Comparison Between Linear and Ridge Regression to Predict Brain Connectivity

Neslihan Çekiç
Faculty of Computer and Informatics Engineering
Istanbul Technical University
Istanbul/Turkey
Email: cekic16@itu.edu.tr

İlgin Balkan
Faculty of Computer and Informatics Engineering
Istanbul Technical University
Istanbul/Turkey
Email: balkani17@itu.edu.tr

Merve Candan
Faculty of Computer and Informatics Engineering
Istanbul Technical University
Istanbul/Turkey
Email: candan16@itu.edu.tr

Buket Akgün
Faculty of Computer and Informatics Engineering
Istanbul Technical University
Istanbul/Turkey
Email: akgunb16@itu.edu.tr

I. Introduction

This paper contains details of the project developed to predict the evolution of connectivity between brain cells over time. To solve this problem, we made 2 regression models, compared the results, and determined the model that fits the data better. In the rest of the paper, the way in which the data is processed, which methods are used, what results are obtained from the project and comparison between linear and ridge regression for this problem will be shared.

Team name: 150160041_150160060_150170901_150160044

Kaggle Names: mervecandan35,neslihaneki,lginbalkan, buketakgun

Final Score: 0.00198

Public Leaderboard Rank: 4

II. Preparation of Dataset

A. Normalizing Data

To achieve better results when training a model, it is necessary to scale and normalize the dataset, as linear regression performs better on scaled and normalized data. Therefore, sklearn.thestandartscaler was used to change the data distribution and sklearn.maxabssaler was used to scale the data.[1]

B. Feature Selection

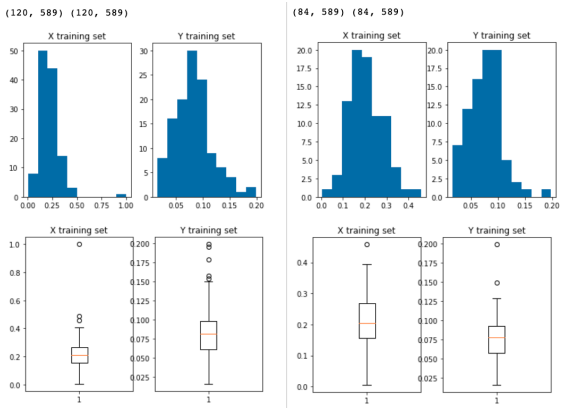
Feature selection is the process of reducing the number of input variables while developing a predictive model. It is preferable to reduce the number of input variables to both reduce the computational cost of the modeling and, in some cases, to improve the performance of the model.

Variance threshold from sklearn.feature_selection library was used while making feature selection. This algorithm selects and removes features with low variance that do not meet a threshold.[1]

Using this function, the feature number has been reduced from 595 to 589.

C. Removing Outliers

Because the linear model is very influenced by outliers, the outliers have been removed from the dataset with Isolation-Forest in sklearn. iForest isolate anomalies in dataset that are both few in number and different in the feature space.[1]



(a) (b)

Fig. 1: (a) Feature 1 (b) Feature 1 After Outliers Removal

Using this function, 120 datas of X train and y train was reduced to 84 datas. It was observed that the outliers were erased to a good extent by plotting the histogram and boxplot of the ”f1” feature.

III. Used Methods

A. Linear Regression

Linear regression, which is a fairly simple regression method, creates a best fit line by deducting the relationship between a dependent variable and one or more independent variables. Assuming that all features in the dataset are independent from each other, the model was created by regression for each feature separately.

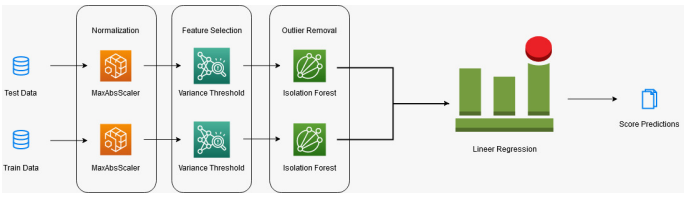


Fig. 2: Learning Pipeline

Since linear regression is a simple model, it is highly affected by outliers and because of we dealing with multiple

features, multicollinearity (existence of near-linear relationships among the independent variables) is also affect the model.[2] Although the outlier problem was solved while preparing the data, Ridge Regression was also implemented to test the multicollinearity condition.

B. Ridge Regression

Ridge regression is a method of model tuning used to analyze any data suffering from multicollinearity. This method performs L2 regulation. When the multicollinearity problem arises, the least squares are neutral and if the variances are large, this causes the predicted values to be far from the true values.[3]

1) GridSearchCV: GridSearchCV in sklearn library used for tuning hyperparameters of ridge regression.[1] It finds optimum parameters to execute an algorithm. To determine cross-validation splitting strategy, repeated 5-fold cross-validation is used in GridSearchCV.[1]

IV. Results and Conclusion

Since we used 2 different regression methods which are Ridge Regression and Linear Regression when predict model, we calculated MSE for both. With Linear Regression, we regressed each feature, so a better result came out. For the MSE calculation, we took the weighted average of the predict values.

The MSE values we found for both methods are as follows:

- MSE for Linear Regression: 0.003859185866312879
- MSE for Ridge Regression: 0.004846453803806332

A. k-fold Cross Validation Results

We measure the brain connectivity which spaced out by 6 months, and we used 5-fold cross validation to test our methods.

```
Ridge 1 - fold MSE: 0.0027508091477733768 MAE: 0.03842905446562381
Linear 1 - fold MSE: 0.0024439728780453153 MAE: 0.035249612694487654

Ridge 2 - fold MSE: 0.004620197521767629 MAE: 0.04249809194344948
Linear 2 - fold MSE: 0.004452496764841197 MAE: 0.0397526135239887

Ridge 3 - fold MSE: 0.004218108158703233 MAE: 0.039577782157840205
Linear 3 - fold MSE: 0.004126789866215357 MAE: 0.0373597480950762

Ridge 4 - fold MSE: 0.006674935597915899 MAE: 0.04009550333341659
Linear 4 - fold MSE: 0.006154153728732497 MAE: 0.037123103332970246

Ridge 5 - fold MSE: 0.0033978251032277507 MAE: 0.03831945765847103
Linear 5 - fold MSE: 0.003045128196368309 MAE: 0.03632246912742595
```

Fig. 3: Mean Square Error and Mean Absolute Error results for CV

B. Correlation Matrices

The correlation matrix displays the correlation coefficients (each cell in the table contains the correlation coefficient) for different variables and shows the correlation between all the possible pairs of values in a table. [4]

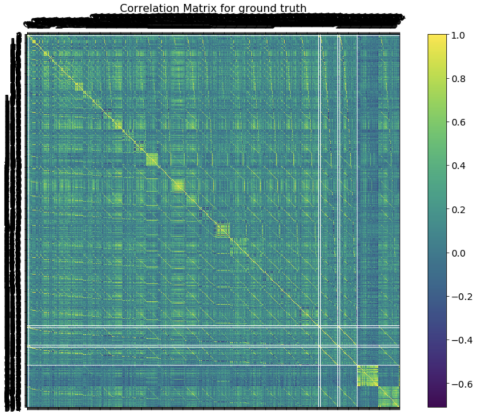
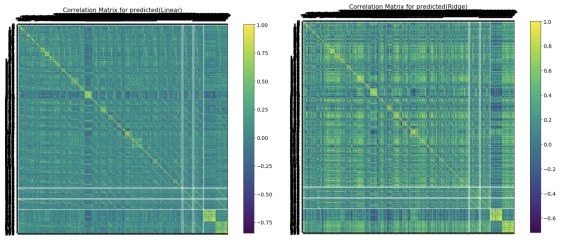


Fig. 4: Correlation Matrix for Ground Truth

Correlation measures the direction and strength of the relationship of two numerical variables while linear regression relates it through an equation. When we look at the correlation matrix of the data estimated by 2 regression models, we observe that the correlations between variables increase in the ridge regression. Therefore, we can say that the predicted results of the Ridge model are higher than linear as we calculated.



(a) (b)

Fig. 5: (a) Correlation Matrix for Linear Predicted (b) Correlation Matrix for Ridge Predicted

C. Kaggle Score and Ranking

Final Score: 0.00198
Public Leaderboard Rank: 4

References

[1] Pedregosa et al., *Scikit-learn: Machine Learning in Python*. JMLR 12, pp. 2825-2830, 2011.

[2] T. Gohmann, *NCSS Statistical Software "Multiple Regression"* . vol. 305, pp. 1-10, http://www.ncss.com/.

[3] T. Gohmann, *NCSS Statistical Software "Ridge Regression"* . vol. 335, pp. 1-25, http://www.ncss.com/.

[4] https://corporatefinanceinstitute.com/resources/excel/study/correlation-matrix/