REPORT:

Programming Assignment – 1

Modelling slump flow of concrete using MLE, Ridge and LASSO regression

- **MACHINE LEARNING ENVIORMENT USED :**

The machine learning environment used here is the Jupyter Notebook from the Anaconda Navigator GUI platform. The language used here is Python 3.

First, I started off by creating a new iPynb Notebook and importing all the necessary libraries from the **sklearns** library in python. After the environment variables were set up, I created a .csv file at the root directory of my Anaconda Command prompt, for all of the slump data provided in the project description.

Further, I started the implementation by reading the necessary data (i.e. the exploratory variables and the response variable) from the csv by using the package "pandas" which was used to manipulate and access certain columns from the data set. After this, I used the Lasso and Ridge functions as the fitting models to return a specific model as per requirement (in each of the tasks of the project) and then used this model to test and predict the data for the set of 18 remaining data values.

While doing these operations many data structures and methods were used such as:

- **train_test_split** : For randomly selecting, splitting and testing the 4x1 data sets and performing cross-validation for the trained models.
- **numpy :** It was used to store the data values in the form of an array and then using a numpy variable to perform functions like finding mean squared errror,average, max, min, std deviation, etc.
- **Linear Models :** Lasso, LassoCV, Ridge, RidgeCV, linear regression,etc were imported from the sklearns library.
- **Matplotlib :** This library was used to plot the 3 graphs in the project scatter plot and the regularization path line graphs.

After deriving the minimum value of alpha from the above models we used that specific model to test on the remaining 18 values so as to achieve the optimized prediction and fit.

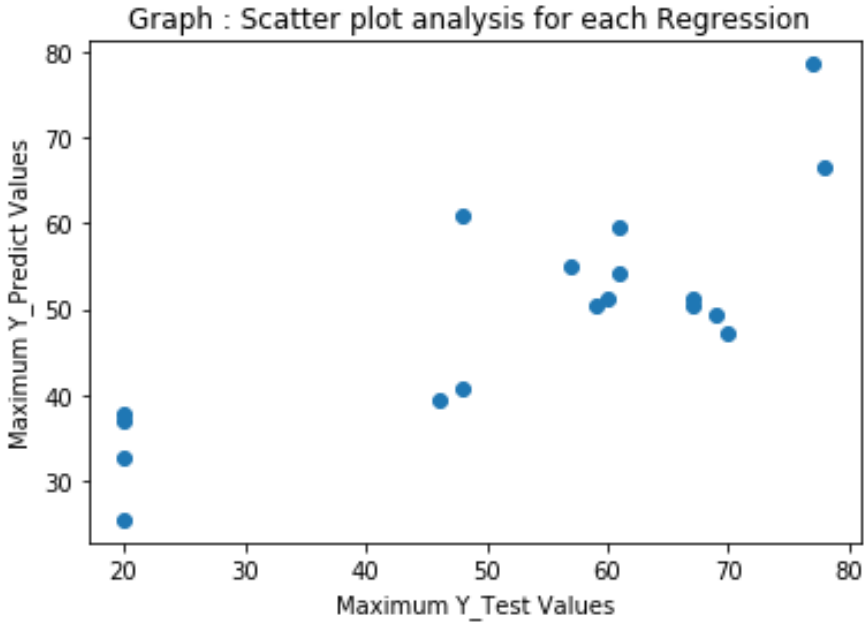
TASK 1: Inferences from the MSE test values

In order to determine the best model out of the three classes of linear regression, I performed 5 fold cross validation and noted down the MSE values (average and standard deviation). This process was repeated 10 times where for every iteration there was a different value of the testing and training data. The table below shows the comparison between the MSE values for each of the tasks 1.1, 1.2 and 1.3.

Type of Regression	Average MSE for 10 iterations	Standard Deviation of MSE Values	Number of variables used
Un-regularized Regression	185.176	43.682	----
Ridge Regression	183.401	39.925	3
Lasso Regression	165.49	19.157	7

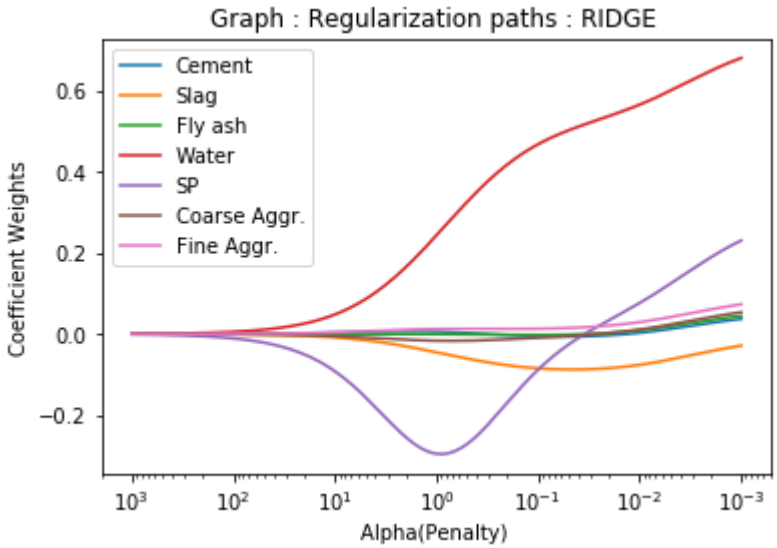
From the above data we can make the inference that Mean Squared Error (MSE) values for the Un-regularized and Ridge regression are almost the same. While in the case of Lasso Regression the average value of MSE is less than the other two regressions. Therefore, we can say that Lasso gives us the best performance out of the three. This could be due to the fact that Lasso takes into account all the 7 explanatory variables, while Ridge regression takes only 3 and Un-regularized takes none.

Here, a similar process was followed except that the r2.score function was used to get the R^2 values and then using the maximum R^2 values index to get the corresponding y_predict and y_test values. These were required for the following plot.

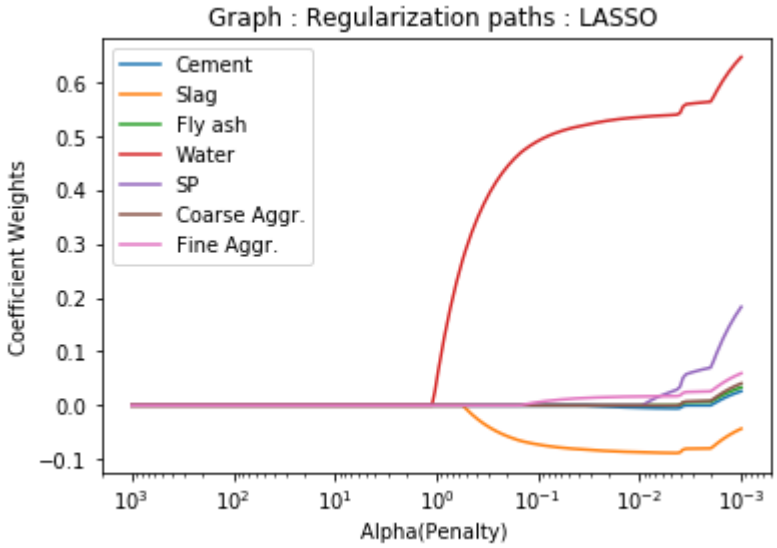


TASK 2:

The process was similar to the one in Task 0 was used to implement the Ridge regression model and the following curve was observed when the Alpha values were plotted against the coefficient weights of the explanatory variables.



Here the Lasso model was used to implement the data relations and the following curve was observed when the Alpha values were plotted against the coefficient weights of the explanatory variables.



Libraries Used :

Sklearns, glmnet_py, numpy, pandas, matplotlib,etc

Time Dedicated :
A total of 3 weeks were dedicated to the project including the problem statement understanding and the implementation part.

Languages used :

Python 3 was the only language used

SUMMARY :
After running the tests for the 10 MSE values, I found that the Ridge regression and the Unregularized regression had similar error values whereas the Lasso regression had a much lesser MSE and hence was the most accurate model.

Collaborators:

Shubham Badola, Utsav Mathur, Prajin Jonchhe and Dhairay Desai

References:

http://scikit-learn.org/stable/auto_examples/linear_model/plot_ridge_path.html

http://nbviewer.jupyter.org/github/ubdsgroup/ubmlcourse/blob/master/notebooks/LinearSystems.ipynb

http://scikit-learn.org/stable/

https://drsimonj.svbtle.com/ridge-regression-with-glmnet

http://openclassroom.stanford.edu/MainFolder/DocumentPage.php?course=MachineLearning&doc=exercises/ex5/ex5.html