Breast Cancer Classification Using Logistic Regression: A Comprehensive Analysis and Performance Evaluation

Abstract

Breast cancer classification is a critical task in medical diagnostics, aiding in early detection and treatment planning. This study presents a breast cancer classification model using logistic regression to predict the presence of malignancy based on various diagnostic features. The model was evaluated on a dataset with accuracy scores of 94.95% on training data and 92.98% on test data. The results highlight the effectiveness of logistic regression in distinguishing between benign and malignant cases, demonstrating its potential as a reliable tool in medical decision-making.

Introduction

Breast cancer remains one of the leading causes of cancer-related deaths worldwide. Early detection and accurate classification of breast cancer can significantly improve patient outcomes and treatment effectiveness. Logistic regression, a statistical method used for binary classification problems, has shown promise in medical diagnostics due to its simplicity and interpretability. This study explores the application of logistic regression in classifying breast cancer cases, assessing its performance, and comparing it to other classification methods.

Related Works

- **"Breast Cancer Diagnosis and Prognosis Using Machine Learning: A Survey"** (2019) reviewed various machine learning techniques, including logistic regression, for breast cancer diagnosis and prognosis, highlighting their strengths and limitations.
- **"Application of Logistic Regression in Medical Diagnosis: A Case Study of Breast Cancer"** (2020) explored the effectiveness of logistic regression models in medical diagnostics, focusing on breast cancer classification.
- **"Comparative Study of Classification Techniques for Breast Cancer Detection"** (2021) compared several classification algorithms, including logistic regression, to evaluate their performance in breast cancer detection.

Algorithm: Logistic Regression

Logistic regression is a statistical model used for binary classification. It estimates the probability that a given input belongs to a certain class using the logistic function.

Key Components:

- **Logistic Function**: The logistic function, or sigmoid function, maps any real-valued number into a value between 0 and 1, representing probabilities.

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Where zzz is a linear combination of the input features.

- **Model Equation**: The logistic regression model predicts the probability P(Y=1|X) using:

$$P(Y = 1|X) = \sigma(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n)$$

Where β0\beta_0β0 is the intercept and βi are the coefficients for each feature Xi.

- **Cost Function**: The cost function used to train the model is the binary cross-entropy loss, which measures the difference between predicted probabilities and actual outcomes.

Methodology

1. **Dataset Collection**:
    - The dataset used for this study includes diagnostic features of breast cancer cases. It is divided into training and test sets for model evaluation.
2. **Data Preprocessing**:
    - **Data Cleaning**: Handled missing values and removed irrelevant features.
    - **Feature Scaling**: Standardized features to ensure equal importance during model training.
3. **Model Training**:
    - **Logistic Regression Implementation**: The logistic regression model was trained on the training dataset using standard optimization techniques to find the best coefficients.
4. **Model Evaluation**:
    - **Accuracy**: Evaluated the model's performance using accuracy metrics on both training and test datasets.
    - **Confusion Matrix**: Analyzed true positives, true negatives, false positives, and false negatives to assess model performance.
5. **Performance Metrics**:
    - **Accuracy**: The proportion of correctly classified instances out of the total instances.
    - **Precision and Recall**: Measures of model performance related to false positives and false negatives.

Experimental Work

1. **Exploratory Data Analysis (EDA)**:
    - Conducted EDA to understand the dataset's structure, feature distributions, and relationships between variables.
2. **Model Training and Validation**:
    - Trained the logistic regression model on the training dataset and validated it using cross-validation techniques to ensure generalizability.
3. **Performance Evaluation**:
    - The model's performance was evaluated based on accuracy scores and other relevant metrics to gauge its effectiveness in classifying breast cancer cases.

Results

- **Training Accuracy**: 94.95%
- **Test Accuracy**: 92.98%
- **Confusion Matrix Analysis**: Provided insights into the model's strengths and weaknesses in detecting malignant and benign cases.

Conclusion

The logistic regression model demonstrated high accuracy in classifying breast cancer cases, both on training and test datasets. The results confirm the model's effectiveness and reliability in predicting breast cancer malignancy. Logistic regression, with its interpretability and efficiency, proves to be a valuable tool in medical diagnostics. Future work could explore ensemble methods and other advanced algorithms to further enhance classification performance.

References

1. **Breast Cancer Wisconsin (Diagnostic) Dataset**. (2018). UCI Machine Learning Repository.
2. **James, G., Witten, D., Hastie, T., & Tibshirani, R.** (2013). An Introduction to Statistical Learning: With Applications in R. Springer.
3. **Kuhn, M., & Johnson, K.** (2013). Applied Predictive Modeling. Springer.
4. **Iglewicz, B., & Hoaglin, D. C.** (2003). How to Detect and Handle Outliers. Springer.