CSC 510 Assignment 2 Report

Justin Oakley

Contents

I Introduction 1

II Comparison of Regression Models 1

II-A Generated Set of Points . . . 2

II-B Graduate Student Datasets . 4

I. Introduction

This report is about knowing machine learning techniques that will help in selecting suitable techniques which will provide excellent quality results for applications using the computing environment utilized in an experiment. The reason for why understanding these methods are so important in big data is because they provide higher quality data not only for machines to processes, but also for them to produce. The steps taken in order to obtain the desired data output are based on the type of learning that the machine must go through and the appropriate selection of models and algorithms that follow with that type of learning. In this assignment, supervised learning is the main type of learning that will be focused upon and the modeling strategies that will be heavily utilized in data analyzation are regression and classification.

As for the information that was processed, five different datasets were used in this assignment; one computer-generated dataset of fifty observations alongside two datasets based on recent graduate and graduate students, called *recent-grads.csv* and *grad-students.csv*, are experimented on in the regression model section, and two datasets about captured network normal and attack traffic, named *nslkdd-version1.csv* and *nslkdd-version2.csv*, are utilized in the classification portion. All data analyzation for the assignment is coded in the Python 3 language via the Apache Spark 2.4.0 computational engine on an Apache Hadoop cluster. With this in mind, the goal of the comprehending supervised machine learning techniques through model and algorithm selection can be achieved.

II. Comparison of Regression Models

The first step in developing a working knowledge of the functionality of machine learning techniques is to understand the significance of regression. Regression modeling is about predicting values based on the actual responses of y and the domain values of x being utilized in the processes of an error factor minimization and regularization to paramterize the model:

$$y = A\mathbf{x}' - b$$

In order to accurately predict values using regression, the optimization process must take place, which means that several different models must be parametrized and compared together to see which has the lower error factor.

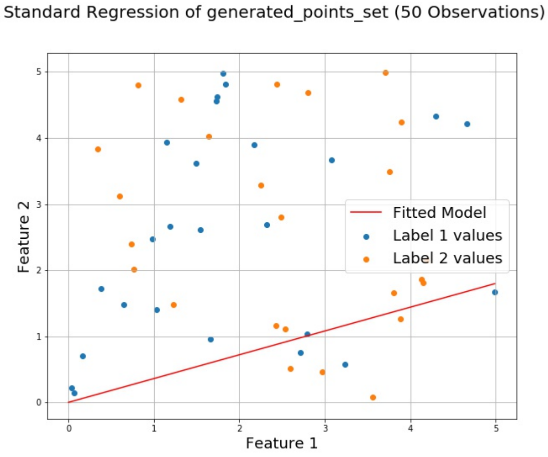
For this experiment, four different types of regression were performed on a generated dataset of fifty observations and the *recent-grads.csv* and *grad-students.csv* files. The three datasets where processed through all four types of regression to find the optimal parameters needed to make the most accurate predictions and the best fitting line or hyperplane. The four regression models used in this portion of the assignment were standard regression, ridge regression, lasso regression, and elastic-net regression. In the following subsections, each dataset's regression models will be compared and contrasted together, respectively.

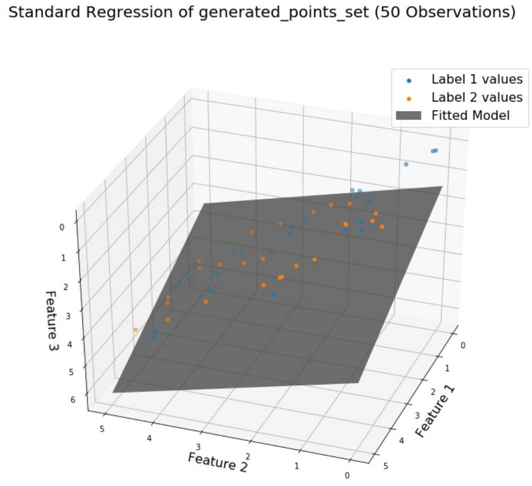
A. Generated Set of Points

In standard regression, the estimated parameter is found by calculating:

$$A = y\mathbf{x}'(\mathbf{x}\mathbf{x}')^{-1}$$

then plugging the matrix A into the model equation. When using the one-dimensional domain formed by the generated set of points was applied this algorithm, the following figure spawned: while the two-dimensional domain of the dataset resulted in:

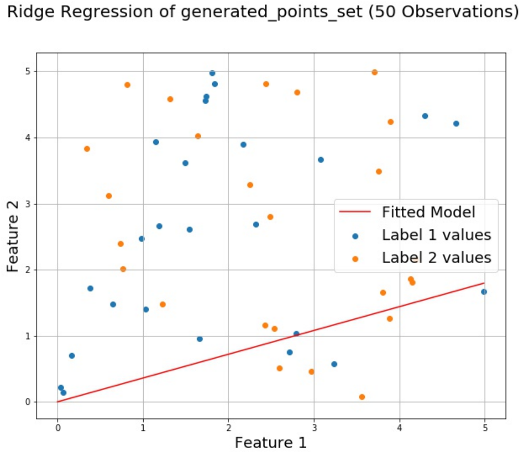


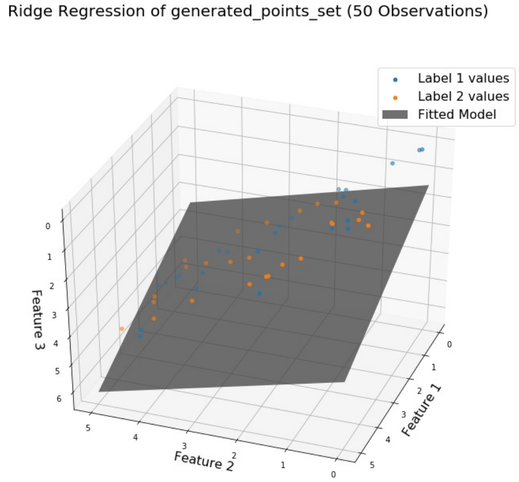


Now in the processing of the ridge regression model, where the regularixation parameter is λa^2 which gives the estimate for the vector (matrix) model is

$$A = y\mathbf{x}'(\mathbf{x}\mathbf{x}' + \lambda I)^{-1}$$

the following plots were created from one-dimensional domain and the two-dimensional domain of the dataset:



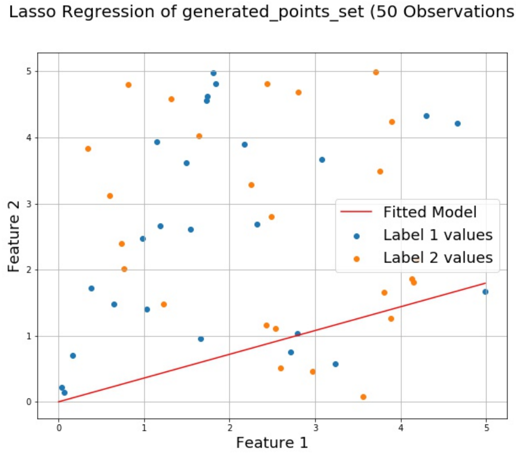


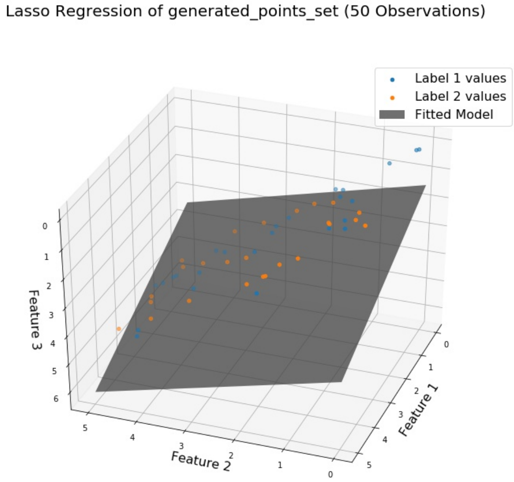
When comparing the two regression models together, the line of best fit does not appear to change in slope value, at least visually, meaning that a change in regression paramters does not effect the response variable Y.

As for the lasso regression model, which is

$$A = (y\mathbf{x}' - \frac{\lambda}{2})\mathbf{s})(\mathbf{x}\mathbf{x}')^{-1}$$

,using the generated set of points' one-dimensional domain and two-dimensional domain, the plots came out to be



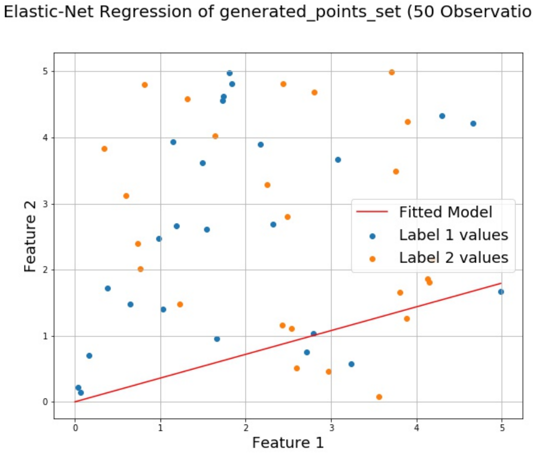


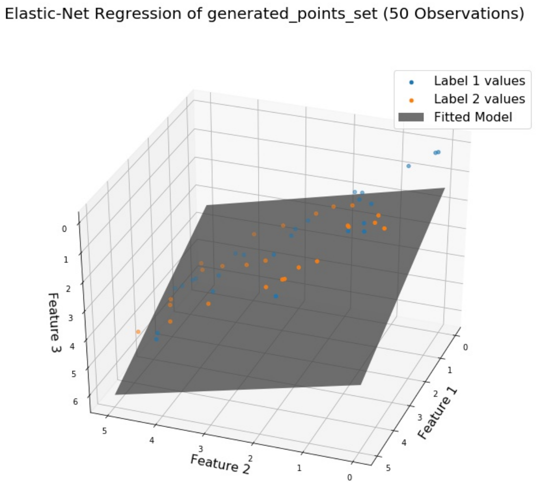
Notice that still the change in parameters does not visually effect the regression line and regression plane, meaning that adding certain regularization parameters to the error factor does not affect the model.

Finally, the last regression model in the optimization process is the elastic-net regression technique. Elastic-net regression uses both regularization parameters (that were previously used in ridge and lasso regression) to form the estimate:

$$A = (y\mathbf{x}' - \frac{\lambda_2}{2})\mathbf{s})(\mathbf{x}\mathbf{x}' + \lambda_1 I)^{-1}$$

When the domains formed from the generated set of points are used using this type of modelling, they are visualized as such:





Again, there seems to be no distinguishable difference in graphic visualizations. This means that the generated set of points has an optimized model regardless of whatever type of regression would be used on the dataset.

B. Graduate Student Datasets