Polynomial Regression report

This report discusses the design of a polynomial regression model and the issue of overfitting. The model was developed from scratch and overfitted on the given dataset, which resulted in a large mean square error (MSE). To prevent overfitting, different techniques such as L1 norm, L2 norm, and a combination of both were applied. The best performing model was the Ridge regression model with an MSE of 9.724.

Introduction:

Polynomial regression is a powerful tool in data analysis and can be used to predict the outcome of a variable based on its relationship with other variables. However, the polynomial model can be overfitted on the data, leading to poor predictions on new data. This report presents a solution to the problem of overfitting by applying regularization techniques such as L1 and L2 norms.

Design of Polynomial Regression Model:

The polynomial regression model was developed from scratch by using the polynomial function to fit the dataset. The model was then overfitted on the given dataset by setting a high degree of 15, which resulted in a very large mean square error (MSE) of 1.8648754662337333e+37. This high MSE indicates that the model was too complex and overfitted on the data, which led to poor generalization on new data.

Preventing Overfitting:

To prevent overfitting, different regularization techniques were applied to the polynomial regression model. The L1 norm was applied to the model using scikit-learn's built-in function, sklearn.linear\_model.Lasso. The Lasso model had an MSE of 51.98. The L2 norm was applied to the model using sklearn.linear\_model.Ridge, which had the best performing model with an MSE of 9.724. An ensemble model comprising of the Lasso, Ridge, and LinearRegression with PolynomialFeatures was also developed, which had an MSE of 111.264.

Conclusion:

In conclusion, overfitting is a common problem in polynomial regression models, which can lead to poor predictions on new data. Regularization techniques such as L1 and L2 norms can be used to prevent overfitting. The Ridge model was found to be the best performing model with an MSE of 9.724.