Assignment 4 report

Performance Comparison and Feature Analysis in Linear Regression Models

In this report, I compare the performance of unregularized (Linear Regression) and regularized (Ridge Regression and Lasso Regression) models for predicting disease progression using the diabetes dataset. I also interpret the coefficients of the models to identify the most important features for prediction. Additionally, I discuss the strengths and limitations of linear regression for this task.

1. Performance Comparison

I evaluate the performance of the models using Mean Squared Error (MSE) and R-squared metrics. The results are as follows:

- Linear Regression:
  - MSE: 0.45655844989286215
  - R-squared: 0.49648680873991846
- Ridge Regression:
  - MSE: 0.4538881699787352
  - R-squared: 0.49943171351921944
- Lasso Regression:
  - MSE: 0.9353291468042304
  - R-squared: -0.03152304747942458

Based on the MSE and R-squared values, both Linear Regression and Ridge Regression perform better than Lasso Regression. However, Ridge Regression slightly outperforms Linear Regression in terms of both metrics.

2. Interpretation of Coefficients

I analyze the coefficients of the models to determine the most important features for predicting disease progression. The coefficients represent the relationship between each feature and the target variable.

- Linear Regression Coefficients:
    - [0.01352522, -0.01787645, -0.13537619, 0.28798257, 0.2314548, -0.55763991, 0.28243652, 0.14633152, 0.21855619, 0.43763473, 0.07599883]
- Ridge Regression Coefficients:
    - [0.01347934, -0.01716758, -0.13364481, 0.28866895, 0.2294461, -0.42512131, 0.18070361, 0.08770086, 0.19869395, 0.38786468, 0.07691122]
- Lasso Regression Coefficients:
    - [4.85115384e-04, -4.19393335e-04, 2.72609196e-04, 2.01691563e-04, 4.17269671e-04, -1.58782240e-04, -2.65640671e-04, -3.34430014e-04, -8.71494306e-05, 4.99946716e-04, 6.08148304e-04]

Important Features:
Based on the absolute values of the coefficients, I identify the top 3 features for each model:

- Linear Regression Top 3 Features:
    - s2
    - s6
    - bp
- Ridge Regression Top 3 Features:
    - s2
    - s6
    - bp
- Lasso Regression Top 3 Features:
    - s6
    - age
    - sex

3. Strengths and Limitations of Linear Regression

Strengths:

- Simplicity: Linear regression is a straightforward and interpretable model that assumes a linear relationship between the features and the target variable.
- Speed: Training and prediction in linear regression are computationally efficient, making it suitable for large datasets.
- Interpretability: The coefficients provide insights into the importance and direction of each feature's impact on the target variable.

Limitations:

- Linearity Assumption: Linear regression assumes a linear relationship between the features and the target variable. If the relationship is nonlinear, linear regression may not capture it effectively.
- Limited Flexibility: Linear regression models are less flexible in capturing complex relationships compared to more advanced models like decision trees or neural networks.
- Sensitivity to Outliers: Linear regression models can be sensitive to outliers, which can disproportionately affect the model's performance and coefficient estimates.
- Multicollinearity: Linear regression can struggle with highly correlated features (multicollinearity), leading to unstable coefficient estimates and reduced interpretability.

In the context of predicting disease progression, linear regression can provide valuable insights into the importance of different features. However, its limitations in handling nonlinear relationships and complex interactions among variables might hinder its predictive performance. Regularized variants like Ridge and Lasso Regression aim to address some of these limitations by introducing regularization terms that help control overfitting and handle multicollinearity.

In my analysis, both Ridge and Lasso Regression models outperformed the basic Linear Regression model, indicating the benefits of regularization. Ridge Regression, in particular, demonstrated slightly better performance than Linear Regression, suggesting that it effectively handled multicollinearity in the dataset.

Overall, while linear regression provides a useful starting point for understanding the relationship between features and the target variable, more advanced models should be considered for improved predictive accuracy and flexibility when dealing with complex datasets or nonlinear relationships.