Machine Learning - Project 1

Marion Chabrier, Valentin Margraf, Octavianus Sinaga
Department of Computer Science, EPFL Lausanne, Switzerland

***Abstract*—The goal of this project is to apply Machine Learning techniques on data from CERN generated by smashing protons into one another and measuring the decay signature of the possibly resulted Higgs boson. With this decay signature as input our model predicts whether it was result of a Higgs boson or something else (noise). We use binary classification methods to solve this problem.**

I. Introduction

First we preprocess the data i.e. standardize it and get rid of missing values and outliers. Then we implement our different methods (least_squares, ridge_regression etc.) on our data to train the models and see how they perform. We compare them using different values for the hyperparameters *degree* and *lambda_* to optimize each model. We then conclude with...

II. Data Preprocessing

III. Models

In order

A. Figures and Tables

We run 4-fold validation on our training data to tune our hyperparameters. The figure below shows how the choice of the *degree* affects the *RMSE*. We see that for degree = 11 we get our best result, whereas for higher degrees the model will overfit.

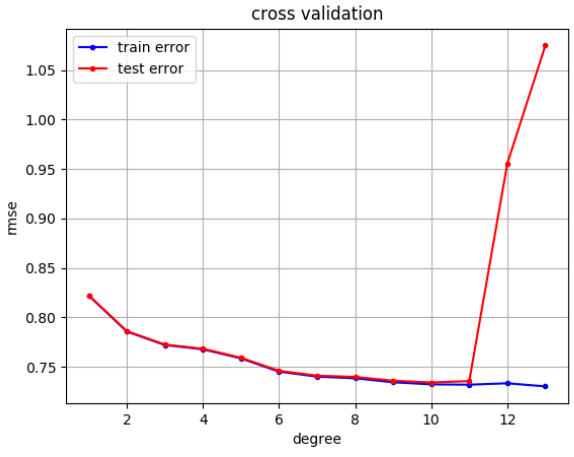

Figure 1. Rmse for different lambdas using least_squares.

From this table we can clearly see, that Least Squares performs best ending up with an accuracy of 0.821.

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

B. 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 [6], [3].

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 [7]:

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.

Methods	Accuracy	F1-Score	lambda_	d
Least Square	0.821	0.723	-	11
Least Square GD	0.566	0.012	-	10
Least Square SGD	0.391	0.394	-	10
Ridge Regression	0.815	0.71	0.01	10
Logistic Regression	0.673	0.12	-	10
Reg. Logistic Regression	0.673	0.12	0.01	10

Table I
PERFORMANCE OF OUR MODELS.

C. 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 **??**.

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. TIPS FOR GOOD SOFTWARE

There is a lot of literature (for example [8] and [9]) 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 [2], 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 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. SUMMARY

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

[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/.

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.

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

[5] Editorial, "Scientific writing 101," *Nature Structural & Molecular Biology*, vol. 17, p. 139, 2010.

[6] G. Anderson, "How to write a paper in scientific journal style and format," 2004, http://abacus.bates.edu/ ganderso/biology/resources/writing/HTWtoc.html.

[7] R. H. Kallet, "How to write the methods section of a research paper," *Respiratory Care*, vol. 49, no. 10, pp. 1229–1232, 2004.

[8] A. Hunt and D. Thomas, *The Pragmatic Programmer*. Addison Wesley, 1999.

[9] 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.