Recognition of Higgs Boson based on Machine Learning

Jun Qing, Lingjun Meng, Aibin Yu

***Abstract*—In this project, we need to deal with the binary classification problem in order to analyze a real-worl dataset. The dataset was obtained from CERN particle accelerator experiments and aiming to recognize the Higgs particle.**

We firstly analyze and clean the data. Afterwards we compare six basic machine learning algorithms learnt in class, and make further improvement using polynomial feature expansion. Then, we estimate the generalization error of different models through cross validation and select the best one for further processing. Finally we split the data into groups and train the model on each group. The final accuracy on test data achieved 79.5%.

I. Introduction

Due to the rapid decay rate of Higgs boson particles, scientists don't observe it directly, but rather measure its "decay signature", or the products that result from its decay process. Thus, we employ machine learning methods to extract the relationship of the particle types and the events signature, and obtain a model that can recognize the Higgs boson recording to CERN particle accelerator data.

For this project, we visualize and clean the data at the beginning. Then we divide the data into training set and validation set. We apply six regression models on the training set and predict the labels of validation. Next we try to make an improvement by using polynomial feature expansion. We find that the ridge regression model with feature expansion performs the highest prediction accuracy on the testing set. Furthermore, we split the data into different groups by the PRI jet num feature, and then apply ridge regression model on each group. By doing this, we obtain our best model with highest prediction accuracy on the testing set.

II. Models and Methods

A. Data Analysis and Processing

In order to get some intrinsic structure of the dataset that we could exploit, we did some exploratory data analysis. By inspection, We found out some data points had meaningless values, which are equal to -999 in the dataset. These missing values must be removed or replaced. Otherwise they will destroy the stability of the model. Then we visualize the data point in 2D space by two randomly selected features. One of the visualization is shown in Figure 1. It is evident that there exist some data outliers which may cause a large bias in the model. These data outliers can be regarded as noise and may be generated by measurement error or some other practical reasons. Thus, it is essential to do data cleaning. We implement three methods for data cleaning:

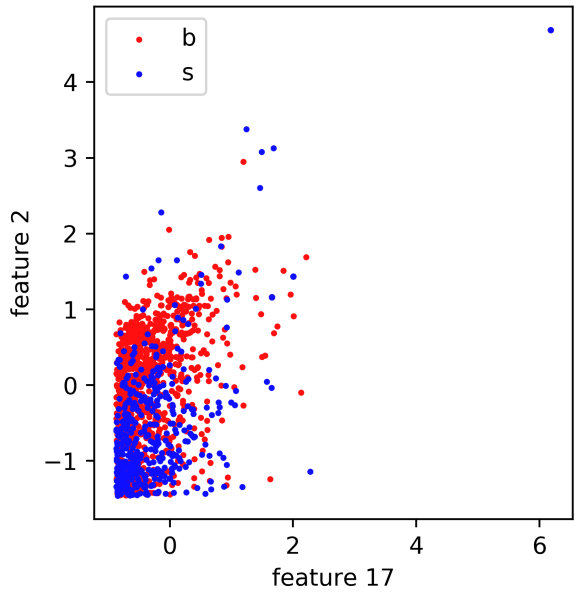

Figure 1. Figure 1. Data visualizing.

1) **Fixing Null value.** We observed that there are many columns containing Null value (-999). They are missing values. We detected these abnormal values and replaced them with the mean of non-NULL values.
2) **Bounding data outliers.** If the value of a sample feature deviates from the average value (μ) of this feature by more than 3 times the standard deviation (σ), we judge it as an outlier, and then substitute it by $\mu + 3\sigma$ or $\mu - 3\sigma$.
3) **Normalization.** Since the learning convergence is sensitive to the feature ranges and the difference among the ranges of different features may cause divergence in the training process. Thus, we perform normalization using the following equation:

$$x_{nom} = \frac{x - \mu}{\sigma}$$

With the methods described above, we obtain a normalized dataset without outliers and Null value for training.

B. Implement of Regression Algorithms

We compare six basic machine learning algorithms in order to select a best model for further improvement. First, we randomly divide the whole training data into training set

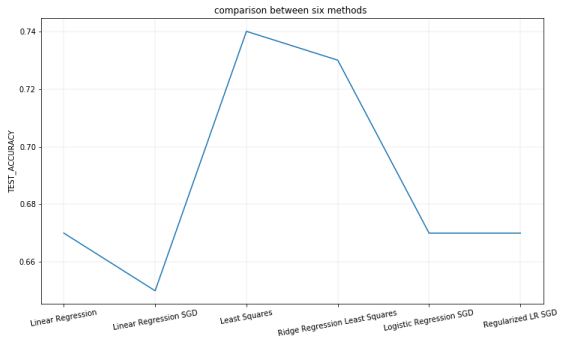

Figure 2. Comparison of six learning methods.

and validation set. The ratio of the data set size of training set and validation set is 3:1. Then, we train our models on the training set and test them on the validation set. The learning rate we choose is 0.001.

The accuracy on validation set of the six models are compared in Figure 2. As is shown is Figure 2, Least Square algorithm performs a highest prediction accuracy, which reaches 74%, while Ridge Regression follows at 73%.

III. Model Improvement

A. Feature Augmentation

Although the linear model has reached a relatively high prediction accuracy, it still performs some large bias because of underfitting. Thus, we impose on polynomial feature expansion. As is shown in Figure 3, the higher orders polynomial feature expansion evidently enhance the prediction accuracy comparing with no feature expansion, which can be represented by the case with polynomial degree equal to one. It is worth noting that when the polynomial degree goes to 8, there exist overfitting in Least Square model, while ridge regresion model still perform a good prediction accuracy on testing set. Thus, we select ridge regression model with 8_{th} order polynomial feature expansion for the next cross validation.

B. Cross Validation

In order to select the best hyperparameter λ, the regularized coefficient in ridge regression model, and estimate the generalization error of our model, we implement 4-ford cross validation. The results are shown as Figure 4. Although the validation accuracy increase monotonically with λ decreasing, we observed the optimum of the prediction accuracy on testing set at $\lambda = 0.001$, which reaches 76.5%.

IV. Training by Group

We noticed that the feature, PRI_jet_num is a categorical feature with values 0, 1, 2, 3, and holds strong relations with the feature that exists large amount of Null value. Inspired by this, we spilt the data into four groups by PRI_jet_num,

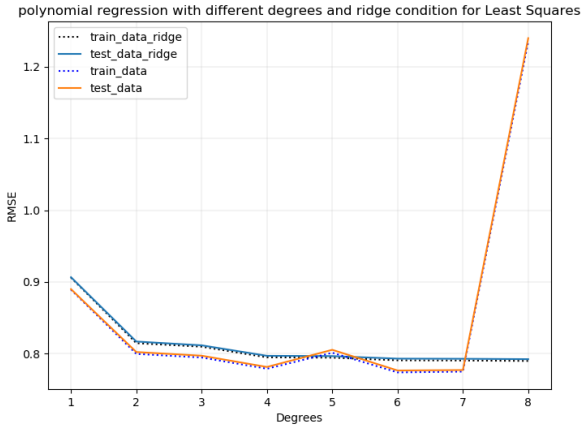

Figure 3. Root Mean Square Error versus polynomial degree of feature expansion.

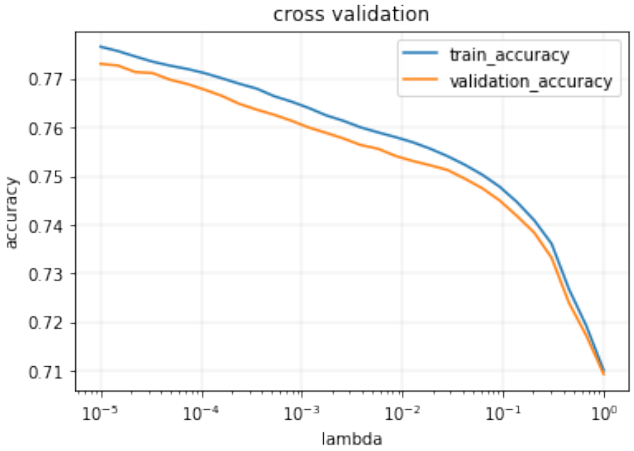

Figure 4. Cross Validation of Ridge regression with feature expansion in different λ.

and training the model by group. By doing this, our model performs 81% accuracy in our own validation set and 79.5% on testing set.

V. Summary

Our best machine learning methods can realize approximately 80% Higgs particle recognition accuracy. Data cleaning, augmentation, and grouping can evidently improve the model performance. Polynomial feature expansion is a efficient way to tackle the underfitting problem of linear model. Besides, regularized item is necessary to avoid overfitting especially after data augmentation. The best model we choose is ridge regression with 8_{th} order polynomial feature expansion and λ