Homework 3: Regularization

Stefanela Stevanović
MLDS1 23/24, FRI, UL
63220492

I. Introduction

Linear regression is a popular method for modeling relationships between variables. However, sometimes we encounter situations where the number of features in our dataset is large or the features are highly correlated. In such cases, traditional linear regression methods may result in overfitting or poor model performance. Ridge and Lasso regression are two methods used to address this issue by introducing a regularization term that penalizes the magnitude of the coefficients. In this homework, we implemented both Ridge and Lasso regression models as classes in Python. We used a closed-form solution for Ridge regression and the Powell method from scipy for Lasso regression. Finally, we applied these models to a dataset containing information about superconducting materials to predict their critical temperatures. We used the root mean square error (RMSE) as a performance metric of the models.

II. Methodology

Data pre-processing. The data used in this project consists of 300 observations and 82 features, and it aims to predict the critical temperatures of different superconducting materials. We standardized the feature data to have a mean of zero and a standard deviation of one. Standardization is used to ensure that all features are on the same scale, which helps prevent some features from having a larger influence on the model than others. This is particularly important when using regularized linear models like Ridge and Lasso regression, which are sensitive to the scale of the features. Finally, we divided the dataset into training and test sets, using the first 200 observations as training data and the remaining 100 observations as test data. The training data was used to fit the models, while the test data was used to evaluate their performance.

Ridge Regression algorithm. Ridge Regression model was implemented as a class that takes a regularization weight as its parameter and provides methods fit(X, y) and predict(X). In the fit(X, y) method, a column of ones was added to the feature matrix in order to compute an intercept value. In order to avoid intercept penalization I[0] was set to 0 and the Ridge Regression formula was solved:

$$\beta = (X.T * X + *I)^{-}1 * X.T * y$$

where X is the feature matrix, y is the target variable, is the vector of coefficients, is the regularization weight, I is the identity matrix, and * denotes matrix multiplication. The predict(X) method takes a feature matrix X as input, adds a column of ones, and returns the predicted target values using the learned coefficients: y_pred = X * β.

Lasso Regression algorithm. Lasso Regression model was implemented as a class that takes a regularization weight as its parameter and provides methods fit(X, y) and predict(X). The fit method uses the Powell optimization algorithm to find the optimal coefficients that minimize the loss function. The loss function is defined as the sum of squared errors (SSE) between the predicted values and actual values, with an added L1 penalty term on the magnitude of the coefficients. The predict method computes the predicted values of the dependent variable using the estimated coefficients and intercept.

Determining the regularization weight λ. The regularization weight or parameter is denoted by lambda (λ) and controls the amount of shrinkage of the coefficients towards zero. A higher value of lambda results in stronger regularization. In order to determine the optimal lambda the 10-fold cross-validation was performed on the training set, with different regularization weights, and the average RMSE across 10-folds was computed. The regularization weight that resulted in lowest RMSE was chosen as the optimal parameter for the ridge regression. However, with Lasso regression, increasing lambda causes some coefficients to became equal to zero, thus simplifying the model. In this case, the "one standard error away" rule was followed when choosing the optimal λ. The rule suggest choosing the simplest model (highest λ) that is no more than one standard error away from minimum RMSE. In order to find this λ, the range of acceptable RMSE scores was defined and the highest lambda that has the RMSE score within the defined range was chosen as the optimal parameter.

Model evaluation. Root Mean Squared Error (RMSE), which is is the square root of the average of the squared differences between the predicted and actual value, was used for the performance evaluation of the regression models. Lower RMSE values indicate that the model is better at predicting the target variable.

III. Results and Discussion

In order to determine the optimal λ for the ridge regression, we ran the 10-fold cross-validation on the training set with lambda values in range from 0.001 to 100. We used the mean RMSE across 10 folds as the performance metrics. The results are shown in Fig.1.

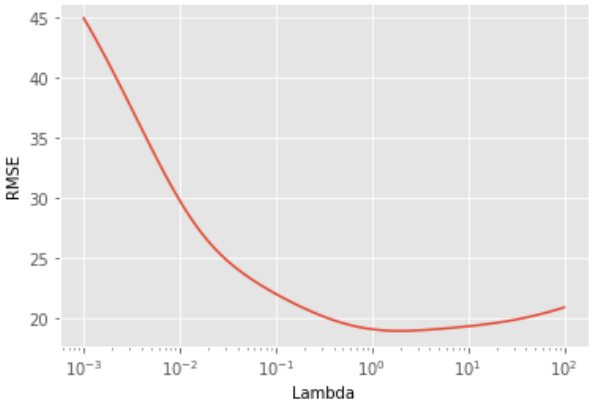

Fig. 1. Ridge regression mean RMSE across 10-folds for different λ values

From the Fig.1 we see that the RMSE first decreases with larger λ as the model becomes less complex and overfitting is reduced, but then at certain point starts to increase as the model becomes too simple and underfitting occurs. The lowest RMSE of 18.9 is reached with the regularization weight of 1.96, which is chosen as the optimal λ parameter for ridge regression.

For the comparison purposes, the optimal λ was also found for the Lasso regression with 10-fold cross-validation on the training set. The experiment was arranged with the λ values from 0 to 1000 with the step size of 10 and RMSE was computed. Fig.2 and Fig.3 show how the RMSE and complexity of the model change with the different regularization weights.

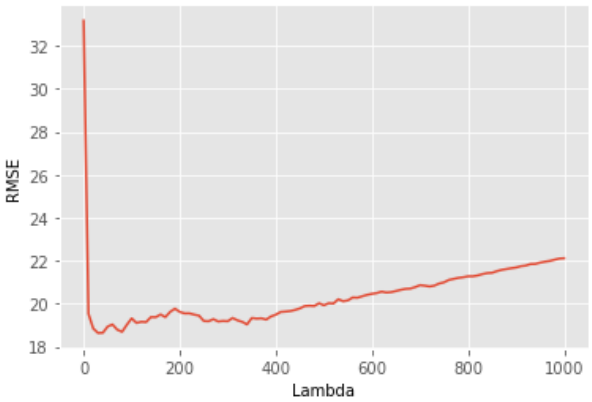

Fig. 2. Lasso regression mean RMSE across 10-folds for different λ values

From the Fig.3 we can see that as the λ increases, more and more coefficients are shrunk to zero, so the number of non-zero coefficients (the complexity of the model) decreases, resulting

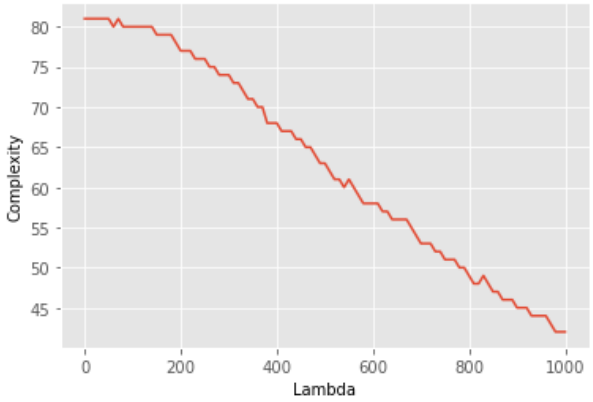

Fig. 3. Lasso regression model complexity for different λ values

in a simpler and more interpretable model. The lowest RMSE of 18.6 on the training set was reached with the λ value of 30, having 81 non-zero coefficients. However, according to "one standard error away" rule, the λ value of 550 was chosen as the optimal parameter, with the RMSE of 20.2 and 60 non-zero coefficients.

Ridge and Lasso regression models with chosen optimal λ were tested on the test set. The linear regression model without regularization was also implemented for the comparison purpose. The performance of these regression models is shown in the table 1.

TABLE I
Performance comparison of models on the test set

Model	RMSE
Basic Linear Regression	40.7
Ridge Regression	20.1
Lasso Regression	14.8

From the table 1 we can see that with the Ridge regression, by adding a regularization term that penalizes large coefficient values, we can achieve better performance on the test set than with the basic linear regression without regularization. The Lasso Regression model has the lowest RMSE of 14.8 among the three models, indicating the best performance on the test set. This suggests that the Lasso Regression model is better at capturing the underlying patterns in the data compared to the other models. This could be because the Lasso Regression model uses L1 regularization which can lead to feature selection and sparsity, allowing the model to focus on the most important features and reducing the impact of irrelevant or noisy features in the data.