BLG454E Learning From Data Term Project 2021 Fall

Cihat Akkiraz, Ramazan Yetişmiş, Ayşe Betül Çetin, Ayberk Bozkuş, Serhat Vural

Abstract: In this project, the best results were tried to be obtained by using various machine learning methods with the measured results regarding the brain connections of an elderly population. The model created with the training data sets is intended to predict the brain connection for the next time point. The best success in this regard has been achieved with the custom multi output Bayesian ridge method.

Index Terms: Machine Learning, Brain Connectivity

1. Introduction

In this project, we have trained a model to predict the future of the brain connectivity by using the data of previous time point. This is because we can observe changes in the brain by measuring the brain connection at two different time points, 6 months apart, at t0 and t1. The data used in project is driven from brain morphology and this is a many to many regression problem.
Kaggle competition information:
Final score: 0.00188
Rank: 1
Team name: 150180704_150190708_ 1501800730_150140009_150160067

2. Datasets

Outliers are serious source of problems. In preprocessing part, to remove outliers, LocalOutlierFactor used. LocalOutlierFactor is a function of sklearn library. It extracts samples from both t0 and t1. While doing this, the local density of a sample to the local densities of its neighbors compered. If a sample have important lower density than we consider that as an outlier.

3. Methods

In addition, we used python with libraries random, numpy, sklearn and csv.
CSV: Used for reading the data sets provided
Numpy and Pandas: Used to create train and test variables.
Sklearn: Briefly used for model building and training
Random: Used to declare seed of Cross Validation as 1

For training, we have used Bayesian Ridge Regressor. In Bayesian regression, regularization parameters estimated from the data. In other words, the parameters are also part of the estimation.

A list of learning model is trained for each feature by using train data set. For doing this work easier seperate custom regressor is created using BaseEstimator and RegressorMixin. For each feature seperate model is created via fit function and output is created for each feature via predict function.

GridSearchCV is a function of SK-learn (modelselection package). By using GridSearchCV, hyperparameter tuning have done automatically for finding the optimal values.

Learning pipeline is as shown in Figure 1.

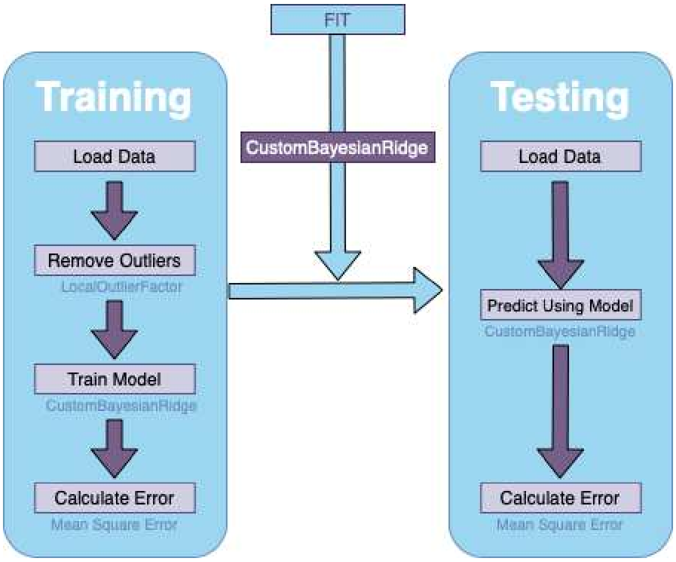
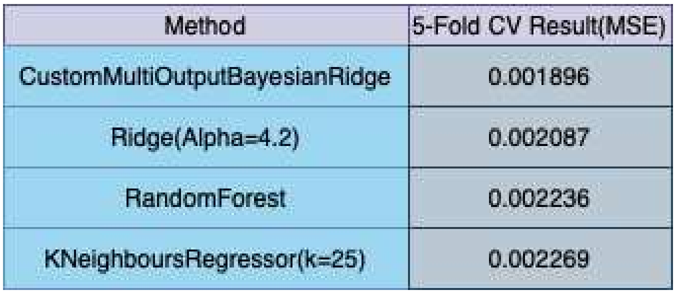


Fig. 1. Pipeline of Proposed Solution

4. Results and Conclusions

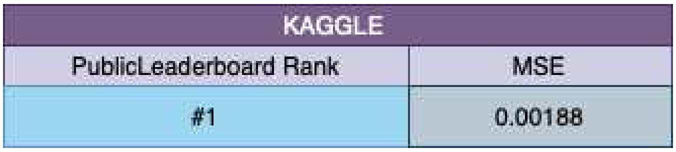
As a result, the methods shown in Figure 2 were tried from the test and train data. The results regarding the methods are next to the methods. The main method was designated as custom multi output bayesian ridge because the best results were obtained with it. As seen in Figure 2, this is the method where the smallest MSE value is obtained.

Kaggle score and ranking of the team within the scope of the project is as shown in Figure 3.



Method	5-Fold CV Result(MSE)
CustomMultiOutputBayesianRidge	0.001896
Ridge(Alpha=4.2)	0.002087
RandomForest	0.002236
KNeighboursRegressor(k=25)	0.002269

Fig. 2. Results of Different Methods



KAGGLE	
PublicLeaderboard Rank	MSE
#1	0.00188

Fig. 3. Kaggle Results