Data Science
Assignment 1:Linear Regression

Adarsh Kumar
2018CSB1066
Dr.Ramanathan Subramanian
https://sites.google.com/site/raamsubram/

Indian Institute of Technology Ropar
Ropar,Punjab
India

Abstract

This document contains the observations and inferences,achieved by using different forms of Linear Regression(OLS,Ridge and Lasso) on the Boston house pricing dataset.The models were run on varying parameters,and their results recorded both in graphical and tabular manner.
It also contains the expected and achieved behaviour of the models along with reasoning explaining the observed behaviour.

1 Introduction

We used Linear Regression and its various types to predict house prices,using the Boston house price data set available in the sk-learn library.We then divided the data set into two parts,one for training and one for testing,the size of the two sets were in the ratio of 7:3.

2 Dataset Exploration

The Dataset used is the Boston Housing Prices dataset.It has 506 data points each containing 13 features which are: CRIM: Per capita crime rate by town
ZN: Proportion of residential land zoned for lots over 25,000 sq. ft
INDUS: Proportion of non-retail business acres per town
CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
NOX: Nitric oxide concentration (parts per 10 million)
RM: Average number of rooms per dwelling
AGE: Proportion of owner-occupied units built prior to 1940
DIS: Weighted distances to five Boston employment centers
RAD: Index of accessibility to radial highways
TAX: Full-value property tax rate per 10,000 dollars
PTRATIO: Pupil-teacher ratio by town
B: $1000(B_k - 0.63)^2$, where Bk is the proportion of [people of African American descent] by town
LSTAT: Percentage of lower status of the population
MEDV: Median value of owner-occupied homes in 1000s of dollars

3 Training and Testing

3.1 OLS

We first trained the model using OLS regression, which has the following loss and cost functions:
$L(pred, act) = (act - pred)^2, J = \frac{1}{n}\sum L(pred^{(i)}, act^{(i)})$
for all i in the training set consisting of n training samples.

3.2 Ridge Regression

Since the OLS model is very prone to overfitting,we used the Ridge and Lasso variants of Linear Regression,which generalise on the training set by giving very small coefficient values to factors which are non essential hence reducing the extent of overfitting. The loss and cost function for Ridge variant of Linear Regression is:
$L(pred, act) = (act - pred)^2, J = \sum_{i=1}^{n} L(pred^{(i)}, act^{(i)}) + \lambda * \sum_{i=1}^{m} w_i^2$

3.3 Lasso Regression

Lasso regression is another way to reduce the variance of Linear Regression by preventing overfitting.The loss and cost function is defined in this way:
$L(pred, act) = (act - pred)^2, J = \sum_{i=1}^{n} L(pred^{(i)}, act^{(i)}) + \lambda * \sum_{i=1}^{m} abs(w_i)$

3.4 Varying λ

For the Ridge and the Lasso models we varied the value of λ from 1 to 200 and plotted the variation of coefficients of the model.

4 Results

Model(λ)	Train error	Test error
OLS	20.920389	25.650789
Ridge(50)	21.896319	29.176430
Lasso(50)	58.887501	77.426224
Ridge(100)	22.460152	30.197292
Lasso(100)	59.748839	77.749073
Ridge(150)	21.896319	29.176430
Lasso(150)	61.184401	78.541483

4.1 Explanation:Train Error

The OLS model works moderately well with mean square error around 20 and hence the absolute error between 4 and 5.
We use the Ridge model to prevent overfitting(if any in the OLS model) and to create a more generalised model,hence we expect the **Bias** to increase which it did,hence the Ridge model decreases overfitting to some extent.

Coming to the Lasso model,we use it for similar purposes as the Ridge model but in this case we see that using the Lasso model,increases the Mean Square Error significantly hence in this cases,the Lasso model under-fitted the data to a very high extent.

4.2 Explanation:Test Error

The OLS model works moderately well in this case too,with the error increasing by only 5,hence the absolute error increasing by around 3.
The ridge model,which was expected to reduce the **Variance** of the OLS model,didn't do so since the test error increased.
Similarly for the lasso model,the test error is very high,hence it underfitted the data even more than the Ridge model did.

4.3 Conclusion

From all this,we can conclude that the OLS model didn't overfit the data,and the Lasso and the Ridge model underfitted the data for the three values of λ taken.Hence the best performance was of the OLS model followed closely by Ridge and the worst was of the Lasso model.