Machine Learning: Projet 1
Higgs boson detection challenge

Mathieu Marchand, Wanting Li, Kieran Vaudaux
EPFL, Switzerland

***Abstract*—In the first class project of Machine Learning, we applied concepts we have learned in the lectures and practiced on Higgs-Boson dataset to predict whether the decay signature resulted from a Higgs boson (signal) or other particles (background). We developed six different machine learning algorithms and evaluated their performances.**

I. Introduction

The Higgs boson is an elementary particle in the Standard Model of physics produced by the quantum excitation of the Higgs field, the existence of which explains why other particles have mass. The faint signal of the Higgs boson was initially discovered in 2012 by statistical analysis of enormous amounts of data from the international ATLAS and CMS collaborations at the Large Hadron Collider (LHC) at CERN near Geneva, Switzerland.

In this project, the objective is to build a binary classifier to label collision events as either a Higgs boson signal or background noise based on the 30 parameters provided. After conducting exploratory data analysis, data transformation and feature augmentation, we applied several machine learning algorithms – Gradient Descent, Stochastic Gradient Descent, Least Squares, Ridge Regression, Logistic Regression and Regularized Logistic Regression – to generate predictions and compare the test performance of these models. Further implementation details and discussions are presented in Section II and Section III.

II. Methodology

The first crucial step in this project is to understand the data we have so that we can understand how to use this dataset to build the best possible model. The dataset we have at our disposal is made up of $250,000$ observations, each grouping 30 variables. Part of these variables are "raw" (primitive) quantities about the bunch collision as measured by the detector and the other part of the variables are values derived from the primitive variables.

A. Data cleaning

1) Data splitting: All these variables have continuous values, except for the variable named *PRI_jet_num* which can take the value $0, 1, 2$ or 3. Depending on the value of this variable some variables are meaningless and take the value -999. Moreover, the variable *DER_mass_MMC* contains missing values, signified by the value -999. We have therefore chosen to separate the dataset into three sub-datasets, one for each of the following cases:

- **Subset 0:** *PRI_jet_num* $= 0$,
- **Subset 1:** *PRI_jet_num* $= 1$ and
- **Subset 2:** *PRI_jet_num* $\in \{2, 3\}$.

This choice comes from the fact that for each of these three sub-datasets, there is a set of variables that don't have a physical meaning.

2) Missing values: However, for the variable *DER_mass_MMC* we cannot act in the same way because it has missing values in each of the sub-datasets that we can consider, however as this value is not meaningless but only could not be calculated for some observations. By trying to estimate these missing values by the mean of this variable in each of the three sub-datasets after deleting them, we noticed that there were a large number of observations that could be considered as outliers, so we preferred to use the median as an estimate of the missing values.

3) Outliers: We then looked at the presence of outliers in our data. Considering the classical characterisation of outliers, it turned out that there were on average more than 5000 outliers per variable. Moreover, as the majority of outliers appeared on different observations and variables, removing these extreme data from the sub-dataset did not seem to be an acceptable solution.

B. Data transformation and data augmentation

1) Data transformation: Unable to simply remove the outliers, we deepened our data analysis and by looking at the distribution of our data we noticed that a large number of the variables were positively skewed with a heavy distribution tail (see Fig. 1), which explains the presence of so many outliers. To deal with this, we applied the following logarithmic transformation to the variables that had a skewed distribution:

$$\mathbf{x} \to \log\left(\mathbf{x} - \min_{i=1,\dots,n}\{\mathbf{x}_i\} + 1\right)$$

. Figure 1 shows the impact of this transformation on the variable *DER_sum_pt*. On the one hand, it allows us to transform the extreme values so that we no longer have outliers, but it also allows us to make the distribution of the data closer to a normal distribution, which will help us when training our models.

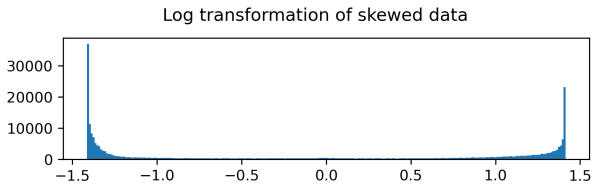

Fig. 1: Transformation of positively skewed data with a $\log$ transformation.

Moreover, by analysing the distribution of the data we were able to observe the presence of two variables, *DER_met_phi_centrality* and *DER_lep_eta_centrality* which seemed to have a bimodal distribution, i.e. **??**. We therefore chose to introduce an indicator variable which takes the value 1 if the variable is less than its median and 0 otherwise. In this way, we hope to succeed in capturing an important aspect of these variables.

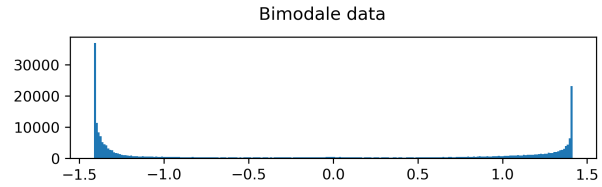

Fig. 2: Bimodal data

It turned out that we obtained better results by normalising the data we had just transformed. To do this, we applied the following normalisation:

$$\mathbf{x} \to \frac{\mathbf{x} - \bar{\mathbf{x}}}{\text{std}(\mathbf{x})}$$

2) Data augmentation: Another important point of this project is the increase of the dimension of the space of our features for, in this way we hope to succeed in making our data almost separable in this new space and thus to succeed in classifying them at best. The **Subset 2** being the only one to have two possible values for the variable *PRI_jet_num*, the addition of an indicator variable for each possibility seemed appropriate.

Then, by reading the documentation associated with the dataset, and particularly appendix A.3, we were able to see how to reconstruct variables that made physical sense from the ones we had, such as the moment vectors for the different particles. This seemed to be an interesting and logical way to add new variables.

The number of these new variables varied according to the sub-dataset, but this added a minimum of 16 variables. However, even if this allowed for better results on intermediate tests, these were still unsatisfactory. To further increase the size of the feature space, the variables in each dataset were separated into three groups, one for variables that were angles, one for variables in the initial dataset that were not angles and one for variables constructed from the initial dataset. Thus, with the size of these subgroups being more reasonable, it was now possible to construct polynomial permutations between the variables in these groups of degrees higher than two, which was not possible without having divided the variables into groups. Before considering polynomial permutations between the variables of the angle group, it was more natural to first transform these angles by considering the differences and sums of these angles and then to take the cosines and sines of these new angles, and do this for different frequencies.

Once the data augmentation was complete, the data was normalised using the same transformation as before so that the variables were unit-less and of the same order of magnitude.

III. Results

A. Optimization

1) Method: Among the six methods implemented, we had to choose which method to use for the final training of our models. With the increase in data, it turned out that the least squares and ridge regression methods were not usable due to the presence of too many relationships between the variables. A solution would have been to perform a Principal Component Analysis (PCA) to keep only the variables explaining the majority of the variance present in the observations, but we preferred to focus on the other methods. Secondly, the stochastic gradient descent method with a batch size of 1 was too unstable on the tests we performed. Finally, on these same tests, the logistic regression methods gave slightly worse results than the gradient descent method, moreover the gradient descent method converged to a vector of weights whose Euclidean norm was of the same order of magnitude as our observations, so it did not seem imperative to us to regularise the weights with the logistic regression method.

For the management of the learning rate, the implementation of a scheduler was necessary. In this way, the learning rate was divided by 5 if the model loss increased more than a certain number of times consecutively. A patience of 2 was chosen experimentally.

2) Model: After several tests, the ones that gave the best results for the subdatasets were:

- **Subset 0 & Subset 1**: The variables reconstructed from the dataset documentation are present up to the 13th power, with cross-terms up to degree 2. Then, the transformed variables of the initial dataset that are not angles are present up to the 16th power, with cross-terms up to degree 2. And finally, the sines and cosines of the sums and differences of the variables of the initial dataset that are angles are present up to the 7th power with cross-terms up to degree 2.
- **Subset 2**: The variables reconstructed from the dataset documentation are present up to the power of 2 with all cross-terms. Then, the transformed variables of the initial dataset which are not angles, are present up to the 3rd order with all cross-terms. And finally, the sines and cosines of the sums and differences of the variables of the initial dataset that are angles are only present to the power of 1.

IV. Conclusions

This project allowed us to see how crucial the data cleaning and data processing of a dataset could be, even if it was already quite well pre-processed. This could only be done thanks to the visualization of our data on different graphs, an expertise in the studied domain (here physics) can even be important to interpret the available variables and to know how to make the most of them without making our model grow exponentially. Finally, we managed to obtain an accuracy of 0.795 on our best submission.

To improve our results, after having proceeded with the data augmentation, it would have been interesting to use a PCA to reduce the space of our features while keeping a maximum of information. This would have allowed us to use the least squares and ridge regression methods which, on small models, gave better results than the gradient descent methods. Another aspect that we would have liked to have had more time to investigate was the generation of new observations from the labeled data that we had. If this had worked, we would have had a way of artificially increasing our number of observations, which would have allowed us to train our model more accurately.