Recursive Feature Elimination (RFE) with Linear Regression

Introduction

This document provides an overview of the implementation of Recursive Feature Elimination (RFE) using a Linear Regression model. The process involves iteratively removing the least important features to improve model interpretability and performance.

Dataset Description

The Diabetes dataset from `sklearn.datasets.load_diabetes()` is used. It contains:

- **Features**: 10 variables including age, sex, BMI, blood pressure, and six blood sample-derived measurements.
- **Target**: A continuous measure of diabetes progression one year after baseline.
- **Number of Instances**: 442
- **Number of Attributes**: 10 predictive features, one target variable

Implementation Steps

1. Data Loading and Exploration

- The dataset is loaded and converted into a Pandas DataFrame.
- Basic statistics and descriptions of features and target variable are printed.
- Data is split into training (80%) and testing (20%) sets.

2. Train Linear Regression Model

- A `LinearRegression` model is trained on the training dataset.
- The model's performance is evaluated using the R^2 score on the test set.
- **R^2 Score of Linear Regression Model**: 0.4526

3. Implement Recursive Feature Elimination (RFE)

- `RFE` is applied with `LinearRegression` as the estimator.
- The model iteratively eliminates features until only one remains.
- The R^2 score is recorded at each step.

- Feature importance rankings and coefficients are tracked.

4. Visualization

- A plot of the R^2 score vs. the number of retained features is generated.
- The optimal number of features is determined based on a significance threshold of 0.01 in R^2 improvement.
- **Optimal Number of Features Identified**: 2

5. Feature Importance Analysis

- A DataFrame is created to show the ranking and coefficient values at each iteration.
- The top three most important features are identified and analyzed:
    - **s1 (Serum Cholesterol Level)**: 931.49
    - **s5 (Log Serum Triglycerides Level)**: 736.20
    - **BMI (Body Mass Index)**: 542.43
- Initial and final feature rankings are compared:
    - **Initial Ranking**: BMI, s5, s1, s2, bp, sex, s4, s3, s6, age
    - **Final Selected Features**: All features retained after final iteration

Results

- The optimal number of features is identified as 2.
- The most significant features contributing to diabetes progression are determined as s1, s5, and BMI.
- The feature elimination process helps improve model interpretability without significant loss of predictive power.

Conclusion

Recursive Feature Elimination (RFE) effectively selects the most relevant features in a dataset while maintaining model accuracy. This technique enhances model simplicity and interpretability, making it useful for feature selection in regression-based models.