BLG454E Learning From Data
Term Project

Batuhan Özdöl 150180701 - Ezgi Alçiçek 150160032
Yusuf Utku Gül 150150006 - Onur Yavuz Ergut 040100126

June 13, 2020

1 Introduction

Brain connectivity refers to a pattern of anatomical links ("anatomical connectivity"), of statistical dependencies ("functional connectivity") or of causal interactions ("effective connectivity") between distinct units within a nervous system. The units correspond to individual neurons, neuronal populations, or anatomically segregated brain regions. Neural activity, and by extension neural codes, are constrained by connectivity. Brain connectivity is thus crucial to elucidating how neurons and neural networks process information [1].
In this challenge of the course which is the term project, it is expected to apply the tools of machine learning to predict the brain connectivity by creating the best fit model and predict the next point results of the test set using that model. The main goal of this project is to predict brain connectivity features at t1 from the same features measured at the previous time point t0.In this project, a solution is proposed using Sample Selection, Pairwise Correlation, Ordinary Least Squares, RidgeCV, and K-Fold. The given training datasets contain 150 subjects and 595 features. Performance evaluation for this project has been determined by the Kaggle competition. Our team name is 040100126_150150006_150160032_150180701, our team is in the 5th place with the score 0.00199.

2 Datasets

There is a given data set to train our defined model, regarding that ach brain is encoded in symmetric connectivity matrix X $\in R^{(35x35)}$, where an element X(i, j) denotes the strength of the connectivity between two brain regions i and j. It was vectorized by the off-diagonal upper triangular part of X, to generate a feature vector X $\in R^{(1xd)}$ that represents a sample of the brain where d is equal to 595. By stacking the samples vectors vertically across N=150 subjects, the data matrix X $\in R^{(Nxd)}$ was constructed.
Finally, the given training datasets contain 150 subjects and 595 features which corresponds to a matrix $\in R^{(150x595)}$. The brain connectivity is measured at two different time points t0 and t1 provide to make tracks in the brain as a network.

2.1 Preprocessing

First, to avoid column names in input training and test data sets, ID and our feature columns are got separated by dropping the ID column. After that, the pairwise correlation is applied to the training datasets. With this approach, the aim is to drop the features with high correlation since they are more linearly dependent and hence have almost the same effect on the t1 results.
Also, Ordinary Least Squares Regression is used to provide a good basis for learning more advanced concepts and techniques in our backward elimination function. Backward elimination is a feature selection technique while building a machine learning model. It is used to remove those features that do not have a significant effect on the dependent variable or prediction of output [2]. By using this function, necessary features are selected that affect our prediction accuracy.
Finally, Outlier samples are found in the dataset which is an observation that lies an abnormal distance from other values. It has been evaluated one by one via applying generalized cross-validation and determined by the min error value. Min error value is determined through analyzing the data. To prevent the deviation of our regression curve, outlier samples are dropped to reach a reliable curve before using our model.

3 Methods

RidgeCV method is selected as our training and test (prediction) model because, in Ridge regression, you can tune the lambda parameter so that model coefficients change. The lambda value that minimizes MSE was supposed to be selected as the final model [3].
RidgeCV method is used with negative mean square error as scoring, and the parameter corresponds to normalization is set as True. The scoring parameter is selected regarding the given datasets and the problem description. Alphas parameter is created considering that the RidgeCv method chooses the best alpha value for the model in the given alphas set, as alphas parameter. CV parameter of the RidgeCV is left as default which corresponds to generalized cross-validation [4].
In the end, 5-Fold Cross-Validation is used to compute our Mean Squared Error (MSE), by taking one fold for test and others for training by using our model. Corresponding pipeline that illustrating the key steps of the proposed solution that have been explained in preprocessing and methods sections can be seen below in Figure 1.

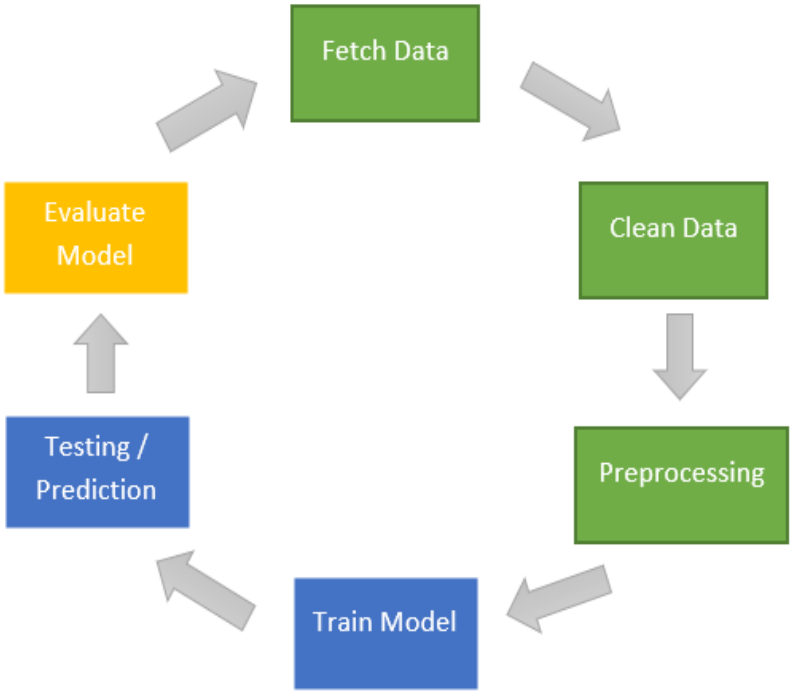

Figure 1: Main figure represents learning pipeline

4 Results and Conclusions

The results are evaluated by the MSE as expected. By applying the 5 Fold Cross Validation method, the average error determined by the model is around 0.00207% for the training datasets. To see the performance of our model on test data, predictions of test data are submitted to Kaggle in the defined format. Kaggle calculated and ranked the submission scores using the public test data throughout the competition by looking at Mean Squared Error (MSE).
In conclusion, it can be said that dimension reduction through feature selection and sample selection using pairwise correlation and backward elimination have enormous effects over the results. Through this project, the importance of the feature and sample selection methods is obtained. Our team name is 040100126_150150006_150160032_150180701, our team is in the 5th place with the score 0.00199.

5 References

[1] Sporns, O., Indiana University, Scholarpedia, 10.4249/scholarpedia.4695.
[2] https://www.javatpoint.com/backward-elimination-in-machine-learning
[3] Hoerl, A.E. and Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12: 55-67
[4] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.RidgeCV.html