Higgs Boson Challenge : Finding evidence of the particle's presence using Machine Learning

Niccolo Sacchi, Antonio [???] Valentin Nigolian
Department of Computer Science, EPFL Lausanne, Switzerland

***Abstract*—A critical part of scientific discovery is the communication of research findings to peers or the general public. Mastery of the process of scientific communication improves the visibility and impact of research. While this guide is a necessary tool for learning how to write in a manner suitable for publication at a scientific venue, it is by no means sufficient, on its own, to make its reader an accomplished writer. This guide should be a starting point for further development of writing skills.**

I. Introduction

The aim of writing a paper is to infect the mind of your reader with the brilliance of your idea [1]. The hope is that after reading your paper, the audience will be convinced to try out your idea. In other words, it is the medium to transport the idea from your head to your reader's head. In the following section, we show a common structure of scientific papers and briefly outline some tips for writing good papers in Section **??**.

At that point, it is important that the reader is able to reproduce your work [2], [3], [4]. This is why it is also important that if the work has a computational component, the software associated with producing the results are also made available in a useful form. Several guidelines for making your user's experience with your software as painless as possible is given in Section **??**.

This brief guide is by no means sufficient, on its own, to make its reader an accomplished writer. The reader is urged to use the references to further improve his or her writing skills.

II. Models and Methods

There were two main parts of developing our ML system, data analysis and algorithmic design. While the first one focused on the nature, intricacies and interconnexions of the raw data and its preparation, the second one focused on the treatment of said data after refining. Let us now delve a bit further into those two aspects.

A. Data Analysis and Exploration

With a dataset of 250'000 items and 30 features, there was a lot of data to work with. To get a better sense of its nature, we used various mathematical tools. Those were applied on the whole dataset (train + test) when possible and only on the training set when the analysis was related to the prediction. The following are the tools we used :

1) Outliers analysis The first thing to do to get a better sense of the data was to plot it. We thus plotted the whole dataset by feature and observed if could find any outliers or discrepancies. There were indeed a few outliers but considering the vast amount of data otherwise available, we decided to simply ignore them.
2) Distribution analsysis : We wanted to see how features values were distributed over both the signal data and the background data. Indeed, we made the assumptions that if two features had a similar distribution between the two sets, then this feature would only have limited prediction power. Among all features, four were spotted to have the almost exact same distributions. Those are "PRI_tau_phi", "PRI_lep_pt", "PRI_lep_phi" and "PRI_met_phi". We decided to try to drop those columns to see to what extent it changed the resulting predictions, if any.
3) PRI_jet_num-based split : After looking further at the data, we noticed that there were a lot of -999 values spread on the dataset. More importantly, we noticed that the ammount of those values depends greatly on one particular feature, which happens to be the same categorical feature : "PRI_jet_num", representing the number of "jet events" (a physics term unknown to us) occurring at every event. For instance, if this feature has value 0, then 10 other features will have only -999 values and 26% of the first feature will be -999, as per if it has a 3 value, then no feature is all -999 and only 1.4% of the first feature will be -999. For this reason, we decided to drop the following features depending on the "PRI_jet_num" feature's value. This lists the features dropped by their index in the feature list.
   a) jet = 0 : [4, 5, 6, 12, 22, 23, 24, 25, 26, 27, 28, 29]
   b) jet = 1 : [4, 5, 6, 12, 22, 26, 27, 28]
   c) jet = 2 : [22]
   d) jet = 3 : [22]

   Note that we also dropped column 22 which corresponds to the jet feature itself. Indeed, we figured that by giving as much importance to this feature, we would not need it any more to classify our data.

4) Correlation analysis : Making the assumption that two (or more) highly-correlated features must have roughly the same impact when predicting the label of an item, we computed the correlations between each pairs of features for different correlation values (e.g. ¿0.6, ¿0.8, ¿0.9 and =1) and identified the more or less correlated features.

B. Algorithms and Techniques

III. Results

Acknowledgements

The author thanks Christian Sigg for his careful reading and helpful suggestions.

References

[1] S. P. Jones, "How to write a great research paper," 2008, microsoft Research Cambridge.

[2] M. Schwab, M. Karrenbach, and J. Claerbout, "Making scientific computations reproducible," *Computing in Science and Engg.*, vol. 2, no. 6, pp. 61–67, 2000.

[3] J. B. Buckheit and D. L. Donoho, "Wavelab and reproducible research," Stanford University, Tech. Rep., 2009.

[4] R. Gentleman, "Reproducible research: A bioinformatics case study," *Statistical Applications in Genetics and Molecular Biology*, vol. 4, no. 1, 2005. [Online]. Available: http://www.bepress.com/sagmb/vol4/iss1/art2