

Writing Scientific Papers and Software

Cheng Soon Ong

Department of Computer Science, ETH Zurich, Switzerland

Abstract—The Higgs Boson challenge involves using a large dataset to classify whether or not a particle is a Higgs Boson particle. We assessed various regression and classification algorithms that could be used to predict the outcome and implemented some on the dataset. To further improve the results we preconditioned the data through various methods. Finally, a testing system was written to test all the combinations of preconditioning algorithms with the machine learning algorithms and output the one with least errors.

I. INTRODUCTION

The Higgs Boson is a recently discovered very unpredictable particle. Through the particle collider at CERN, much data has been collected regarding the features of a decay processes when a Higgs Boson particle was present and was not present. We have analyzed and automated several machine learning algorithms to predict whether or not the Higgs Boson particle is present given these features.

When creating an algorithm there are many choices for how to precondition the dataset and predict the output. We attempted to discover the optimal model by running various algorithms with various preconditioned datasets, and choosing the best one. We describe each of the techniques we use in the Precondition section and Classification section.

II. PRECONDITIONING

A. Filling in Missing Data

—maybe use numbers for these using means learning what to fill deleting those features deleting those data sets.

B. Feature Extraction and Augmentation

—using Variance method and stuff —using feature visualization (can include graphs)

C. Normalization

list several normalization methods we tried.

- 1) Gaussian Normalizer
- 2) Decimal Scaling
- 3) Minmax Normalizer
- 4) Gaussian Outlier removal

III. CLASSIFICATION MODELS

We used linear models and ran various classification algorithms

A. Stochastic Gradient Descent

detail parameters and stuff

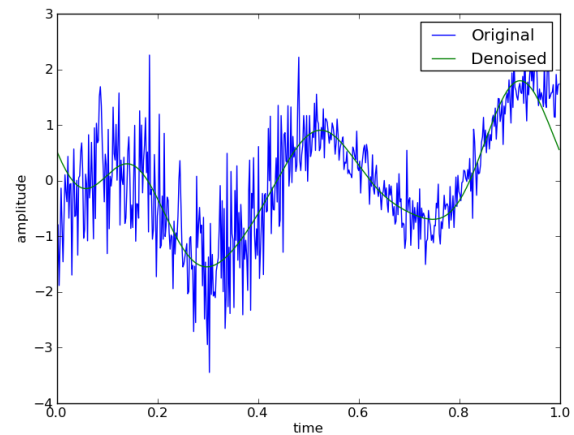


Figure 1. Signal compression and denoising using the Fourier basis.

B. Least Squares

C. Ridge Regression

This was the given figure example

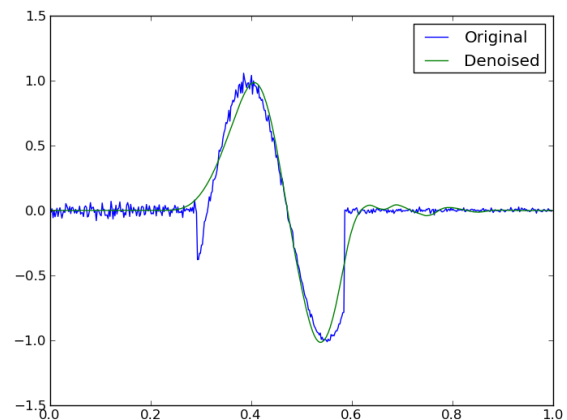


Figure 2. Signal compression and denoising using the Daubechies wavelet basis.

Referring to Figure 1 and Figure 2.

D. Logistic Regression

this is a citation [1], [2].

E. Testing and Results

Describe how testing algorithms works Use figures and maybe tables in results: See this example: Table I.

IV. TIPS FOR GOOD SOFTWARE

There is a lot of literature (for example [3] and [4]) 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 [5], 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 toolboxes¹.
- A list of authors and contributors can be included in a file called `AUTHORS`, acknowledging any help that you 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 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 $\text{Mac}\TeX$ (<http://www.tug.org/mactex/>). On Windows, use for example $\text{Mik}\TeX$ (<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
```

¹For 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/>.

```
$ 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. SUMMARY

Do this too. 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

probs acknowledge class and stuff The author thanks Christian Sigg for his careful reading and helpful suggestions.

oh and see the code for how to do the bibliography

REFERENCES

- [1] G. Anderson, “How to write a paper in scientific journal style and format,” 2004, <http://abacus.bates.edu/ganderso/biology/resources/writing/HTWtoc.html>.
- [2] J. B. Buckheit and D. L. Donoho, “Wavelab and reproducible research,” Stanford University, Tech. Rep., 2009.
- [3] A. Hunt and D. Thomas, *The Pragmatic Programmer*. Addison Wesley, 1999.
- [4] J. Spolsky, *Joel on Software: And on Diverse & Occasionally Related Matters That Will Prove of Interest etc.: And on Diverse and Occasionally Related Matters ... or Ill-Luck, Work with Them in Some Capacity*. Apress, 2004.
- [5] M. Schwab, M. Karrenbach, and J. Claerbout, “Making scientific computations reproducible,” *Computing in Science and Engg.*, vol. 2, no. 6, pp. 61–67, 2000.

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

Table I
CHARACTERISTICS OF FOURIER AND WAVELET BASIS.