Higgs Boson Classification by Machine Learning

Qianqing Wang[1], Xinyu Liu[2] and Tianyu Gu[1]
1. *Department of Civil Engineering, EPFL, Switzerland*
2. *Department of Electrical Engineering, EPFL, Switzerland*

Abstract

This paper shows the results of different machine learning methods for binary classification on CERN's Higgs boson dataset. The best model proposed achieved an accuracy of 82.0% and a F1 score of 0.733 in the test set, and effects of different feature engineering methods and different classification models are investigated and discussed in this paper.

I. Introduction

The discovery of Higgs boson [1], [2] was awarded the 2013 Nobel price in physics. The Higgs boson is generally produced by collision between protons at high speed, and it is very unstable, decaying into other particles almost immediately, thus it can only be observed indirectly by measuring its decay signature. This paper aims to offer a machine learning-based estimation method to predict if the decay signature indicates the existence of a Higgs boson or it just originates from other processes or particles.

The decay signature dataset with 250000 samples in this paper are generated by the particle accelerator from CERN. Each sample has 30 features and 1 binary label denoting if it is a Higgs boson signal or a background signal. The best model in this paper achieved an accuracy of 82.0% in the Alcrowd platform leader board, and it was accomplished by outlier removal, missing data imputation, one-hot encoding, polynomial feature expansion and feature interaction under the Ridge Regression model.

II. Methodology

A. Data Cleaning

Before feature engineering, prominent errors in the dataset should be eliminated first, and this is essential for prediction accuracy. During the exploratory analysis, some problems are found and their corresponding solutions are listed below,

1) *There are NULL (-999) values in the data*, which can be divided into two cases. First, samples can be separated into four groups according to the discrete feature PRI_jet_num. NULL values in feature DER_mass_MMC are distributed uniformly in four groups, so the NULL values can be directly replaced with the median of non NULL in this feature. For the second case, groups with labels of 0 or 1 tend to have several features with all NULL values. For these NULL values, they were replaced with 0 after data normalization.

2) *There are many outliers in the data*. Outliers are values diverged from overall distribution of data so that they should be removed to guarantee the reliability of features. Given a feature vector x, a sample x_i is not considered as a outlier if

$$\overline{x} + k \times \sigma(x) > x_i > \overline{x} - k \times \sigma(x) \tag{1}$$

where $\overline{x}$ is the mean and $\sigma(x)$ is the standard deviation of x, and k is a parameter controlling the extent of outlier removal. Outlier removal was done before missing value imputation, so the bias of imputation was not incorporated into this step.

B. Feature Engineering

Feature engineering might be the most influential part in this project, and the authors have tried many ways to fusion and augment different features to achieve higher training as well as test accuracy, which are listed below,

1) *One-hot encoding*. The discrete feature PRI_jet_num set the data into four groups. One way to consider this is to train the data separately, and in theory this way can achieve high accuracy but the training process can be computation-expensive, so the authors chose one-hot encoding method for convenient and fast training of the dataset.

3) *Normalization*. Features were normalized, with a mean of 0 and a standard deviation of 1, except the one-hot ones.

2) *Principal component analysis (PCA)*. PCA can remove some redundant features to improve training efficiency and also prevent overfitting efficiently.

4) *Nonlinear feature expansion*. Considering the machine learning methods in this paper mostly have a linear prediction function while in reality the features usually have a nonlinear complex relationship, the authors expanded each feature using a polynomial approach, which transforms each feature vector x_i into a new feature matrix with a polynomial degree of m,

$$x_i \Rightarrow [x_i, x_i^2, x_i^3, ..., x_i^m] \tag{2}$$

Moreover, several cross-terms, which are derived from the product of different features, were added to augment the difference between four sample groups and also add more degrees of freedom for nonlinear fitting.

C. Regression Models

Six basic machine learning regression techniques were used to examine which one has the best accuracy. Normally, to test and compare the accuracy of these methods, the dataset was split into 2 parts, with 80% of it as training set and the rest 20% as test set. Moreover, as some models have hyper parameters, they were determined by grid search under a cross-fold validation algorithm with 5 folds.

III. Results and Discussions

A. Comparison of Data Cleaning Parameters

Comparison of the results of different cleaning parameter k as mentioned in Section II A is shown in Table I. The results were obtained by logistic regression with 4 polynomial degrees added in the features. It is shown that the best test accuracy is achieved when k equals to 4, thus this is chosen as a standard data cleaning parameter in this paper.

TABLE I
COMPARISON OF LOGISTIC REGRESSION MODELS WITH DIFFERENT OUTLIER REMOVAL PARAMETERS

Cleaning parameter k	Training samples	Training accuracy	Test accuracy
3	179924	81.21%	78.66%
4	189748	80.95%	79.52%
5	194313	80.81%	77.18%
-	200000	61.69%	67.90%

B. Comparison of Regression Techniques

Training and test accuracy of six machine learning models are shown in Table II. Learning rates γ were carefully chosen to make sure the models reach their minimum loss, and hyper parameters λ were determined by cross validation. The least square model did not work because of singular matrix problem, and among the other models, regularized logistic regression has the best accuracy on test set, and the overall performance outcomes go along with theoretical expectations. To achieve the best prediction, regularized logistic regression should be used, but the authors find that this algorithm is very time-costly to converge with added features after feature engineering, so the second best option, ridge regression was chosen to proceed with in the end.

TABLE II
COMPARISON OF DIFFERENT REGRESSION MODELS

Regression model	γ	λ	Training accuracy	Test accuracy
Least squares	-	-	-	-
Least squares (GD)	1e-6	-	77.48%	76.32%
Least squares (SGD)	1e-6	-	76.89%	76.70%
Ridge regression	-	1e-2	81.07%	79.68%
Logistic regression	1e-2	-	80.95%	79.52%
Reg logistic regression	1e-3	1e-1	81.13%	79.75%

C. Comparison of Feature Engineering Methods

Three different feature engineering methods are compared in this section by ridge regression. Firstly, the authors tried to use PCA on the dataset and changed the original feature number from 30 to 27 based on the analysis on cumulative explained variance. However, the results after PCA got worse, so there might be some important information removed during PCA process, and thus this method is abandoned. Secondly, 406 cross-term features were added by multiplying different features vectors, and this greatly improved test accuracy, and thus it is known the problem for ridge regression is underfitting. Moreover, the authors tried to change the polynomial degree from 4 to 5 and 3, and the results both got worse.

TABLE III
COMPARISON OF DIFFERENT FEATURE ENGINEERING METHODS

PCA	Polynomial degree m	Feature interaction	λ	Training accuracy	Test accuracy
/	4	/	1e-2	81.07%	79.68%
✓	4	/	1e-2	79.87%	78.41%
/	4	✓	1e-2	82.87%	82.18%
/	5	✓	1e-2	82.90%	81.52%
/	3	✓	1e-2	81.83%	80.65%

Finally, the ultimate model was chosen to be the third model in Table III, and its λ was determined from a 5 fold cross validation, as shown in Fig. 1, and it finally achieve a test accuracy of 82.0% on the competition platform.

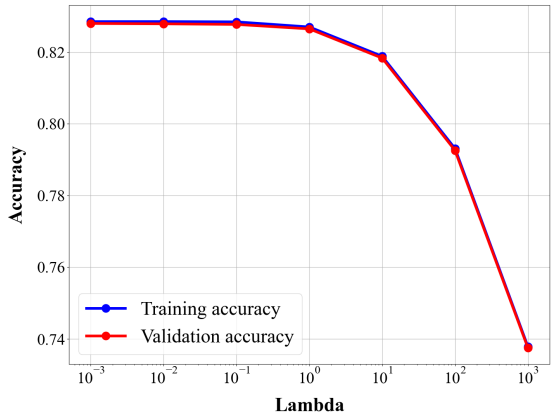

Fig. 1. Cross validation result of ridge regression model

IV. Conclusions

- For basic machine learning methods like ridge regression and logistic regression, the underfitting problem is much more significant than overfitting.
- Accuracy on the test set improved about 15% by proper data cleaning and feature engineering methods in this study. Therefore, compared with prediction algorithms, the data itself can be more important, as it provides an upper bound for prediction accuracy, and prediction models are adopted just to approach this bound.
- In practice, when choosing prediction models, not only accuracy but also time expense should be considered. Moreover, in theory, ridge regression should have much worse performance than regularized logistic regression, but this is not the case in this project. As is said in the No Free Lunch Theorem, we can not have a best universal algorithm, but only try to figure out what is the best for a specific dataset. Therefore, try more and do more.

REFERENCES

[1] P. W. Higgs, "Broken symmetries and the masses of gauge bosons," Physical Review Letters, vol. 13, no. 16, p. 508, 1964.

[2] ——, "Spontaneous symmetry breakdown without massless bosons," Physical review, vol. 145, no. 4, p. 1156, 1966.