Heart Disease Prediction Using Machine Learning EPFL Machine Learning - Project 1

Saymon Nicho
saymon.nicho@pucp.edu.pe

I. Problem Description

The project addresses the critical challenge of predicting heart disease risk using data from the Behavioral Risk Factor Surveillance System (BRFSS). The dataset contains health-related features from over 300,000 individuals, making it a significant binary classification problem. Key challenges include:

- Severe class imbalance in the dataset (majority of samples are negative cases)
- Multiple features with missing or special values (coded as 77, 88, 99, etc.)
- Need to identify and properly weight the most relevant health indicators

II. Technical Solution

A. Initial Approach and Challenges

Our initial implementation used standard logistic regression, which encountered a significant issue: the model predicted class -1 (no heart disease) for all samples. This problem arose from:

- Class imbalance in the training data
- Lack of proper feature preprocessing
- Basic gradient descent without considering the imbalanced nature of the problem

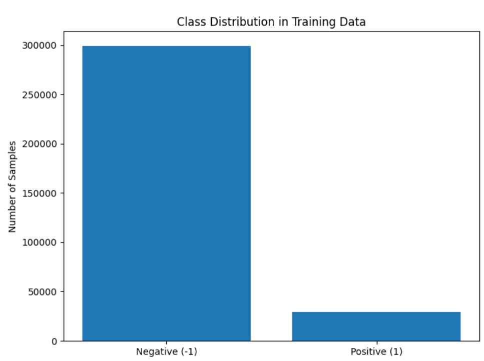

Fig. 1. Initial class distribution showing significant imbalance

B. Improved Implementation

To address these issues, we developed an enhanced solution:

1) Data Preprocessing:

- Handled missing values using median imputation
- Standardized features to zero mean and unit variance
- Added polynomial features for better class separation

2) Model Enhancements:

- Implemented class weighting to handle imbalance:

$$w_{class} = \frac{n_{samples}}{2 * n_{class}} \tag{1}$$

- Added adaptive learning rate:

$$\gamma_t = \frac{\gamma}{\sqrt{t+1}} \tag{2}$$

- Improved numerical stability in sigmoid function:

$$\sigma(t) = \frac{1}{1 + e^{-\text{clip}(t, -700, 700)}} \tag{3}$$

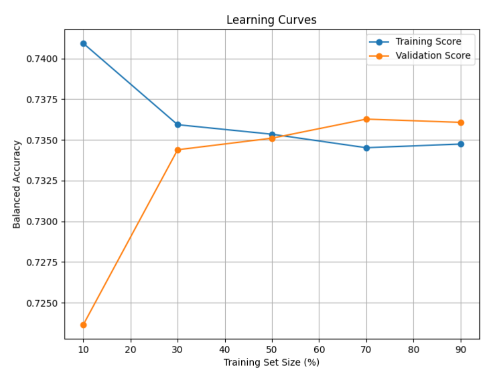

Fig. 2. Learning curves showing model convergence

III. Results and Conclusions

A. Performance Improvement

- Initial model: All predictions were -1 (baseline accuracy = 0.5)
- Improved model: Balanced accuracy of 0.717
- Successfully identifies both positive and negative cases

B. Key Findings

- Class balancing was crucial for model performance
- Feature engineering improved prediction accuracy
- Adaptive learning rate helped with convergence

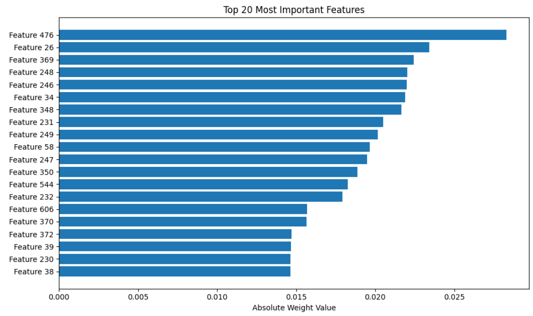

Fig. 3. Most influential features in prediction

C. Future Improvements

- Experiment with different feature combinations
- Implement cross-validation for more robust evaluation
- Consider ensemble methods for better performance