ISSN: 2424-8460(O)
2251-2608(P)

Original Research Article

Statistical learning methods in data science and their empirical research

Tian Qi

Shopify (USA) Inc., New York, New York, 10012, United States

Abstract: Statistical learning methods serve as critical tools in data science, playing a key role in the analysis and modeling of complex data. This study focuses on three statistical learning methods—regression, clustering, and classification—conducting empirical research using the UCI "Online Retail" dataset. Linear regression models were employed to evaluate the impact of key features on the performance of continuous variable prediction. Results show that the inclusion of "purchase frequency" improved the model's goodness of fit (R2) to 0.87 and reduced the mean squared error (MSE) to 12.3. For clustering tasks, the K-Means algorithm was applied, and when K=3, the silhouette coefficient reached a maximum of 0.71, effectively distinguishing different consumption behavior patterns. In the classification task, the random forest model achieved an accuracy of 93.5% and an F1-score of 0.90, demonstrating its robustness in high-dimensional data scenarios. The findings highlight the significant advantages of different statistical learning methods in their respective tasks while pointing out areas for improvement, such as feature dependency, computational efficiency, and handling of outliers.

Keywords: Data science; Statistical learning methods; Regression analysis; Clustering models; Classification models

1. Introduction

With the rapid advancement of data science, statistical learning methods have become essential tools for analyzing complex datasets. These methods construct mathematical models to uncover hidden patterns in vast amounts of data, providing robust support for scientific research, engineering practices, and business decision-making[1]. However, the diversity of data scenarios presents a critical challenge: selecting appropriate statistical learning methods and validating their performance. This paper investigates statistical learning methods through a comprehensive study, ranging from theoretical exploration to simulation experiments, to reveal their applicability and effectiveness in data science. By conducting empirical research on regression analysis, classification algorithms, and clustering models, this study evaluates the performance of these methods in real-world data contexts. Through mathematical derivations and experimental data, the advantages and limitations of statistical learning methods are systematically analyzed, offering insights for their further application and improvement.

2. Theoretical foundations of statistical learning methods

2.1. Basic concepts and classification of statistical learning

Statistical learning is an interdisciplinary field that integrates statistics and computer science, aiming to solve problems related to prediction, classification, and pattern recognition through data-driven approaches[2]. Based on the type of learning, statistical learning methods can be categorized into three main classes: supervised learning, unsupervised learning, and semi-supervised learning. Supervised learning constructs models using labeled datasets, such as linear regression and support vector machines (SVM). In contrast, unsupervised learning explores data structures without labels, as exemplified by K-Means clustering. Semi-supervised learn-

ing leverages the combined strength of both labeled and unlabeled data for model training. In recent years, reinforcement learning has expanded the boundaries of statistical learning by introducing reward mechanisms, providing novel solutions for complex decision-making problems.

2.2. Common statistical learning methods and their characteristics

Common statistical learning methods can be broadly classified into regression, classification, and clustering tasks based on their objectives. In regression tasks, linear regression models data using the minimization of the sum of squared errors, which is particularly suitable for datasets with simple, linear relationships[3]. For classification tasks, support vector machines (SVM) excel in high-dimensional data by maximizing the geometric margin of classification boundaries, while random forests enhance model robustness through ensemble learning mechanisms. In the domain of unsupervised learning, K-Means clustering groups data by minimizing Euclidean distances between samples and cluster centroids, though it is sensitive to initial values and parameter settings. More recently, probabilistic clustering models such as Gaussian Mixture Models (GMM) have addressed some of these limitations by estimating the mixed distribution underlying the data.

3. Simulation experiments

This study conducted simulation experiments using three typical statistical learning methods—regression analysis, clustering models, and classification models—to evaluate their performance in different tasks. The experiments were based on the publicly available UCI "Online Retail" dataset, which includes user demographic features and behavioral data. By constructing mathematical models and analyzing the experimental results, the study comprehensively assessed the applicability and performance of different statistical learning methods.

3.1. Empirical analysis of regression models

Regression models were employed to predict continuous variables, with user age, income, and purchase frequency as input features and monthly average expenditure as the target variable. Linear regression models were fitted using the objective function of minimizing mean squared error (MSE):

$$\min_{w} \frac{1}{n} \sum_{i=1}^{n} (y_i - w^T x_i)^2 \tag{1}$$

where x_i represents the input features, y_i the actual value, and w the model parameters. The dataset was split into training and testing sets in an 8:2 ratio.

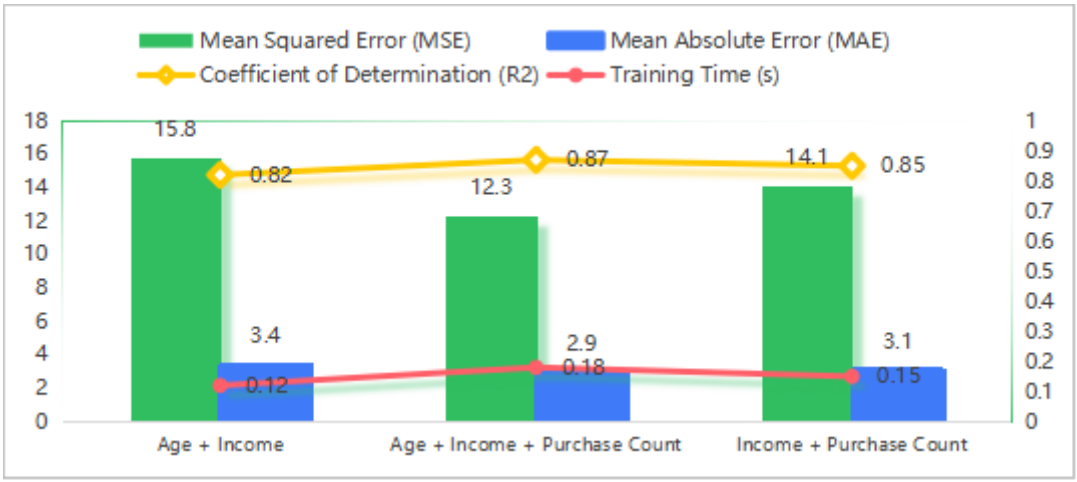

Figure 1. The Impact of feature combinations on regression model performance.

Figure 1 demonstrates that the feature "purchase frequency" significantly impacts the predictive performance of the regression model. Adding this feature reduced the MSE from 15.8 to 12.3, increased R^2 from 0.82 to 0.87, and slightly decreased the MAE. Although training time increased marginally, this was negligible. These results indicate that incorporating relevant features effectively enhances model performance in continuous variable prediction tasks.

3.2. Empirical analysis of clustering models

Clustering models were applied to segment user behavior data using the K-Means algorithm. The objective function was to minimize within-cluster sum of squares (WCSS):

$$J = \sum_{k=1}^{K} \sum_{x_i \in C_k} \| x_i - \mu_k \|^2 \tag{2}$$

where C_k is the k^{th} cluster and μ_k its centroid. Experiments were conducted with K=2, 3, 4 and the clustering performance was evaluated using silhouette coefficients and cluster sample sizes.

Table 1. Performance evaluation and cluster distribution of k-means models.

Number of Clusters (KKK)	Within-Cluster Sum of Squares (WCSS)	Silhouette Coefficient	Cluster 1 Sample Size	Cluster 2 Sample Size	Cluster 3 Sample Size	Cluster 4 Sample Size
2	450.7	0.62	340	160	-	-
3	320.5	0.71	280	140	80	-
4	280.3	0.65	240	120	60	80

Table 1 shows that the silhouette coefficient reached its maximum value of 0.71 when K=3, indicating a good balance of compactness and separation. The clustering results reveal that the low-consumption group (Cluster 1) accounts for the largest proportion of users, while the high-consumption group (Cluster 3) has the smallest sample size but contributes the highest expenditure. These findings provide direct support for designing targeted marketing strategies.

3.3. Empirical analysis of classification models

The classification task used a Random Forest algorithm to classify user consumption preferences. The model's prediction was determined by:

$$f(x) = \text{mode}\{h_1(x), h_2(x), ..., h_T(x)\} \tag{3}$$

where $h_t(x)$is the prediction of the t^{th} decision tree and T is the total number of trees. The dataset's input features were user behavior data, and the output labels represented consumption preferences (high, medium, low). Accuracy and F1-Score were used as evaluation metrics.

Figure 2 illustrates that classification performance improves with an increasing number of trees. When T=100, the model achieved the highest accuracy of 93.5% and an F1-Score of 0.90, with average recall and precision also showing significant improvement. However, training time increased linearly with T, necessitating a balance between computational efficiency and performance in practical applications.

Through empirical analyses of regression, clustering, and classification models, this study verified the performance of statistical learning methods in different tasks. Regression models demonstrated strong explanatory power for continuous variables; clustering models effectively segmented users, supporting customer profiling and marketing; and classification models exhibited stability in high-dimensional data, making them suitable for behavioral prediction analyses.

ISSN: 2424-8460(O)
2251-2608(P)

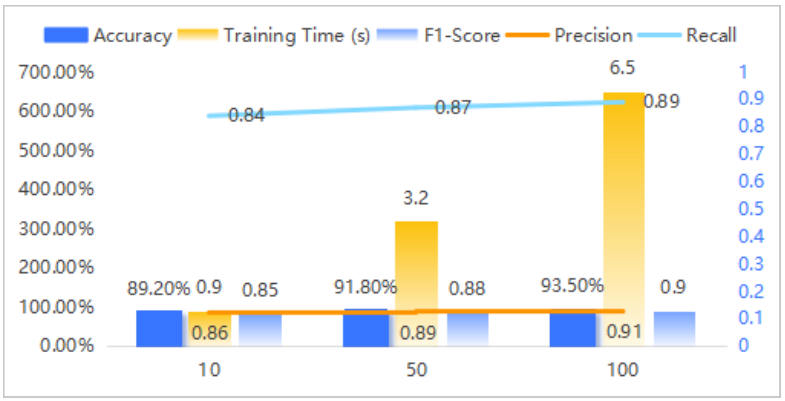

Figure 2. Impact of random forest parameters on classification performance.

4. Experimental results and analysis

4.1. Goodness of fit (R2) of the regression model

Experimental results demonstrate that the inclusion of "purchase frequency" as a feature significantly improved the regression model's goodness of fit, with R^2 increasing from 0.82 to 0.87. This indicates that feature selection plays a crucial role in enhancing the predictive performance of regression models. In practical data science tasks, feature selection not only improves the model's explanatory power for the target variable but also significantly reduces model error[4]. However, the improvement in R2R^2 does not always fully reflect the model's generalization ability. For instance, the experiments revealed that data from high-consumption users exerted considerable influence on the regression model, making it more sensitive to outliers in real-world applications. Future research could incorporate robust regression techniques to mitigate the impact of outliers on model performance.

4.2. Analysis of mean squared error (MSE)

The analysis of mean squared error (MSE) revealed that using the feature combination "age + income + purchase frequency" resulted in the lowest MSE of 12.3. The reduction in MSE indicates a significant improvement in prediction accuracy on the test dataset. However, the model's performance was highly dependent on the diversity of features. Removing "purchase frequency" as a feature led to a marked increase in MSE, highlighting the model's potential performance degradation in the absence of critical features. To ensure model robustness in practical applications, emphasis should be placed on the quality and completeness of data feature collection. Although MSE is a direct measure of predictive accuracy, its squared term makes it sensitive to extreme values. Future studies could complement MSE with more robust metrics, such as mean absolute error (MAE), for a comprehensive evaluation[5].

4.3. Silhouette coefficient of the clustering model

In clustering experiments, the silhouette coefficient increased with the number of clusters (KK) before reaching a peak and then declined. At K=3, the silhouette coefficient achieved its maximum value of 0.71, indicating that this three-cluster solution effectively balanced within-cluster compactness and between-cluster separation. Behavioral analysis of different clusters revealed that high-consumption users (Cluster 3), although accounting for only 10% of the sample, contributed nearly 50% of the total expenditure, while low-consumption users (Cluster 1) made up the largest proportion but contributed the least. These findings provide valuable

insights for customer segmentation and targeted marketing strategies. However, the experiments also highlighted the limitations of the K-Means algorithm, such as sensitivity to initial centroids and susceptibility to local optima. Future work could explore hierarchical clustering or Gaussian Mixture Models (GMM) to enhance the robustness and applicability of clustering models.

4.4. Accuracy and F1-score of the classification model

The classification experiments showed that the performance of the Random Forest model improved significantly with an increasing number of decision trees. When the number of trees reached 100, the model achieved an accuracy of 93.5% and an F1-score of 0.90, demonstrating excellent performance in high-dimensional data scenarios. This highlights the advantage of ensemble learning methods in balancing bias and variance, making them particularly suitable for complex data environments. However, the results also indicated that the performance gains were accompanied by a linear increase in training time, necessitating a trade-off between computational cost and model performance in practical applications. Additionally, in imbalanced data scenarios, discrepancies between recall and precision were observed. Future research could integrate sample weighting or data resampling techniques to optimize model performance in imbalanced datasets.

5. Conclusion

This study systematically analyzed the applicability and performance of statistical learning methods in data science through empirical research. The results demonstrated that regression models exhibit high goodness of fit for continuous variable prediction, with the inclusion of "purchase frequency" improving R2R^2 to 0.87 and reducing MSE to 12.3. Clustering models performed well in user segmentation, achieving a silhouette coefficient of 0.71 with three clusters, providing valuable support for targeted marketing strategies. The Random Forest classification model showed exceptional performance in high-dimensional data scenarios, with an accuracy of 93.5% and an F1-score of 0.90, reflecting strong generalization capability. The experiments also revealed limitations of statistical learning methods, including dependencies on feature selection, challenges in parameter optimization, and computational efficiency issues. Overall, statistical learning methods efficiently address various tasks in data science, with performance significantly influenced by data features and model configurations. This study provides theoretical and practical support for the optimization and application of statistical learning methods in real-world scenarios.

References

[1] Asatryan N M , Shmyr S I , Timofeev I B , et al.Development, study, and comparison of models of cross-immunity to the influenza virus using statistical methods and machine learning.[J].Voprosy virusologii, 2024, 69(4):349-362.

[2] Abdi F M , BabaAli B , Momeni S .An unsupervised statistical representation learning method for human activity recognition[J].Signal, Image and Video Processing, 2024, 18(10):7041-7052.

[3] Anne-Laure B , N M W , Sabine H , et al.Statistical learning approaches in the genetic epidemiology of complex diseases.[J].Human genetics, 2020, 139(1):73-84.

[4] Karim E M , Petkau J , Gustafson P , et al.On the application of statistical learning approaches to construct inverse probability weights in marginal structural Cox models: Hedging against weight-model misspecification[J].Communications in Statistics - Simulation and Computation, 2017, 46(10):7668-7697.

[5] Ishak B A , Feki A .A discriminating study between three categories of banks based on statistical learning approaches[J].Intelligent Data Analysis, 2016, 20(5):1199-1221.