A New Regularized Logistic Regression Method to detect Higgs Boson production

Marbot Tanguy, Pecchini Elena, Echeverry Hoyos Mateo
Machine Learning, EPFL Lausanne, Switzerland

***Abstract*—The Higgs Boson is an elementary particle able to explain why fundamental particles in our universe have mass. At CERN, physicists recreate the process of discovering the Higgs Boson in order to study it. However, estimating whether a Higgs Boson is produced is not a trivial task due to its instability. Machine learning algorithms are used to solve this problem. In the context of the Higgs Boson machine learning challenge, we developed a model based on regularized logistic regression which is able to perform binary classification with an accuracy of 79%. We propose to handle missing data by substituting them with the mean of the feature vector to which they belong; and to remove features that are correlated. Results show that our model is generalizable as it performs well on unknown data, and represents a good trade-off between accuracy and complexity.**

I. Introduction

The Higgs Boson is an elementary particle in the Standard Model of particle physics produced by the quantum excitation of the Higgs field [1]. It is central in the Higgs mechanism, which explains why gauge bosons have mass: this mechanism is of key importance to reconcile theory and experimental evidence, as measurements show that bosons have masses. At CERN, where the Higgs boson was discovered in 2013, physicists recreate the process of discovering this elementary particle by smashing protons into one another at high speed. To estimate whether a Higgs boson is produced, the decay signature of the collision event is analyzed [2]. However, this is not a trivial task as the decay signature of a Higgs boson can look similar to that of some other particles. For this reason, machine learning algorithms are needed to predict whether an event is the result of a Higgs boson or the result of something else (binary classification task). To address this need, in 2014 the Higgs boson machine learning challenge was launched by ATLAS. The dataset consists of 250000 samples, each described by 30 features. Notably, variables that are meaningless or cannot be computed are marked with an extreme value of -999 [2]. In the context of this challenge, several models have been developed over the years. However, none of them is free from limitations. The goal of our analysis is to find a model that represents a good trade-off between accuracy and complexity.

II. Models and methods

The model we developed to perform binary classification is regularized logistic regression. The method used to find the optimal weights is gradient descent.

The model selection process is detailed hereafter. The first step consists of exploratory data analysis: the distribution of features can be seen in Fig.1. The second step consists of handling missing values, which are set to -999 in the dataset. In our feature matrix each missing value is replaced with the mean of the feature vector to which it belongs, computed considering only defined values. This step is crucial to maintain the distribution of defined values after standardization.

Then, data cleaning and feature selection is performed. Model selection is performed by means of backward elimination, meaning that the starting model includes all the given features, and then the least significant variables are removed one after the other [3]. More in detail, linear correlation between features is checked using cross-correlation (see Fig.2). Then, non-linear relationships between features are checked using Spearman correlation. Features that are linearly correlated are removed in order to obtain a simpler model: 13 features are eliminated according to this criterion. Finally, polynomial expansion is performed for each of the remaining features up to degree 3. Then, features are standardized, meaning that they are rescaled such that they have the properties of a standard normal distribution with a mean of zero and a standard deviation of one. This step is necessary for many machine learning algorithms to work properly (e.g., linear regression, logistic regression, ...) [4]. Another reason why standardization is performed in the context of this analysis is that it enables a faster convergence of gradient descent.

The next step consists of choosing the function that better approximates the distribution of the data. As previously mentioned, our choice is logistic regression as this algorithm performs well in binary classification tasks. Then, a ridge regularization term is added, as it increases the stability of the model. Model parameters are optimized. Finally, the performance of the prediction model is assessed by means of 8 fold cross validation (loss and accuracy are computed), using 1000 training samples.

III. Results

As shown in Fig.1, many features have undefined values (marked with a value of -999): more in detail, there are 7 features with 70.5% of undefined values and 4 features with a percentage between 15% and 40%.

In addition, features differ significantly in terms of mean and standard deviation. For example, the minimum mean

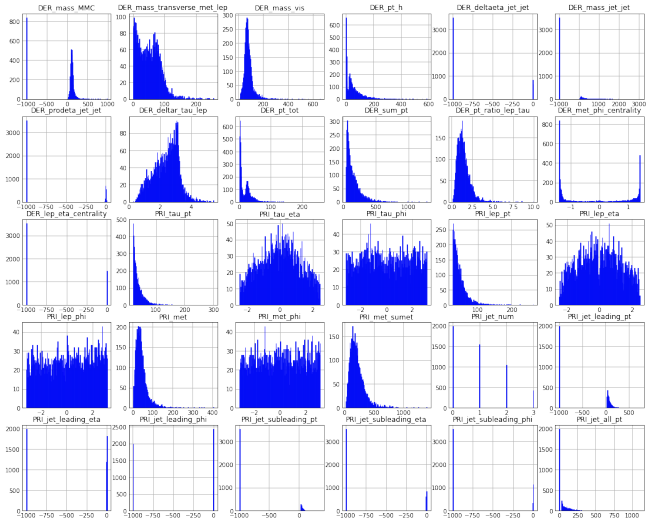

Figure 1. Features distributions. Undefined values are set to -999.

is -0.796 for the feature DER_prodeta_jet_jet; whereas the maximum mean is 366.297 HeV for the feature DER_mass_jet_jet.

As visible in Fig.2, there are clusters of correlated features. For example, features 4,5,6,12,26,27,28 show a correlation coefficient between 0.95 and 1.

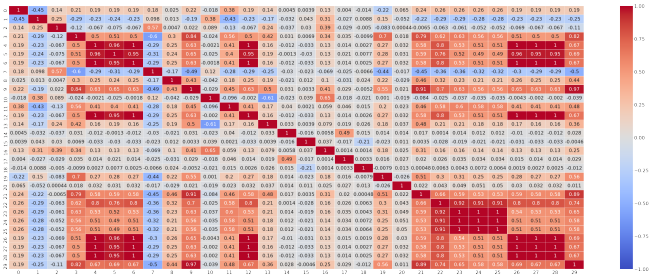

Figure 2. Correlation heat map of the features using cross-correlation. Red = strong correlation, white = no correlation, blue = strong anticorrelation.

The step size γ for gradient descent is set to $2 * 10^{-6}$, which is small enough to avoid divergence. This value of γ is found by means of cross-validation using logistic regression as a model on unprocessed data. Weights are first set to 0.

The test accuracy of several logistic regression models resulting from 8-fold cross-validation is reported in Table I. As we can see, handling the undefined values before standardization as described in Section II improves the accuracy of the model. The best mean accuracy is achieved when correlated features are removed and polynomial expansion of degree 3 is applied to the remaining 17 features. Applying a regularization term has both pros and cons: it decreases the mean accuracy (from 0.733 to 0.728) and it decreases the variance of the prediction (from 0.0395 to 0.0120). Indeed, regularized logistic regression seems to be more stable than logistic regression. As only with regularized logistic regression the mean accuracy is higher than 70% for all runs of the cross-validation, we choose to include a ridge penalty in the final model. The value of $\lambda = 0.278$ is found to be optimal when $\gamma = 2 * 10^{-6}$.

Raw data	0.711 +/- 0.0551
S	0.694 +/- 0.0526
NaN + S	0.714 +/- 0.0537
NaN + S + Poly	0.727 +/- 0.0441
NaN + S + Poly + Feature selection	0.733 +/- 0.0395
NaN + S + Poly + Feature selection + Reg	0.728 +/- 0.0120

Table I
TEST ACCURACY OF LOGISTIC REGRESSION OVER 8 FOLD CROSS-VALIDATION ($\gamma = 2 * 10^{-6}$). S = STANDARDIZATION, NAN = MISSING VALUES HANDLING, POLY = POLYNOMIAL EXPANSION ($d = 3$), REG = REGULARIZATION WITH RIDGE PENALTY ($\lambda = 0.278$)

Finally, we found that the combination of $\gamma = 0.02$ and $\lambda = 8.1 * 10^{-5}$ gives a test accuracy of 79%, keeping the other parameters equal and applying the feature processing described above (standardization, handling missing values, feature selection, polynomial expansion of degree 3). This is our best score on the submission platform.

IV. DISCUSSION

The presented model can be used to predict whether an event is the result of a Higgs boson or the result of something else. Some strengths of our model lie in the choice of logistic regression. Logistic regression is easy to implement, easy to interpret and very efficient to train. Moreover, it makes no assumptions about distributions of classes in feature space. It is very fast at classifying unknown records and not inclined to over-fitting. It handles well extreme values, as the linear function is mapped into a sigmoid function. However, logistic regression has also limitations, as it assumes a linear relationship between the dependent variable and the independent variables, which is often not the case in real-word scenarios, and it constructs linear boundaries [5]. Another limitation of our model lies in the fact that its accuracy is 79%, meaning that on average 20% of the data are misclassified. Other strengths of our model are that it standardizes features, meaning that model performance is not affected by the different values range of features, and that it is able to efficiently handle undefined values. Moreover, the model complexity represents a good trade-off between bias and variance. Notably, adding the ridge regularization term to the loss enables a low variance and a good accuracy. Another advantage of the regularization term is that it avoids the model to diverge when data are linearly separable.

V. SUMMARY

A new model for binary classification in the context of the Higgs boson machine learning challenge is here presented. The model is based on regularized logistic regression, which makes it simple and efficient, and uses a new approach to handle missing data. Notably, this new model enables a good trade-off between accuracy and complexity.

REFERENCES

[1] Wikipedia. Higgs boson. [Online]. Available: https://en.wikipedia.org/wiki/Higgs_boson

[2] C. G. I. G. D. R. Claire Adam-Bourdariosa, Glen Cowanb. (2014) Learning to discover: the higgs boson machine learning challenge. [Online]. Available: https://higgsml.lal.in2p3.fr/files/2014/04/documentation_v1.8.pdf

[3] S. Srinidhi. Backward elimination for feature selection in machine learning. [Online]. Available: https://towardsdatascience.com/backward-elimination-for-feature-selection-in-machine-learning-c6a3a8f8cef4g

[4] Wikipedia. Feature scaling. [Online]. Available: https://en.wikipedia.org/wiki/Feature_scaling

[5] G. Chauhan. All about logistic regression. [Online]. Available: https://towardsdatascience.com/logistic-regression-b0af09cdb8ad