Machine Learning Project I

Andres Montero, Elias Poroma, Jonas Jäggi
School of Computer and Communication Sciences, EPFL, Switzerland

***Abstract*—Machine learning is a field of artificial intelligence that uses statistical techniques to give computer systems the ability to "learn" from data, without being explicitly programmed. And in this project it is applied to estimate the likelihood that a given data set is the result of a specific particle, for example the Higss Boson. This report gives an overview of six machine learning methods implemented to obtain the predictions and the results. It describes the pre-processing, data cleaning and algorithms used in order to obtain a more significant information. Comparing these results, the machine learning method that gives the lowest error and therefore the best approximation is BEST METHOD !.**

I. Introduction

The Higgs boson is an elementary particle in the Standard Model of particle physics, produced by the quantum excitation of the higgs field and explains why some particles have mass [**?**]. To confirm the existence of this particle, the CERN made several experiments from which we obtained real data and the objective of this work is to present a machine learning model that will give the best fit to the data provided. However the data also contains other possible particles, so the main objective is to determine if the data of each experiment belongs to the higss boson or to other particle depending on the values provided on the data set. For this purpose two data sets are provided, one to be used as the training data and the other as the testing data. The train data contains N=250000 measurements, where each of this is described by 30 features and the output, which is 's' correct positive or 'b' false positive. The test data contains 568.268 measurements with the same 30 features and the output to be determined by the machine learning model. The results of this work will be uploaded to kaggle [**?**], which will evaluate the results presented and provide a score of the model.

II. Pre-Processing of Data

Before apply any machine learning model it is really important that the data is understood and that the noise existing in the data is clean, this means that if there are incorrect or missing values, correlated or categorical values, it is needed to filter/correct them otherwise they may impact in a negative way in our prediction.

A. Understanding the Data

In the data presented there are 30 features, which are explained in the paper of the Higgs Challenge [**?**]. After a close analysis of this information, the results are:

- The feature in column 22 `PRI_jet_num` is a categorical feature, with discrete values of (0,1,2,3). For this reason it should be considered for categorization and extracted from the data.
- Depending on the values of `PRI_jet_num` we have other columns where the value is undefined, therefore this columns should be dropped on each categorical training.
- Even after the removal of the undefined columns depending on the categorical values we still find values that are undefined -999, which delivers noise to the model, so they are replaced with the mean of the column is each case.
- After that we need to remove the outliers values, as we are using models depending on MSE which penalizes heavily the outliers values, for this we use the IQR technique [**?**].

B. Polynomials, Standardization, Offset

- Polynomials are used in the models because linear models may cause underfiting, however the degree of the polynomials should be calibrated carefully to avoid over-fitting.
- Once the polynomials are ready it is needed to standardize the data as the models we used converge faster with this feature, for this purpose the values (mean, standard deviation) from the train data are used.
- Adding one vector of "ones" as the offset in the data set also know as the "bias" term.

III. Models-Machine Learning

For this project we implemented 6 different models to make the predictions, and for each one of them we implemented cross validation to assure that the model will work as expected with new values. For this we implemented k-fold cross validation with a values of k=5 to split the data in 5 even groups, where 4 groups are used for training and one group is used for test. The results found for the different six models are summarized in Table II. As it can be understood from the table the best model is "Best Model", so this will be the model used to describe in detail the process to achieve the result obtained. To begin with the training of our model first we analyzed different scenarios:

1) Training the model - Standardization
   First the model trained with the entire set of data, without applying categorical training. In this case is

only applied standardization as explained in II-B The results obtained are: 1. RMSE: 2. Kaggle: With these results we clearly deduct that more cleaning stages were necessary to understand the data and achieve that our model behaves as expected.

2) Training the model - Removing Outliers
   Analyzing the previous step, we realized that the model showed really high values, to fix this problem we proceed with the removing outliers step explained in II-B. The results obtained are: 1. RMSE: 2. Kaggle: And it shows and improvement of 00%
3) Training the model - Categorical values
   Finally to obtain the official result presented to kaggle the model will trained depending on the categorical values as explained II-A The results obtained are: 1. RMSE: 2. Kaggle: Which compared to the step 1 shows and improvement of 00%.

Table I
MODELS TO USE.

Model	RMSE	Hyper Param	gamma, iterations
Least Squeares GD	8	degree= lambda=	1, 1
Least Squeares SGD	8	degree= lambda=	1, 1
Least Squeares	8	degree= lambda=	-
Ridge Regression	8	degree= lambda=	-
Logistic Regression	8	degree= lambda=	1, 1
Regularized Logistic Regression	8	degree= lambda=	1, 1

IV. BEST MODEL DETAILED DESCRIPTION

As we observed in the previous section the best model that fits the data is BEST MODEL, therefore a detailed explanation of the method and the code is explained:
To start the weights are initialized as a column of ones, then we define that all the data processing functions are true, as discussed before gives a better result. Then we input the value for the selected model and define the maximum iterations and start the cross validation step to assure that the models will work as expected. Once this is completed a grid search method is applied for the degree of the polynomials and also for the best lambda used to penalize the model and avoid over-fitting in case that the degree found in the grid search is to high. The result shows that the best degree is 2 and the best lambda is VALUE as you can see in the figure 1. With these results it is deduced that the lambda value is almost zero which could cause that the model over-fits if the degree of the polynomial was higher, however with just a value of two in the degree, the risk of over-fitting is low, therefore the lambda value will be omitted. Once the best hyper-parameters are found the model can start with the training for each of the categorical values, giving a prediction to each categorical value and appending this to final result that will be saved as an excel file.

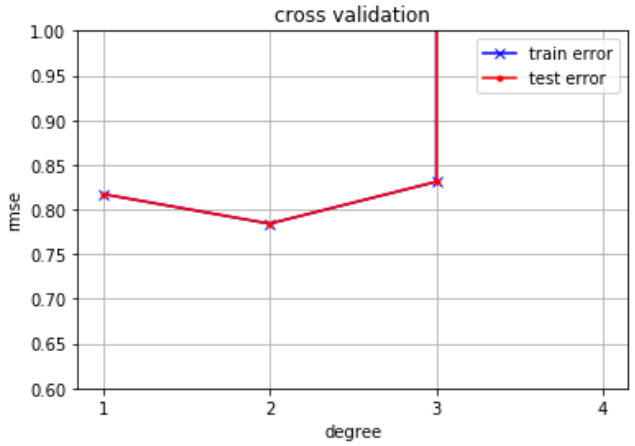

Figure 1. RMSE for different degrees of polynomial expansion - Ridge regression.

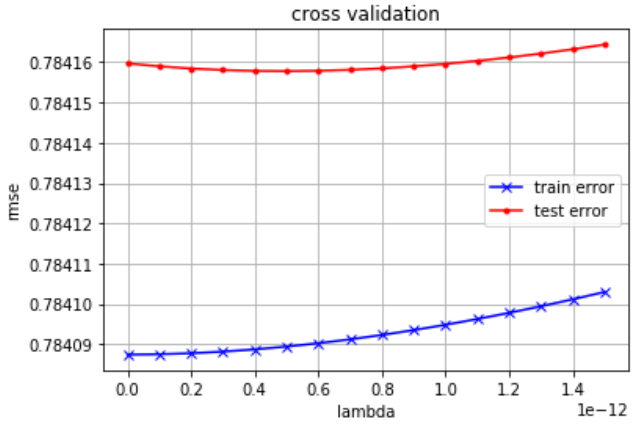

Figure 2. RMSE for different regularization values - Ridge regression

Table II
SIGNIFICANCE OF FEATURE ENGINEERING

Data treatment	Training set division	Kaggle score
only standardization & offset	no division	???
+ outlier replacement	no division	0.65773
+ outlier replacement	division into jet categories	???

Scientific papers usually begin with the description of the problem, justifying why the problem is interesting. Most importantly, it argues that the problem is still unsolved, or that the current solutions are unsatisfactory. This leads to the main gist of the paper, which is “the idea”. The authors then show evidence, using derivations or experiments, that the idea works. Since science does not occur in a vacuum, a proper comparison to the current state of the art is often part of the results. Following these ideas, papers usually have the following structure:

Abstract
: Short description of the whole paper, to help the reader decide whether to read it.

Introduction
: Describe your problem and state your contributions.

Models and Methods
: Describe your idea and how it was implemented to solve the problem. Survey the related work, giving credit where credit is due.

Results
: Show evidence to support your claims made in the introduction.

Discussion
: Discuss the strengths and weaknesses of your approach, based on the results. Point out the implications of your novel idea on the application concerned.

Summary
: Summarize your contributions in light of the new results.

V. Tips for Good Writing

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 3 and Figure 4, we can see the two different situations where Fourier and wavelet basis perform well.

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.

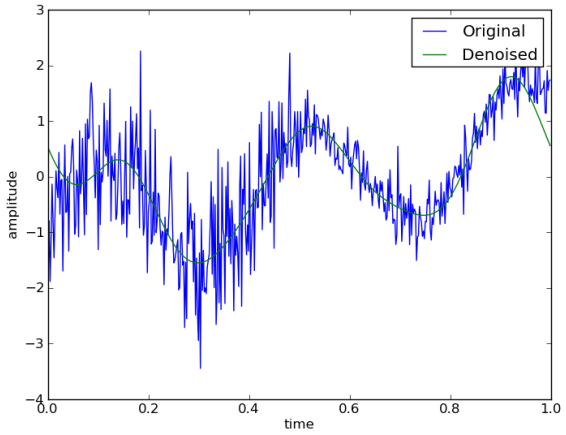

Figure 3. Signal compression and denoising using the Fourier basis.

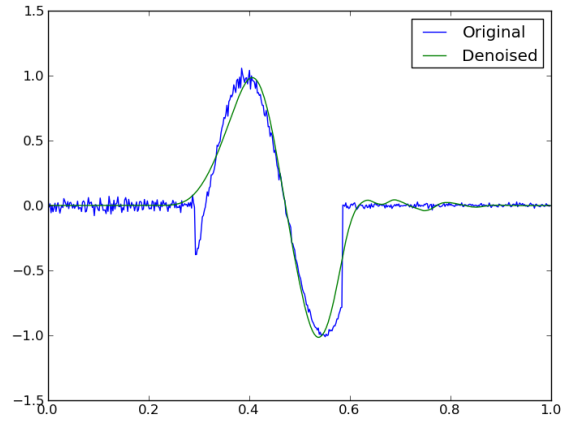

Figure 4. Signal compression and denoising using the Daubechies wavelet basis.

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 III.

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).

VI. Tips for Good Software

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 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.

[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/.

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.

VII. 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,

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

Table III
CHARACTERISTICS OF FOURIER AND WAVELET BASIS.

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.