The Report of Choosing Logistic Regression as a Machine Learning Model

Qzxl76 Heria CHEN

January 2025

Abstract

This report details the process of selecting and implementing a machine learning model for a multi-modal classification task involving brain, image, and text features. Through exploratory data analysis, I identified key characteristics of the data, including varying numerical scales and non-normal distributions across modalities. Based on these findings, I chose logistic regression as my base model due to its robustness to non-normality, ability to handle multi-modal features, and inherent interpretability. I implemented a custom logistic regression model with L2 regularization and compared its performance to the `sklearn` implementation. Furthermore, I proposed a multi-modal fusion paradigm incorporating normalization, regularization, and dimensionality reduction techniques. Experimental results demonstrate the effectiveness of my approach and highlight the importance of careful data preprocessing and model selection.

1 Data Exploration

KNOW: Preliminary exploratory data analyses revealed variations in the magnitude of feature values across different modalities (brain, image, text), with distributions deviating from strict normality. Specifically, brain features ranged approximately between $[-9.12, 6.92]$, image features between $[-278.45, 619.51]$, and text features between $[-6.81, 14.92]$. This disparity suggests the need for normalization or standardization. Sample statistics indicated near-zero means for all modalities, potentially indicating prior centering. However, image features exhibited a significantly larger standard deviation (4.5185) compared to brain (≈ 1) and text (0.7) features, suggesting a wider spread. Shapiro-Wilk tests confirmed the non-normality of these distributions (p 0).

SEE: Histograms of randomly sampled data from each modality (brain, image, text) visually confirmed the approximate centering around zero. Brain and text features showed relatively compact distributions, while image features exhibited a broader spread, with some dimensions reaching values of 15-20. Box plots and histograms further illustrated the differences in numerical scales across modalities.

FIND: Based on the observed data characteristics, logistic regression was chosen due to its:

- Robustness to non-normal feature distributions.
- Ability to handle varying numerical scales across modalities.
- Capacity for regularization (e.g., L2) to mitigate multicollinearity.
- Interpretability through feature coefficients.

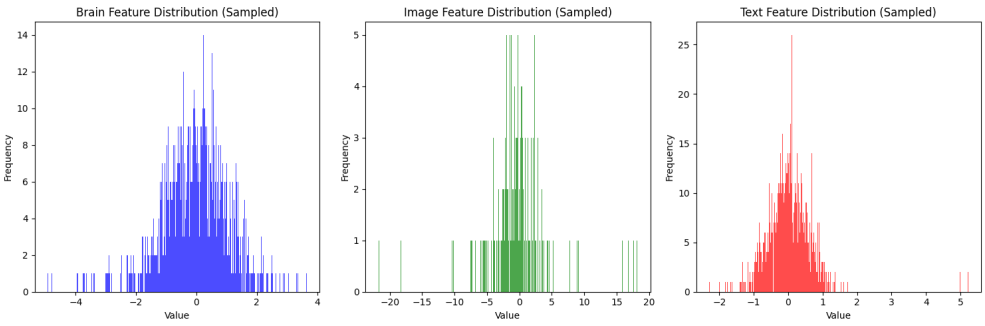

Figure 1: Distributions of brain, image, and text features.

2 Model Implementation

Implement: Implemented a custom logistic regression model with L2 regularization from scratch, avoiding direct use of `sklearn`'s `fit()` method. This allowed for greater control over the training process and facilitated comparison with the `sklearn` implementation.

Compare: Compared the custom implementation with the `sklearn` logistic regression model across various feature combinations (Brain, Text, Image, and their combinations). Analyzed performance discrepancies between the two implementations, focusing on potential differences in optimization algorithms and numerical stability.

Improve: To enhance model performance, we implemented several improvements:

- Consistent normalization using our custom `standard_scale` function to address scale differences across modalities.
- Optional dimensionality reduction using threshold filtering for text and PCA for image features.
- L2 regularization with hyperparameter tuning to prevent overfitting, especially in high-dimensional feature spaces.
- Manual k-fold cross-validation for robust performance evaluation.

3 Results

Performance: Optimized model achieved cross-validation accuracies ranging from 0.5 to 0.62 across different feature combinations, Suspect that the low performance is caused by data imbalance.. While this represented a significant improvement over the initial implementation, it remained slightly below some `sklearn`-based benchmarks. Investigated the impact of training set size by testing different data splits (50-50, 40-60, 30-70, 20-80), observing a clear correlation between training set size and accuracy.

Visualisation: Visualize the results of the model using figure for each feature combination provided further insights into model performance, revealing varying performance across modalities.

Ablation: Ablation studies using different feature combinations demonstrated that the combination of Brain+Text+Image generally yielded the best performance. Using Brain features alone resulted in significantly lower scores. The performance of intermediate combinations varied depending on data quality and feature alignment.

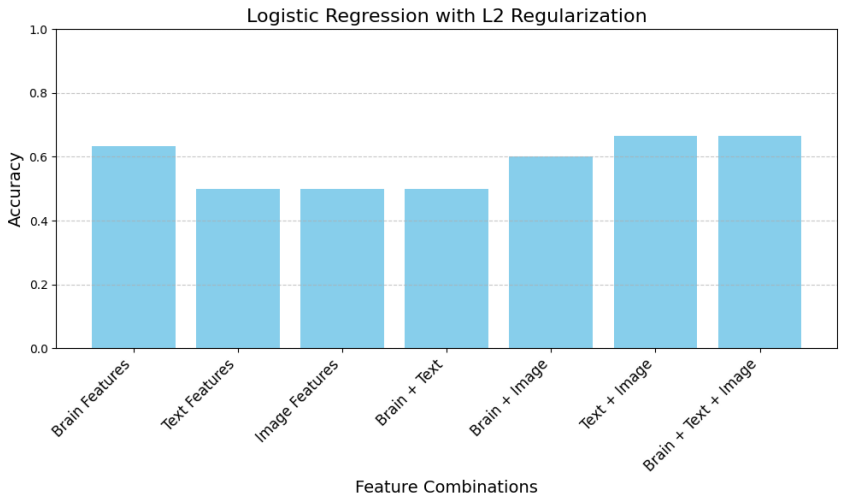

Figure 2: The Accuracy of different combination.

4 Proposed Paradigm

Paradigm: Our proposed multi-modal fusion paradigm for binary classification consists of the following steps:

1. **Data Preprocessing:** Cleaning, noise reduction, and outlier removal for each modality.
2. **Feature Normalization:** Z-score standardization (mean 0, variance 1) applied to all modal features.
3. **Multimodal Fusion:** Concatenation of normalized Brain (EEG), Text, and Image features.
4. **Logistic Regression (L2):** Training of a logistic regression model with L2 regularization.

This paradigm emphasizes normalization and L2 regularization to address scale differences and redundancy among features, contrasting with simpler concatenation methods.

Adjustment: To address potential issues with direct concatenation, such as unstable models and overfitting, we implemented the following adjustments:

- Sub-modal normalization: Computing mean and standard deviation separately for each modality before concatenation.
- Regularization tuning: Adjusting the L2 regularization strength based on feature dimensionality and data characteristics.
- Dimensionality reduction: Applying threshold filtering for text features and PCA for image features to reduce noise and improve efficiency.

Reflection: Our paradigm offers the following advantages:

- **Interpretability:** Logistic regression provides readily interpretable feature weights.
- **Simplicity:** The concatenation and logistic regression approach is straightforward to implement and train.

Limitations of our approach include:

- Inability to capture non-linear interactions between modalities.
- Potential scalability issues with very large datasets or a large number of modalities, which may necessitate deeper models.
- Sensitivity to hyperparameter tuning, especially regularization strength and normalization parameters.

References

Hackeling, G. 2017. Mastering Machine Learning with scikit-learn. 2nd ed. Birmingham: Packt Publishing.