Statistical Analysis of Income Prediction

Logistic Regression and Probabilistic Modeling

Javkhlan

2025-12-01

Table of contents

1 Introduction

This report analyzes the Adult Income dataset to predict whether an individual earns more than $50,000 annually. We employ **Logistic Regression**, a fundamental statistical method for binary classification that models the probability of class membership.

The focus is on the probabilistic interpretation of the model, the derivation of the loss function via Maximum Likelihood Estimation (MLE), and the evaluation of the model using statistical metrics.

1.1 Educational Context

Beyond the application, this project emphasizes the educational value of implementing statistical algorithms from first principles. Rather than relying solely on "black-box" implementations from libraries like Scikit-Learn, we utilize a custom implementation of Logistic Regression. This allows for a transparent examination of the optimization process—specifically how Gradient Descent navigates the loss landscape to find optimal coefficients.

2 Mathematical Framework

2.1 The Logistic Model

We model the probability that the target variable Y takes the value 1 (income $>$ 50K) given input features $X = x$ using the sigmoid function $\sigma(z)$:

$$P(Y = 1|X = x) = \sigma(w^T x + b) = \frac{1}{1 + e^{-(w^T x + b)}}$$

where w are the weights and b is the bias. The linear combination $z = w^T x + b$ represents the log-odds (logit).

2.2 Maximum Likelihood Estimation

To estimate parameters w and b, we maximize the likelihood of the observed data. Assuming independent and identically distributed (i.i.d.) samples, the likelihood L is:

$$L(w, b) = \prod_{i=1}^{m} P(y^{(i)}|x^{(i)})^{y^{(i)}} (1 - P(y^{(i)}|x^{(i)}))^{1-y^{(i)}}$$

Taking the negative logarithm gives us the **Binary Cross-Entropy** loss function, which we minimize:

$$J(w, b) = -\frac{1}{m} \sum_{i=1}^{m} \left[y^{(i)} \log(\hat{y}^{(i)}) + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)}) \right]$$

Unlike Mean Squared Error (MSE), which is non-convex for logistic regression, the Log-Loss function is convex, guaranteeing that Gradient Descent will converge to the global minimum.

2.3 Regularization

To prevent overfitting, we introduce an L_2 regularization term (Ridge), which corresponds to placing a Gaussian prior on the weights $w \sim \mathcal{N}(0, \tau^2)$. The objective function becomes:

$$J_{reg}(w, b) = J(w, b) + \frac{\lambda}{2m} ||w||^2$$

3 Data and Methodology

We utilize the Adult Income dataset. The data is split into training (80%) and validation (20%) sets.

3.1 Data Characteristics and Preprocessing

The dataset contains a mix of numerical (e.g., Age, Capital Gain) and categorical (e.g., Education, Occupation) features. A critical challenge in this dataset is **class imbalance**: approximately 76% of individuals earn $\leq$ 50K, while only 24% earn $>$ 50K. This imbalance implies that a naive model predicting the majority class for every instance would achieve 76% accuracy but have zero predictive power for the target class.

To prepare the data for our gradient-based optimization: 1. **One-Hot Encoding**: Categorical variables are transformed into binary vectors. 2. **Standard Scaling**: Numerical features are normalized to have mean 0 and variance 1. This is crucial for Gradient Descent, as it ensures the loss landscape is symmetric, preventing the optimizer from oscillating or converging slowly.

4 Results

4.1 Performance Metrics

Metric	Value
Accuracy	0.85
Precision	0.79
Recall	0.72
F1 Score	0.75

The F1 Score, which balances precision and recall, is particularly important given the class imbalance. A naive accuracy measure would be misleadingly high due to the 76% majority class.

4.2 Confusion Matrix

The confusion matrix is displayed using a plot.

4.3 ROC Curve

The ROC curve demonstrates the model's ability to distinguish between classes. The area under the curve (AUC) provides a single metric for model performance across all classification thresholds.

4.4 Feature Importance

The coefficients of the model indicate the relative importance of each feature. A higher absolute value of a coefficient implies a greater impact on the model's predictions. Features like `age`, `education-num`, and `hours-per-week` show significant importance.

5 Discussion

- Quarto enables reproducible, documented analysis with code and narrative.
- Equations like **?@eq-logreg** are first-class citizens alongside figures and tables.
- Use citations such as Roback and Legler (2021) to anchor methods in literature.

5.1 Statistical Interpretation

The default threshold for classifying an instance as income >50K is 0.5. However, this can be adjusted depending on the desired balance between precision and recall. For instance, in scenarios where false negatives are costly, lowering the threshold might be beneficial.

5.2 Limitations and Future Work

- The assumption of linearity between the log-odds of the outcome and the predictors may not hold in all cases.
- The model does not account for potential interactions between features.
- Future work could explore non-linear models or ensemble methods for comparison.

6 Conclusion

This template can be cloned for new reports. Replace text, update references, and embed your analysis code.

7 Appendix

7.1 Additional Math

A simple linear model:

$$y = X\beta + \varepsilon \tag{1}$$

Refer to Equation Equation 1 in text.

7.2 Re-usable Blocks

- Use sections and sub-sections to structure content.
- Add callouts, code-folding, and filters as needed.
- Keep references in references.bib and cite with `@key`.

References

Roback, Paul, and Julie Legler. 2021. *Beyond Multiple Linear Regression: Applied Generalized Linear Models and Multilevel Models in r*. Chapman; Hall/CRC. https://bookdown.org/roback/bookdown-BeyondMLR/ch-MLRreview.html.