MachineLearning

MUSTAFA SARIGÜL - 20160602112

December 24, 2019

Abstract

How to find the best regression model for six variables which relation unknown? I analyzed with four model which are Linear Regression, Random Forest Regression, Decision Tree Regression and Lasso Regression with two methods that are score of regression given library and Mean Square Error with using cross-validation. I worked with my default set.

1 Introduction

If we want to get the best regression model about unknown relationship in any data set, we need to analyze on different model. Before the analysing, I create two algorithms with using specific libraries that one of them for Linear Regression and Random Forest and one of them for Lasso Regression and Decision Tree Regression. If we get result that gives us satisfaction,we can say that this model might be the best model for just given data set.In detail, in our default data set, we have six variables that are "x1,x2,x3,x4,x5,x6". They all can be a predictor but also someone of these variables might be useless, which can be a negative effect on our regression models, to predict "Y" value. After each testing, we calculate of scores of models or mean square error that are with cross-validation or without cross-validation for finding more effective variables in the data set and for the satisfied result. On the based,We have a different approximation for every regression model which have analyzed.

2 Linear Regression

Simple linear regression is not efficient for our data set but we testing for each variable makes sense for the understanding data set. What does say us the data set and its variables. They could help us to decide. Simple linear regression says that some part of variables test,their predicted values do not closed values.As result of this section, one variable never be satisfied for the data set. satisfied result. On the based,We have a different approximation for every regression model which have analyzed.

3 Random Forest Regression

The random forest regression more complicated than linear regression, but it gives you results which are closed perfect prediction and has lower error estimations.In random forest regression, creates some pattern for yourself with decisions, but these decisions always are changing and tries the all decisions so that it provides give the result more satisfied. As my works, I analysed these difference that is not bigger, but I could observed this difference.

4 Decision Tree Regression

Decision tree regression, works as Random Forest Regression on based, but the main difference is categorising the outputs as learned during learning.Decision tree regression, categorises the result by grouping and gives the result groups which are created yourself,when new data comes. It provides first check chance to regression so that it affects the own results to increase error estimations and to give more rational results. As my works,decision tree regression looks the best model which are talked. I observed this idea with mean square error with cross-validation as well as without cross-validation in my outputs which are attached end of my report.

5 Lasso Regression

Lasso regression is an advanced version of linear regression, but unlike the simple linear regression, lasso regression is eliminating least important features. This eliminating called features selection. It provides that these features does not affect your result. As my researches and works I observed results which are more then all methods which are talked about and I tried. In my outputs of lasso regression satisfied me and I stopped my researching of best model and decided the lasso regression enough for my data set.

6 Implementation

During my researches,I used more than libraries. Some of them are regression libraries some of them specific basic Python libraries such as numpy,pandas and sklearn to implemented in almost every code.

6.1 Linear Regression

I used pandas and sklearn.linear_model libraries to implement the linear regression code. In additional I used score sub-function which is belongs to RandomForestRegressor (in sklearn.linear_model library) to understand the efficiency of the regressions.

6.2 Random Forest Regression

I used pandas and sklearn.ensemble libraries to implement the random forest regression code. In additional I used score sub-function which is belongs to LinearRegression (in sklearn.ensemble library) to understand the efficiency of the regressions.

6.3 Decision Tree Regression

I used pandas,numpy and sklearn libraries to implement the linear decision tree code.I used train_set_split from sklearn.model_selection to split my variables. I used DesicionTreeRegressor from sklearn.tree to create regressor for m model. In additional I used make_scorer and mean_squared_error sub-functions which are derived the sklearn.metric and cross_val_score which is derived sklearn.model_selection to understand the efficiency of the regressions.

6.4 Lasso Regression

I used pandas, numpy and sklearn libraries to implement the lasso code.I used train_set_split from sklearn.model_selection to split my variables. I additional I used LassoCV from sklearn.linear_model and cross_val_score to understand the efficiency of the regressions.

7 Results

As I am not expert, my results are not totally scientific for this project and researches. I know that I did not find more efficient model, but there can be more efficient my models. I find four models and applied to my data set which is given by my assistant. After applying, I observed two table of model is more useful that are lasso regression and decision tree regression.I decided the lasso regression more useful and efficient for my data set because it gives me at least error estimations.

8 Conclusion

In this project, I saw that there is competition about machine and statistical learning. They tries to answer some question such as "What is best regression model ?","Which classification is model more useful ?". I solved some these questions such that "Which libraries needs you during overcoming the models ?". This project is also showed me different regression can give you similar results but, we can not say that they do not have differences. However I gained experience in machine learning and statistical learning and developed my skills related in this field. Lastly, I met some websites and platforms, StackExchange is most useful.

8.1 Output

Lasso regression is my decision of best model for my data set and me. My lasso output is below:

```
MSE with CV: 955.2094513564064
MSE without CV: 984.4684357244176
####################################
My predictions in LASSO Regression:
          x1  x2  x3  x4  x5  x6  Y
SampleNo
101       50  22  17   0  -8  22 [[ 885.18202316]
102       15   3  -5  94 -15   3  [-253.45736331]
103       35  31  14  77   0  31  [1482.8795692 ]
104        8  20  10  92  -7  20  [1611.72581895]
105       43  18   5  35   6  18  [2292.51794169]
106        5  17   9  35  -4  17  [1089.67541948]
107       45  16  18  25   5  16  [1685.42343348]
108       34  25  -1  35 -11  25  [1056.63602407]
109       43   5   9  17  17   5  [  52.13980135]
110       45  19   8  12  -7  19  [ 692.70659558]
111       20  31  18   3 -11  31  [1601.63052672]
112       12  45  11  19  -4  45  [-167.48403844]
113       46   2  -4   4   5   2  [ 976.21415085]
114       37  14  12  62   8  14  [1322.78722956]
115        4   9  -7  29  -2   9  [1514.27383109]
116        8  39  18  86 -16  39  [ 805.92745722]
117        6  48  13  96   1  48  [1618.55649791]
118       34  46   6  58   5  46  [ 763.95160725]
119        9  37  18   4 -20  37  [1361.03171423]
120       27  38  19  82  -7  38  [ 496.13835686]]
####################################
```

Figure 1: My Lasso Output

Note: All outputs were attached project RAR files.