Machine Learning Course, Class Project 1

Guerraoui Nada, Peeters Thomas, Petersen Molly
School of Computer and Communication Sciences, EPFL, Switzerland

***Abstract*—In this report we will talk about a Higgs Boson classification. First, we will talk about the exploratory data analysis we performed to analyse the different features to keep the most relevant ones. Then, we will talk about the pre-processing and feature engineering done so far to clean the data and extract useful information. We will also talk about all the models we have made. Finally, we will discuss the results.**

I. Introduction

A. About the Problem

The Higgs Boson is a particle discovered in 2013 at CERN in Geneva Switzerland, although scientists have speculated about their existence and contribution to mass of other particles for decades prior to their discovery. These particles decay quickly into other particles, called ”channels”. These particles make up a unique ”decay signature” that allow scientists to identify Higgs Boson particles, as they are unable to observe them directly.

B. Machine Learning for Classification

The Higgs Boson challenge took place on Kaggle. The goal originally was to encourage cross discipline collaboration between physicists and data scientists. Here we have for our training set 250 000 samples with 30 features and for the testing set 568 238 samples. We use both logistic and linear regression techniques to identify Higgs Boson events. We first implemented 6 basic methods using 6 different methods for the weight computation. We implemented later 6 more advanced models using new methods for the preprocessing.

II. Models and Methods

We first did the classification using 6 basic models. For the pre-processing part of each of these models, we kept the 30 different features, we remove the outliers using interquartile method and then we standardize the features so that they have zero mean and a standard value equal to 1. We then use 6 different methods to compute the weights: least squares gradient descent (model 1), least squares stochastic gradient descent (model 2), least squares (model 3), ridge regression (model 4), logistic regression (model 5) and regularized logistic regression (model 6). We obtained the best results when we used logistic and regularized logistic regression (0.74 on AIcrowd, preform well for binary classification). We then try to improve the pre-processiong of the data to get even better results.

We preformed classification with six more complex algorithms: (A) a linear regression with continuous and standardised variables and γ = 0.05, (B) logistic regression with two separate models based on complete case status with independent variables modeled as binary (0/1) and γ = 0.5, (C) logistic regression with four strata for the `JET number` with independent variables modeled as binary and γ = 0.9, (D) a penalized logistic regression with standardised features all undergoing polynomial and trigonometric expansion, (E) logistic regression with four strata for the `JET number` with independent variables all undergoing polynomial and trigonometric expansion with degree 5, and (F) a penalized logistic regression with four strata for the `JET number` with independent variables all undergoing polynomial and trigonometric expansion.
Model performance was judged on percentage of events correctly classified. All model weights were optimized using gradient decent, with gamma values determined by cross validation.

A. Feature Engineering

All variables were investigated for missingness. Missingness was highly associated with `JET number` since many variables were undefined if `JET number` = 0. Overall variable missingness ranged form 0 to 71%. Missing data also was more common among non-events, likely due to the association with `JET number`.

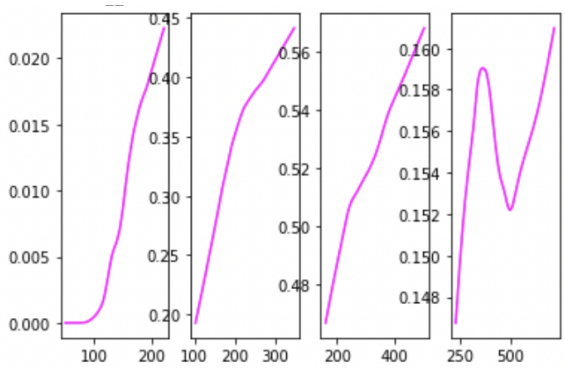

Figure 1. Example of lowess plot used to decide binary cutoff points for feature augmentation. Y-axis is percent with Higgs Boson event.

Three methods were employed for feature engineering: standardised continuous features, converting all independent

Model	Training Score (%)	Test (AICrowd) Score (%)
(A) Two-model linear regression with continuous and standardised variables	79.6	79.9
(B) Two-model Logistic Regression with Binary Predictors	77.4	77.5
(C) Four-model Logistic Regression with Binary Predictors	79.4	79.5
(D) Penalized Logistic Regression with Polynomial and Trigonometric Expansion	80.5	81.3
(E) Four-model Logistic Regression with Polynomial and Trigonometric Expansion	82.5	82.5
(F) Four-model Penalized Logistic Regression with Polynomial and Trigonometric Expansion	82.5	82.5

Table I
PERFORMANCE OF DIFFERENT MODELS ON CLASSIFICATION OF HIGGS-BOSON EVENT.

features to binary values(models A, B and C), and polynomial and trigonometric expansion (models D and E). Cutoffs for each variable were decided by visually inspecting the data using lowess (LOcally WEighted Scatterplot Smoothing) curves, which fit local polynomial linear regressions to individual points allowing for the visualization of local trends [1]. Binary cutoff points were determined separately for each model within the different algorithms (complete case vs missing, and by `JET number`). An example of a lowess curve stratified by `JET number` is shown in Figure 1. The figure shows the distribution of the outcome across values of `PRI_met_sumet`, and additionally demonstrates the variability of the range of values this variable takes across `JET number`. Important to note, the lowess plots do not indicate the distribution of values by density or frequency. If the data are skwewd this would not be apparent by these plots. A polynomial value of five was determined through cross validation.

For the second algorithm using two separate models based on complete case status- first a model using only complete cases was built and classified. Then a separate model using all the data, with indicator variables for missing variables (0 for missing and 1 for not missing), and then interaction terms between the missing indicator and the respective variable was created. Variables with missingess were included only as part of this interaction variable and not on their own. Missing values for these variables were set to 0, but the choice is arbitrary since the imputed values will not count towards the model as a result of only including the interaction.

III. RESULTS

Overall, our best model was the four model penalized logistic regression with polynomial and trigonometric expansion (model F) with an overall training accuracy of 82.5 and a test (AICrowd) accuracy of 82.5. All model results are in Table I. The training accuracy was similar to test accuracy on AICrwod for all models, which suggests the training and test data were separated completely at random and that our models does not overfit. We compute the value of the step size γ using cross-validation. Here are the plot of score in function of gamma for model (F) (See figure 2).

When looking at how the models performed on the training data by outcome, the models generally had higher accuracy with correctly predicting non-events as opposed to events. As overall model accuracy increased, this was often due to an increase in correctly predicting events as opposed to an increase in correctly predicting non-events. For example, in model A, 86.7% of non events were classified correctly, while 60.0% of events were. With model F, 88.1% of non-events were correctly classified while the percentage of correctly classified events jumped to 71.8%.

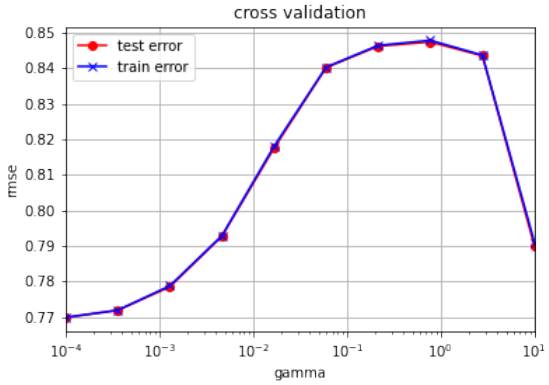

Figure 2. Prediction score obtained for samples with Jet number = 0 using model F

Models B and C, which differ by how the data was separated (missingness vs. `JET number`), performed similarly. This is unsurprising since missingness was associated highly with `JET number`. Even though the different levels of `JET number` had different binary cutoffs (ie `JET number` 3 vs. 4, which were combined in the the "complete case" model in B and therefore would have shared the same cutoff), tailoring the cutoffs seem to have little effect on accuracy.

IV. DISCUSSION

Our results found that, even though logistic regression is a basic technique for classification tasks, it performed overall moderately well. Our feature engineering was pretty basic, more creative ways of modeling the independent variables may have been able to capture the relationship with the outcome better. Additionally, there exists many more sophisticated classification methods that could potentially surpass our models predictions. For example neural networks have been shown to be successful in modeling very complicated relationships.

REFERENCES

[1] W. S. Cleveland, "Robust locally weighted regression and smoothing scatterplots," *Journal of the American Statistical Association*, vol. 74, pp. 829–836, 1979.