Quantitative Foundations (CISC 820) - Project 1: Linear Regression

Aravind Vicinthangal Prathivaathi, Dingrong Wang

September 15, 2020

1 Report

1.1 Predicted Errors

We first used Linear Regression with all the features to estimate the Mean Sum of Squared Errors. In this, we used 10-fold cross-validation for polynomial selection and test error prediction and then used the optimal polynomial value for estimating the training error. The below table shows this training error and test error for the linear regression with all features selected for basis expansion.

Training Error	Test Error
33.5448	52.8229

Table 1: Linear Regression using the entire feature set

Since, the initial Regression model selected all features from the feature set, which may cause over-fitting. So we tried to apply a feature selection method, called Lasso, in an attempt to reduce the errors. The errors linear regression using the lasso method for feature selection are given below.

Training Error	Test Error
37.7210	41.6445

Table 2: Linear Regression using lasso feature selection

1.2 Feature Selection

We used polynomial selection along with 10-fold cross-validation to get the optimum polynomial order for the basis expansion of the input vector, we then used the polynomial order to construct the basis expansion matrix and weight matrix. Here, we selected the polynomial order through trial and error, where we tried orders from 1 to 10 and chose the polynomial order which gave the least test error.

For the project, we first used linear regression with all the features. This resulted in the error shown in Fig 2. To avoid over-fitting we tried another method called Lasso regression. Lasso regression is linear regression but we add some bias to the regression hyperplane we get from the least square error(LSE) method.

This bias uses a value called Lambda which rotates the regression hyper-plane. At an optimal Lambda value, the slope for some of the features vs dependent value will become zero, while other features will still have some slope. The slope of the features which turn to 0 is the features that do not contribute to the model and hence can be discarded. Thus, for some bias value, Lasso regression will turn to discard non-contributing features by making the slope of the non-contributing feature vs dependent attribute parallel to the axis of the non-contributing feature, thus preventing the non-contributing features from affecting the model.

1.3 Test Error Prediction

For predicting the test error, we used k-fold cross-validation. For every train-test split in the 10-fold cross-validation, we calculated the sum of test errors for the test-split across all folds and we then took an average of this error to arrive at the test error. So, for basic polynomial regression, the estimated test MSE should be 52.8229. For lasso regression, the estimated test MSE should be 41.6445.

1.4 Dealing with Over-fitting

For dealing with overfitting, we used 3 methods. K-fold Cross-validation, polynomial selection, and lasso regression.

- **K-fold cross validation:** Here we chose 10 as the K value. We split the data into 10 groups and for each unique group, we did the following: 1. Hold one group as test data 2. Select the remaining data groups for training. 3. Fit a model on the train split and evaluate it on the test split. 4. Store the error score and discard the model for this cross fold. 5. Do the above steps for 10 cross folds.

- **Polynomial Selection:** Selecting the correct order of polynomial for the regression line is important to get a good fit. For our project, we did a trial and error method, where we ran the above k-fold cross-validation algorithm for each polynomial order from 1 to 10 and chose the best polynomial order the resulted in the least test error.

- **Lasso Regression:** Over-fitting is also caused by feature selection. By selecting features that do not contribute to the model, we might arrive at an over-fit hyper-plane. Therefore, we might need to consider which features to choose from. Lasso regression is a technique that deals with over-fitting by introducing bias. For some value of the bias, the slope of the non-contributing features vs the dependent attribute turns zero while some slope for contributing features exists. That is bias value, we have a good to fit graph with optimum selected features. This is also explained in 1.2