Spring 2023

Variable Selection Using Lasso and Elastic Net Regression on High Dimensional Genetic Architecture Data of Maize Flowering Time

Pradip Dhakal
University of Central Florida, dhakalpradip@knights.ucf.edu

Variable Selection using Lasso and Elastic Net Regression on High Dimensional Genetic Architecture Data of Maize Flowering Time

Pradip Dhakal
Statistics and Data Science Department
University of Central Florida
Orlando, USA
dhakalpradip@knights.ucf.edu

***Abstract*—Variable selection is one of the key components in the machine learning area. This method reduces the unwanted and redundant predictors in the model, which prevents the overfitting situation. Since the model contains few significant predictors, the model is less likely to learn the trend from the noise. Further, the time to train the model reduces when we have only a few valuable variables.**

***Keywords*—Variable Selection, High Dimensional Data, Lasso Regression, Elastic Net Regression, Overfitting**

I. Introduction

When dealing with many predictors, it is always best to consider the essential variables. This process of considering a few valuable predictors is known as variable or feature selection. Many methods have been proposed in the literature for variable selection, such as best subset selection, forward selection, backward selection, stepwise selection with both forward and backward selection, lasso shrinkage method, elastic net shrinkage method, and so on. Also, the dimension reduction method, such as principal component analysis and t-distributed stochastic neighbor embedding (t-SNE), have been commonly used to solve this problem. In this paper, we will fit the regression model using the lasso and elastic net shrinkage methods. More details about each technique will be discussed later in the analysis section. The dataset used in this paper is available online [1]. [2] also used Lasso regression to shrink the features when predicting the critical temperature of superconductors. Some other related works on genetic data of maize flowering time are [3] and [4].

II. Data

A. Data Information

The data consists of 4981 rows and 7393 columns. The response variable in this data is the DtoA variable, which is the number of days to anthesis for maize. There are 7389 independent variables in the dataset, which are m1 to m7389. These variables indicate the gene score from the SNP markers. The values are mostly 0, 1, and 2; however, there are intermediate values for some of the columns. The dataset also contains a few other columns. The pop column contains the family label: integer values from 1 to 25, the Entry column includes the order of the data, and the Geno Code column contains the information about the inbred line ID and family ID.

B. Data Preparation

The dataset contains missing values, so I must deal with the missing values before moving to the data modeling process. When diving more into the data, I found that 487 rows only have the Geno Code, pop, Entry, and DtoA values. All the independent column (m1- m7389) values are entirely empty for these 487 rows. The dataset contains missing values, so I must deal with the missing values before moving to the data modeling process. When diving more into the data, I found that 487 rows only have the Geno Code, pop, Entry, and DtoA values. All the independent column (m1- m7389) values are entirely empty for these 487 rows. Hence, I removed these rows from the dataset resulting in 4494 rows of clean data.

Next, the data is then separated into training and test datasets using a 70–30 split ratio. The training dataset has 3145 rows and 7389 columns. The testing dataset has 1349 rows by 7389 columns. The main goal of this split is to test the performance on data points that the model has never encountered before.

III. Analysis

In this section, I will evaluate two different models: the Lasso regression model and the elastic net model. Before diving into these models, I want to introduce multiple linear regression with all the predictors and explain the problems associated with this model.
Mathematically the multiple linear regression is given by the equation (1).

$$Y_i = \beta_0 + \beta_1 * X_{i1} + \beta_2 * X_{i2} + ... + \beta_{7389} * X_{i7389} + \epsilon_i \quad (1)$$

where $\beta_0, \beta_1, ..., \beta_{7389}$ are the regression coefficients and $X_{i1}, X_{i2}, ..., X_{i7389}$ are the features or independent variables.

The above model contains many useless and redundant predictors; these unwanted variables give rise to many issues.

One of the issues is model overfitting since the model will learn the trend from the noise when we have a lot of unwanted variables. Hence these models have low accuracy compared to the model with only valuable variables. Another issue comes while training the model. i.e., the training time is more for the model with many predictors. Therefore, fewer predictors are desired to reduce the computational cost.

A. The Lasso regression model

The Lasso regression is one of the shrinkage methods used for variable selection proposed by Robert Tibshirani. [5] This method shrinks the regression coefficients, reducing the number of features in the model. The objective function for the Lasso regression is given by the equation (2).

$$\frac{1}{2n}\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| \tag{2}$$

where λ is called the tuning parameter; it controls the bias-variance tradeoff.

We need to find the optimal value for the λ that minimizes the objective function. This is done by using the 5-fold cross-validation method. The plot for the cross-validation model is given in the figure (1).

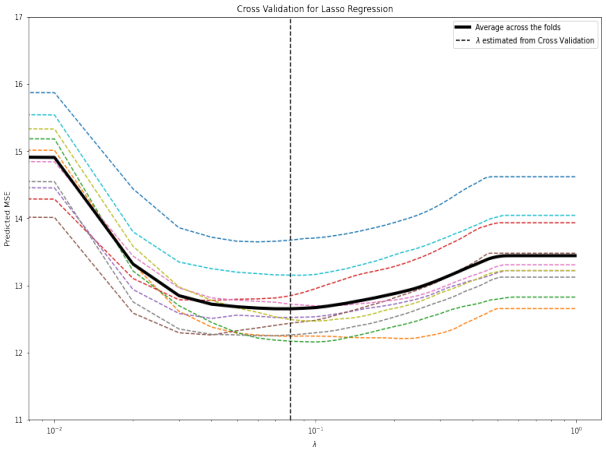

Fig. 1. Optimal value for λ

The best value for λ is 0.08. Now this lambda value is used to perform training on the training dataset and predictions on the test dataset. The root mean square was calculated to evaluate this model, and it was found to be 3.5834.

When using the best λ value, the lasso regression model shrunk the regression coefficients of 7306 variables to zero, leaving 83 variables in the model. Further, I decided to reduce the number of variables in the model by increasing the λ value. I plotted the coefficient path for this model to decide the λ value as given in the figure (2).

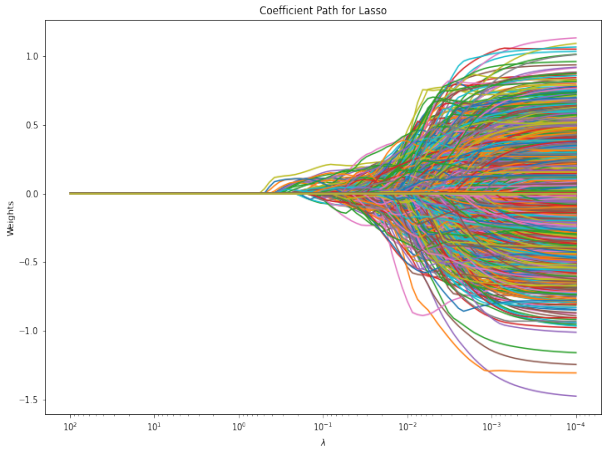

Fig. 2. Coefficient Path for Lasso Model

From figure (2), I decided to use a λ value of 0.2. When $\lambda = 0.2$, the number of variables selected by this model reduced to only 18 variables. The root mean square value only increased by a minimal number (≈ 0.04), making it 3.62.

The 18 variables selected are given in the table (I).

m333	m372	m439	m452	m459	m1447
m1629	m1635	m2427	m2463	m2501	m5815
m5853	m6369	m6466	m6492	m6501	m6513

TABLE I
VARIABLES SELECTED BY LASSO MODEL WHEN $\lambda = 0.2$

B. The Elastic Net Regression Model

The elastic net regression is the variable selection method that combines the Lasso and Ridge regression techniques. When many features are correlated to each other, the Lasso regression selects one of the features from such a group and ignores the rest. To overcome this limitation, elastic net regression was proposed by Zhou and Hastie in 2005. [6] The objective function for the Elastic Net regression is given by the equation (3).

$$\frac{1}{2n}\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j\right)^2 + \lambda\left(\alpha\sum_{j=1}^{p}|\beta_j| + \frac{1-\alpha}{2}\sum_{j=1}^{p}(\beta_j)^2\right) \tag{3}$$

where λ is called the tuning parameter; it controls the bias-variance tradeoff, and α is called the $l1$ ratio.

We need to find the optimal value for λ that minimizes the above objective equation, which is done using the 5-fold cross validation method. The plot for the cross-validation model is given in the figure (3).

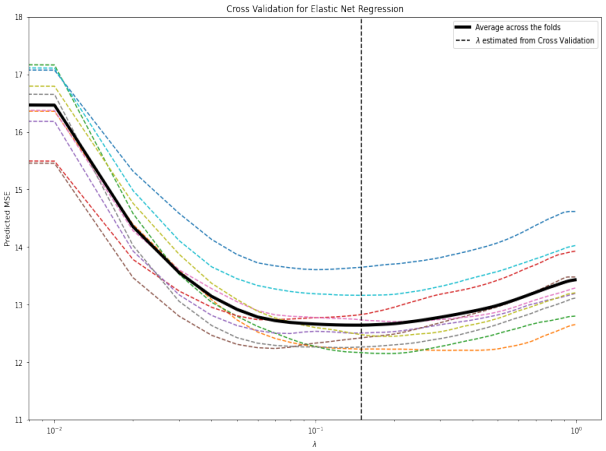

Fig. 3. Optimal value for λ

The best value for λ is 0.15. Now this lambda value is used to perform training on the training dataset and predictions on the test dataset. The root mean square was calculated to evaluate this model, and it was found to be 3.5835. This value is almost the same as the root mean square of the Lasso regression.

The elastic net regression model shrunk the regression coefficients of 7181 variables to zero, leaving only 208 variables in the model. The number of variables selected for this model is larger than the Lasso model, with the best λ value. The model with 208 variables is not desired, so I reduced the number of variables by increasing the λ value. I plotted the coefficient path for this model to decide the value for λ as given in the figure (4).

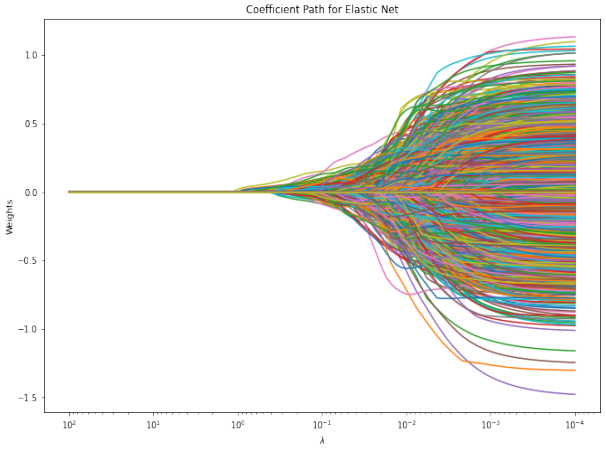

Fig. 4. Coefficient Path for Elastic Net Model

Looking at the plot above, I decided to use the λ value of 0.8, which further reduced the number of variables selected to 21. The root mean square only increased by a small number (≈ 0.1), making it 3.6741.

The 21 variables selected are given in the table (II).

m367	m368	m369	m370	m371	m372	m438
m439	m440	m441	m444	m445	m448	m452
m6492	m6493	m6498	m6499	m6500	m6501	m6502

TABLE II
VARIABLES SELECTED BY ELASTIC NET MODEL WHEN $\lambda = 0.8$

IV. CONCLUSION

Two shrinkage models: the Lasso regression model and the Elastic Net regression model, were used to perform variable selection and fit the model for the given maize dataset. The table (III) summarizes the results for both models above.

Model	λ	Num of Var	RMSE
Lasso Model	0.08 (CV)	83	3.5834
Lasso Model	0.20	18	3.620
Elastic Net Model	0.15 (CV)	208	3.5835
Elastic Net Model	0.80	21	3.6741

TABLE III
RESULTS SUMMARY

* Num of Var is the number of variables selected in the model.

* CV on the λ values indicates the best λ value obtained by 5-fold cross validation.

If we want even fewer variables than proposed above, then instead of increasing λ to reduce the variables from individual models, we can look at the common variables selected by both models. The common variables selected from both models are given in the table (IV).

m372	m6492	m452	m6501	m439

TABLE IV
COMMON VARIABLES FROM BOTH LASSO AND ELASTIC NET MODELS

REFERENCES

[1] Maize Data. (n.d.). https://www4.stat.ncsu.edu/

[2] Dhakal, Pradip, "Machine Learning-based Approaches for Predicting the Critical Temperature of Superconductor" (2023). Data Science and Data Mining. 9. https://stars.library.ucf.edu/data-science-mining/9

[3] Alipour Yengejeh, Amir, "Genome-Wide Association Study of The Maize Crop by The Lasso Regression Analysis" (2023). Data Science and Data Mining. 6. https://stars.library.ucf.edu/data-science-mining/6

[4] Fiagbe, Roland, "Linear Regression with Regularization on the Genetic Architecture of Maize Flowering Time" (2023). Data Science and Data Mining. 8. https://stars.library.ucf.edu/data-science-mining/8

[5] Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[6] Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x