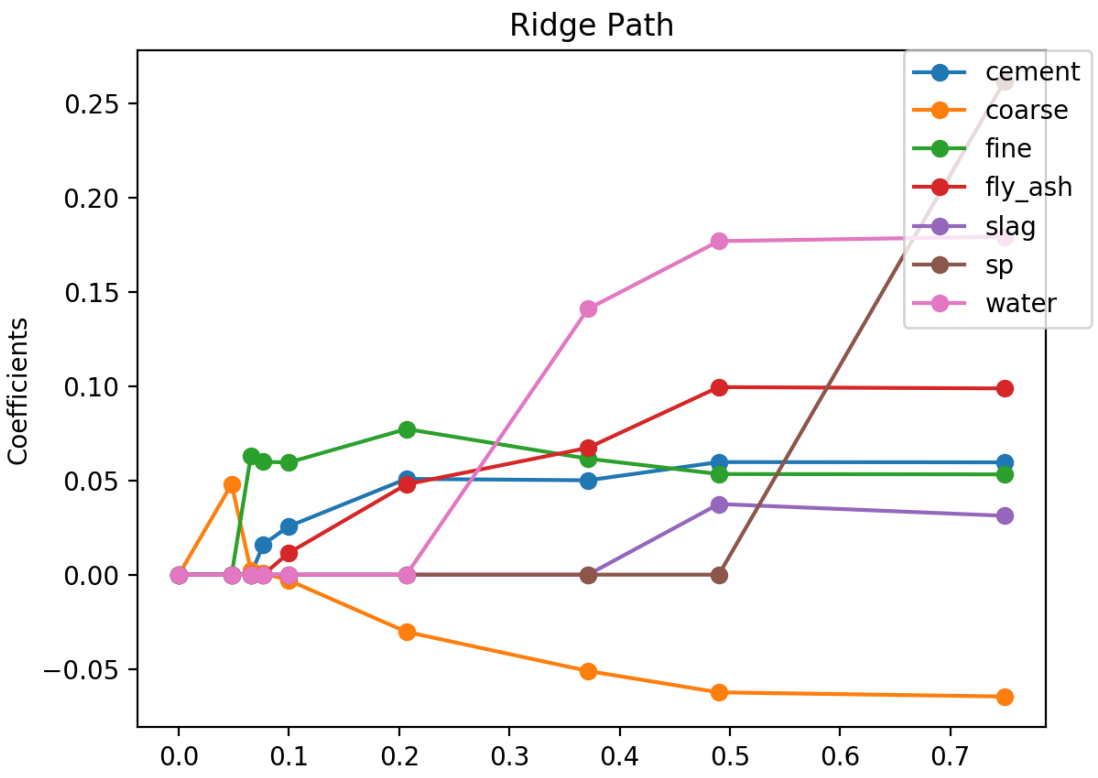
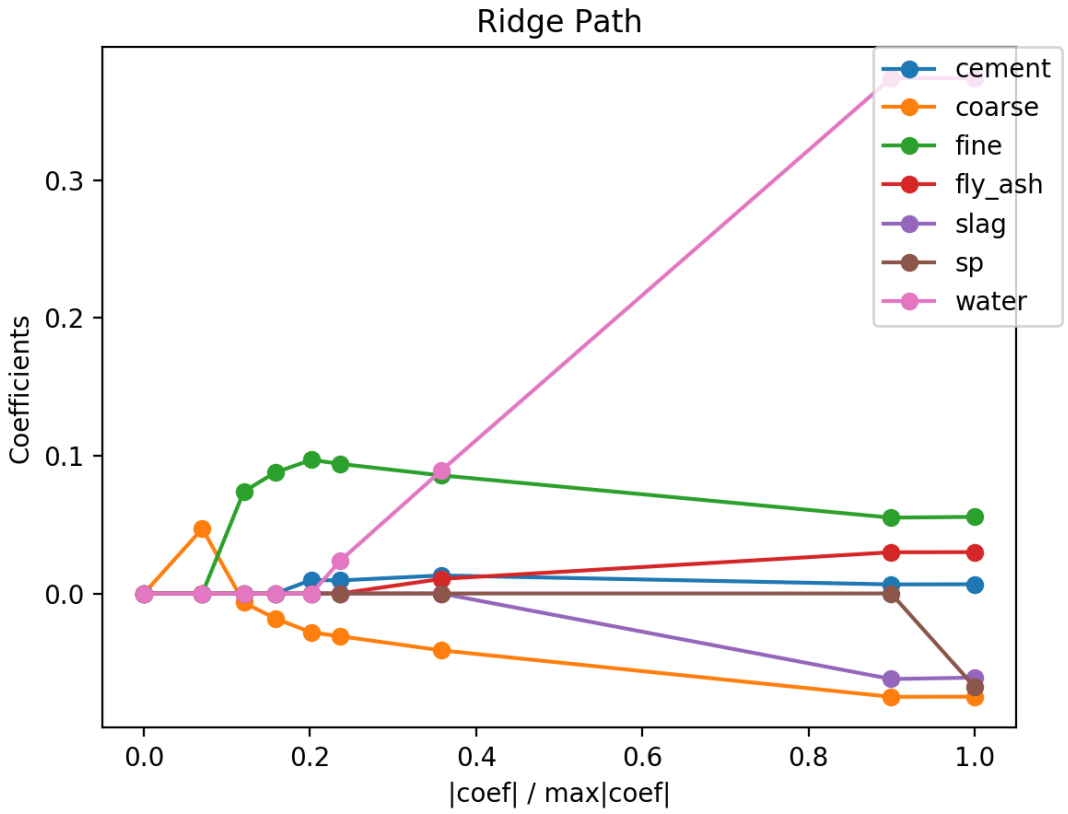
Jimmy Huang
Person Number: 50084050
CSE574 Machine Learning Spring 2018
Programming Assignment #1 Report

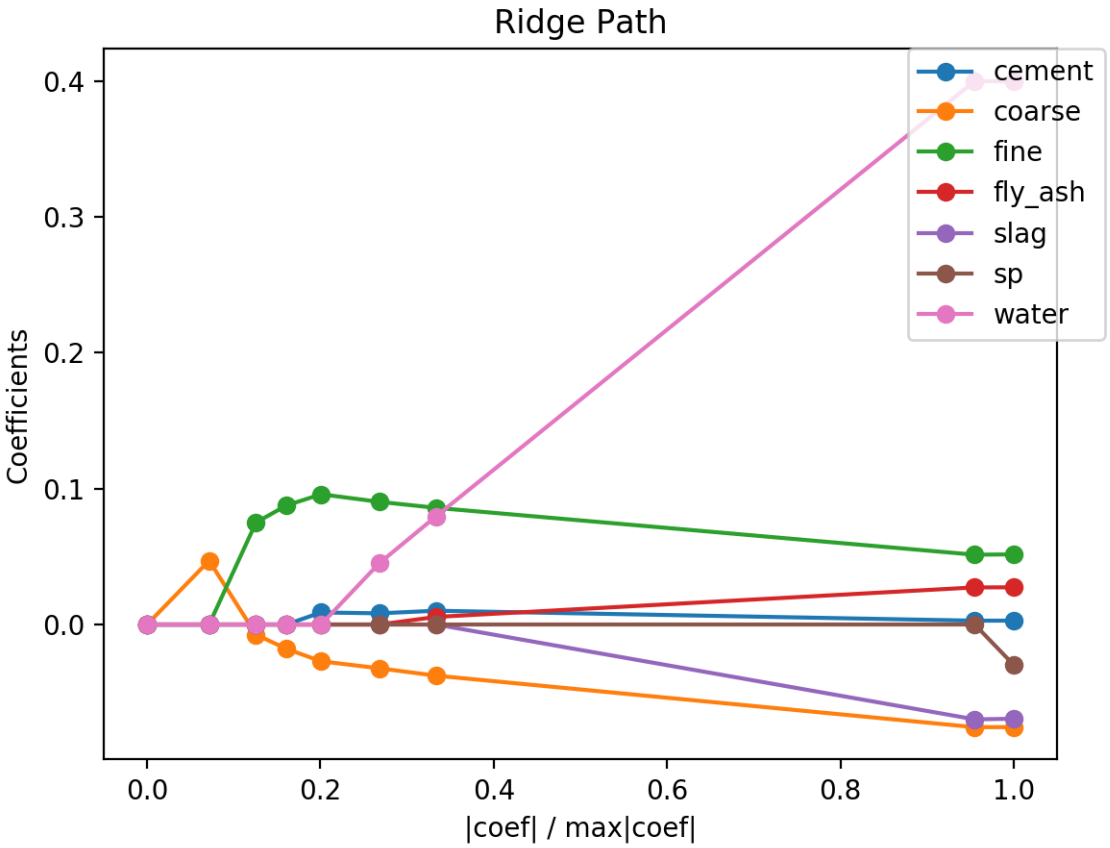
The machine learning environment I used was PyCharm since it is very useful for using the python language. I have spent around five days working on the assignment with the main bulk of it in the last two days. Most of the time came from me trying to understand the statistical and theoretical aspects of the assignment since I am not very strong in probability. I mostly used scikit-learn because of the powerful tools and built in libraries that were extremely relevant to this assignment such as the linear model and model selection. I also made use of pandas and numpy when I needed to create data frames or numpy ndarrays. My main source of help was from scikit-learn documentation from scikit-learn.org.

When I compared my regression models, I noticed striking differences in the MSE, R-squared, and regularization path graphs. For the unregularized regression of slump flow values against the seven explanatory variables, I used an ordinary least squares (OLS) regression model since it is unregularized. By unregularized, I mean that it is not affected by any weights for each of the explanatory variables as it would be with the ridge regression model and the lasso regression model. I noticed the R-squared score is consistently higher for the OLS model than the ridge model and the lasso model. This makes sense because there will be higher bias due to the regularization but there should be a lower variance. If we set the alpha value of the regularization models to be 0 it would be the same as the OLS or an unregularized regression, but we want there to be some bias since it may model the data more accurately, which is the purpose of our ridge and lasso regressions. Generally the R-squared of the OLS is around ~.05 to ~.15 higher than the R-squared of the ridge regression but it is generally around ~.20 to ~.30 higher than the R-squared of the lasso regression. From my results, we consistently see that the R-square of the three models are in this order from highest to lowest: OLS > ridge > lasso. However, there are times when the R-squared for the ridge is greater than the OLS and lasso and at times the R-squared for the lasso is greater than OLS and ridge but generally the OLS > ridge > lasso order holds. When we look at the mean squared error (MSE) of each of the three regression models, we see that the order is the opposite of the R-squared order above. The MSE of OLS regression is consistently the lowest, followed by ridge regression, and then lasso regression (OLS < ridge < lasso).

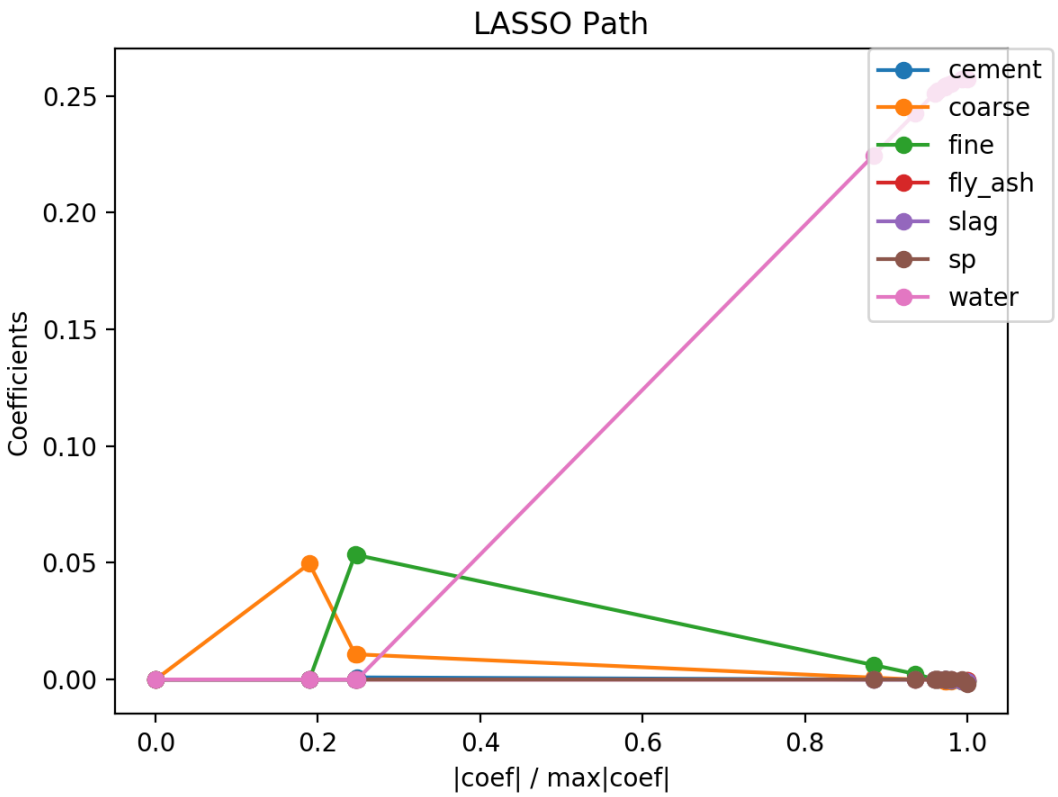
For selecting the model of the OLS regression, ridge regression, and the lasso regression I did a 5-fold cross validation and compared the scores of each iteration. I selected the model data that had the highest cross validation score and then use that data to create the model to be used since it is out best model. I fit this data to each of the regression models and then compared the results of the R-squared, MSE, RSS, and the coefficients of each explanatory variable. I was also able to create regularization paths for the ridge and lasso best fit models. The regularization paths are very interesting for the ridge regression and the lasso regression. I have plotted it several times and saved the images. The results for the ridge are as follows:

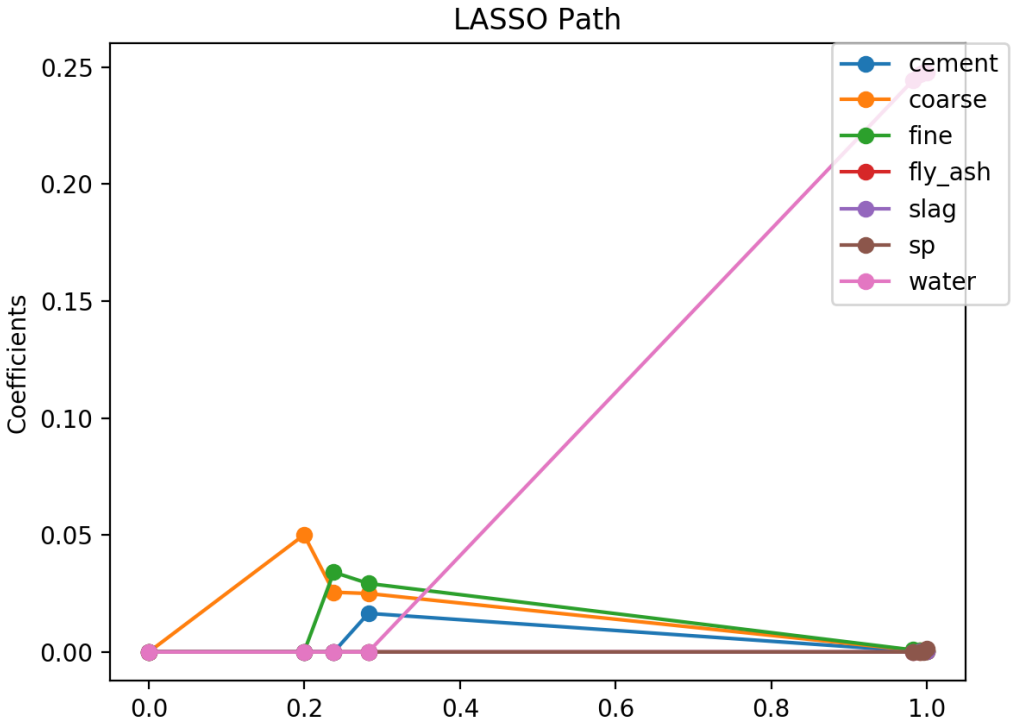

Ridge Path
cement
coarse
fine
fly_ash
slag
sp
water
Coefficients
0.25
0.20
0.15
0.10
0.05
0.00
−0.05
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

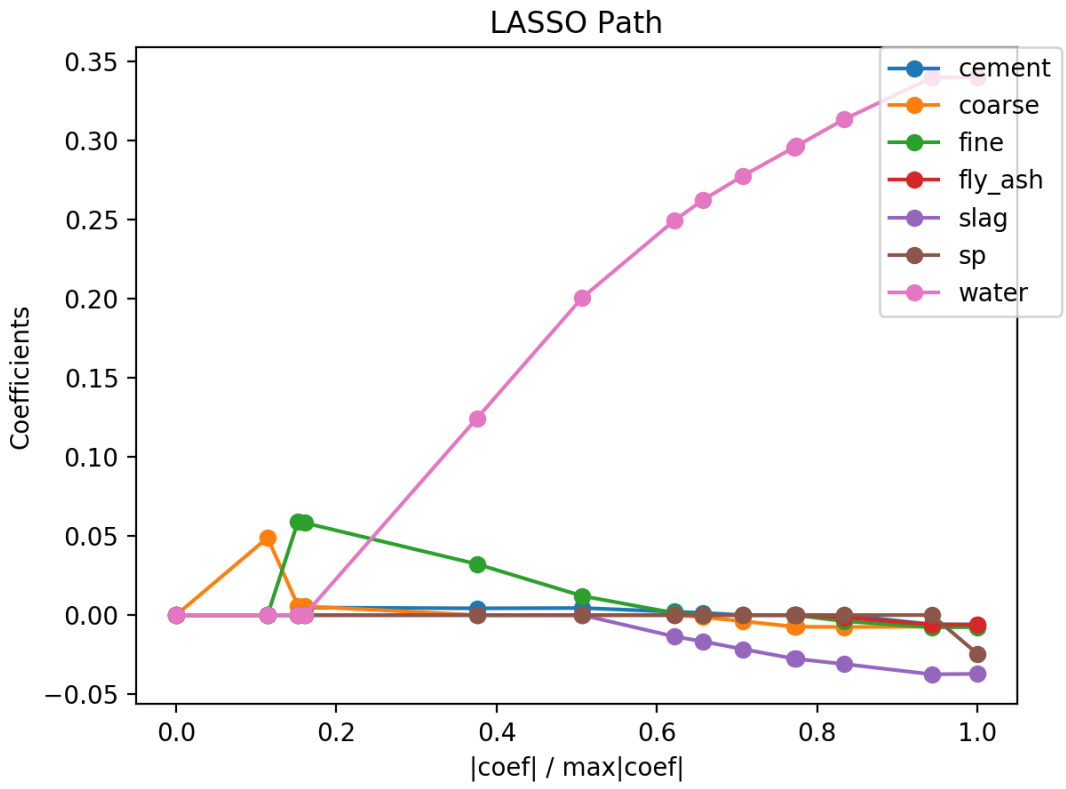

Ridge Path
cement
coarse
fine
fly_ash
slag
sp
water
Coefficients
0.3
0.2
0.1
0.0
0.0
0.2
0.4
0.6
0.8
1.0
|coef| / max|coef|



As we can see with the ridge regression regularization path graphs there is a lot of differing values for the ridge regression, especially the first graph. For all of the graphs, we notice that a lot of them get close to 0 but never go to zero. The ridge regression will put different coefficient weights on each of the explanatory variables based on its model but it will never set one to be equal to zero. We see that there is generally an importance in the water which makes sense because if you look at the article by I-Cheng Yeh we can see in section 8 that the explanatory variables that have the greatest effects on slump flow are fly ash, water, and superplasticizer(sp). We can see these three explanatory variables having the greatest weights in the first graph. The second and third graphs seem to add emphasis on fine aggregate as well, but still gives higher weights to fly ash and superplasticizer. As lambda or in my code's case, alpha gets larger the regularization terms will dominate the optimization terms. We set alpha to be equal to the max cross validation score of ridge as the Professor had instructed that we will be using the cross validation for the model selection to select the complexity of the ridge and lasso regression models. With the cross validation score, we will have a value no larger than 1 but higher than 0, so it will fulfill the requirement that it will not be very low resulting in the same as OLS and it wont be very high resulting in a mode where no explanatory variable impacts the prediction. Because of this I used the cross validation as the alpha or complexity of the regularization regression models. These help us reach the conclusion that the ridge model will place higher weights on explanatory variables the model finds more relevant.


LASSO Path
cement
coarse
fine
fly_ash
slag
sp
water
Coefficients
0.25
0.20
0.15
0.10
0.05
0.00
0.0
0.2
0.4
0.6
0.8
1.0
|coef| / max|coef|


LASSO Path
cement
coarse
fine
fly_ash
slag
sp
water
Coefficients
0.25
0.20
0.15
0.10
0.05
0.00
0.0
0.2
0.4
0.6
0.8
1.0



The lasso regularization path is interesting due to its consistency. The lasso regression model will always make water the heaviest weight in all instances. This is true no matter the value of the coefficient of water because it will always move all the other explanatory variables to 0. If we look closely at the first two graphs, we see that the all the coefficients except for water converge onto one point. That point has a y-value of 0. This is a unique trait of the lasso regression model which can be useful for large data sets with a lot of explanatory variables because it can create an accurate model with much less overhead. By setting the coefficients of explanatory variables to 0, the model predicts that these explanatory variables do not affect the response variable much and by removing them altogether it saves us from a lot of computations. We can see from the graphs above that the water coefficient is extremely higher than the other coefficients. The model is telling us that water is the most important explanatory variable to the slump flow.

How could we improve the modeling process (Extra task)? We noticed that the data is generally consistent, however, there are instances where the R-squares, MSE, and RSS orders will be in an unpredictable order. Sometimes we have a different permutation of the ordering, such as ridge > lasso > OLS. This does not happen that often but it does nonetheless. One way we can improve our accuracy is simply to obtain more data. We only have 103 observations which is not a very good population size. If we could collect much more data, we can definitely create much more accurate models. With good representative input data we can obtain good representative output data. Another way that we can improve the modeling process is implementing an artificial neural network (ANN) with back-propagation to model material behavior. Yeh stated that an ANN slump flow model is much more accurate than one based on regression analysis. (Yeh, 474) We can see – from table 3 of the reading (Yeh, 477)– the extremely low RMSE and extremely high R-squared values of ANN vs regression. The ANN will also be able to “reproduce the experimental results it was trained on.” (Yeh, 476) There are many other ways we can improve the modeling process, we can expand basis functions in the regularized expansions as well.

Sources:
http://scikit-learn.org/stable/modules/classes.html
Scikit-learn

https://www.sciencedirect.com/science/article/pii/S0958946507000261
Yeh, I-Cheng, Modeling slump flow of concrete using second-order regressions and artificial neural networks, Cement and Concrete Composites, Vol.29, No. 6, 474-480, 2007.