squares (OLS), whose goal is to minimize the sum of squares of residuals, that is, to find a set of coefficients that minimizes the difference between the predicted value and the actual value. An estimate of the regression coefficient can be obtained by means of analytical or numerical solutions.

2.4 Model diagnosis and selection

Model diagnosis is the process of evaluating whether the linear regression model is appropriate and whether the assumptions are satisfied. Commonly used diagnostic tools include residual graph, Q-Q graph and influence analysis. Model selection involves the process of selecting an optimal model from several candidate models. Methods such as Akaike Information criteria (AIC), Bayesian information criteria (BIC), or cross-validation (CV) can be used for model selection. In addition, according to the characteristics of specific problems, we can also consider using stepwise regression, principal component regression and other methods to select important predictors.

3 The implementation of linear regression algorithm

The core goal of linear regression algorithms is to find the best parameter estimation, such that the difference between the predicted value of the model and the actual data is minimal. The algorithmic flow typically consists of the following steps:

Data preprocessing: includes data cleaning, missing value processing, outlier detection and processing, and data standardization or normalization.

Feature selection: Determine which independent variables will be included in the model. This can be done through correlation analysis, domain knowledge, or feature selection algorithms.

Model fitting: Use least squares or other optimization techniques to estimate the parameters of a linear model.

Model evaluation: Use test sets or cross-validation to evaluate the performance of the model.

- Model diagnosis: Check whether the model conforms to the basic assumptions of linear regression and adjust for possible model incompatibilities.

Prediction and application: Use well-fitted models to make predictions and apply them to real problems.

4 The application of linear regression model

4.1 Simple linear regression

Simple linear regression is the most basic regression model, which involves only one independent variable and one dependent variable. This model is suitable for exploring the direct relationship between two variables. For example, in economics, simple linear regression can be used to analyze the relationship between consumer spending and personal income. In such applications, the model is often able to reveal how one variable changes as another variable changes, thus providing a basis for decision making.

4.2 Multiple linear regression

When there are multiple factors affecting the dependent variable, a multiple linear regression model is needed [3]. This model can simultaneously consider the influence of multiple independent variables on dependent variables. In medical research, multiple linear regression is often used to analyze how lifestyle, genetic factors, and other environmental factors combine to influence certain health indicators, such as blood pressure or cholesterol levels. Using multiple linear regression, the researchers were able to identify the key factors that have the greatest impact on health.

4.3 Logistic regression

Although not strictly linear regression, logistic regression uses a similar model framework to solve classification problems. Different from linear regression, the dependent variable of logistic regression is categorical variable rather than continuous variable. It is widely used in financial risk assessment, disease diagnosis and other fields. For example, a bank may use a logistic regression model to predict the probability that a loan applicant will default and decide whether to approve a loan application.

4.4 Other Application Cases

In addition to the several common applications mentioned above, linear regression models are used in a variety of other fields. In environmental science, linear regression can help scientists understand the relationship between climate change and carbon dioxide emissions. In market research, it can be used to predict the relationship between product sales and advertising spend. In addition, linear regression is also used in social science research, such as analyzing the relationship between education level and unemployment rate. These cases show that linear regression is a versatile and powerful analytical tool applicable to a wide range of practical problems.

5. Linear regression model optimization

5.1 Feature Selection

Feature selection is one of the key steps in the process of model optimization. Its purpose is to select the most explanatory subset of response variables from all possible predictors. Effective feature selection can improve the interpretability of the model, reduce the risk of overfitting and reduce the computational cost. The methods of feature selection can be divided into three categories: filtering method, wrapping method and embedding method. The filtering method selects characteristics according to statistical tests. The wrapping method evaluates the importance of features by constructing multiple models. The embedding method integrates the feature selection process into the model training process. In practical applications, feature selection usually requires a combination of domain knowledge and data characteristics to determine the best strategy.

5.2 Regularization method

Regularization method is an effective way to avoid overfitting. It penalizes the model complexity by adding a regularization term to the loss function. Common regularization methods include ridge regression (L2 regularization) and lasso regression (L1 regularization).

Ridge regression inhibits the size of the coefficient by introducing a penalty term proportional to the size of the coefficient, while lasso regression enables feature selection while reducing the coefficient, as it reduces some of the coefficients to zero. The choice of regularization parameters is usually determined by cross-validation.

5.3 Model integration

Model integration is the technique of combining multiple different models to improve overall predictive performance. Integrated methods such as Bagging and Boosting have been shown to significantly improve the accuracy and robustness of models in many cases. Bagging reduces variance by building multiple independent models and averaging them; Boosting, on the other hand, builds models iteratively, with each subsequent model focusing on samples misclassified by the previous model. Random forest and gradient lift tree are two widely used integrated models. In the context of linear regression, the ensemble method can be applied by averaging the predictions of multiple linear models.

6. Conclusion and prospect

6.1 Research Summary

This research comprehensively reviews the theoretical basis, implementation and application of linear regression algorithm in various fields. Starting from the definition of linear regression, we discuss its mathematical principles and assumptions, and introduce the process of parameter estimation and model diagnosis in detail. Through the description of algorithm flow and the writing of pseudo-code, this paper shows how to implement linear regression model in Python environment. In addition, we also analyze the application cases of different types of linear models such as simple linear regression, multiple linear regression and logistic regression in practical problems. In terms of model optimization, advanced techniques such as feature selection, regularization methods and model integration are discussed to improve the performance and generalization ability of the model.

6.2 Research Limitations

Although this study covers many aspects of linear regression algorithms, there are still some limitations. First of all, this study mainly focuses on theory and basic applications, and some emerging scenarios of high-dimensional data processing and big data analysis have not been deeply explored. Second, the data set used in the experimental part was limited and did not cover all possible data types and complexities. Finally, strategies for model optimization have been mentioned, but not all possible optimization techniques have been comprehensively compared and evaluated.

6.3 Future research direction

Given the limitations of the current research, future work can be carried out in the following directions: First, the application of linear regression in big data analysis can be explored, especially how to efficiently process high-dimensional data and real-time data streams. Secondly, it is necessary to further study the regression methods under nonlinear and complex data structures. In addition, developing new algorithms and techniques to improve the applicability and accuracy of models in specific fields is also an important research direction in the future. Finally, strengthening interdisciplinary collaborations that combine linear regression with recent research in other fields may yield new insights and approaches. Through these efforts, we can expect to further enhance the theoretical depth and practical application value of linear regression algorithm in the future.

Reference

1. Hall P , Horowitz J L .Methodology and convergence rates for functional linear regression[J].The Annals of Statistics, 2007, 35(1):70-91.DOI:10.1214/009053606000000957.
2. Hoerl A E , Kennard R W .Ridge regression: biased estimation for nonorthogonal problems[J].Technometrics A Journal of Stats for the Physical Chemical & Engineering ences, 2000, 42.DOI:10.2307/1271436.
3. Quiming N S , Denola N L , Saito Y ,et al.Multiple linear regression and artificial neural network retention prediction models for ginsenosides on a polyamine-bonded stationary phase in hydrophilic interaction chromatography[J].Journal of Separation Science, 2015, 31(9):1550-1563.DOI:10.1002/jssc.200800077.