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 feature set is the result of a specific particle, for example the Higgs Boson. This report gives an overview of six machine learning methods that were implemented and evaluated by the accuracy of their predictions. It describes the pre-processing, data cleaning and algorithms used in order to find which model has the best fit for the data provided. After comparing these results, the machine learning method which reaches the highest accuracy is Least Squares.**

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 the data is obtained and the objective of the here presented work is to present a machine learning model that will give the best prediction. However the data also contains measurements of other possible particles, so the main objective is to determine if the data of each experiment belongs to the higss boson or to the other particles depending on the values provided on the data set. The training data set contains N=250000 measurements, and each measurement is described by 30 features and one label, which is 's' correct positive or 'b' false positive. Another 568.268 measurements with the same 30 features and the corresponding labels are available for Kaggle [**?**] in order to evaluate the submissions.

II. Data Pre-Processing

It is crucial that the data is understood and that the noise existing in the data is cleaned. First the models were trained without data pre-processing and this lead to bad results II. To avoid this and also to show the impact that data cleaning has on a machine learning model the different steps of the process are described.

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 categorical extraction from the data.
- Depending on the values of `PRI_jet_num` we have other columns where the value is undefined, therefore these columns should be dropped on each categorical training.
- Some feature may have the value -999, which means that independently of the category they belong to, they are undefined too. This delivers noise to the model, so we replace them with the mean of the column in 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, We use IQR technique [**?**] to replace outliers.

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 of them we implemented different data cleaning stages to avoid noise in the model. Then to assure that the model will work as expected with new values, k-fold cross validation was used 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 I. The best results were obtained with "Ridge Regression" model, so this will be used to describe in detail how the results were obtained. To begin with the training of our models first we analyzed different scenarios:

1) Training the model - Standardization
   First each 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 for the accuracy is : `00%` 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 for the accuracy is : 00% which shows and improvement of 00% compared to the previous result.
3) Training the model - Categorical values
Finally to obtain the official result presented to kaggle the model will also be trained depending on the categorical values as explained in II-A. The results obtained for the accuracy is : 00% which shows and improvement of 00% compared to the previous result.

Table I
MODELS TO USE.

Model	Accuracy	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

After the data cleaning stage is completed 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 II. 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 figures 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.

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	???

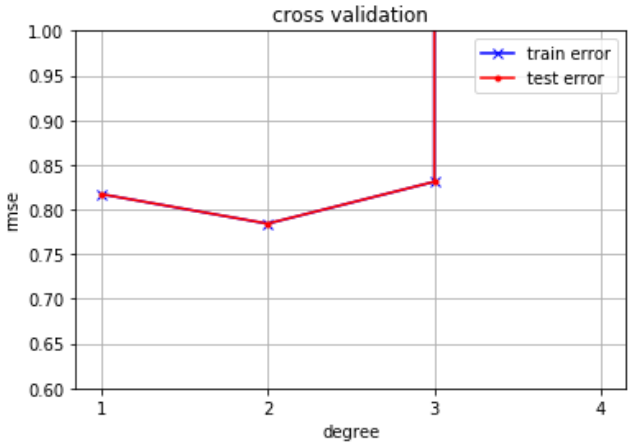

Figure 1. RMSE for different degrees of polynomial expansion - Ridge regression.

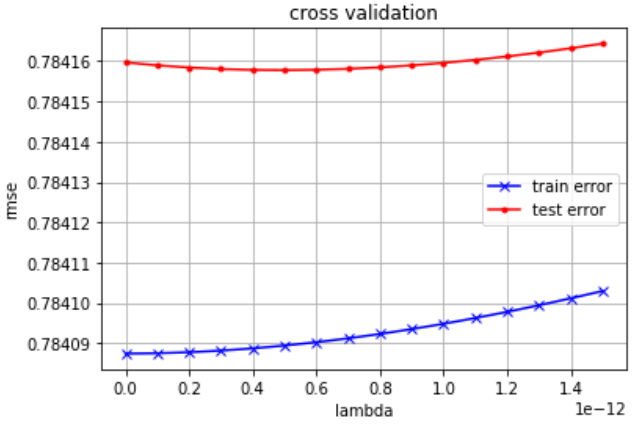

Figure 2. RMSE for different regularization values - Ridge regression

IV. RESULTS AND SUMMARY

To predict if the data of each row corresponds to the Higss boson 6 different models are proposed, which are trained with k-fold cross validation method and grid search to find the best hyper-parameters. Different data pre processing stages are applied to evaluate the best results II. **Ridge Regression** has the best behavior for the data provided and the highest accuracy (0.80 KAGGLE) . A detailed explanation of how to install and run the program is explained in the "read me" file presented with the results.