Study of Supervised Logistic Regression Algorithm

Prof.Dr.Neelam A.Kumar[1], Laxmi Jangale[2],Vaishnavi Sathe[3],Aishwarya Shelke[4],Tanvi Redij[5]
[1]*(Professor, SRCOE, Department of Computer Engineering Pune)*
[2],[3,4,5]*(Student, SRCOE, Department of Computer Engineering Pune)*

Abstract: *Logistic regression is a widely used supervised learning algorithm, primarily applied to binary classification problems. Its simplicity, interpretability, and solid mathematical foundation have made it a popular choice in various domains such as medical diagnosis, credit scoring, and marketing. This review paper provides a comprehensive overview of logistic regression, covering its mathematical formulation, optimization techniques, and evaluation metrics. The paper discusses how the model is trained using maximum likelihood estimation and optimized through gradient descent. Furthermore, it explores regularization methods like L1 and L2 to prevent overfitting.Despite its limitations, such as its assumption of linearity in the log-odds and sensitivity to multicollinearity and imbalanced data, logistic regression remains a powerful and interpretable model for classification tasks. Recent advancements have focused on addressing these challenges by enhancing scalability, handling imbalanced datasets, and exploring non-linear extensions. The review highlights the algorithm's robustness and future potential in large-scale and complex data environments.*

Key Word*: Logistic regression; Supervised learning; Binary classification; Feature engineering*

I. Introduction

Logistic regression is a widely used supervised learning algorithm for binary classification tasks, where the goal is to predict the probability of a data instance belonging to one of two classes. The algorithm models the relationship between the input features and a binary target variable using the logistic (sigmoid) function, which maps predicted values to probabilities between 0 and 1. A threshold, often set at 0.5, is then used to classify observations into one of the two classes. Logistic regression's popularity stems from its simplicity, interpretability, and effectiveness across various applications, including medical diagnosis, risk assessment, and customer behavior analysis. The model coefficients offer insights into the impact of each feature on the likelihood of a given outcome, making logistic regression particularly valuable in fields where understanding feature importance is crucial. Despite its limitations, especially in modeling complex, non-linear relationships, logistic regression remains a cornerstone in machine learning and statistical analysis.[5]

II. Literature Review

Agarwal A. et al. in "Communication-Efficient Secure Logistic Regression".(2024)proposes a novel approach for securely training logistic regression models on secret-shared data. The approach presents several key techniques, such as secure protocols for evaluating the sigmoid function and improving communication in comparison protocols. The system achieves a significant reduction in communication and processing time compared to traditional protocols, showcasing its efficiency for secure machine learning.[6]

Chen L. and Zhang H. et al. in "Regularized Logistic Regression for Feature Selection in High-Dimensional Biomedical Data"(2021,Journal of Biomedical Informatics) explore regularization techniques in logistic regression to handle high-dimensional biomedical datasets, often characterized by thousands of genetic or clinical variables but relatively few samples. Their research focuses on L1 regularization (Lasso) and the elastic net to enhance feature selection and interpretability. Through simulations and real-world testing on genomic datasets, the authors demonstrate that regularization helps identify the most influential genes linked to specific diseases, reducing model complexity and improving generalization. This approach holds significant potential for early diagnosis and personalized medicine.[2]

Mandal D. and Ray S. et al. in "Ensemble Methods with Logistic Regression for Improved Classification in Credit Scoring"(2022,Journal of Financial Economics and Machine Learning) proposes an ensemble learning method combining logistic regression with boosting techniques for credit scoring. Recognizing logistic regression's limitations with complex, non-linear relationships, the authors integrate logistic regression with gradient boosting to create a hybrid model that captures more intricate patterns in financial data. Experimental results on multiple credit scoring datasets reveal that the hybrid model improves both accuracy and robustness compared to traditional logistic regression, providing a more reliable tool for credit risk assessment.[3]

Patel A. et al. in "Kernel-Based Logistic Regression for Non-Linear Decision Boundaries in Medical Image Analysis"(2023) tackle the limitation of logistic regression in modeling non-linear decision boundaries, particularly in medical image analysis, where data often contains non-linear structures. By applying kernel functions to logistic regression, they extend the model's capability to detect subtle patterns in image-based data, such as MRI scans. The study's results show that kernel-based logistic regression performs comparably to more complex models like support vector machines, but with the added benefit of probabilistic outputs, useful in medical decision-making. This advancement suggests a promising alternative for scenarios where interpretability and probabilistic confidence are critical. [9]

Johnson T. et al. in "A Comprehensive Review on the Application of Logistic Regression in Social Media Sentiment Analysis"(2024, Journal of Data Science Applications) review logistic regression's application in sentiment analysis, particularly within social media data. They focus on logistic regression's strengths and limitations when applied to text classification, including handling imbalanced sentiment classes and the need for dimensionality reduction. The authors highlight recent advances in combining logistic regression with natural language processing techniques, such as word embeddings, to enhance feature representation. Their findings indicate that logistic regression, when combined with text preprocessing methods, can achieve competitive results, making it a practical choice for large-scale sentiment analysis in social media monitoring and customer feedback.[4]

III. Logistic Regression Archietecture

Logistic Regression :

Logistic regression is a supervised learning algorithm. In supervised learning, the model is trained on a labeled dataset, meaning that each training example includes both the input data (features) and the corresponding output label (target class). Logistic regression is a statistical method used for binary classification problems, where the goal is to predict the probability of a binary outcome (i.e., two possible classes, such as "yes" or "no", "success" or "failure"). Logistic regression is a classification algorithm, not a regression technique. Logistic regression models the relationship between a dependent variable (which is categorical, typically binary) and one or more independent variables (which can be continuous or categorical) using a logistic function. The logistic function, also known as the sigmoid function, maps the output to a probability between 0 and 1.

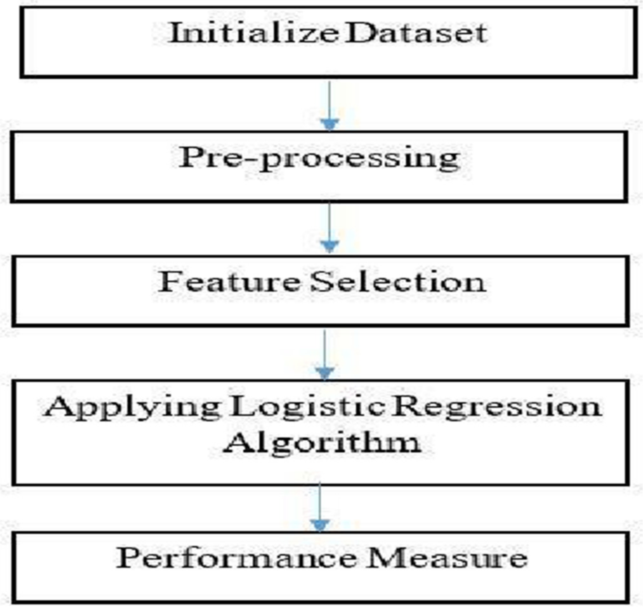

fig.1.Logistic Regression Architecture

Steps of Logistic Regression:

1. **Data Collection and Preparation:**
   Collect relevant data (e.g., age, cholesterol, etc.) for heart disease prediction. Preprocess the data (handle missing values, scale features, encode categorical variables, and split into training and testing sets).
2. **Initialize Logistic Regression Model:**
   Set up the logistic regression model to predict the probability of the outcome (e.g., heart disease or not).
3. **Train the Model:**
   Train the model using Maximum Likelihood Estimation (MLE) to find the best coefficients by minimizing the log-loss function.

Use optimization algorithms like gradient descent to adjust the coefficients.

4. **Make Predictions:**
   The model predicts probabilities for the test data.
   Apply a decision threshold (usually 0.5) to classify the outcome (1 for heart disease, 0 for no heart disease).
5. **Model Evaluation:**
   Use metrics like accuracy, precision, recall, F1-score, and ROC-AUC to evaluate model performance on the test data.

IV. Advantages

1. **Simple and Interpretable:** Logistic Regression is easy to understand and interpret, making it ideal for medical applications where explainability is important.
2. **Efficient:** It is computationally efficient, making it suitable for large datasets with many features.
3. **Probabilistic Output:** The algorithm provides probabilities, which can be useful in medical decision-making when assessing risk levels**.**
4. It makes no assumptions about distributions of classes in feature space.
5. Good accuracy for many simple data sets and it performs well when the dataset is linearly separable.

V. Disadvantages

1. **Limited to binary outcomes:** Logistic regression is only designed to model binary outcomes, so it may not be suitable for non-binary outcomes withoutmodifications.
2. **Assumes linearity:** Logistic regression assumes a linear relationship between the dependent variable and the independent variables.
3. **Sensitive to outliers:** Logistic regression can be particularly sensitive to outliers, which can skew the coefficients and influence predictions.
4. **Overfitting:** Logistic regression is vulnerable to overfitting, especially if thenumber of observations is fewer than the number of features.
5. **Can't handle large numbers of categorical features:** Logistic regression is not ableto handle a large number of categorical features or variables.

VI. Application

1. **Healthcare**
   Disease Prediction: Logistic regression can predict the likelihood of diseases like diabetes, cancer, or heart disease based on risk factors (e.g., age, gender, family history, lifestyle).
   Medical Diagnosis: Used to classify patients as likely or unlikely to develop a certain condition based on symptoms, test results, and medical history.
2. **Finance**
   Credit Scoring: Helps predict whether a borrower is likely to default on a loan. This prediction is based on income, employment history, and credit history.
   Fraud Detection: Identifies the likelihood of fraudulent transactions by analyzing patterns and irregularities in transaction data.
3. **E-commerce**
   Purchase Prediction: Logistic regression helps predict whether a visitor will complete a purchase based on their browsing behavior and demographic data.
   Recommendation Systems: Used to model and predict whether a user will click on or purchase recommended items.
4. **Education**
   Student Performance Prediction: Predicts the probability of students passing or failing a course based on various factors like attendance, grades, and engagement levels.
   Admissions Decisions: Helps predict the likelihood of applicants being successful based on academic records, extracurricular activities, and other criteria.

VII.Conclusion

Logistic regression survival a powerful and interpretable algorithm for binary classification problems, particularly in supervised learning. While it has some limitations related to its linear nature and sensitivity to imbalanced data, its simplicity and robustness have led to widespread use across industries. Recent advances in regularization, scalability, and handling imbalanced data are further strengthening its utility in practical applications. Future research will likely focus on enhancing its adaptability to more complex and large-scale datasets.

References

[1]. M. Chen, L. Liu, and H. Zhang"Logistic Regression in Rare Event Data: A Case Study in Credit Card Fraud Detection" 2018International Conference on Machine Learning and Applications (ICMLA).

[2]. Chen L. and Zhang H. "Regularized Logistic Regression for Feature Selection in High-Dimensional Biomedical Data" 2021 Journal of Biomedical Informatics.

[3]. Mandal D. and Ray S."Ensemble Methods with Logistic Regression for Improved Classification in Credit Scoring" 2022 Journal of Financial Economics and Machine Learning.

[4]. Johnson T."A Comprehensive Review on the Application of Logistic Regression in Social Media Sentiment Analysis" 2024Journal of Data Science Applications.

[5]. Neelam Labhade-Kumar "To Study Different Types of Supervised Learning Algorithm" May 2023, International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), Volume 3, Issue 8, May 2023,PP-25-32, ISSN-2581-9429 DOI: 10.48175/IJARSCT-10256

[6]. Agarwal A."Communication-Efficient Secure Logistic Regression" 2024IEEE European Symposium on Security and Privacy (Euro S&P)

[7]. X. Zou"High Performance Computing Applied to Logistic Regression: A CPU and GPU Implementation Comparison" 2023IEEE International Conference on High Performance Computing and Communications (HPCC).

[8]. Neelam Labhade-Kumar, Study on Object Detection Algorithm, Indian Journal of Technical Education UGC Care Group I, ISSN 0971-3034 Vol47,Special Issue,PP- 14-17, April 2024

[9]. Patel A."Kernel-Based Logistic Regression for Non-Linear Decision Boundaries in Medical Image Analysis" 2023IEEE International Symposium on Biomedical Imaging (ISBI).

[10]. Neelam Labhade-Kumar "Combining Hand-crafted Features and Deep Learning for Automatic Classification of Lung Cancer on CT Scans" 2023, Journal of Artificial Intelligence and Technology.

[11]. T. Kim, J. Park, and L. Chen"Improving Logistic Regression for Imbalanced Data Classification Using Synthetic Sampling Methods" 2019IEEE Access Journal,118234–118244,DOI: 10.1109/ACCESS.2019.2929734.

[12]. R. White, M. Green, and S. Thomas"Regularized Logistic Regression for Feature Selection in High-Dimensional Omics Data" 2020Bioinformatics, 3245–3252, DOI: 10.1093/bioinformatics/btz952.

[13]. M. K. Sharma, P. Patel, and R. Lopez"Logistic Regression with Elastic Net Regularization for Credit Scoring in Finance" 2020Proceedings of the International Conference on Financial Engineering (ICFE), 12-19, DOI: 10.1109/ICFE49509.2020.00010.

[14]. Neelam Labhade-Kumar "Enhancing Crop Yield Prediction in Precision Agriculture through Sustainable Big Data Analytics and Deep Learning Techniques", Carpathian Journal of Food Science and Technology,2023, Special Issue, 1-18

[15]. J. Evans, A. Liu, and K. Roberts"Ensemble Approaches Combining Logistic Regression for Predicting Loan Defaults" 2021IEEE Transactions on Knowledge and Data Engineering, 2095–2104, 2317–2327, DOI: 10.1109/TKDE.2021.3057459.

[16]. R. Johnson, T. Lee, and S. Martinez"Weighted Logistic Regression for Imbalanced Sentiment Analysis in Social Media"2022 Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 1323–1334, DOI: 10.18653/v1/2022.emnlp-main.141.

[17]. H. Garcia, K. Nguyen, and Y. Kim"Deep Logistic Regression: Combining Neural Networks with Logistic Regression for Image-Based Classification" 2023 IEEE Transactions on Neural Networks and Learning Systems,1223–1235,DOI: 10.1109/TNNLS.2023.3172793.

[18]. L. Chen, F. Davis, and M. Choi"Bias Detection in AI: An Explainable Logistic Regression Approach for Fairness in Recruitment" 2023 Journal of Ethics in Artificial Intelligence, 58–71, DOI: 10.1080/26421821.2023.1948129.

[19]. D. Patel, S. Liu, and J. Harris"Adaptive Logistic Regression Models for Real-Time Analytics in E-commerce"2024 ACM Conference on E-commerce, 30-39, DOI: 10.1145/3318508.3342404.