OPTIMAL ALPHA DETERMINATION ELASTIC-NET REGRESSION

Aishwarya Gouru
880297521
UNCG
A_gouru@uncg.edu

Abstract—**The supervised learning is a machine learning technique that helps to map the output to the labels. There are two types of supervised models, predictive and classification model. One such model is elastic-net regression. The goal of the paper is to determine the optimal value of the alpha. By using the scikit learn library, the optimal alpha value determined is 0.1.**

BUILD THE PROGRAMMING ENVIRONMENT

https://docs.anaconda.com/anaconda/ is the website that is used to install the anaconda. After installing Anaconda, necessary libraries like openCV are installed. Python is the programming language that is used for the entire execution. Sklearn is the necessary library used for the implementation of entire paper.

INTRODUCTION

The parametrized mapping that projections a data domain into a response set is provided by supervised learning models, which aids in the retrieval of knowledge (known) from data (unknown). Supervised learning model allows you to collect data or produce a data output from previous experiences. At their most simple level, these learning models can be divided into predictive models and classification models. The predictive models are standard regression, lasso regression, ridge regression and elastic-net regression. The classification means to group the output inside a class. For example, it determines the class of the output generated like a cat or dog. The different types of classification models are random forest, deep learning and etc. The different types of predicted models are lasso, ridge and elastic net regression. The lasso regression can be briefly explained as the linear regression that uses shrinkages. The ridge regression is is defined as the model tuning technique to analyze data. The elastic net regression can be termed as

LASSO REGRESSION

Lasso regression is defined by the linear regression which uses shrinkages. The word shrinkage means that the data is shrinked towards a central point like the mean or standard deviation. In Lasso, both variable selection and regularization are used. The regularization parameter is λ|a|.

The error factor is lasso regression is determined by

$$E = (y - ax)^2 + \lambda|a|$$

RIDGE REGRESSION

Ridge regression is a model tuning technique that can be used to analyze data with multicollinearity. L2 regularization is performed with this process. Where there is a problem with multicollinearity, least-squares are unbiased, and variances are high, the expected values are far from the actual values.

The error factor is ridge regression is determined by

$$. E = (y - ax)^2 + \lambda|a|^2$$

ELASTIC NET REGRESSION

Elastic-net regression is a supervised learning model. It combines the penalties of lasso(L1) and ridge(L2) regression functions. This model combines both lasso and ridge regression and learns from their shortcomings. Elastic-net Regression's main goal is to find the coefficients that decrease the number of error squares by adding a cost to them. Elastic-net integrates the methods of L1(ridge) and L2 (Lasso and Ridge). As a result, the smoothing process is more effective.

The weighting of the number of all penalties to the loss function is regulated by the hyperparameter "λ." The completely weighted penalty is applied by default with a value of 1.0; the penalty is not applied with a value of 0. Lambada values as low as 1e-3 or even lower are very common. The parameter alpha determines the mix of the penalties and is often pre-chosen on qualitative grounds. The advantage is that elastic net requires a combination between both penalties, which can result in greater results on certain problems than a model of just one penalty.

For Elastic-net, the penalty is determined by adding the penalties of both lasso and ridge regression:

$$\text{E}=(y - ax)^2 + \lambda_1(a)^2 + \lambda_2|a|$$

STEPS INVOLVED

- Build the environment.
- Download the libraries.
- Import the libraries into the environment.
- Load the dataset.
- Save the features to X variable and save the labels to Y variable.
- Generate test and train datasets of both x and y datasets.
- Use the X train dataset and y train dataset to perform elastic net regression.
- This generated model is used to predict the output.
- The R^2 value is generated.
- Find the optimal alpha and l1 ratio.
- Generate output.

DATASET DESCRIPTION

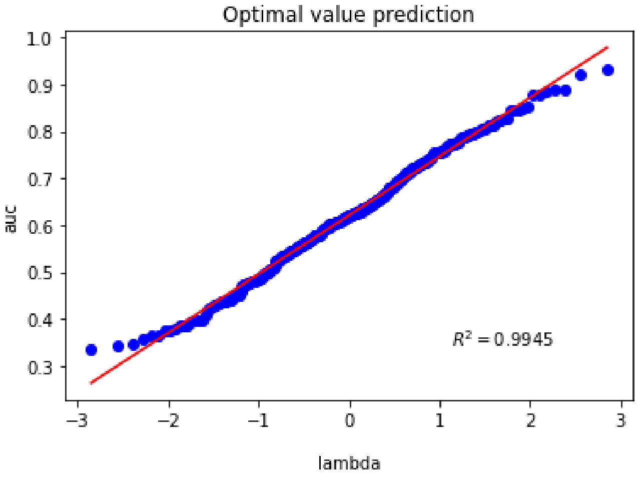
The important library used for the implementation of elastic net regression is Scikit learn. By using the scikit learn library, we determine the alpha values. The alpha value is the sum of λ_1 and λ_2. By implementing the model, the best alpha value determined is **0.1**. The Elastic Net regression model will be easily introduced in this article using Python and the Mango-Cantaloupe merged dataset. The dataset generated is developed by converting the images of mango and cantaloupe to dataframes and are labeled 0 and 1 respectively. These two datasets are merged together to obtain a merged dataset with 0 and 1 labels.

The dataset selected has both the features and the labels. The features are stored in the variable x and the labels are stored in the variable y. The chosen dataset has 1271 observations. The generated X and Y variables are now converted to training and testing dataset. The training and testing dataset is in 80:20 ratio. The X training dataset has and x testing dataset has number of observations. Similarly, for both y training and testing dataset. The imported elastic net library is used to fit the x training and y training dataset. This generated model is used to predict on the y testing dataset. The response set R is generated and the value of R^2 is also estimated.

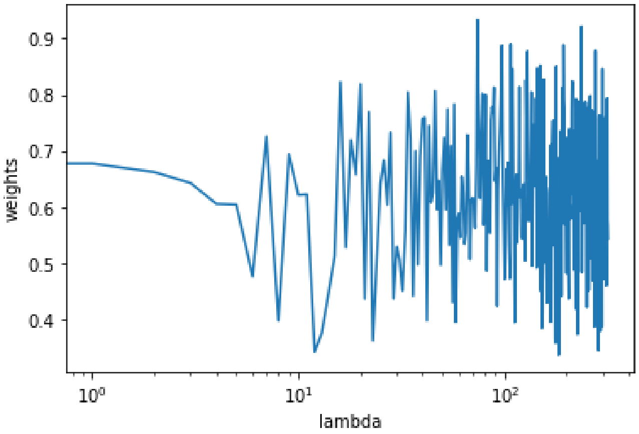
By generating the parameters from X and y dataset, the best alpha value is determined and the l1 ratio is also estimated.

Results

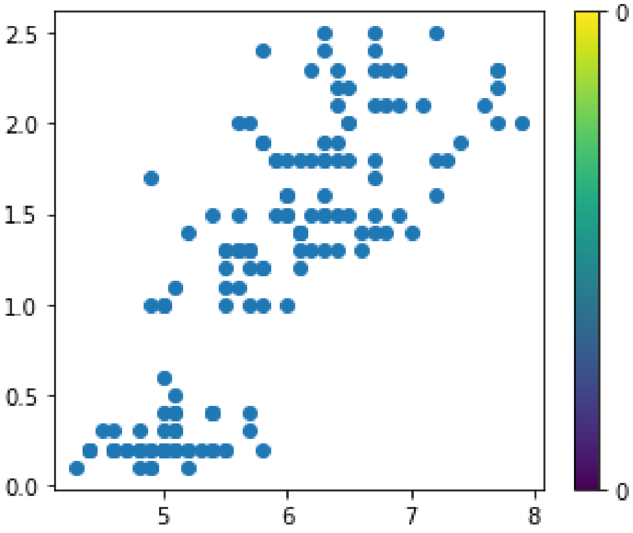
The predicted model is having R^2 of 0.9945. the predicted value using a probability plot is as below:



The gca graph generated having the log(α) values is plotted below:



The scatter plot generated using the y_pred values is shown below:



By performing the elastic net regression, the optimal alpha (Lambda) value determined is 0.1 and l1_ratio is 4.4 for the given dataset.

.

CONCLUSION

With the use of Elastic-Net regression, we have determined the optimal value of alpha for the Mango-Cantaloupe merged dataset. I have used the features in x and labels in y to train and fit the data using Elastic-Net library and I have also generated the responsive set R and estimated the R^2 using the probability plot. In conclusion, by using the elastic-net regression, I have successfully obtained an optimal alpha value of 0.1 and the l1_ratio is 4.4 for the Mango-Cantaloupe merged dataset.

REFERENCES

[1] Suthaharan, Shan. Machine Learning Models and Algorithms
for Big Data Classification: Thinking with Examples for Effective Learning. Springer, 2016.

[2] https://machinelearningmastery.com/elastic-net-regression-in-python/

[3] JOURNAL OF MULTIMEDIA, VOL. 1, NO. 5, AUGUST 2006