Bayesian Ridge Regression to Predict the Brain Connectivity

Melih Kağan Özçelik
dept. of Computer Engineering
Istanbul Technical University
Istanbul, Turkey
ozcelikm16@itu.edu.tr

Gözde Bayar
dept. of Economics
Istanbul Technical University
Istanbul, Turkey
bayarg17@itu.edu.tr

Furkan Uysal
dept. of Computer Engineering
Istanbul Technical University
Istanbul, Turkey
uysalf15@itu.edu.tr

Berfin Tutku Doğan
dept. of Economics
Istanbul Technical University
Istanbul, Turkey
doganber16@itu.edu.tr

Çağan Kiper
dept. of Computer Engineering
Istanbul Technical University
Istanbul, Turkey
kiper19@itu.edu.tr

***Abstract*—This paper focuses on forming a machine learning model that predicts the brain connectivity for the next time point. On this purpose, we used Bayesian Ridge Regression algorithm. When the training data part is over, we tested our predictive model and evaluated it with mean squared error and mean absolute distance metrics. Then, we explained our evaluation results with plots and tables.**

***Index Terms*—machine learning, regression, bayesian ridge, feature selection, local outlier factor**

I. Introduction

This project deals with forming a machine learning model to predict the brain connectivity at the next time point using the given connectivity matrix. For given an elderly population, the 35 by 35 symmetric connectivity matrix is constructed using the strength of the connectivity between to brain region. 35 regions of each brain were taken into account. The gap between the two time points for this project is 6 months.

This project is the term project for the BLG454E-Learning From Data course at Istanbul Technical University in Spring-2021 semester. There is also a Kaggle competition for this project. Our final score is 0.00256. Our team name in kaggle is "150160050_150150044_070170461_070160422_150190706".

II. Datasets

A. Given dataset, train and test

In our dataset, we have the connectivity matrices in vectorized version which is $X \in \mathrm{R}^{35\times35}$. In this matrices, each element $\mathrm{X}(i,j)$ signifies the strength of the connectivity between a region in the brain and another(i,j). To obtain a feature vector $x \in \mathrm{R}^{1\times d}$, we vectorize the off-diagonal upper triangular part of X. 'd' is the number of relations between regions and we can calculate it as $\frac{34\times35}{2}$ = 595.

Train and test datasets were uploaded to Kaggle. Initially, 150 samples for both t_0 and t_1 exists for the training test. For the testing step the test data matrix $D \in \mathbb{R}^{80\times596}$ for timestamp t_0 is used to predict timestamp t_1. Predicted matrix for timestamp t_1 is submitted to Kaggle after melting.

B. Data preprocessing

In order to select features to be used in our model, we created 35×35 symmetric connectivity matrices from the feature vectors by processing them and for all samples, we compared changes in different features by using vectorized version. Feature dropping is used as it can be seen in the correlation figure. We removed the features with a standard deviation of 0 from the model such as feature 4. Additionally, local outlier factor is used to eliminate outliers data points.

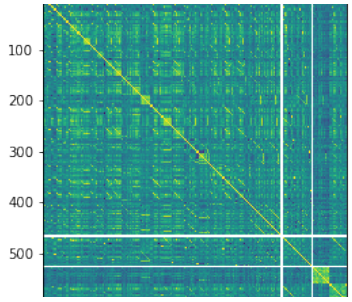

Fig. 1. Pairwise correlation matrix of all columns in the training dataframe at t_0

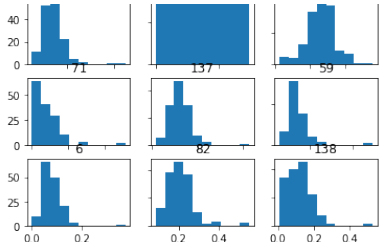

Fig. 2. Histogram to show same valued features along the samples

III. Method

In order to choose the most suitable model to given train and test data, we have tried different regression models as follows: Linear Regression, Ridge Regression, Stochastic Gradient Descent, Bayesian Ridge, Huber Regression, Lasso, Bagging Regression, ElasticNet, Random Forest Regression, Ada Boost Regression, Support Vector Regression.

	0	1	2	3	4	Var	Mean
LinearRegression	0.003221	0.004635	0.006969	0.002299	0.003247	0.000003	0.004074
Ridge	0.003199	0.004622	0.006966	0.002286	0.003236	0.000003	0.004062
SGDRegressor	0.003541	0.004919	0.007542	0.002536	0.003358	0.000004	0.004379
BayesianRidge	0.002730	0.004353	0.006959	0.002007	0.002978	0.000004	0.003805
HuberRegressor	0.002852	0.004553	0.006838	0.002093	0.002895	0.000004	0.003846
Lasso	0.003206	0.004625	0.006964	0.002287	0.003238	0.000003	0.004064
BaggingRegressor	0.003072	0.004698	0.007705	0.002347	0.004473	0.000004	0.004459
ElasticNet	0.002891	0.004846	0.007701	0.002388	0.003248	0.000005	0.004215
RandomForestRegressor	0.003065	0.004910	0.008030	0.003282	0.003986	0.000004	0.004655
AdaBoostRegressor	0.002751	0.004577	0.007286	0.002140	0.002983	0.000004	0.003948
SVR(kernel="linear")	0.004184	0.005788	0.008691	0.003685	0.004711	0.000004	0.005412
SVR(kernel="rbf")	0.004334	0.006074	0.009110	0.004105	0.005196	0.000004	0.005764

Fig. 3. Calculated Mean Squared Errors of 5-Fold CV

As can be seen in the Figure 3, the model with the lowest mean value is the Bayesian Ridge.

A. Bayesian Ridge Model

Bayesian is essentially the means through which statistical models are defined and estimated. If we have inadequate data in the dataset or the data are poorly distributed, Bayesian regression may be quite helpful.The objective of Bayesian Linear Regression is not to identify model parameters but rather to discover the posterior distribution.The output y is not the only one to be assumed to come from the distribution, but also the parameters of model. Bayesian Ridge Regression, which measures the probabilistic model of the regression issue, is one of the most helpful sort of Bayesian regression. [1]Probabilistic models can be predicted by Bayesian Ridge.For instance, the weight determined is slightly different than that determined on ordinary least squares due to the Bayesian context. However, the Regression of Bayesian Ridge is more resilient in the face of ill posed problems. [2]

IV. Results and Conclusions

We evaluated our model with mean squared error(Table 1) and Pearson correlation coefficient (Figure 4) evaluation metrics on training set using 5 fold cross validation.

TABLE I
Results of the model on the training set using 5-CV

Fold	MSE
1	0.002730
2	0.004353
3	0.006959
4	0.002007
5	0.002978

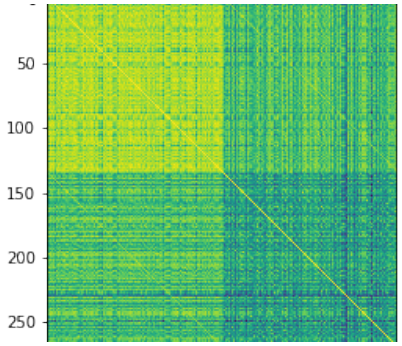

Fig. 4. Pearson correlation between predictions and ground truth.

After evaluating our model using cross validation, we chose the number of neighbours $k = 10$ for our Bayesian Ridge regression model. Training the model with this parameters and testing it on the given test set resulted in an MSE of 0.00256 and 11th ranking on Kaggle.

References

[1] Available: https://www.geeksforgeeks.org/implementation-of-bayesian-regression/
[2] Available: https://scikit-learn.org/stable/modules/linear_model.html#bayesian-ridge-regression

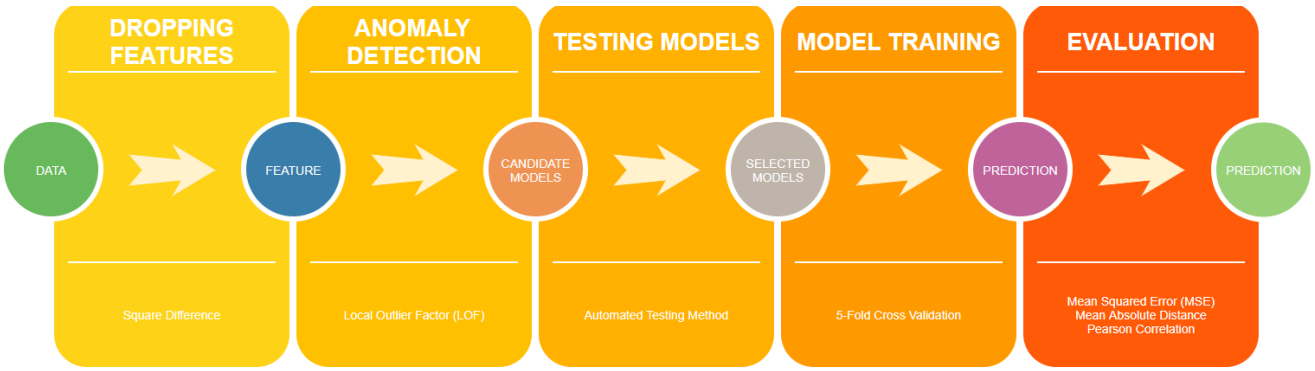

Fig. 5. Learning Pipeline