CE 475 Machine Learning Project Report

Created by Mehmet Aydın KICIRTI

20160602105

Specification of the problem

I am taking a dataset which contains 120 values on the columns names x1 to x6 and also another column Y. The aim is to estimate Y values as accurately as possible for last undefined 20 data points. First of all, I thought which X variables were more suitable for predicting Y values. Then I have to find a good model to predict new Y values.

Methodology

First of all, When I was defining our data's. I noticed that some data point repeats again. After according to my research, I found Feature Selection to estimate our data points which columns gave us the best estimations and also reduces training time. When I use these eliminations methods. There are more eliminations techniques such as Pipeline, Tree-Based Feature Selection, and Recursive Feature Elimination. I preferred Recursive Feature Elimination because it provides either for small scale dataset or easily find repeated data and this process is applied until in dataset are exhausted. Then, By using this algorithym I got below these results;

```
Which column give best accuracy sorted order=>  [(1, 'x1'), (2, 'x3'), (3, 'x2'), (4, 'x5'), (5, 'x6'), (6, 'x4')]
```

According to these results, I tried to check really is it true or not. Then, I noticed while I was trying x1,x3,x2,x5 columns, it gave us the best estimations. Therefore I configurated again our data points by these results. After I already decided on the regression algorithm, I started to implement each of these. Firstly, I created my train and test data using train_test_split(). Then, while using these each regression technique, fitted our train data. Then, I started to make assumptions like with the r2_score, mean_absolute_error, and mean_squared_error values of my data that I fit and predicted together with the methods provided by the Sklearn metrics library. Finally, according to the results I obtained after calculating the accuracy of these models, the random forest had the best accuracy. After these trials, I also applied cross-validation to select which model I should use. He showed me that the random forest was the best choice in cross-validation accuracy. After these studies, I found unknown data using the random forest to estimate the last 20 y value.

Implementation

I used Python language when I estimating data and also used the following libraries in my project.

- Pandas
- Numpy
- Sklearn.ensemble
- Sklearn.tree

- Sklearn.linear_models
- Sklearn.preprocessing
- Sklearn.model_selection
- Sklearn.feature_selection
- Sklearn (imported only metrics to measure classification performance)

Results

After I predicted our data's, we need to measure performance between each regression models, so I used that can be able to show better performance using sklearn library metrics. Therefore, this library can easily find their mean square error, root mean square error,r-square score and also cross-validation score.

Linear Regression :

Mean Absolute Error Linear Regression : 1335.6996534782027

MSE for Linear Regression : 3086026.0433522174

Root MSE for Linear Regression : 1756.7088669874179

R2 Score for Linear Regression : 0.42761685900599433

CVScore for Linear Regression : -0.28280572254618835

Polynomial Regression :

Mean Absolute Error Polynomial Regression : 1031.1934726289433

MSE for Polynomial Regression : 2223966.4875362637

Root MSE for Polynomial Regression : 1491.296914613674

R2 Score for Polynomial Regression : 0.5875080424730796

CVScore for Polynomial Regression : -0.28280572254618835

Lasso Regression :

Mean Absolute Error Lasso Regression : 1335.6991836797101

MSE for Lasso Regression : 3086024.71925884

Root MSE for Lasso Regression : 1756.7084901197582

R2 Score for Lasso Regression : 0.42761710459326907

CVScore for Lasso Regression : -0.2828048293556982

Random Forest Regression :

Mean Absolute Error Random Forest Regression : 401.6337500000001

MSE for Random Forest Regression : 439676.07734375016

Root MSE for Random Forest Regression : 663.0807472274777

R2 Score for Random Forest Regression : 0.9184507289845913

CVScore for Random Forest Regression : 0.6161659602723354

Decision Tree Regression :

Mean Absolute Error Decision Tree : 1366.605

MSE for Decision Tree : 3443231.813523669

Root MSE for Decision Tree : 1855.594733104098

R2 Score for Decision Tree : 0.3613638339700086

CVScore for Decision Tree : 0.04876002056381431

As you can see below the picture as the predicted y values.

Sample No	x1	x2	x3	x4	x5	x6	Y
101	40	17	31	60	17	31	436,1
102	25	11	11	41	11	6	1288,7
103	34	-1	16	20	-1	15	4445,475
104	29	-6	10	14	-6	42	2555,4
105	4	4	8	95	4	29	2793,025
106	21	19	4	38	19	32	323,875
107	13	-12	6	44	-12	42	830,7
108	35	11	30	47	11	13	1485,05
109	40	9	35	86	9	44	397,75
110	20	-10	28	20	-10	11	708,125
111	18	16	20	81	16	21	362,05
112	33	16	14	16	16	3	218,7
113	30	-12	8	17	-12	36	3386,025
114	11	6	35	67	6	10	1763,25
115	37	-19	3	71	-19	15	4343,2
116	30	-17	34	83	-17	4	57,3
117	0	17	19	95	17	36	686,875
118	2	-7	24	77	-7	9	62,375
119	3	13	8	40	13	35	8624,5
120	5	9	21	13	9	49	4434,85

As seen above from the results, I got the best score from Random Forest Regression with smaller errors than previous models, so the best accuracy belongs to Random Forest.

Conclusion

As a result, In my Project, I have used more approaches from our laboratory courses. While I was implementing the project, I had the opportunity to use the information already learned in the lesson and also I noticed that our laboratory homeworks are mainly covering what I have used. Based on what I learned, I narrowed down the regression methods that I will use based on their characteristics and also the Sklearn library documentation provided guidance on how to proceed for me. Then I tried to use different regression methods such as Linear, Random Forest, Lasso regression after that random forest regression gave us good prediction values and also the best accuracy in that. By the way, Polynomial regression also gave low error, there can be more pluggable. Therefore, I chosen Random Forest to predict the data’s.