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 actually was result of a Higgs boson or something else (noise). We use different regression methods to tackle this problem. We got F-1 score 0.728 and accuracy 0.822 as the best result by implementing Least Squares amongst other methods that we've implemented.**

I. Introduction

First we preprocess the data i.e. standardize it and get rid of missing values and outliers. Then we implement the six different methods: Least Squares, Least Squares GD, Least Squares SGD, Ridge Regression, Logistic Regression, Regularized Logistic Regression. We use each method to learn a model on the training data and see how well they perform. For each model we additionally vary the hyperparameters to optimize the performance. Finally we compare their performances on the test data from CERN by submitting it on AICrowd.

II. Data Preprocessing

The preprocessing deals with:

- Substitution of the -999 values for each entries using the mean of 'clean' data in train and test dataset
- Standardization of the value for all entries with standard deviation and mean
- Deletion of outliers in the train data (set the treshold to cut off the entries).

In order to find out, which approaches of preprocessing yield good results, we evaluate their effect on the MSE. In figure 1 we used Least Squares to demonstrate this. In can be observed, that if we just standardize the data, the MSE is quiet high. By substituting the -999 values by the mean of the data, a big reduction in the loss can be achieved. Furthermore the loss can be reduced by doing both, standardizing the data and removing outliers. For removing outliers we plotted the boxplot to observe the quantiles of each feature. We use this value as referenced threshold by adding some bias (small integer) to filter out the data points in which the values overly exceeding the quantiles. In this particular experiment, we set the treshold to 8.5 since the maximum quantile value for all 30 features is around 6-7 as shown in figure 2.

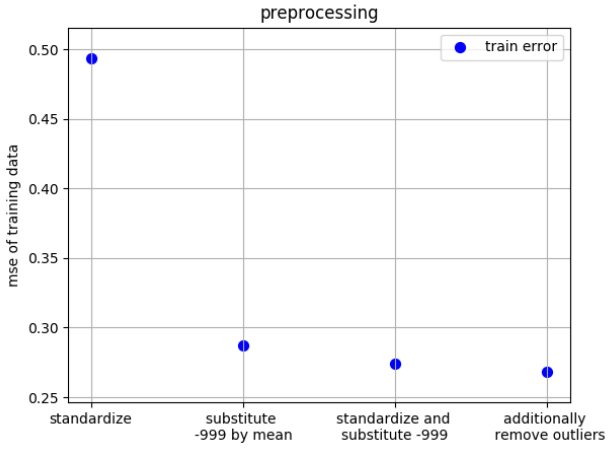

Figure 1. MSE for different approaches of preprocessing using Least Squares to evaluate their effect.

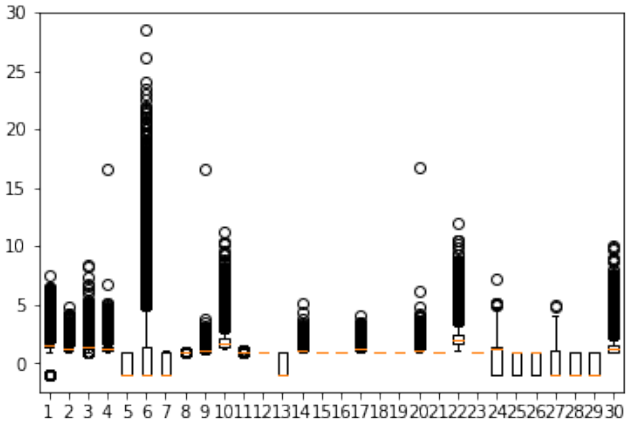

Figure 2. Boxplot for train dataset.

III. Methods

For each model we run 4-fold cross validation on our training data to tune our hyperparameters in order to optimize our model. The hyperparameters in this case are the *degree* for all the models and the constant *lambda* for the Ridge Regression and the Reg. Logistic Regression.
Figure 3 shows how the choice of the *degree* affects the *RMSE* in the case of Least Squares. We run the cross validation for degrees between 1 and 13 and find out, that for degree = 11 we get our best result. For higher degrees the model overfits whereas for lower degrees it underfits.

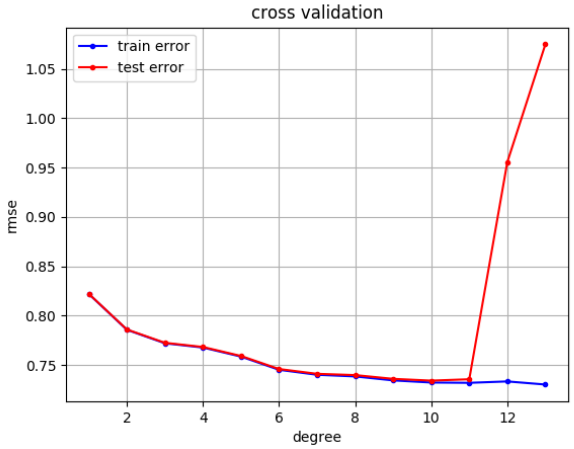

Figure 3. RMSE for different degrees using Least Squares.

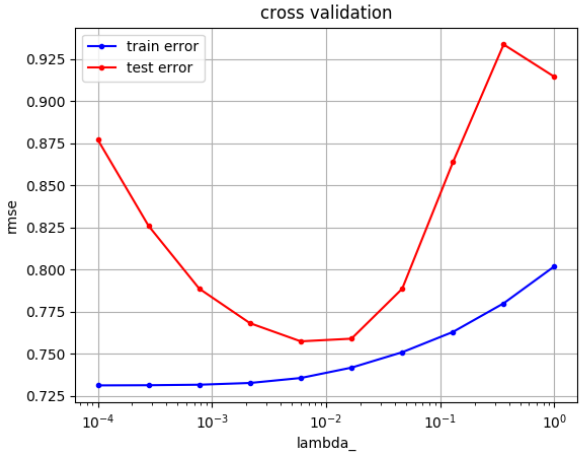

Figure 4. RMSE for different lambdas using Ridge Regression (deg. 12).

For the Ridge Regression a degree of 12 gives the best result. We then again run cross validation to optimize the second hyperparameter *lambda*. A value of 0.00599 gives the best result, which can be checked in Figure 4. When we choose this value too small, the test error gets much bigger, whereas the training error reduces. If *lambda* is too big, both, the test and training error augment. The final used hyperparameter for each method can be found in table 1.

Methods	degree	lambda
Least Squares	11	-
Least Squares GD	10	-
Least Squares SGD	10	-
Ridge Regression	12	0.00599
Logistic Regression	10	-
Reg. Logistic Regression	10	100

Table I
OPTIMIZED HYPERPARAMETERS COMPUTED THROUGH 4-FOLD CROSS VALIDATION.

IV. RESULTS

After having optimized the hyperparameters for each model we want to see how the different models perform on the test data from CERN. We therefore submit each prediction on *AICrowd* and see what result it gives us. In table 2 they can be compared.

Methods	Accuracy	F1-Score
Least Squares	0.822	0.728
Least Squares GD	0.682	0.511
Least Squares SGD	0.391	0.394
Ridge Regression	0.815	0.713
Logistic Regression	0.819	0.718
Reg. Logistic Regression	0.819	0.717

Table II
PERFORMANCES OF OUR MODELS SUBMITTED ON AICROWD.

Ending up with an accuracy of 0.822 and F1-Score of 0.728, Least Squares performs best amongst all methods. Both Least Squares Descent methods in contrast perform not very well. Ridgre regression performs quite good as well, it gives an accuracy of 0.815 and F1-Score of 0.713. Logistic and Regularized Logistic Regression give both an accuracy of approx. 0.819 and also perform quite good when taking the F1-Score as accuracy measure.

V. DISCUSSION

Least Squares Gradient performs not as good as the Least Squares method, even though in theory it would converge to the same optimum. We had to choose a quite small *gamma* (approx 10^{-19}) for not making explode the loss. The stepsize can be concerned as an additional hyperparameter which to choose correctly. The Least Squares Stochastic Gradient Descent performs even worse. Reasons for that might be also badly chosen stepsize *gamma* and batchsize. Since Least Squares already gave us a good result, we did not focus much on optimizing the stepsize in order to make the both Gradient methods perform better.
Logistic Regressions and Reg. Logistic Regression perform quite similar which is actually surprising, given the fact that the additional penalty term is supposed to support simpler models. But since we ended up with an optimal degree of 10 for both methods, we might have chosen *lambda* to small for having a real impact. For both of those methods we set the initial weight as the weight we got from Least Squares. By doing so we got better results compared to initiating the weight with zero or random values.

VI. SUMMARY

In this project we used different regression methods to predict the Higgs Boson. After preprocessing the data and optimizing the hyperparameters for each method using 4-fold cross validation, we have chosen the Least Squares Method to tackle this task. This method performed best with Accuracy = 0.822 and F1-Score = 0.728.