BENGAL COLLEGE OF ENGINEERING AND TECHNOLOGY

NAME : ANIRBAN BANERJEE
UNIVERSITY ROLL NO :12500122052
STREAM : CSE
YEAR : 4th
SEMESTER : 7th
SECTION : A
SUBJECT : Machine Learning
SUBJECT CODE : PEC-CS 701E

Topic name:- Linear Regression & Gradient Descent

ACKNOWLEDGEMENT

First and foremost, I would like to express my sincere gratitude to my Sir Soumyajit Pramanick sir and your department HOD Abdul Rahim sir for their continuous support in my MACHINE LEARNING study and research for the report. His immense knowledge, motivation and guidance helped me a lot in completion of this project. Then a special acknowledgement goes to my principal Sir Prof.
Santanu Koley sir who gave me the golden opportunity of this wonderful project. Finally, I would like to thank my fellow batchmates for helpful discussions and ideas along the way.

~ Anirban Banerjee (CSE 4th year)

INTRODUCTION

Logistic regression is a fundamental statistical and machine learning technique used for classification problems. Unlike linear regression, which predicts continuous outcomes, logistic regression predicts categorical outcomes, often binary in nature, such as “yes/no,” “spam/not spam,” or “disease/no disease.” It models the probability of an event occurring by fitting data to a logistic curve, a type of sigmoid function that maps values to a range between 0 and 1. This makes logistic regression highly interpretable and suitable for decision-making in real-world applications.

The core principle of logistic regression lies in estimating the odds of an event and then transforming them into probabilities using the logistic function. The relationship between input features and the log-odds of the outcome is modeled linearly, but the output is a probability score. A classification boundary, also known as a decision boundary, is determined by applying a threshold (commonly 0.5) to these probabilities. This boundary separates different classes in the feature space and serves as the rule for prediction.

Logistic regression remains widely used due to its simplicity, efficiency, and interpretability. It is less computationally intensive than many modern machine learning algorithms and provides coefficients that have clear statistical meaning. Despite its limitations in handling complex nonlinear relationships, logistic regression serves as a baseline model and a starting point for understanding more advanced classification methods. Its strength lies in combining statistical rigor with practical applicability, making it an essential tool for researchers, analysts, and data scientists.

METHODOLOGY

The methodology of logistic regression involves several structured steps that transform raw data into a predictive classification model. The first step is **data preprocessing**, which includes handling missing values, scaling or normalizing features, and encoding categorical variables into numerical form. Logistic regression assumes that the predictor variables are independent and linearly related to the log-odds of the outcome. Therefore, ensuring data quality is crucial for reliable model performance.

Next, the model is constructed by defining the **logit function**. The logit is the natural logarithm of the odds of the dependent variable belonging to a particular class. Mathematically, it is expressed as:

log⁡(p1−p)=β0+β1x1+β2x2+…+βnxn\log \left(\frac{p}{1-p}\right) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_nx_nlog(1−pp)=β0+β1x1+β2x2+…+βnxn

where ppp is the probability of the positive class, β\betaβ values are coefficients, and xxx values are the independent features.

The coefficients are estimated using **Maximum Likelihood Estimation (MLE)**, which identifies the parameter values that maximize the probability of observing the given dataset. Unlike linear regression, which minimizes squared errors, logistic regression optimizes likelihood functions.

Once the model is trained, the output probabilities are converted into class predictions using a threshold. The most common threshold is 0.5, but this can be adjusted depending on the problem's sensitivity and specificity requirements. The classification boundary is then visualized in the feature space as a line (in two dimensions) or a hyperplane (in higher dimensions). Model evaluation is conducted using performance metrics such as accuracy, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC).

DISCUSSION

The discussion around logistic regression and classification boundaries centers on its advantages, limitations, and role in modern data science. One of the main strengths of logistic regression is its interpretability. Each coefficient provides insight into the relationship between predictors and the log-odds of the outcome, allowing analysts to explain why a particular decision is made. This transparency is valuable in fields such as healthcare, finance, and social sciences, where interpretability is critical.

The concept of a classification boundary is integral to logistic regression. In simple two-dimensional cases, this boundary is a straight line that divides the data into two regions based on the predicted probability. For higher dimensions, it generalizes into hyperplanes. While the boundary is linear by default, logistic regression can handle non-linear separations through feature engineering (e.g., polynomial features, interaction terms). This makes the method flexible yet still computationally efficient.

Despite these strengths, logistic regression also has limitations. It assumes a linear relationship between predictors and the log-odds, which may not hold true in complex real-world data. It is sensitive to multicollinearity, outliers, and imbalanced datasets, which can distort predictions. Moreover, logistic regression struggles with large, high-dimensional datasets where more advanced models like support vector machines, random forests, or deep neural networks perform better.

Nevertheless, logistic regression remains a robust baseline model and is often used as a benchmark for evaluating more complex algorithms. Its balance of simplicity, speed, and interpretability ensures its continued relevance, particularly when quick and understandable results are needed.

APPLICATION

Logistic regression has extensive applications across diverse fields due to its ability to handle binary and multiclass classification problems. In healthcare, it is widely used for disease prediction and medical diagnosis. For example, logistic regression can model the probability of a patient developing diabetes based on features such as age, body mass index, and blood sugar levels. Its interpretability allows medical professionals to identify risk factors and design preventive measures.

In finance, logistic regression is employed in credit scoring and fraud detection. By analyzing applicant profiles, income levels, and past repayment behavior, banks can estimate the likelihood of loan defaults. Similarly, logistic regression models transaction data to identify suspicious activities that may indicate fraud.

In marketing, businesses use logistic regression to predict customer behavior, such as whether a user will click an advertisement or purchase a product. By modeling consumer features and past interactions, companies can optimize targeted marketing campaigns. Logistic regression also finds applications in churn prediction, where companies identify customers at risk of leaving their services.

In the domain of social sciences, logistic regression assists researchers in analyzing survey data and studying social outcomes like voting patterns, education levels, and employment probabilities.

Furthermore, logistic regression plays a foundational role in machine learning pipelines as a baseline model for binary and multiclass classification tasks. It is often used for text classification, spam detection, sentiment analysis, and natural language processing tasks. Its ability to handle both interpretability and predictive accuracy makes it an essential tool in practical applications.

CONCLUSION

Logistic regression and classification boundaries represent a cornerstone in the field of statistical modeling and machine learning. As one of the most widely used classification techniques, logistic regression provides a structured approach to predicting categorical outcomes based on input features. Its strength lies in combining mathematical rigor with real-world interpretability, making it both accessible to beginners and valuable to professionals across disciplines.

The concept of classification boundaries highlights how logistic regression translates probability outputs into actionable decisions. A simple linear decision boundary can effectively separate classes in many cases, while feature transformations extend its applicability to more complex problems. Although its assumptions, such as linearity of log-odds and independence of predictors, may restrict performance in certain datasets, logistic regression remains an efficient and reliable modeling technique.

From healthcare to finance, marketing, and social sciences, the applications of logistic regression demonstrate its adaptability and effectiveness. It helps professionals make data-driven decisions, identify risks, and predict future events. Its role as a baseline model in machine learning ensures that it continues to be used for comparison with more advanced methods, providing a benchmark for evaluating predictive performance.

In conclusion, logistic regression is not just a statistical technique but also a bridge between theory and practice. While newer and more sophisticated models have emerged, logistic regression remains indispensable due to its simplicity, transparency, and efficiency. By mastering logistic regression and understanding classification boundaries, one gains a strong foundation for tackling more advanced challenges in machine learning and predictive analytics.

BIBLIOGRAPHY:-

- ✓ *WWW.WIKIPEDIA.COM*
- ✓ *WWW.GEEKSFORGEEKS.COM*
- ✓ *CLASS NOTES*

THANK YOU