The Higgs Boson Machine Learning Challenge

Matteo Calafà, Giulia Mescolini, Paolo Motta
First Project for the course "Machine Learning" at EPFL Lausanne, Switzerland

***Abstract*—The report contains a proposal of solution for the Higgs Boson Machine Learning Challenge, proposed in the framework of the course "Machine Learning" at EPFL Lausanne. Several algorithms are presented to approach this classification problem on CERN particle accelerator data.**

I. Introduction

The goal of the challenge is to estimate the likelihood that a given event's signature is the result of a Higgs boson or of some other process/particle, because, rather than observing the boson directly, scientists measure the products that result from its decay process, which may be similar to other particles' ones.
In section II, we present the analysis of the database and the meticulous preprocessing; afterwards, in section III, we present the models built with the 6 requested algorithms and the selection of the best hyper-parameters; afterwards, in section IV, we illustrate their performance.

II. Preprocessing

A. First Analysis of the Dataset

The dataset contains 250000 points for training and 568238 for testing, with 30 features and their corresponding binary labels ("-1" for "background" and "1" for "signal"), which clearly have to be predicted in the case of the test set.
First of all, we noticed that one feature, *PRI_jet_num*, is the only one to be categorical; it represents the number of jets (showers of hadrons originating from a quark and a gluon, clustered together after being produced in a particle collision) and it ranges from 0 to 3. Inspired by the challenge documentation, we noticed that some features are meaningless for some numbers of jets, therefore we split the dataset into 4 subclasses, each one characterized by a different *PRI_jet_num*.

B. Management of Missing Values

From the documentation, we know that each "-999" present in the dataset represents a missing value; if for a feature more than the 70 % of data is missing, we decide not to consider it, while the remaining missing values are substituted by the median of the feature, which is, according to theory, a robust estimator.

C. Standardization

In order to ensure a good functioning of the numeric optimization, it is a good practice to standardize the dataset: we subtract from each feature its mean and divide by its standard deviation. This helps the feature matrix in having a better condition number.

D. Feature Engineering

We plot the features and ideate strategies to deal with their peculiarities.

- **Logaritmic transform:**
- **Useless features:**
- **Angles:**

E. Polynomial Feature Expansion

F. Management of Outliers

To deal with the presence of outliers, we fix $\alpha = 0.1$ and decide to cap the extreme values of each feature to the α-quantile (for the lower tail) and to the 1-$alpha$-quantile (for the upper tail).

III. Models and Methods

The ideas for good writing have come from [**?**], [**?**], [**?**].

A. Getting Help

One should try to get a draft read by as many friendly people as possible. And remember to treat your test readers with respect. If they are unable to understand something in your paper, then it is highly likely that your reviewers will not understand it either. Therefore, do not be defensive about the criticisms you get, but use it as an opportunity to improve the paper. Before your submit your friends to the pain of reading your draft, please *use a spell checker*.

B. Abstract

The abstract should really be written last, along with the title of the paper. The four points that should be covered [**?**]:

1) State the problem.
2) Say why it is an interesting problem.
3) Say what your solution achieves.
4) Say what follows from your solution.

C. Figures and Tables

Use examples and illustrations to clarify ideas and results. For example, by comparing Figure 1 and Figure 2, we can see the two different situations where Fourier and wavelet basis perform well.

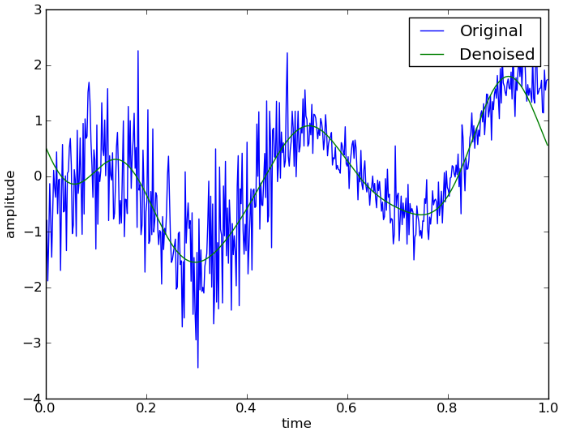

Figure 1. Signal compression and denoising using the Fourier basis.

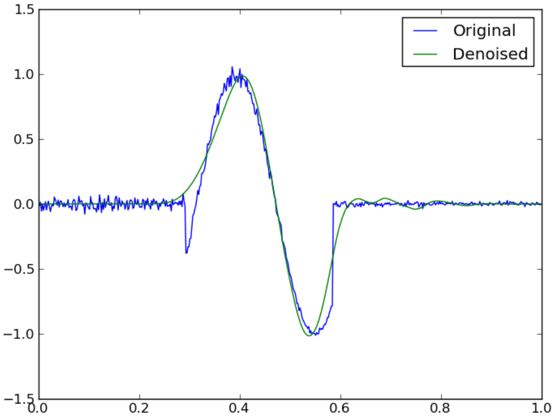

Figure 2. Signal compression and denoising using the Daubechies wavelet basis.

D. Models and Methods

The models and methods section should describe what was done to answer the research question, describe how it was done, justify the experimental design, and explain how the results were analyzed.

The model refers to the underlying mathematical model or structure which you use to describe your problem, or that your solution is based on. The methods on the other hand, are the algorithms used to solve the problem. In some cases, the suggested method directly solves the problem, without having it stated in terms of an underlying model. Generally though it is a better practice to have the model figured out and stated clearly, rather than presenting a method without specifying the model. In this case, the method can be more easily evaluated in the task of fitting the given data to the underlying model.

The methods part of this section, is not a step-by-step, directive, protocol as you might see in your lab manual, but detailed enough such that an interested reader can reproduce your work [**?**], [**?**].

The methods section of a research paper provides the information by which a study's validity is judged. Therefore, it requires a clear and precise description of how an experiment was done, and the rationale for why specific experimental procedures were chosen. It is usually helpful to structure the methods section by [**?**]:

1) Layout the model you used to describe the problem or the solution.
2) Describing the algorithms used in the study, briefly including details such as hyperparameter values (e.g. thresholds), and preprocessing steps (e.g. normalizing the data to have mean value of zero).
3) Explaining how the materials were prepared, for example the images used and their resolution.
4) Describing the research protocol, for example which examples were used for estimating the parameters (training) and which were used for computing performance.
5) Explaining how measurements were made and what calculations were performed. Do not reproduce the full source code in the paper, but explain the key steps.

E. Results

Organize the results section based on the sequence of table and figures you include. Prepare the tables and figures as soon as all the data are analyzed and arrange them in the sequence that best presents your findings in a logical way. A good strategy is to note, on a draft of each table or figure, the one or two key results you want to address in the text portion of the results. The information from the figures is summarized in Table I.

When reporting computational or measurement results, always report the mean (average value) along with a measure of variability (standard deviation(s) or standard error of the mean).

IV. Results

There is a lot of literature (for example [**?**] and [**?**]) on how to write software. It is not the intention of this section to replace software engineering courses. However, in the interests of reproducible research [**?**], there are a few guidelines to make your reader happy:

- Have a `README` file that (at least) describes what your software does, and which commands to run to obtain results. Also mention anything special that needs to be set up, such as toolboxes[1].
- A list of authors and contributors can be included in a file called `AUTHORS`, acknowledging any help that you

[1] For those who are particularly interested, other common structures can be found at http://en.wikipedia.org/wiki/README and http://www.gnu.org/software/womb/gnits/.

Basis	Support	Suitable signals	Unsuitable signals
Fourier	global	sine like	localized
wavelet	local	localized	sine like

Table I
CHARACTERISTICS OF FOURIER AND WAVELET BASIS.

may have obtained. For small projects, this information is often also included in the `README`.

- Use meaningful filenames, and not `temp1.py`, `temp2.py`.
- Document your code. Each file should at least have a short description about its reason for existence. Non obvious steps in the code should be commented. Functions arguments and return values should be described.
- Describe how the results presented in your paper can be reproduced.

A. LaTeX Primer

LaTeX is one of the most commonly used document preparation systems for scientific journals and conferences. It is based on the idea that authors should be able to focus on the content of what they are writing without being distracted by its visual presentation. The source of this file can be used as a starting point for how to use the different commands in LaTeX. We are using an IEEE style for this course.

1) Installation: There are various different packages available for processing LaTeX documents. On OSX use MacTeX (http://www.tug.org/mactex/). On Windows, use for example MikTeX (http://miktex.org/).

2) Compiling LaTeX: Your directory should contain at least 4 files, in addition to image files. Images should be in `.png`, `.jpg` or `.pdf` format.

- IEEEtran.cls
- IEEEtran.bst
- groupXX-submission.tex
- groupXX-literature.bib

Note that you should replace groupXX with your chosen group name. Then, from the command line, type:

```
$ pdflatex groupXX-submission
$ bibtex groupXX-literature
$ pdflatex groupXX-submission
$ pdflatex groupXX-submission
```

This should give you a PDF document `groupXX-submission.pdf`.

3) Equations: There are three types of equations available: inline equations, for example $y = mx+c$, which appear in the text, unnumbered equations

$$y = mx + c,$$

which are presented on a line on its own, and numbered equations

$$y = mx + c \tag{1}$$

which you can refer to at a later point (Equation (1)).

4) Tables and Figures: Tables and figures are “floating” objects, which means that the text can flow around it. Note that `figure*` and `table*` cause the corresponding figure or table to span both columns.

V. CONCLUSION

The aim of a scientific paper is to convey the idea or discovery of the researcher to the minds of the readers. The associated software package provides the relevant details, which are often only briefly explained in the paper, such that the research can be reproduced. To write good papers, identify your key idea, make your contributions explicit, and use examples and illustrations to describe the problems and solutions.

ACKNOWLEDGEMENTS

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