Machine learning report

FENG Wentao, WANG Yunbei, ZHUANG Ying
Department of Electrical Engineering, EPFL, Switzerland

***Abstract*—In this paper, we show the methods of processing the feature and the model we developed which is used to predict whether or not a given observation is a Higgs boson. We used the logistic regression model and the ridge regression model to predict on the test set and achieved the highest accuracy of 0.831 and 0.822 respectively. Finally, we chose to submit the ridge regression model since it has reasonable computing time and acceptable accuracy.**

I. Introduction

In this project, we aim to simulate the process of discovering the Higgs particle since Higgs particles are essential for explaining why other particles have mass. Our goal is to train a binary classifier by using given training set and then use the obtained model to predict whether an event was signal (a Higgs boson) or a background (something else). The methods we used to process the data, and the models we chose to train are shown in Section II. The result is written in Section III. Also, we analyze the strengths and the weaknesses of our approaches and compare our models in Section IV. Section V is the summary of the project.

II. Models and Methods

First, we began with observing the features of data based on the original CSV file, and we found a large amount of -999 appeared dispersedly in the data set. Then we found that the column *PRI_jet_num*, which consists of the integer number from 0 to 3, has a strong connection with the appearance of the value -999. -999 appears in some columns when the value of *PRI_jet_num* is equal to 0 or 1 and disappears when this value is 2 or 3.

So, in our first test, we decided to drop all the rows whose *PRI_jet_num* is 0. However, there are still many invalid data as we chose to reserve the samples with their *PRI_jet_num* equal to 1. We hence plotted the distribution of value for every feature. According to these figures, we dropped the column where more than half of their data are -999. Because they have little useful data to analyze (ex: *DER_deltaeta_jet_jet*). We also dropped the features which have the uniform distribution or Gaussian distribution for that we could not distinguish two types of labels from their distribution (ex: *PRI_lep_phi*, *PRI_lep_eta*).

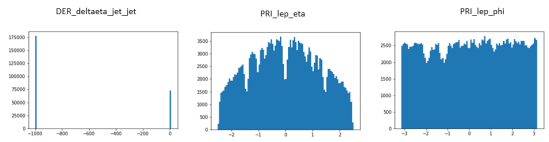

Fig. 1: three types of features that we decided to drop

As for the rest of the features, there are some features strongly polarized which is quite valuable for the binary classifier. Furthermore, some features concentrate around one central value and have little invalid data. We thus decided to replace their -999 by their median.

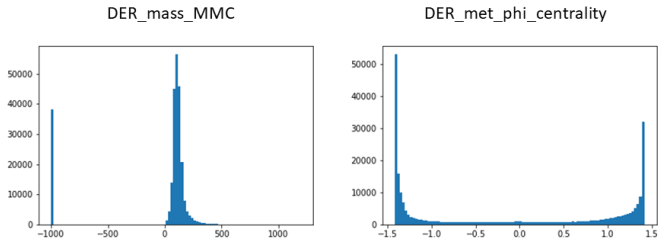

Fig. 2: two types of features that we decided to use in the test

Then we used python to perform the data processing: At the beginning, we removed the label *s* and *b* in the original file and replaced them by 1 and -1 respectively. After that, we trained the logistic regression model [1], [2] with the processed data and utilized the obtained model to predict on the test set. The result of the first attempt is 0.740.

The second time, we normalized our data (used in the first experiment) utilizing the standard normalization method, trained a new model and tested the model on the test set. This time, we noticed that even though the accuracy of the prediction did not improve much, the computation efficiency and space efficiency improved dramatically. Thus, we decided to keep normalizing the data in every following experiment.

The third time, we classified all samples into four categories according to the value of *PRI_jet_num*, and each type of samples corresponds to the value of *PRI_jet_num* of 0, 1, 2, and 3. This time, we processed the four types of samples individually. Furthermore, for each group, we divided them into two subgroups according to their labels and drew the distribution of the feature. So we obtained, for each feature, eight pictures. This time, we tried to study the correlation between two features. We picked every pair of two features and calculated their coefficient of correlation [3]. Then we listed all pairs of features which have a coefficient higher than 0.8. That means they have a strong correlation. So for each pair, we saved only one feature out of two and dropped the other one. In our case, we left the column *DER_mass_vis*, *DER_pt_h* and *DER_sum_pt*. Furthermore, unlike the first attempt (dropped the column which has too many invalid data or has the uniform distribution), this time, we kept the features with the Gaussian distribution. Then we trained four

models separately based on the data with distinct values of *PRI_jet_num* and developed four logistic regression models which have different parameters. We used these models later to predict the labels of our test data. The final score improved to 0.758.

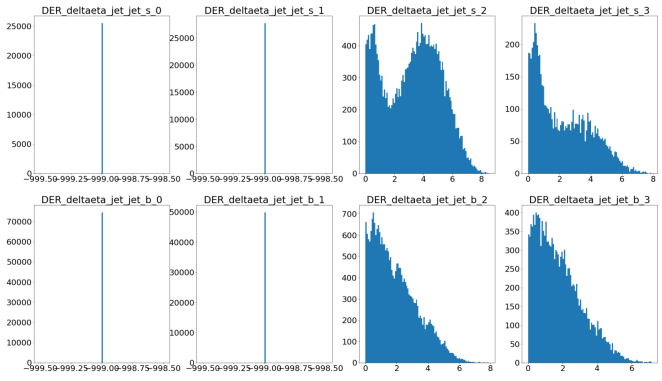

Fig. 3: feature distribution based on the value of *PRI_jet_num*

In the fourth experiment, we continued to analyze our features. First, we calculated the log of the features whose values are all positive or calculated the normalized values otherwise. These processes significantly saved our computed cost and storage cost. Then, we used the polynomial with a degree three to augment the four categories of features that we prepared during our third test, and each term in the polynomial has a weight 1 [4]. Next, we fed these data to four models respectively and used the obtained models to make the prediction. This time, the accuracy has significantly improved to about 0.831.

For the last experiment, we changed the logistics regression model to the ridge regression model and then used the prepared samples in the fourth experiment (feature augmentation through polynomial with a degree three) to train the model and obtained four ridge regression models with different parameters. We performed these models on the test set and got an accuracy of 0.822. Although the correct rate decreased, the calculation speed was considerably improved.

III. Results

In the fourth and the last experiments, we fed two different types of models with the same prepared samples. The accuracy of the logistics regression model is 0.831 which is a little bit higher than the correct rate 0.822 of the ridge regression model. Nevertheless, the computation efficiency and storage efficiency are improved dramatically with using the ridge regression model. The calculation time required for the ridge regression model is reduced by at least a factor of ten compared to the logistics regression one.

IV. Discussion

In our last two experiments, we used the nonlinear logistic regression model as well as the linear ridge regression model to perform a fit over the same processed features. From the result, the logistic regression model is more accurate than the linear ridge regression model.

For samples, we used a polynomial with a degree of 3 to augment features, which improved the accuracy of the model. However, feature augmentation increases the amount of training data, which increases computational complexity and hence dramatically lengthens the time required to train the model. Furthermore, if we continued to increase the degree, for instance to 4, the precision of prediction of logistic regression model increased slightly. However, the computing time extended significantly. This implies that a higher degree of the polynomial is no longer very suitable for our case of studying.

For models, the iteration step size of the logistic regression model cannot be set too large; otherwise, the model is prone to diverge, which results in numerous steps of iterations and slow convergence for training model. We can solve this problem by Newton iteration. However, the Newton method needs to calculate the Hession matrix for each iteration. When the sample is large, the Hession matrix is hence large and is easy to cause the kernel to collapse. For ridge regression, although the accuracy of linear regression is lower than the logistic regression model, the computation is simpler, and training model consumes less time and less memory. Furthermore, we can use a polynomial to enhance the feature, but we need to pay attention to that when the degree is high, the model may overfit.

As for future studying on this project, we need to understand the realistic meaning and connection of the features. Then, we need to train the data with a suitable model based on physics knowledge to find the best solution to the specific subject.

V. Summary

By analyzing a real-world data set, we have had a deep understanding and learning of the data processing, as well as feature analyzing. Additionally, we studied both the advantages and disadvantages of different models of machine learning. Finally, after tens of tests, we achieved an accuracy of 0.822 with a short computing time using the ridge regression model.

This project gave us an opportunity to use the theoretical from our textbook to the real-world application. The experience and the lessons that we have gained during the project are beneficial in our future studying.

References

[1] M. E. Khan, “Logistic regression.” 2018.
[2] F. Pedregosa, “sklearn.linear_model.logisticregression @ONLINE,” 2011.
[3] J. Brownlee, “Discover feature engineering, how to engineer features and how to get good at it @ONLINE,” September 2014.
[4] F. Pedregosa, “sklearn.preprocessing.polynomialfeatures @ONLINE,” 2011.