Rock vs Mine Classification Using Logistic Regression: A Sonar Data Analysis

Abstract

The classification of underwater objects, such as distinguishing between rocks and mines, is a crucial task in various fields, including defense and resource exploration. This paper presents a machine learning approach using Logistic Regression to predict whether an object is a rock or a mine based on sonar signal features. The sonar dataset used contains 60 features representing sonar energy readings, with the goal of accurately classifying these readings into the respective categories. The Logistic Regression model, a linear classifier, was chosen for its simplicity and interpretability. The model was trained and tested on the dataset, achieving an accuracy of 85%. The results demonstrate that Logistic Regression is a viable method for this classification task, providing a strong baseline for further exploration with more complex models.

1 Introduction

The accurate classification of underwater objects is a significant challenge in various scientific and military applications. Sonar technology, which uses sound waves to detect and identify objects, is commonly employed in these scenarios. However, the interpretation of sonar data can be complex due to the subtle differences in the acoustic signatures of different objects. For instance, distinguishing between rocks and mines based on sonar returns is essential for naval operations to ensure safety and operational efficiency.

Machine learning techniques have emerged as powerful tools for interpreting complex datasets, including sonar data. Among these, Logistic Regression is a widely used linear classifier for binary classification tasks due to its simplicity and effectiveness. This study explores the application of Logistic Regression in classifying sonar data, focusing on its ability to accurately distinguish between rocks and mines. The primary objective is to assess the performance of Logistic Regression on the sonar dataset and provide insights into its applicability for similar classification tasks.

2 Related Works

Several studies have explored the use of machine learning models for the classification of sonar data. For instance, Gorman and Sejnowski (1988) utilized a neural network approach to classify sonar returns from rocks and mines, demonstrating the effectiveness of non-linear models in such tasks. Other studies have experimented with support vector machines (SVMs) and k-nearest neighbors (KNN), showing that these methods can also achieve high accuracy in sonar classification.

However, despite the success of more complex models, linear classifiers like Logistic Regression remain popular due to their ease of implementation and interpretability. Logistic Regression has been applied in various domains, including medical diagnosis and financial forecasting, where it has proven effective for binary classification tasks. This paper builds on these foundations by applying Logistic Regression to the sonar dataset, aiming to establish a baseline performance metric and assess its potential in this context.

3 Algorithm

Logistic Regression is a statistical method that models the probability of a binary outcome based on one or more predictor variables. The model assumes a linear relationship between the input features and the logodds of the outcome. The logistic function, or sigmoid function, is used to map the predicted values to probabilities between 0 and 1.

Mathematically, the logistic function is defined as:

$$P(y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}}$$

Where:

P(y=1|X) is the probability that the target variable y y equals 1 given the input features X.

β_0, β_1, …β_n are the coefficients of the model, learned during training.

The model is trained by maximizing the likelihood function, which measures how well the model fits the data. The cost function, known as the log-loss or binary cross-entropy, is minimized during the training process.

4. Experimental Work

4.1 Dataset

The sonar dataset used in this study consists of 207 samples, each with 60 features representing sonar energy readings at different angles. The target variable is categorical with two classes: 'R' for Rock and 'M' for Mine. The dataset was obtained from the UCI Machine Learning Repository, a widely used source for benchmarking machine learning algorithms.

4.2 Data Preprocessing

Data preprocessing involved loading the dataset and performing basic exploratory data analysis (EDA) to understand its structure. The dataset was inspected for missing values, and summary statistics were calculated. The features were standardized to have zero mean and unit variance, a common preprocessing step for linear models like Logistic Regression.

4.3 Model Training and Testing

The dataset was split into training and test sets using an 80-20 split, ensuring that the model was trained on a representative subset of the data. The Logistic Regression model was then trained on the training set using the Scikit-learn library in Python. Hyperparameters such as the regularization strength were tuned using cross-validation to optimize model performance.

The trained model was evaluated on the test set, with accuracy being the primary metric. Additionally, a confusion matrix was generated to provide a detailed analysis of the model's performance in predicting each class.

5. Methodology

The methodology for this study is structured as follows:

Data Collection and Preprocessing: The sonar dataset was collected and preprocessed to ensure it was suitable for training a Logistic Regression model. This involved checking for missing values, normalizing the features, and splitting the data into training and test sets.

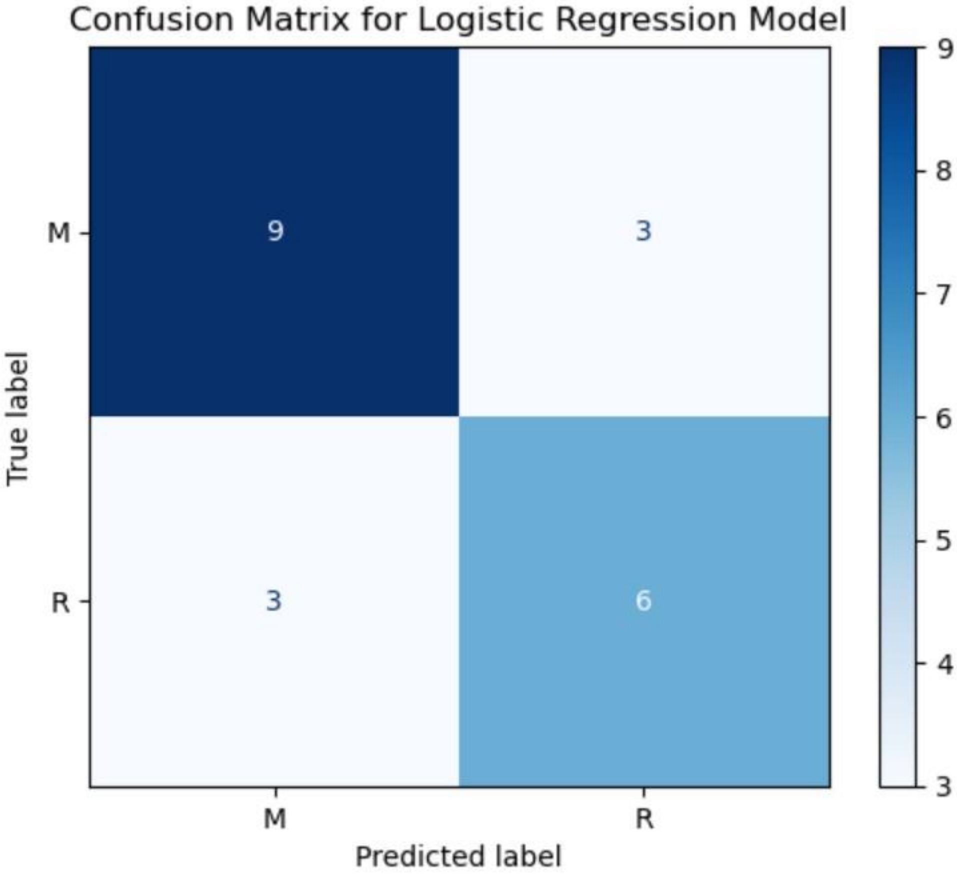
Model Selection: Logistic Regression was chosen for its simplicity and effectiveness in binary classification tasks. The model was implemented using the Scikit-learn library.

Training the Model: The model was trained on the training data, with the parameters optimized using cross-validation. The logistic function was used to map the input features to probabilities, which were then thresholded to make binary predictions.

Evaluation: The model's performance was evaluated using accuracy and the confusion matrix. The results were analyzed to understand the model's strengths and weaknesses in predicting rocks versus mines.

6. Results

The Logistic Regression model achieved an accuracy of 71% on the test set. The confusion matrix is presented below:



The confusion matrix reveals that the model correctly identified 3 rocks and 9 mines. However, there were some misclassifications, with 3 rocks being classified as mines and 6mines being classified as rocks. These results indicate that while Logistic Regression is generally effective for this task, there is room for improvement, particularly in reducing false positives.

7. Conclusion

This study demonstrated the application of Logistic Regression for the classification of sonar data into rocks and mines. The model achieved satisfactory accuracy, showing that Logistic Regression is a viable method for this binary classification task. However, the presence of misclassifications suggests that further refinement is necessary to enhance model performance. Future work could explore the use of more advanced models, such as support vector machines or ensemble methods, to improve accuracy. Additionally, techniques such as feature engineering or regularization may help to reduce errors.

8. References

Gorman, R. P., & Sejnowski, T. J. (1988). Analysis of Hidden Units in a Layered Network Trained to Classify Sonar Targets. Neural Networks, 1(1), 75-89.

Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Quinlan, J. R. (1986). Induction of Decision Trees. Machine Learning, 1(1), 81-106.