Name - Sagar Suman

Roll No. 2019197

Linear Regression on Abalone DataSet from scratch-

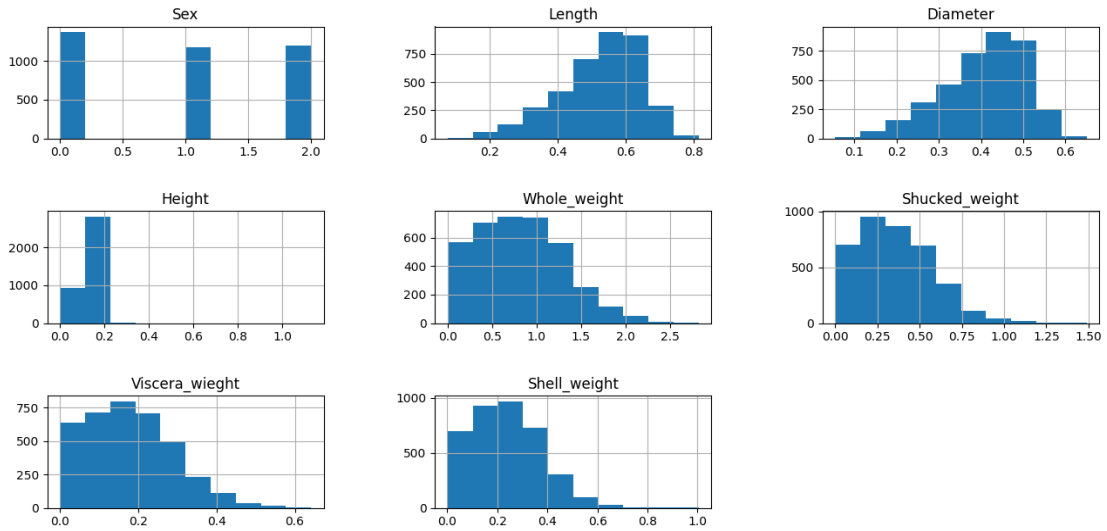
PROBLEM STATEMENT –

Predicting the age of abalone from physical measurements. The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope -- a boring and time-consuming task. Other measurements, which are easier to obtain, are used to predict the age. Further information, such as weather patterns and location (hence food availability) may be required to solve the problem. So we will be implementing Simple linear Regression, Ridge Regression and Lasso Regression from scratch. Further using K—fold for getting approximate of true error rate.

Data Set - http://www.cs.toronto.edu/~delve/data/abalone/desc.html

There are 9 columns in dataset. We firstly add the column names in Dataset.data file. We used columns names mentioned in Dataset.spec file. Then using pandas, we have read the dataset. We are using first 8 columns as input variables (X) to predict the last column (y). Out of 8 input features, 1st feature is of type string and rest 7 features are of floating data type. 1st column represents sex, which have only three classes {M,F,I}. We converted this column to float by mapping - M to 0, F to 1 and I to 2.

After this, we have splitted our data set into 90% (for training + validation)[X_train , y_train] and 10% (for testing)[X_test, y_test] using scikit-learn. We then visualize various attributes of X_train and we get the following graphs -



We can see some of attributes are not scaled properly, so it requires normalization.

Gradient descent -

We have defined function for gradient descent in which we have implemented its functionality from scratch. We have also customized this function to take into account for L1 and L2 regularization.

Part a) -

Now, let's move to first part -

1. Firstly, we have used KFold implementation of scikit-learn to do 5 splits.

2. For each of 5 splits, we are using 4 splits for training and 1 spilt as validation set.

3. Now, we will perform following for each of val set -

    3.1. In training set we are performing normalization, using formula: X -min/max

    3.2. We then store the min, max value of above.

    3.3. We initialize our parameters as 9x1 vector equal to 1.(We have also added x0 = 1 in X_train).

    3.4. We decide some learning rate and no. of iterations and perform gradient descent to minimize parameter.

    3.5. We get some parameters from gradient descent.

    3.6. We normalize validation set using values of step 3.2. and formula 3.1.

    3.7. We calculate the RMSE on validation set and see the result.

We use step 3 to tune the hyperparametes, to get minimum avg RMSE on val sets.

Finally, we come to conclusion that

learning rate = 0.2

and iterations = 200.

For above iteration vs RMSE graph is as follows -

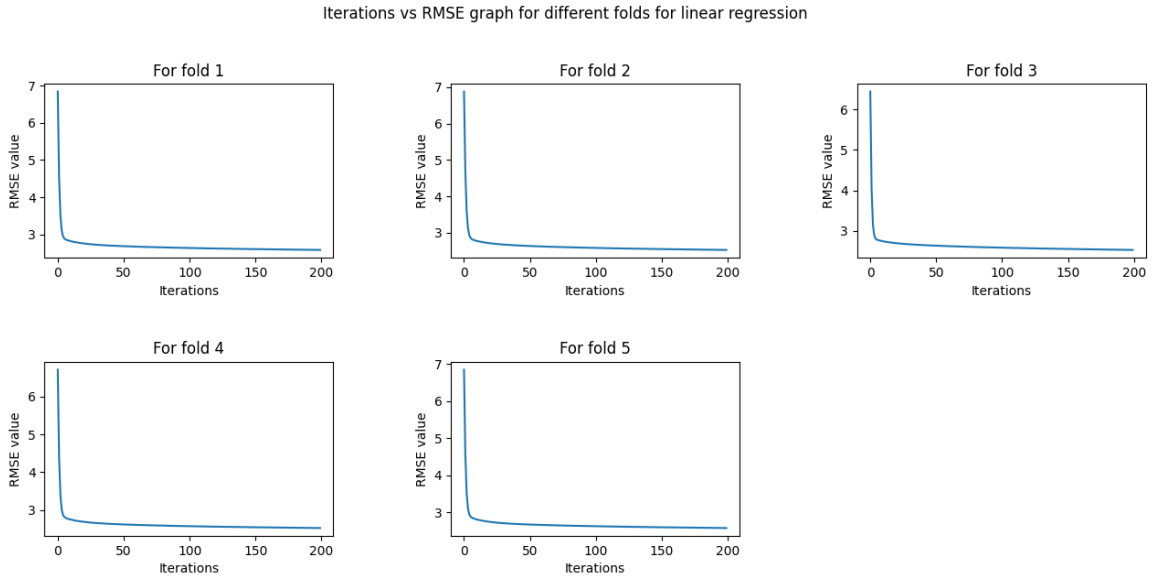


Fig. Iteration vs RMSE graph for linear regression for different folds

RMSE values on validation set is as follows -

```
File  Edit  Tabs  Help

For Linear regression, RMSE value on fold 1 validation set is: 2.4521359054929794
For Linear regression, RMSE value on fold 2 validation set is: 2.668021658014849
For Linear regression, RMSE value on fold 3 validation set is: 2.4802293373337196
For Linear regression, RMSE value on fold 4 validation set is: 2.7190512349404257
For Linear regression, RMSE value on fold 5 validation set is: 2.4931281210550225
For Linear regression, Average RMSE value of all folds is: 2.562513251367399

For Linear regression with L1 reg, RMSE value on fold 1 val set is: 2.4589203631612833
For Linear regression with L1 reg, RMSE value on fold 2 val set is: 2.649870100864001
For Linear regression with L1 reg, RMSE value on fold 3 val set is: 2.4861357765910856
For Linear regression with L1 reg, RMSE value on fold 4 val set is: 2.7292018038649593
For Linear regression with L1 reg, RMSE value on fold 5 val set is: 2.506713934491885
For Linear regression with L1 reg, Average RMSE value of all folds is: 2.566168395794643

For Linear regression with L2 reg, RMSE value on fold 1 val set is: 2.394285845818699
For Linear regression with L2 reg, RMSE value on fold 2 val set is: 2.497212356480969
For Linear regression with L2 reg, RMSE value on fold 3 val set is: 2.4663989187174384
For Linear regression with L2 reg, RMSE value on fold 4 val set is: 2.6015284384688138
For Linear regression with L2 reg, RMSE value on fold 5 val set is: 2.398323613901638
For Linear regression with L2 reg, Average RMSE value of all folds is: 2.4715498346775115

For Only Linear regression, RMSE value on testing set is: 2.711478525392836
For Linear regression + L1, RMSE value on testing set is: 2.7143461458954548
For Linear regression + L2, RMSE value on testing set is: 2.5777440746170472

By using sklearn Linear Regression on fold 1 we get RMSE: 2.127930147301517
By using sklearn Linear Regression on fold 2 we get RMSE: 2.265360779050746
By using sklearn Linear Regression on fold 3 we get RMSE: 2.199253201290417
By using sklearn Linear Regression on fold 4 we get RMSE: 2.305151064795437
By using sklearn Linear Regression on fold 5 we get RMSE: 2.150131620247328
By using sklearn Linear Regression, AVERAGE RMSE on all folds: 2.2095653625370892

By using sklearn Lasso Regression(L1) on fold 1 we get RMSE: 2.1836499705586228
By using sklearn Lasso Regression(L1) on fold 2 we get RMSE: 2.3250455922086832
By using sklearn Lasso Regression(L1) on fold 3 we get RMSE: 2.307863340415771
By using sklearn Lasso Regression(L1) on fold 4 we get RMSE: 2.4186447050465065
By using sklearn Lasso Regression(L1) on fold 5 we get RMSE: 2.2210484398250925
By using sklearn Lasso Regression(L1), AVERAGE RMSE on all folds: 2.291250409610935
```

Fig. Follow first five lines, which corresponds to Linear regression

Part b) -

In this we follow the same procedure as part a. In step 3.4 use gradient descent of L1 and L2 respectively.

At the end of tuning, we come to conclusion that

For L1 -

learning rate = 0.2

iterations = 200

regularization parameter(lambda) = 0.01

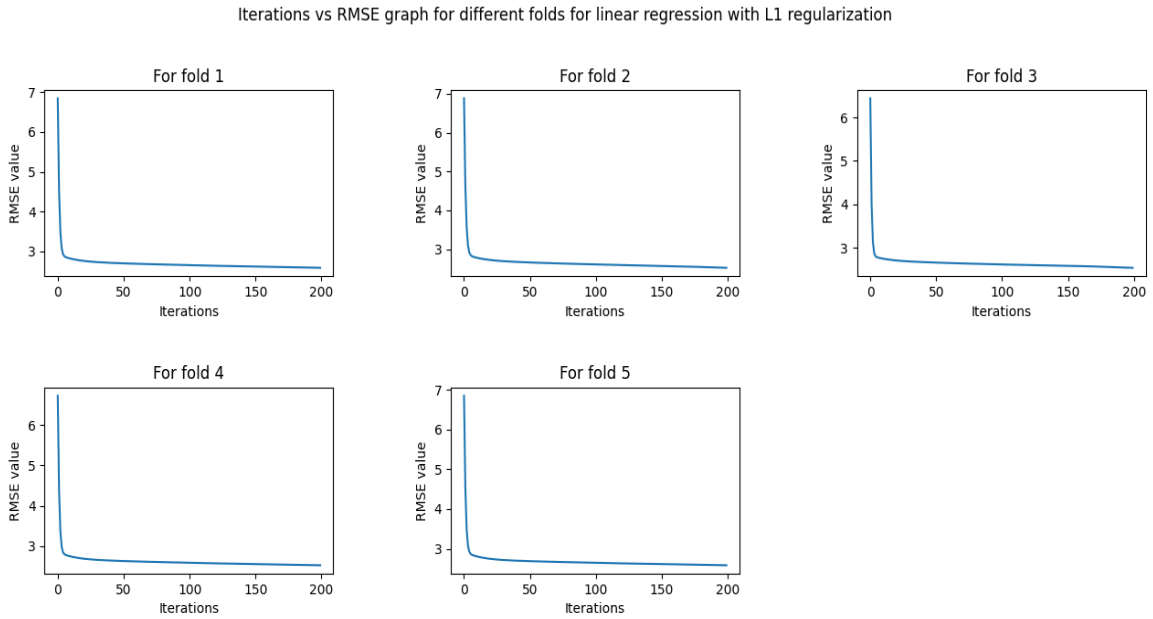
For L2 -

learning rate = 0.2

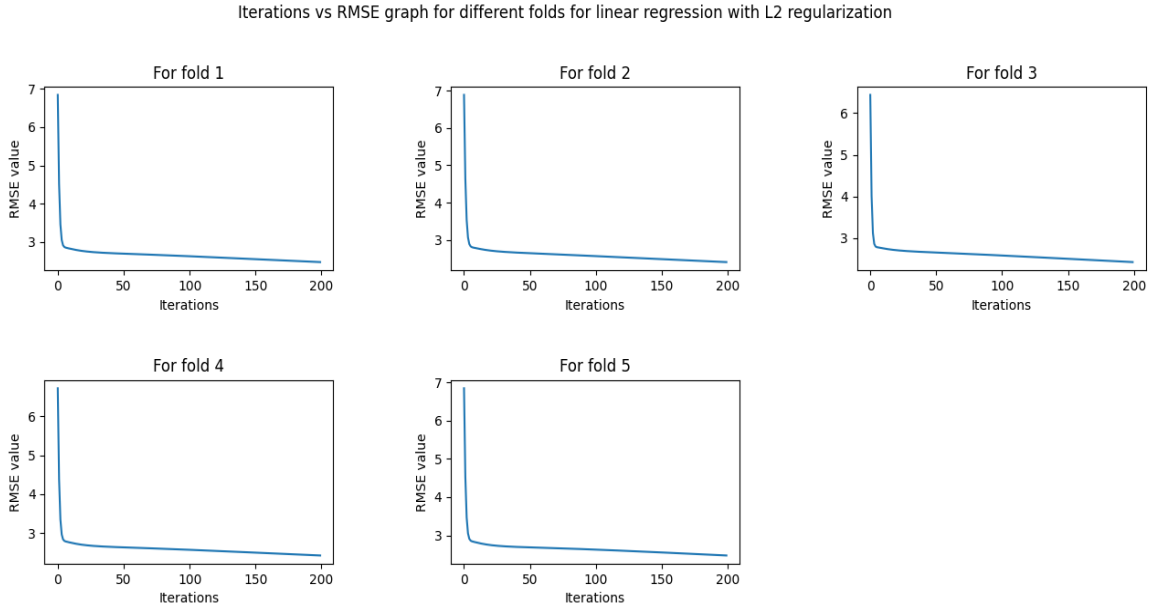
iteration = 200

regularization parameter(lambda) = 0.005

For L1 iteration vs RMSE graph is as follows -



For L2 iteration vs RMSE graph is as follows -



RMSE values on validation set is as follows -

```
File  Edit  Tabs  Help
For Linear regression, RMSE value on fold 1 validation set is: 2.4521359054929794
For Linear regression, RMSE value on fold 2 validation set is: 2.668021658014849
For Linear regression, RMSE value on fold 3 validation set is: 2.4802293373337196
For Linear regression, RMSE value on fold 4 validation set is: 2.7190512349404257
For Linear regression, RMSE value on fold 5 validation set is: 2.4931281210550225
For Linear regression, Average RMSE value of all folds is: 2.562513251367399

For Linear regression with L1 reg, RMSE value on fold 1 val set is: 2.4589203631612833
For Linear regression with L1 reg, RMSE value on fold 2 val set is: 2.649870100864001
For Linear regression with L1 reg, RMSE value on fold 3 val set is: 2.4861357765910856
For Linear regression with L1 reg, RMSE value on fold 4 val set is: 2.7292018038649593
For Linear regression with L1 reg, RMSE value on fold 5 val set is: 2.506713934491885
For Linear regression with L1 reg, Average RMSE value of all folds is: 2.566168395794643

For Linear regression with L2 reg, RMSE value on fold 1 val set is: 2.394285845818699
For Linear regression with L2 reg, RMSE value on fold 2 val set is: 2.497212356480969
For Linear regression with L2 reg, RMSE value on fold 3 val set is: 2.4663989187174384
For Linear regression with L2 reg, RMSE value on fold 4 val set is: 2.6015284384688138
For Linear regression with L2 reg, RMSE value on fold 5 val set is: 2.398323613901638
For Linear regression with L2 reg, Average RMSE value of all folds is: 2.4715498346775115

For Only Linear regression, RMSE value on testing set is: 2.711478525392836
For Linear regression + L1, RMSE value on testing set is: 2.7143461458954548
For Linear regression + L2, RMSE value on testing set is: 2.5777440746170472

By using sklearn Linear Regression on fold 1 we get RMSE: 2.127930147301517
By using sklearn Linear Regression on fold 2 we get RMSE: 2.265360779050746
By using sklearn Linear Regression on fold 3 we get RMSE: 2.199253201290417
By using sklearn Linear Regression on fold 4 we get RMSE: 2.305151064795437
By using sklearn Linear Regression on fold 5 we get RMSE: 2.150131620247328
By using sklearn Linear Regression, AVERAGE RMSE on all folds: 2.2095653625370892

By using sklearn Lasso Regression(L1) on fold 1 we get RMSE: 2.1836499705586228
By using sklearn Lasso Regression(L1) on fold 2 we get RMSE: 2.3250455922086832
By using sklearn Lasso Regression(L1) on fold 3 we get RMSE: 2.307863340415771
By using sklearn Lasso Regression(L1) on fold 4 we get RMSE: 2.4186447050465065
By using sklearn Lasso Regression(L1) on fold 5 we get RMSE: 2.2210484398250925
By using sklearn Lasso Regression(L1), AVERAGE RMSE on all folds: 2.291250409610935
```

Fig. Follow linear regression with L1 and L2 on validation set

Part c)-

Using parameters founded out in part(a) and part(b), we then trained the models on 90% data (train+val) and test them on testing set.

We got the following RMSE values –

```
File  Edit  Tabs  Help

For Linear regression, RMSE value on fold 1 validation set is: 2.4521359054929794
For Linear regression, RMSE value on fold 2 validation set is: 2.668021658014849
For Linear regression, RMSE value on fold 3 validation set is: 2.4802293373337196
For Linear regression, RMSE value on fold 4 validation set is: 2.7190512349404257
For Linear regression, RMSE value on fold 5 validation set is: 2.4931281210550225
For Linear regression, Average RMSE value of all folds is: 2.562513251367399

For Linear regression with L1 reg, RMSE value on fold 1 val set is: 2.4589203631612833
For Linear regression with L1 reg, RMSE value on fold 2 val set is: 2.649870100864001
For Linear regression with L1 reg, RMSE value on fold 3 val set is: 2.4861357765910856
For Linear regression with L1 reg, RMSE value on fold 4 val set is: 2.7292018038649593
For Linear regression with L1 reg, RMSE value on fold 5 val set is: 2.506713934491885
For Linear regression with L1 reg, Average RMSE value of all folds is: 2.566168395794643

For Linear regression with L2 reg, RMSE value on fold 1 val set is: 2.394285845818699
For Linear regression with L2 reg, RMSE value on fold 2 val set is: 2.497212356480969
For Linear regression with L2 reg, RMSE value on fold 3 val set is: 2.4663989187174384
For Linear regression with L2 reg, RMSE value on fold 4 val set is: 2.6015284384688138
For Linear regression with L2 reg, RMSE value on fold 5 val set is: 2.398323613901638
For Linear regression with L2 reg, Average RMSE value of all folds is: 2.4715498346775115

For Only Linear regression, RMSE value on testing set is: 2.711478525392836
For Linear regression + L1, RMSE value on testing set is: 2.7143461458954548
For Linear regression + L2, RMSE value on testing set is: 2.5777440746170472

By using sklearn Linear Regression on fold 1 we get RMSE: 2.127930147301517
By using sklearn Linear Regression on fold 2 we get RMSE: 2.265360779050746
By using sklearn Linear Regression on fold 3 we get RMSE: 2.199253201290417
By using sklearn Linear Regression on fold 4 we get RMSE: 2.305151064795437
By using sklearn Linear Regression on fold 5 we get RMSE: 2.150131620247328
By using sklearn Linear Regression, AVERAGE RMSE on all folds: 2.2095653625370892

By using sklearn Lasso Regression(L1) on fold 1 we get RMSE: 2.1836499705586228
By using sklearn Lasso Regression(L1) on fold 2 we get RMSE: 2.3250455922086832
By using sklearn Lasso Regression(L1) on fold 3 we get RMSE: 2.307863340415771
By using sklearn Lasso Regression(L1) on fold 4 we get RMSE: 2.4186447050465065
By using sklearn Lasso Regression(L1) on fold 5 we get RMSE: 2.2210484398250925
By using sklearn Lasso Regression(L1), AVERAGE RMSE on all folds: 2.291250409610935
```

Fig. Follow RMSE values on testing set for Linear regression, Linear regression +L1, Linear regression+L2

We can see that linear regression + L2 is performing slightly better with these set of parameters.

Part d) -

In this part, we perform the same steps as part(a) and part(b), but here we will use inbuilt libraries to train the models.

Inbuilt sklearn linear regression does not require any parameters.

By testing on validation set, we set

Parameter for lasso regression as 0.01 and

Parameter for ridge regression as 0.05.

This is result of three models on validation set -

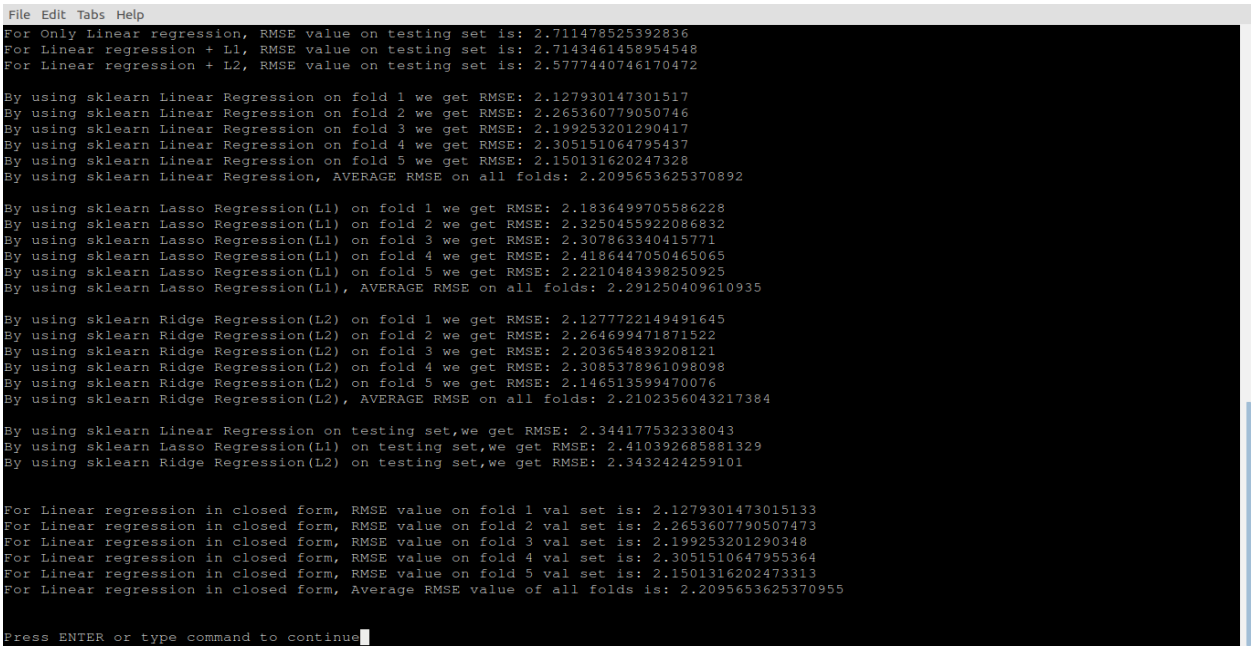


Fig. Follow Sklearn Linear regression, Sklearn Ridge Regression and Sklearn Lasso Regression on validation set

With parameters describe earlier, we then trained the model on validation set and tested it on testing set.

We got following RMSE values -

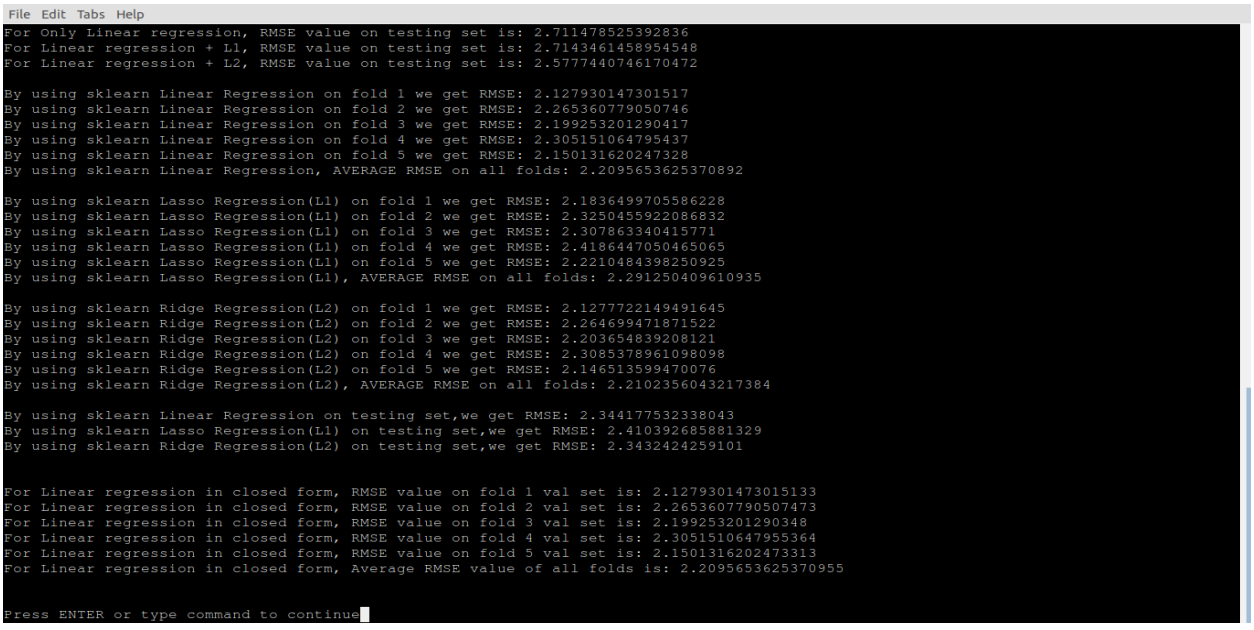


Fig. Follow Testing result of Sklearn Linear regression, Sklearn Ridge Regression and Sklearn Lasso Regression

We can see that only linear regression and linear regression+L2 is almost giving same RMSE values.

Comparing it with our above result part(d), we can see that there is slight difference in decimals. And inbuilt functions are slightly giving better RMSE values. This could be due to implementation of inbuilt functions. As they implement closed form solution and we are training with gradient descent. Also we are only doing 200 iterations in gradient descent.

Part e)-

Now for implementing closed form solution, we used formula derived in theory part of this assignment.

We get following RMSE value on validation sets -

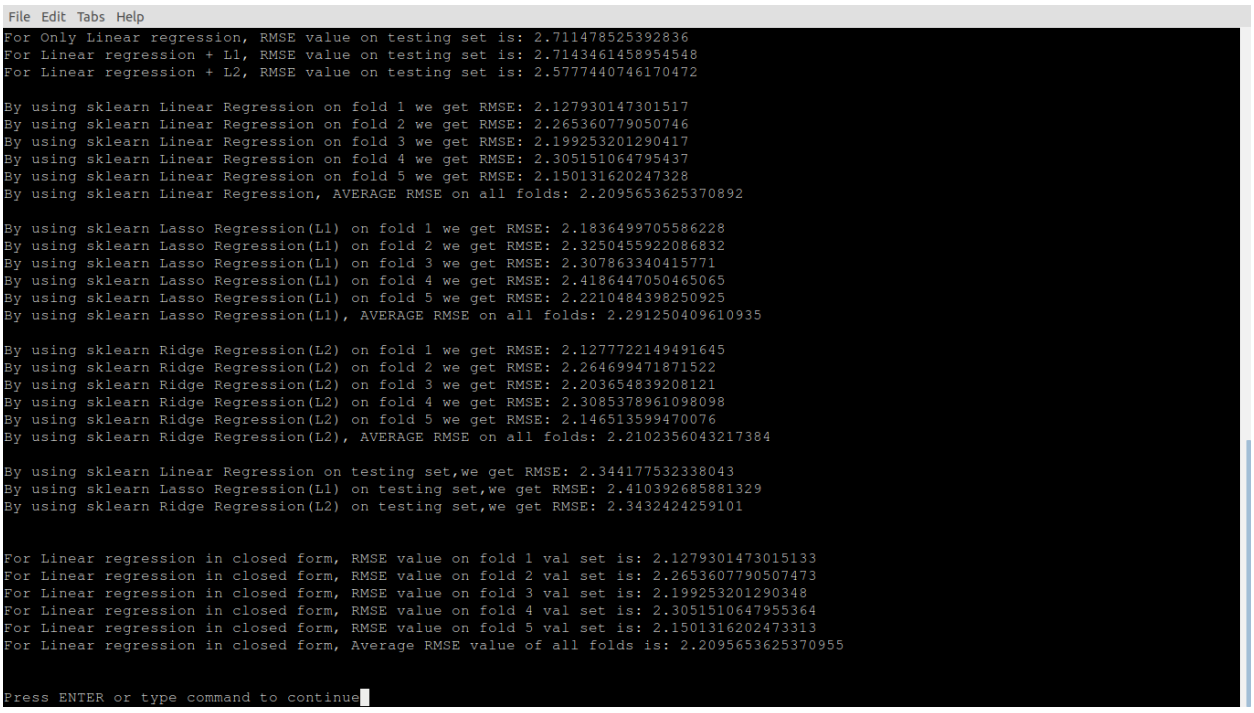


Fig. Follow the last set of results for closed form solution

Note that, result we got almost same result for this as we got by using scikit-learn linear regression.

This is because of the fact that scikit-learn implements closed form solution in its implementations.

For RUNNING python code -

Code.py file should be executed for this part.

Dataset - should be manipulated as discussed at start of this question and should be placed in same folder as code.

There are functions calls at the end of the file Code.py, for each part. Uncommenting any part will not run that part.