Project 4: Compare regression methods

Understanding and comparing several regression algorithms

I have designed a single module to perform all the regression functions named main.py that would take regression functions name and datapath for the dataset file as the input. If housing dataset i.e housing.data.txt is passed as the datapath, then we need to include all the features of the dataset wheras in the case of the California Renewable Production 2010-2018 dataset, I have ignored the first column which is a timestamp.

I have also replaced the NAN values in the second dataset with 0. The dataset values are standarized and then splitted to training and testing dataset. The slope, intercept, mean square errors, r^2 errors and time taken are printed as the ouputs for the program except for normal equation solution and non linear regression.

For the regression functions, we can pass Linear, RANSAC, Ridge, Lasso, Nonlinear and Normal as the arguments.

Analysis

Each dataset is passed through the regression functions except normal equation, for which only the housing dataset is tested. I have used Decision Tree Regressor as the approach to conduct non-linear regression.
The tabular form for accuracy and time taken by each classifiers are:

i. Housing dataset(housing.data.txt)

Functions	MSE train	MSE test	R^2 train	R^2 test	Time taken
Linear Regression	0.236	0.322	0.765	0.673	0.0699
RANSAC Regressor	0.236	0.322	0.765	0.673	0.071
Ridge	0.236	0.323	0.765	0.673	0.0701
Lasso	1.004	0.992	0	-0.006	0.068
DecisionTreeRegressor	0.149	0.333	0.851	0.663	0.0078

This dataset has 506 rows of data and here, Linear Regression, RANSAC regressor and Ridge typically show similar prediction and have similar Mean Squared Errors and R^2 scores. Lasso performs better in terms of time taken than these three and the best seems to be the Decision Tree Regressor.

If we look at the train and test MSE, test MSE is greater than that of train. It means that the models are overfitting. The fastest regression was with Decision Tree Regressor and RANSAC took the longest time among all.

ii. the California Renewable Production 2010-2018 (all_breakdown.csv)

Functions	MSE train	MSE test	R^2 train	R^2 test	Time taken
Linear Regression	0.882	0.888	0.116	0.115	0.2926
RANSACRegressor	0.882	0.888	0.116	0.115	0.2822
Ridge	0.882	0.888	0.116	0.115	0.2953
Lasso	0.892	1.004	0	0	0.2754
DecisionTreeRegressor	0.892	0.9	0.106	0.104	0.2549

This dataset has 67584 rows and is very large than the housing dataset. The Linear Regression, RANSAC Regressor and Ridge showed same pattern here with similar values of MSE and R^2 score.

If we look at the MSE here, it has increased in test dataset as compared to the train dataset which indicates that the models are overfitting. The fastest regression algorithm was Decision Tree Regressor and RANSAC took the longest time among all.

By comparing the MSE and time taken from both the dataset we can choose Decision Tree Regressor, a non linear regression model or any from first three for the Linear regression model.

Normal Equation Solution

Normal equation solution is an approch to do machine learning without the use of scikit learn functions. We use linear algebra and matrix inverse to find the w and then use that to predict the target. For the normal equation solution in housing dataset, I used the training dataset X_train and y_train. It gave MSE value as .236 which is equal to the MSE train value in Linear Regression, RANSAC Regressor and Ridge. This proves that it is possible to perform regression without the use of machine learning functions defined in python.

Conclusion

This project helped me in understanding and comparing several regression algorithms. It helped me differentiate between classification algorithms and regression algorithms. We use MSE and R^2 scores in regression to evaluate the models. The models are better with low MSE and high R^2 score. By testing the two datasets, it can be concluded that Decision Tree Regressor is the best of all.