Assignment 1

Name: Preetesh Verma
EntryNo.:2018EEB1171

Foundations Of Data Science
Course Code:CS521

Abstract

This document demonstrates the various observations made from the first lab assignment given to us in this course.Our assignment was to implement Linear Regression on the classic `Boston House Pricing Dataset`. We were asked to use Linear,Lasso and Ridge Regressions and compare the results obtained from them respectively.The libraried permitted to use were **sklearn,Pandas,Matplotlib and Numpy**.

1 About Linear Regression and Regularization

In statistics, linear regression is a linear approach to modeling the relationship between a dependent variable and one or more independent variables. The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression. By modelling the relationship between the variables we try to predict the values of new data by using the regression equation formed by using the Training Set. This model is valid only when there is a linear relationship between the Dependent and the Independent Variable (high correlation).
Regularization is used along with Regression in order to avoid the overfitting of the data by penalizing in an efficient way the parameters whenever they predict a wrong output.We add the Regularization term in the Cost Function(here it is generally OLS between the Y-predicted and Y-actual) itself. But while performing the Regularization we need to be careful not to give the Lambda(Regularization parameter) too high value or else it will underfit the data.

2 Observations and Conclusions

The following are the observations:
1.Lasso drops the unwanted parameters more severely than Ridge Regression.This observation is justified since the way Lasso is defined as it penalizes the absolute value of the coefficient as compared to the Ridge where we penalize the square of the coefficents.

2.Graph of the Predictor Variables in both the cases are approximately same initially but as we increase the lambda to approx. 5 there is a steep decline in the regression coefficient of ROOM AVAILABILITY predictor variable in the case of Lasso while in Ridge it follows a rather smooth exponential type decline.

3.Four of the five variables always had their values close to 0, which implies that they hardly

affected the price of houses and can be called REDUNDANT VARIABLES.

4.In case of the Linear Regression we had comparatively smaller training set error as compared to the other two but the test set error was higher than the other two.(BIAS VARIANCE TRADE OFF)

5.R-2 score was also calculated by me just to know the degree of explainability of the residuals by the regression model(it turned out to be approx 0.7).

LEARNING– So we should try to use Lasso or Ridge as compared to Linear Regression to obtain better results by avoiding overfitting and selecting an optimum value of lambda.Try to plot the graph of all the predictor variables and also compute the correlation coefficients to get a better idea as to which parameters are good for only if they are linearly correlated it is fruitful to use Linear Regression.

****