Report on Project 1 - Linear Feature Engineering

By

Kaleem Nawaz Khan (kk5271@rit.edu)
Muhammad Raees (mr2714@rit.edu)

A. Train and Test Error

After we tried to fit a linear regression model to the given dataset, the optimum values for training and test error are as follows. Moreover, we also report the average error values after applying K-fold cross-validation (K=5, flexible in code).

1) The Training Error=45.19 (Avg K-Fold=42.73)
2) The estimation of Test Error=67.87

B. Feature Selection

Original Features

We performed the initial experiment with the base input features (926x8) and recorded the training error of 109.39 with respect to the output feature (926x1). Subsequently, we did the experiment with K-Fold cross-validation at various values of K (folds). At K=5, we measured the following values of training and test errors.

1) Average Train Error=107.42 (Min=91.3, Max=118.69)
2) Average Test Error=132.47 (Min=81.52, Max=217.8)

Feature Transformation

On the observation of the above-mentioned training and test errors, we applied feature transformation of higher order polynomials and various combinations of functions (Notably, X^2, X^3, X^4, X^5, log(X), log(X^2), sin(X), and $(X)^{1/2}$. Based on the results of running experiments on various combinations of functions (transformations), we finally restricted our model up to degree-3 polynomial (X, X^2, and X^3) having superior cross-validation results. Again, we reported our results through K-Fold cross-validation (by varying K). At K=5, we report the values of training and test errors.

1) Average Train Error=42.73 (Min=36.1, Max=48.7)
2) Average Test Error=67.87 (Min=39.82, Max=113.63)

Report on Project 1 - Linear Feature Engineering

Additional Analysis

We performed additional analysis for further experiments to identify useful features. We did an MRMR (Maximum Relevancy, Minimum Redundancy) analysis and reduced the number of features. However, for the given dataset, the estimations of MRMR analysis were not profitable. Hence, we can not report an improvement in the results by reducing the features. MRMR analysis also calculates a correlation between features. Although we separately performed a correlation analysis on our features to identify the correlation of input features to the output features, the correlation does not provide any meaningful feature reduction results. We also performed random feature selection by using combinations of features (i.e., randomly dropping a subset of features) for model approximation. Unfortunately, the feature reduction did not help us to improve the results further.

C. Prediction of Test Error

We estimate the test error by performing various cross-validation on the training set. The test error is an approximation of cross-validation performed on various Folds (K=[3,5,7,9]). Based on the average of test errors on the cross-validation (K=5), we estimate and predict the test error on the testing dataset will be around 67.87.

D. Overfitting

Currently, cross-validation training and test errors are on the higher side with respect to the distribution (and mean) of all labels (outputs) of the training dataset. Using higher order polynomial features, we decreased the training error (around 30 for 5-degree polynomial) but, the testing error on the cross-validation test sets rose significantly (around 400) depicting the overfitting. Hence, by decreasing the degree of the polynomial we reduced the overfitting of the model. Likewise, K-Fold estimation provides a good indicator of the pattern in testing and training errors. Therefore, we consolidate the fitting of the model with the degree of polynomial features through K-Fold cross-validation.

E. Conclusion

The model performs a regression analysis on the given dataset. The model estimated on the linear features does not provide a good approximation of the training and test error. Thus requiring a need for the feature transformation to higher order polynomials. The test error can also be predicted based on the cross-validation of the training set. The feature transformation to degree-3 polynomial reports an average training error of 42.73 and a test error of 67.87 through cross-validation using K-folds.