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 regression methods to tackle this problem.**

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

In order to deal with the data, we need to standardize it. The standardization helps us to scale the data in a bounded interval. We standardize both, the test and the train data set. We afterwards delete outliers: values, which are further away from the mean than a certain threshold.
The preprocessing deals with:

- Deletion of outliers in the train data
- Deletion of the entries in the sample that contains the -999 value
- Deletion of the feature in the sample that contains the -999 value
- Substitution of the values -999 by the mean of the corresponding values (mean = 0 after standardizing)

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 just standardizing does not help us to preprocess the data in a good way because the MSE is still 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 standardizing the data and removing outliers.

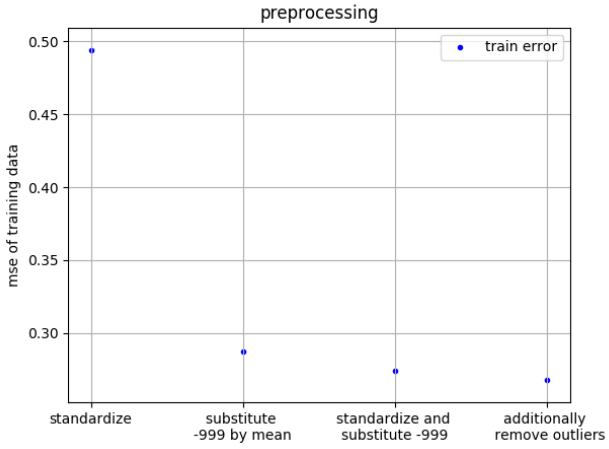

Figure 1. MSE for different approaches of preprocessing using least squares to evulate their effect.

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 Regularized Logistic Regression. Figure 2 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 will overfit and lower degrees instead give a bigger *RMSE*, hence the model underfits.

For the Ridge Regression a degree of 12 gives the best result. Using cross validation we computed the *RMSE* for different values of *lambda* in order to optimize this second hyperparameter. We find out, that a value of approximately 0.0059 gives the best result, which can be checked in Figure 3. 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.

IV. Results

After having optimized the hyperparameters for each model we want to see how the different models perform

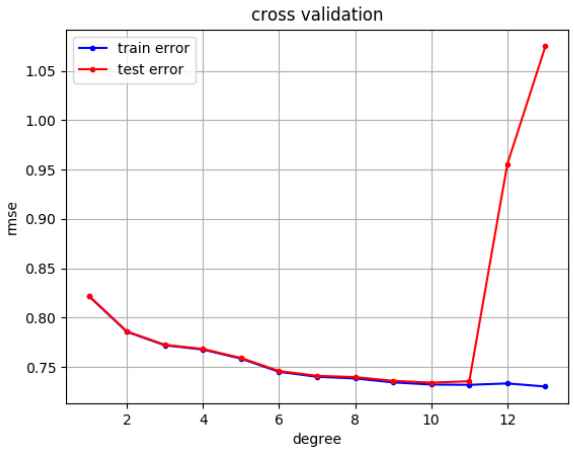

Figure 2. RMSE for different degrees using Least Squares.

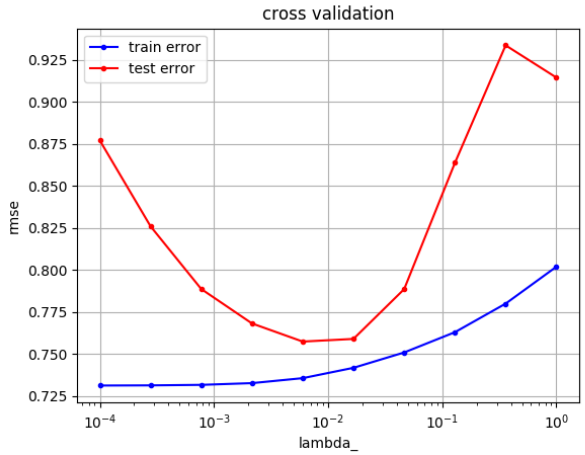

Figure 3. RMSE for different lambdas using ridge_regression (degree 12).

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.

Ending up with an accuracy of 0.821 and F1-Score of 0.723, Least Squares performs best amongst all methods. Least Squares Gradient and Stochastic Gradient 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. 0.67 but perform quite bad when taking the F1-Score as accuracy measure.

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	0.01

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

Methods	Accuracy	F1-Score
Least Squares	0.821	0.723
Least Squares GD	0.682	0.511
Least Squares SGD	0.391	0.394
Ridge Regression	0.815	0.713
Logistic Regression	0.674	0.141
Reg. Logistic Regression	0.673	0.12

Table II
PERFORMANCES OF OUR MODELS SUBMITTED ON AICROWD.

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.
We did not take computational cost into account and Least Squares gave us a good result. That's why we did not focus that much on optimizing the stepsize in order to make the both Gradient methods perform better.
Logistic Regressions and Regularized Logistic Regression perform quite similar which is actually surprising, given the fact that the additional penalty term is supposed to avoid overfitting, hence 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 to have a real impact.

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 concerning Accuracy and F1-Score.