Project 1: Classification of Higgs Boson using Linear Methods

Jangwon Park, Jean-Baptiste Beau, Frédéric Myotte

I. INTRODUCTION

The project presents a challenge of correctly classifying Higgs bosons based on the 30-feature particle accelerator data from CERN. This report summarizes the application of linear methods for binary classification.

II. METHODS

The problem is tackled by linear and logistic regression. Within linear regression, we seek to minimize mean squared error (MSE) by use of normal equations, gradient descent (GD), and stochastic gradient descent (SGD) as well as ridge regression. Within logistic regression, we also employ gradient descent and regularization to optimize our model.

A. Initial results

The following box plots summarize the 10-fold cross validation (CV) results of the methods described above prior to any feature processing steps.

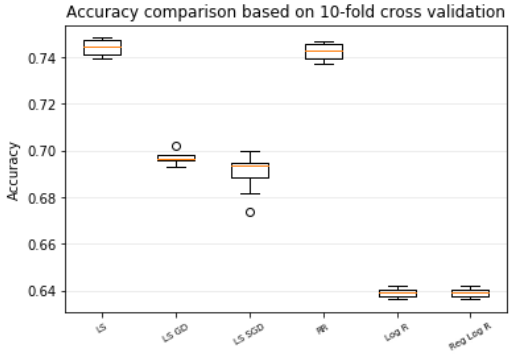

Fig. 1. Model accuracy from 10-fold cross validation. From left to right: least squares, least squares using GD, least squares using SGD, ridge regression, logistic regression, and regularized logistic regression.

B. Optimizing hyperparameters

Among the listed methods above, regularized linear and logistic regression require optimization of λ (i.e. coefficient of the penalty in the loss function). Figure 2 describes this process. For each of the 30 different values of λ on the x-axis, mean model accuracy from 10-fold CV is computed. Averaged over ten seeds, the optimal λ corresponds to the highest accuracy on the y-axis. The same process was also used in regularized logistic regression, where optimal λ was approximately 0.0001.

III. EXPLORATORY DATA ANALYSIS

In this section, we describe important findings from exploratory data analysis which establish the motivation for some feature processing methods.

The first observation is many features have undefined values (-999). According to the CERN Open Data Portal, they arise due to their dependence on another common feature known as *PRI-jet-num*, whose value is either 0, 1, 2, or 3. For instance, *PRI-jet-leading-eta* and *PRI-jet-leading-phi* are undefined when *PRI-jet-num* = 0. This implies that the real number of features may be less than 30 depending on the value of *PRI-jet-num*, hence pointing to the possibility of developing four instances of the model by manually clustering the data based on *PRI-jet-num*.

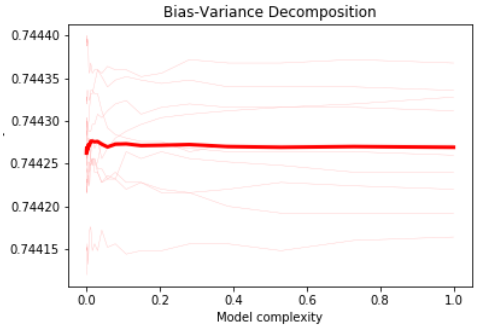

Fig. 2. Bias-variance trade-off curve for ridge regression with various values of λ. The optimal λ is approximately 0.0117.

The second observation is non-linearity between the features and the response. If there is non-linearity, the errors will not have constant variance, violating an important assumption of linear regression. This is verified by a standardized residual plot, which we omit in this report for brevity. Though this phenomenon will persist for a binary classification problem, it can be alleviated by feature augmentation.

Lastly, there may be interactions among features which can be visualized in a pair plot. Possible interactions can be identified when there is a clear non-linear pattern in a scatter plot between two features. For example, features (0,6), (5,6), and (6,7) in figure 3 appear to develop some clear non-linearity as demonstrated by their contour.

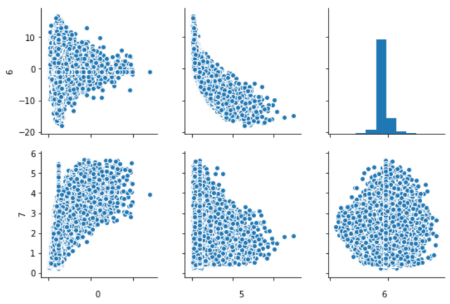

Fig. 3. Pair scatter plot for some features

IV. FEATURE PROCESSING

In this section, we describe various feature processing approaches which directly improved model accuracy.

A. Manual clustering

PRI-jet-num is central to deeming certain features undefined. Since it only takes on four values, it is possible to

manually cluster in such a way that only the samples corresponding to the same value of *PRI-jet-num* are grouped. All undefined features in each cluster can then be removed manually. This is expected to improve model accuracy by the effect of both data cleaning and a form of feature selection. Each cluster still had sufficiently many samples so as not to suffer from high dimensionality (smallest cluster has 22,164 samples).

B. Degree transformation

By using polynomial basis functions, we seek to expand the feature space. Polynomial expansion helps capture complex relationships such as non-linearity.

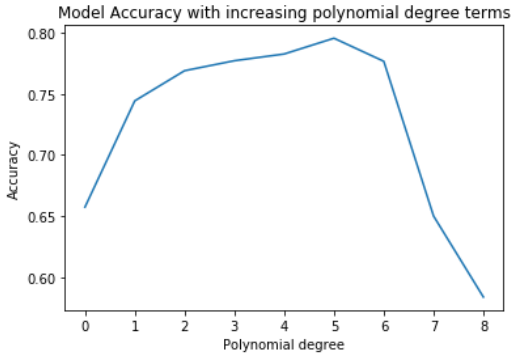

Fig. 4. Bias-variance trade-off curve based on 10-fold CV. Higher polynomial terms, as evaluated by ridge regression, improve accuracy only up to degree 5.

10-fold CV estimates model accuracy for subsequently higher-ordered polynomial terms. Figure 4 illustrates including up to fifth degree terms is optimal, after which the model becomes too complex and overfits the training data.

C. Backward selection

Although including up to fifth order polynomial terms yields the highest accuracy, not all fifth order terms, for example, may be necessary. Therefore, backward selection, which eliminates all non-useful features, is aptly applied after polynomial expansion. At each step, the algorithm removes exactly one feature whose removal results in the largest improvement based on 10-fold CV; this continues until no improvements can be made, which ensures all remaining features are necessarily useful.

D. Interaction terms

Higher-ordered terms can also be products of different features. Linear regression assumes that a feature contributes to the response independently of the magnitudes of other features. If features truly interact, including their products in the model can relax the additive constraint of linear regression and capture more complex relationships. In a similar procedure as figure 4, we rigorously tested for increasingly higher-ordered interaction terms and concluded that terms higher than third order barely improved model accuracy (on the order of 10^{-5}) despite significantly increasing computational time.

E. Forward selection

From over 4,000 interaction terms, certainly not all are useful. Therefore, forward selection, which avoids adding any non-useful features, is particularly appropriate since applying backward selection after adding all terms is computationally costly. At each step, the algorithm adds (and keeps) the first term whose addition improves model accuracy and continues until the search list is exhausted. Since there are many interaction terms, forward selection is based on 5-fold CV to hasten the process.

V. IMPROVEMENTS

The model with the highest accuracy was achieved by least squares after completing all of the feature processing steps in the same order as presented in the previous section. Test data is processed in exactly the same fashion, and predictions are made in each of the four, manually clustered instances independently and later consolidated.

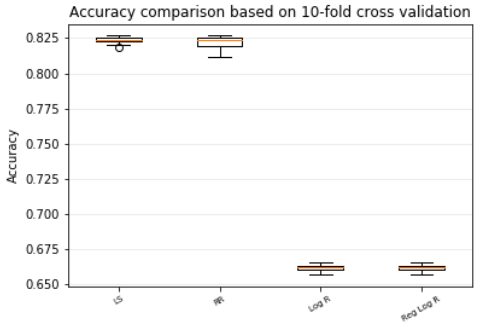

Fig. 5. Model accuracy from 10-fold CV after feature processing. From left to right: least squares, ridge regression, logistic regression, and regularized logistic regression. Tailored backward and forward selection were not as effective in logistic regression (roughly 2% improvement).

Since we did not run into a singular matrix issue, we did not turn to more computationally intensive GD or SGD.

VI. DISCUSSIONS

The score of our model on Kaggle is 82.321% with least squares as our optimal model. There was virtually no difference in performance between least squares and ridge regression, which can be explained by the rigorous procedure of backward or forward selection that continuously prevents overfitting with CV. Therefore, regularization appears to have minimal impact on improvement. In this case, we avoid ridge regression since it has an extra parameter λ to optimize.

Logistic regression in general performed poorly compared to linear regression. We therefore conclude that logistic function was not an effective way of estimating the probabilities of observing Higgs boson given the linear model.

Finally, we note that not performing manual clustering or backward/forward selection resulted in a lower model accuracy by approximately 2% for the least squares method.

VII. CONCLUSIONS

This project only explored linear methods for classification. Compared to the most basic least squares model, we achieved an overall improvement of 7.858% on Kaggle. This verifies that steps taken to non-linearize the model and select useful features via clustering and feature selection were certainly effective. We conclude by conjecturing that non-parametric methods in general (e.g. support vector machines) may produce higher accuracy given the complex, non-linear nature of the problem.