Machine Problem 2: Logistic Regression

Introduction

In this experiment, we explored Logistic Regression, a commonly used supervised machine learning algorithm for binary classification tasks. The dataset used was the Breast Cancer dataset from scikit-learn, which contains medical data such as mean radius, texture, and smoothness of tumors. The goal is to predict whether a tumor is malignant (cancerous) or benign (non-cancerous) based on these features. Logistic Regression is ideal for this type of problem because it outputs probabilities and helps us understand how different features affect the likelihood of an outcome.

Through this task, we aimed to train a model, evaluate its accuracy, and visualize its performance using a confusion matrix and learning curve. These tools help assess how well the model can generalize to new data and whether it suffers from underfitting or overfitting. Overall, this experiment helped us understand how machine learning models can support real-world medical decision-making.

Methodology

The Breast Cancer dataset from scikit-learn was imported and analyzed to understand its structure and features. The dataset was then divided into training (80%) and testing (20%) sets to ensure fair model evaluation.

A Logistic Regression model was trained using the training data to learn the relationship between input features and the target class (malignant or benign). After training, the model's performance was tested using 5-fold cross-validation, which helps verify the model's consistency and prevents overfitting.

Next, a confusion matrix was generated to visualize how many predictions were correct or incorrect, giving us a better idea of the model's accuracy. Finally, a learning curve was plotted to show how the model's performance changes as the training data size increases, helping us determine whether the model is learning efficiently.

Results & Interpretation

Confusion Matrix:

The confusion matrix shows that the logistic regression model correctly classified most samples, with only a few mistakes. Most of the values fall on the diagonal, meaning the model accurately identifies whether a tumor is malignant or benign. This indicates strong predictive ability and reliability when applied to medical data.

Cross-Validation:

The model achieved an average accuracy of around 95%, with a low standard deviation across folds. This result means the model performs consistently even when the dataset is split differently each time. Such stability suggests that the model has learned the underlying patterns well and does not depend too heavily on any single data split.

Learning Curve:

The learning curve shows that both training and validation accuracies are high and converge closely together. This means the model is neither overfitting (memorizing training data) nor underfitting (failing

to learn). It demonstrates that the logistic regression model is well-trained and can generalize effectively to unseen data.

Discussion & Possible Improvements

Although the model performs very well, there are still ways to improve its accuracy and robustness. Adjusting the regularization parameter (C) can help balance simplicity and performance by controlling how much the model penalizes complex relationships. Using other solvers like lbfgs or saga may improve convergence speed and numerical stability.

It may also be useful to compare Logistic Regression with other models such as Decision Trees, Random Forests, or K-Nearest Neighbors (KNN) to check which performs best on this dataset. Additionally, performing feature scaling or feature selection could make the model even more efficient and reduce noise in the data.