A method to solve the Higgs boson challenge

CS-433 Machine Learning - Project 1

C. Tsai, G. Orsi, V. Rossi

***Abstract*—In this paper, we propose a method to solve the Higgs boson classification challenge. We combined statistical feature engineering with physics knowledge to enhance the provided dataset, and we applied logistic and least squares regressions to build an effective classification model.**

I. Introduction

The Higgs boson challenge is a machine learning competition organized by CERN to encourage the cooperation between physicists and data scientists [1]. Data generated from the high-energy particle experiments at CERN is given to the public to solve the classification problem of whether an observation corresponds to a detection of the Higgs boson or background noise.

In this paper, we provide the set of steps that led to training a model able to correctly classify 83% of the observations provided in the test dataset.

We decided to dive deeper into the physical meaning of the features to have a better interpretation of the data and we used statistical methods to improve the performance of the model.

II. Method

Our first step is to transform the raw data into meaningful features. Based on the theoretical prediction of Higgs bosons [2], they are estimated to have a mass around 125 GeV/c^2, no charge nor spin, and positive parity. Since we don't have direct access to all these measurements, we try to transform the given data to the meaning of square of energy. (E_H^2). The idea is to predict the label as a detected Higgs boson if our estimation $E_H^2 - 125^2 > 0$ and vice versa. Although ideally, we want to only identify Higgs bosons in a narrow window centered at 125 GeV, since Higgs bosons are heavier than most known particles, we only consider the case that its total energy is larger than the threshold [2].

A. Energy Representation

We can measure energy by the mass-energy relation [3],

$$E^2 = (pc)^2 + (m_0c^2)^2 \tag{1}$$

where E is energy, p is momentum, m_0 is the rest mass of the particle, and c is the speed of light.

By (1), we can approximate E_H^2 by the linear combination of the square of every data, which has the meaning of either energy (E), mass (M), or momentum (P). Since these terms represent the descendants of the Higgs bosons, we can reconstruct the energy of a Higgs boson by these terms as follows.

$$\tilde{E}_0^2 = \sum_{i,j,k} \text{poly}(E_i, 2) + \text{poly}(M_j, 2) + \text{poly}(P_k, 2) \tag{2}$$

where $\tilde{E}_0$ is a simple estimation of the energy of Higgs bosons. $\text{poly}(x, t) = \sum_{i=0}^{t} w_i x^i$ represents the polynomial expansion of x up to t-th power for some coefficients w_i.

B. Compensation Terms

Let's call the energy, mass, and momentum terms the *intrinsic measurements*. These measurements need to be adjusted to capture the imperfectness of the experiments. We can also approximate this imperfectness by the data other than the intrinsic measurements. We assume these terms linearly or inversely affect the intrinsic measurements. Because most of them estimate the deviation from the ideal measurements, and they are usually small.

So our estimation of the energy of Higgs bosons is

$$\begin{aligned} E_H^2 = &\tilde{E}_0^2 + \sum d \,\text{poly}(\text{related } E \text{ or } M \text{ or } P, 2) \\ &+ \sum d^{-1} \,\text{poly}(\text{related } E \text{ or } M \text{ or } P, 2) \end{aligned} \tag{3}$$

for some extra deviation terms d.

C. Model Categorization

From the documentation, we know that `PRI_jet_num` serves as a categorical indicator that some entries are available or not. So we can divide the data set into three subgroups with `PRI_jet_num=0, =1, or` $\geq$ `2` respectively. We will later train three different models for these different kinds of groups.

D. Data Cleaning

A large portion of values from the first column of the dataset, indicating mass, was missing. We decided to replace those values with the median of the respective feature. We also considered a different approach, which involved computing medians from further subgroups created from a label-based splitting. We ended up discarding this second option because it seemed preferable to apply a method that would handle missing values in both training and testing data sets in the same way, which was not feasible due to the

absence of labels in the latter set and could lead to unknown model behavior.

After all the processing described above, we standardized all values in each feature. During the training of the models, we also computed the cross products of each pair of feature and added that in the training set.

E. Machine Learning Objective

Combining everything above, we can solve the problem in the following manner:

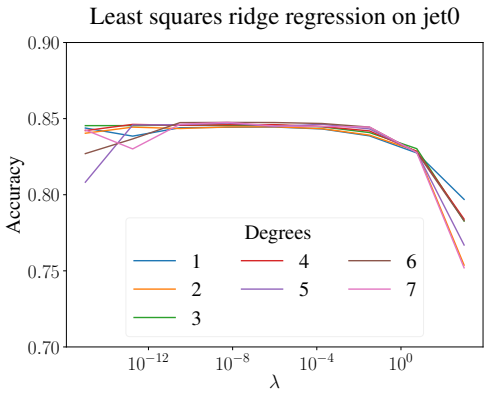
1) Transform the training dataset into polynomial as defined in section II-A and section II-B.
2) Divide the dataset into three subgroups and process the data as mentioned in section II-C and section II-D.
3) Select a model (least square or logistic) and train the best coefficients for the polynomial.

As a result, we have three optimized models for each type of data. We can then predict the label for any new data by applying the corresponding model to each subgroup.

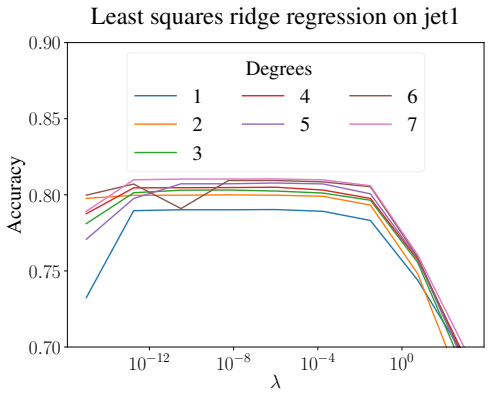
III. Results

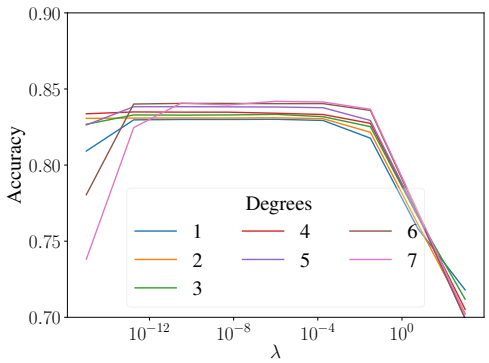
After analysing both the logistic regression and the least squares models, we realized that the least squares would produce better results on average. Therefore, we focused on tweaking the hyper-parameters only in the latter model.

The figures below show the best accuracy obtained with each degree of polynomial expansion and λ for the least squares models trained on the subsets defined by `PRI_jet_num`. The accuracy is computed by using 4-fold cross-validation.



The final result obtained in the provided test dataset was 0.836 accuracy and a 0.751 F1 score using the hyper-parameters shown in the table below. We trained the models also on higher degrees but we realized that the models were overfitting the train dataset, so we chose the degrees shown in table.





`PRI_jet_num`	λ	Degree
0	1e-7	6
1	1e-6	7
2, 3	1e-5	7

Table I
FINAL HYPER-PARAMETERS

References

[1] C. Adam-Bourdarios, G. Cowan, C. Germain, I. Guyon, B. Kégl, and D. Rousseau, "The Higgs boson machine learning challenge," in *Proceedings of the NIPS 2014 Workshop on High-energy Physics and Machine Learning*, ser. Proceedings of Machine Learning Research, G. Cowan, C. Germain, I. Guyon, B. Kégl, and D. Rousseau, Eds., vol. 42. Montreal, Canada: PMLR, 13 Dec 2015, pp. 19–55. [Online]. Available: https://proceedings.mlr.press/v42/cowa14.html

[2] O. Moreira, *Modern physics*. Oakville, Ontario: Arcler Press, 2020.

[3] A. Beiser, *Concepts of Modern Physics: 6th Edition*, ser. Concepts of Modern Physics. McGraw-Hill, 1994.