Regularization

Matej Miočić, 63180206, mm9520@student.uni-lj.si

I. Introduction

In this homework we were tasked with implementation of ridge and lasso regression. We showed our work on the provided superconductivity dataset. Our goal was to minimize the model's root mean square error (RMSE) with the selection of an appropriate regularization weight.

II. Models

A. Ridge regression

Ridge regression (also L2 regularization) is a method of estimating the coefficients where linearly independent variables are highly correlated. We use a closed form solution to compute coefficients,

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

In (1) we do not want to penalize the intercept coefficient, so we set the first element of the identity matrix to 0, since this is where we perform penalization with λ as our penalization parameter.

B. Lasso regression

Lasso regression (also L1 regularization) performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability. We do not have a closed form so we use optimization technique,

$$\beta = argmin(\|\beta^T x_i - y_i\|_2^2 + \lambda\|\beta\|_1). \tag{2}$$

Just like before, we do not want to penalize the intercept coefficient in (2), so we take the norm of β without the first element.

III. Experiments

A. Performance

We were given a dataset with 300 rows and 82 columns. We have a relatively small dataset for this many features, so we expect our penalizing parameter to be high. We selected first 200 instances as training set and the other 100 for the test set. First we standardized the features, so the bigger features would not get penalized more. We used only the information from training set and applied it to the testing set. To obtain the optimal regularization weight (λ) we minimized RMSE on the test set.

Table I
Optimal regularization weight and obtained RMSE for the test set.

	Regularization weight (λ)	RMSE
Ridge	99.64	14.76
Lasso	1.35	14.46

In Table I we show the optimal regularization weight and obtained RMSE for the test set. The problem is, this weight is overfit on the test data which we should not use for training.

To tackle this problem we use 5 fold cross validation on the training set. Since we were getting different weights for each fold, we repeated this process using different folds each time. We decided to repeat 5 fold cross validation 10 times, which is enough to get some insight how our parameter changes and does not allow for too high change for similar folds to appear in each step.

For the 5th fold (we will refer to it as validation set) we once again obtain the optimal regularization weight with minimizing RMSE. We tried different starting points for both methods. For each iteration we obtain 5 optimal parameters for that validation set. We show mean, median and their standard deviations to quantify uncertainty for each method.

Table II
Medium and mean with standard deviation of optimal regularization weight for the corresponding validation set. RMSE is obtained for the test set. In parenthesis we show a starting point for searching for optimal lambda.

	λ median	λ mean	RMSE
Ridge (0)	1.86	9.18 ± 16.25	20.48 ± 4.52
Ridge (1)	2.41	10.56 ± 18.22	20.26 ± 4.53
Ridge (5)	3.80	15.76 ± 30.22	19.54 ± 4.18
Lasso (1)	0.57	0.64 ± 0.71	16.19 ± 1.77

In Table II we show our results for ridge and lasso regression. We show that starting point for minimizing RMSE for validation set changes the outputs for Ridge regression since the function converges differently based on the starting point. After starting point 5, we did not notice big changes. For Lasso regression we did not obtain different results for various starting points. It is worth mentioning that λ cannot go below 0. We see that lasso regression performs much better when obtaining optimal parameter with cross validation. It also came close to the optimal solution.

IV. Conclusion

We learned how to implement ridge and lasso regression. We showed that lasso regression performs better when searching for the parameter with only training set at the cost of computation speed. Meanwhile both methods recieved similar optimal results when overfit on testing data.