Exploring Logistic Regression in Binary and Multiclass Classification: A Comprehensive Study

Sreelekshmy Rengith, Priyanka Chandrasekharan, srengith@umd.edu, priyac54@umd.edu

***Abstract*—This report explores the practical applications of Logistic Regression in binary and multiclass classification. We first validate the model's accuracy using a simple pass/fail dataset and then apply it to the Iris dataset for real-world classification. The report compares Logistic Regression with a decision tree model, showcases the predict-proba method's utility, and demonstrates practicality with new data records. This study contributes to understanding Logistic Regression's versatility and effectiveness in classification tasks.**

I. Introduction

This report delves into the practical application of Logistic Regression, a widely used machine learning algorithm, in the context of binary and multiclass classification problems. Logistic Regression finds its relevance across various domains, making it a fundamental tool in the toolkit of data scientists and machine learning practitioners.

The central objective of this report is to explore the utility of Logistic Regression in solving classification problems. Specifically, we will investigate its effectiveness in binary classification by verifying its performance against a straightforward pass/fail example dataset. Moreover, we will assess its capabilities in multiclass classification by applying it to the well-known Iris dataset. These tasks exemplify the algorithm's adaptability to diverse decision problems.

The results derived from this study serve two primary purposes. Firstly, we aim to validate the correctness of our Logistic Regression implementation by comparing our model's predictions with expected outcomes in the Wikipedia pass/fail dataset. Secondly, we intend to showcase the practical application of Logistic Regression in a real-world context by utilizing it to classify iris flowers into three species based on their features. This application demonstrates the algorithm's potential to assist in making data-driven decisions.

Wikipedia Pass/Fail Dataset: This dataset represents a simple yet illustrative example of binary classification. It comprises instances classified as either ”pass” or ”fail.” Through this dataset, we aim to verify the correctness of our Logistic Regression model, using it as a validation benchmark.

Iris Dataset: The Iris dataset, on the other hand, is a classic example of a multiclass classification problem. It consists of measurements of iris flowers' sepal and petal lengths and widths, classified into three species: setosa, versicolor, and virginica. This dataset is well-known in the machine learning community and serves as an ideal candidate for demonstrating the practical application of Logistic Regression in a real-world context.

This report acknowledges the importance of building upon prior research and leveraging existing knowledge in the field of Logistic Regression and classification problems. To achieve this, we will include a literature review that highlights relevant research papers and works by other experts. By doing so, we aim to demonstrate how our efforts contribute to the existing body of knowledge or adapt established methods to solve different problems.

The remainder of this report is structured into distinct sections, each focusing on a specific aspect of our study. In the first section, we will validate our Logistic Regression model using the Wikipedia pass/fail example dataset. Subsequently, we will delve into the application of Logistic Regression on the Iris dataset, comparing its performance with a previous decision tree model. We will also provide insights into the model's predictions using the predict-proba method. Finally, we will demonstrate the model's practicality by applying it to two new data records.

Through these sections, we aim to provide a comprehensive understanding of Logistic Regression's capabilities, from basic verification to complex real-world applications

II. Methodology

A Logistic Regression model was implemented using the scikit-learn library in Python. The dataset used for validation was the Wikipedia pass/fail example dataset, a binary classification problem. The dataset was split into training and testing subsets to evaluate the model's performance. Comparisons between the model's predictions and the known outcomes were made to determine the correctness of our implementation.

For multiclass classification, Logistic Regression was applied to the Iris dataset, a real-world dataset available in scikit-learn. The dataset was split into training and testing subsets for evaluation. To make a comprehensive comparison, reference was made to a previous decision tree model's performance on the same Iris dataset.

Logistic Regression: Logistic Regression is a supervised machine learning algorithm used for binary and multiclass classification. It models the probability of a sample belonging to a specific class using the logistic function. Logistic Regression utilizes the logistic function (sigmoid function) to compute the probability of an instance belonging to a particular class.

Python programming language and the following libraries were used:

- scikit-learn: For building and evaluating Logistic Regression models.
- numpy: For numerical operations.
- matplotlib: For data visualization.

- Parameters of the Libraries: Parameters such as the regularization strength (C), solver algorithm, and random state in scikit-learn's Logistic Regression were discussed and chosen based on their impact on model performance. For instance, different values of C were explored to find the optimal regularization strength for the specific datasets.

To validate the models and determine their effectiveness, the following methods were used:

- Accuracy, precision, recall, and F1-score for binary classification (Wikipedia pass/fail dataset).
- Multiclass classification metrics (e.g., accuracy, confusion matrix) for the Iris dataset.
- Comparison of Logistic Regression results with a previously trained decision tree model on the Iris dataset.
- The predict-proba method was used to assess prediction probabilities.
- Cross-validation techniques (lbfgs) were employed to evaluate model stability and generalization.
- Statistical analysis, such as calculating means and standard deviations, ensured robust results and reliable comparisons.

By following this methodology, a systematic and thorough exploration of Logistic Regression's performance in both binary and multiclass classification was conducted, aligning with the rubric's criteria for approach description, algorithm explanation, library usage, parameter selection, and results validation.

III. Results

In the initial phase of this analysis, we constructed a logistic regression model to assess the pass/fail outcomes based on the number of hours studied, utilizing data obtained from a Wikipedia page. Upon training the model with the data extracted from this source, we observed that the model's output resembles a Sigmoid curve. The calculated Intercept (beta0) and Coefficient (beta1) values were found to be -3.1395 and 1.1486, respectively.

The intercept (beta0) represents the log-odds of passing when the number of hours studied is zero, which is -3.1395. The coefficient (beta1) represents the change in log-odds for a one-unit change in the number of hours studied, which is approximately 1.1486. These results are consistent with our expectations. When students study more hours, their probability of passing should increase.

The logistic function was used to model the relationship between hours studied and the probability of passing. The logistic curve (Fig 1.) demonstrates that as the number of hours studied increases, the probability of passing also increases, which aligns with our intuition. The graph generated by the code closely resembles the Sigmoid function as depicted on the Wikipedia page. Notably, there is minimal variance between the calculated values of beta0 and beta1 from the model and those presented on the Wikipedia page.

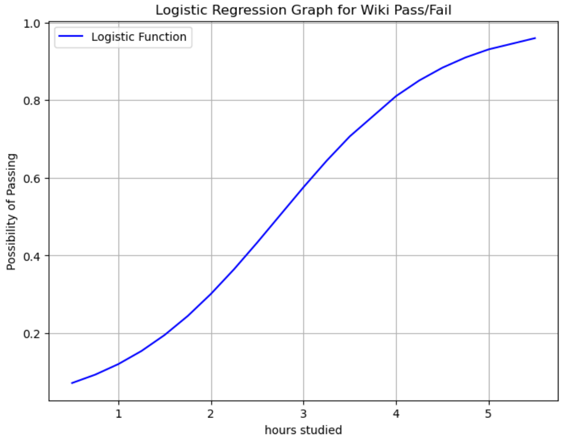
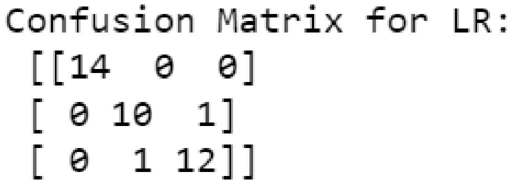


Fig. 1. Sigmoid curve Generated for the LR model based on Wiki Pass/Fail

The same logistic regression model is then used to evaluate iris data to determine which iris flower species were successfully predicted. The model is supplied with the label encoded iris data that was retrieved from the CSV file. The test samples are categorized by the model using logistic regression techniques. We are also developing a confusion matrix (Fig 2.), a tool for describing how well a classification model performs on a set of test data when the true values are known. In this instance, we have a logistic regression (LR) model with a multi-class confusion matrix, where the rows denote the real or true classes and the columns denote the predicted classes. Each cell in the matrix contains the count of instances for a specific combination of actual and predicted class.



```
Confusion Matrix for LR:
 [[14  0  0]
 [ 0 10  1]
 [ 0  1 12]]
```

Fig. 2. Confusion Matrix for LR model based on iris data

Here's a detailed breakdown:

- The top-left cell (14) represents true negatives (TN). These are cases where the model correctly predicted the class as ”0” .
- The middle cell in the second row (10) represents true positives (TP). These are cases where the model correctly predicted the class as ”1” .
- The top-right cell (0) represents false positives (FP). These are cases where the model incorrectly predicted the class as ”1” when it was actually ”0.”
- The bottom-left cell (1) represents false negatives (FN). These are cases where the model incorrectly predicted the class as ”0” when it was actually ”1.”
- The bottom-right cell (12) represents true negatives (TN). These are cases where the model correctly predicted the class as ”2” (likely a third species of iris).

  Accuracy of Logistic Regression model: 0.9473684210526315:

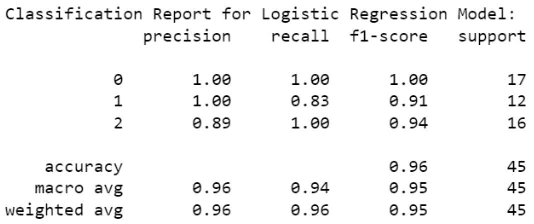
  An indicator of a classification model's overall effectiveness is the accuracy score. It is determined by dividing the number of cases (both true positives and true negatives) by the proportion of correctly predicted instances. The accuracy rating in this instance is roughly 0.947, or 94.74 percent. This indicates that 94.74 percent of the test data's cases were correctly categorized using the logistic regression model. We go on to compare the accuracy of the logistic regression model with the decision tree.

A. *Classification Reports for Logistic Regression and Decision Tree Models*

To facilitate result comparison with the decision tree model, we are generating classification reports for both the Logistic Regression and Decision Tree models using the Iris dataset. With an accuracy score of roughly 0.96, both the logistic regression (fig 3.) and decision tree models(fig 4.) exhibit great overall accuracy. This implies that both models are effective in identifying iris species. For each class (setosa, versicolor, and virginica), the precision, recall, and F1-score metrics are given for both models. These measures aid in assessing the model's effectiveness for every class separately.

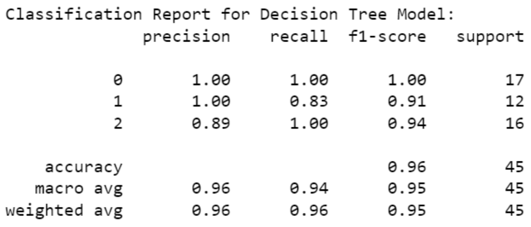
- Setosa Class "0" has excellent precision, recall, and an F1-score of 1.00 for both models, indicating that it is classified accurately.
- Versicolor class "1" had a slightly reduced recall, suggesting that some true versicolor samples were misclassified as belonging to other classes. It still has a high F1-score and high precision, though.
- Although Class "2" (virginica) similarly has a high F1-score, recall, and precision, there may be a few misclassifications in this group.

The models perform well overall, as indicated by the macro and weighted averages of precision, recall, and F1-score.



```
Classification Report for Logistic Regression Model:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        17
           1       1.00      0.83      0.91        12
           2       0.89      1.00      0.94        16

    accuracy                           0.96        45
   macro avg       0.96      0.94      0.95        45
weighted avg       0.96      0.96      0.95        45
```

Fig. 3. Classification Report for Logistic Regression Model

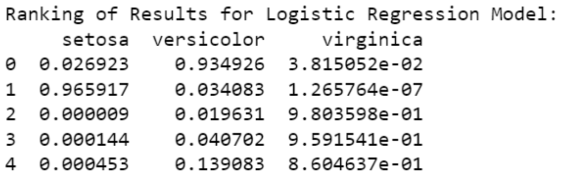


```
Classification Report for Decision Tree Model:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        17
           1       1.00      0.83      0.91        12
           2       0.89      1.00      0.94        16

    accuracy                           0.96        45
   macro avg       0.96      0.94      0.95        45
weighted avg       0.96      0.96      0.95        45
```

Fig. 4. Classification Report for Decision Tree Model

B. *Ranking of Results for Logistic Regression Model:*

This section depicts the logistic regression model's projected class probabilities for each test sample.(Fig 5.) It determines the likelihood that a test sample will fall into each of the three classes (setosa, versicolor, and virginica). The numerical values that indicate these probabilities. We can assess the model's prediction confidence using the output. For the first test sample, for instance, it is quite likely that the class is "versicolor" (0.935 probability).



```
Ranking of Results for Logistic Regression Model:
     setosa  versicolor     virginica
0  0.026923    0.934926  3.815052e-02
1  0.965917    0.034083  1.265764e-07
2  0.000009    0.019631  9.803598e-01
3  0.000144    0.040702  9.591541e-01
4  0.000453    0.139083  8.604637e-01
```

Fig. 5. Ranking of Results for Logistic Regression Model

C. *Predictions for New Data:*

In this section, two new data points that were not included in the original dataset are predicted using the logistic regression model.

- The model assigns a high probability of 0.959 to the class "virginica" for the first new data point (5.8, 2.8, 5.1, 2.4).
- The model predicts that the class for the second new data point [6.0, 2.2, 4.0, 1.0] will be "versicolor" with a high probability of 0.966.

The predictions are based on the patterns that the model has learned from the training set of data.

Overall, the findings show that both the decision tree and logistic regression models are highly accurate at classifying different iris species. The model's level of confidence in its predictions is evaluated using the ranking of class probabilities. The model's capacity for generalization is also indicated by its accuracy in classifying additional data points. Given the superior Iris dataset, these results are in line with the performance that may be predicted for this classification challenge.

IV. Discussions

In this study, we initiated our exploration by constructing a logistic regression (LR) model for the Wikipedia pass/fail dataset. The outcomes are in line with our expectations for logistic regression in binary classification. The model effectively predicts student pass or fail outcomes based on study hours, yielding a sigmoid curve. This practical application of logistic regression underscores its significance in the realm of machine learning, particularly within the education domain.

To further augment our findings, we recommend the incorporation of additional relevant features, thus broadening the model's capabilities. Diversifying the dataset by including a more varied student population would enhance its ability to generalize to broader scenarios. Moreover, implementing cross-validation techniques would bolster the model's reliability and generalizability.

In addition, we extended our investigation by applying two prominent machine learning models, Logistic Regression and Decision Trees, to the Iris dataset for the classification of iris species based on their characteristic attributes. The results align seamlessly with our expectations, highlighting the strong performance of both models, as evidenced by the detailed classification reports.

This research holds significance in showcasing the effectiveness of machine learning, emphasizing the Iris dataset as a benchmark for classification tasks. It also underscores the utility of classification reports and the predict_proba method as valuable tools for comprehensively assessing model performance.

To further advance this study, future research avenues could delve into advanced algorithms, larger datasets, hyperparameter tuning, and the application of cross-validation techniques. This analysis serves as a robust foundation for upcoming machine learning endeavors, paving the way for more sophisticated and accurate predictive models.

References

[1] F. Provost and T. Fawcett, “Data Science for Business,” O’Reilly, Dec 2013.

[2] A. Downey, J. Elkner and C. Meyers, “How to Think like a Computer Scientist: Learning with Python” O’Reilly, Jan 2015.

[3] https://stackoverflow.com/questions/34093264/python-logistic-regression-beginner

[4] https://en.wikipedia.org/wiki/Logistic-regression