Project 1: Report

Machine Learning (CS-433), EPFL, Switzerland

Eloïse Doyard Cyrille Pittet Alessio Verardo

1st November 2021

1 Introduction

The Large Hadron Collider (LHC) at CERN has collected for 20 years data about the decays of the Higgs boson. This boson has been hypothetized as the leading actor in giving mass to the particles and its existence has been claimed thanks to the ATLAS experiment. When the Higgs boson decays, it produces other particules which can be detected by sensors. We call this type of decay an "event". They are either "uninteresting events" (called "background") or "signal" events, i.e. a significant excess of events w.r.t. background events. In case of signal events, if the excess is significant enough, the physicists can then say they have discovered the particle. The goal of this project [3] is, with the help of machine learning, to classify a given event either as a background event or a signal event.

2 Method

We were provided with a dataset of 250'000 data points and 30 features. The description of all the features can be found in the appendix B of the paper [1] associated to this project.

2.1 Data Cleaning

We started the processing of the data with a data cleaning :

1. Drop the columns where more than half of the values where undefined.
2. Try to drop the data points where more than half of the features are undefined. However, none of the data point match this pattern.
3. Replace the undefined value (i.e. -999) by the median of the corresponding feature.
4. Check that the angle features are in the according range, that is $[-\pi, \pi[$.
5. Deal with the outliers using the interquartile range method and clipping extreme values to a lower / upper bound [2].

Thus, we remain with 250 000 data points and only 23 features. The feature expansion will help us augment our number of dimensions.

Improvements	**Local test set**	
	Accuracy	F1
Baseline (raw data)	74.40%	57.04%
Data cleaning	70.91%	65.64%
Bias term, sin / cos	80.81%	72.01%
Bias term, sin/cos w/o clean	78.33%	65.69%
Polyn. expansion	82.63%	73.68%

Table 1: Evolution of the accuracy and the F1-score according to the different steps of data cleaning and feature expansions with the ridge regression model.

2.2 Feature expansion

Since the data are not theoretical, it is very unlikely that the relationships between the features are linear. We implemented the following expansion procedure: Given a row of the dataset $x = [x_1, \ldots, x_D]$ where D is the number of features, we output

$$x' = (1, x, x_1, x_2, x_3)$$

such that

- 1 is a bias term
- x is our original data point
- $x_1 = \{x_i x_j\}, 1 \leq i < j \leq D$
- $x_2 = \{\cos(x_i), \sin(x_i)\}$ for angle features i, j.
- $x_3 = \{x_i \cos(x_j), x_i \sin(x_j)\}$ where x_i is any floating point feature and x_j is any angle feature.

Finally, we apply a polynomial expansion on these already expanded features and our final training vector is x'' defined as

$$x'' = (x', x_5)$$

such that

$$x_5 = \{x_i^j\}, i \in \{1, \ldots, D\}, j \in \{2, \ldots, deg\}$$

where x_i is any of the original features. The choice of the value of deg will be explain later.

The proof of the usefulness of those steps of data cleaning and feature expansion is showed in table 1. We reported the accuracy and the F1-score after the different

steps of data cleaning and feature expansions. The combination of the data cleaning, the adding of a bias term, the use of sine and cosine functions in the feature expansion and the use of a polynomial expansion improved our model by $8,23\%$.

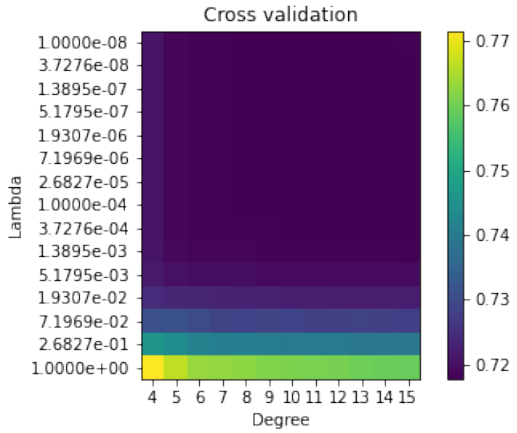

Figure 1: Heatmap of the RMSE loss function values on the local test set depending on the values of the hyperparameters *deg* and λ for the Ridge Regression model during the cross validation.

2.3 Machine learning algorithms

We developed 4 different machine learning algorithms to predict if a set of events that happened in a particular region can be labeled as a *signal* event :

- **Least Squares** We implemented three different ways of minimizing the loss function of this method : using a gradient descent, using a stochastic gradient descent with batches of size 1 and by solving the normal equation.
- **Ridge Regression** This algorithm is a regularized version of the Least Squares method.
- **Logistic Regression** We predict a probability of belonging to the class *signal* and we output the corresponding class.
- **Regularized Logistic Regression** This method is a regularized version of the Logistic Regression.

Cross Validation Hyperparameters play an important role in the performance of our models. In this project, the hyperparameters are *deg*, the optimal degree for the power expansion and λ, the regulariser for the Ridge Regression and the Regularized Logistic Regression.
To find the optimal values, we used a cross validation combined with the K-fold technique. Such techniques are very costly and time-consuming but they are crucial to yield the best predictions.

Method	**Local test set**	
	Accuracy	F1
Least squares	$82,01\%$	$72,65\%$
Least squares GD	71.38%	42.58%
Least squares SGD	69.42%	38.72%
Ridge regression	82.63%	73.68%
Logistic regression	74.82%	62.18%
Reg. log. regr.	69.00%	34.06&

Table 2: Accuracy and F1-score results for each of our implemented machine learning algorithms

An example of the variation of the value of the loss function according to the values of the λ and *deg* can be observed in figure 1.

General Procedure We split the dataset into two different sets : the training set (80% of our initial dataset) and the test set (20% of our initial dataset). We augmented the features and we standardized our data points. Moreover, we performed a cross validation (c.f. Cross Validation) to find the best value for the hyperparameters depending on the model. Finally, we trained on our training set and output predictions for our test set. We reported the results of the performances of our models in table 2.

For the sake of our computers' power, the results that we reported for the regularised logistic Regression were obtained without feature expansion of the data points but with the power expansion. In other words, we used as training vector $x' = (1, x, x_1)$ such that $x_1 = \{x_i^j\}, i \in \{1, \dots, D\}, j \in \{2, \dots, deg\}$ where x_i is any of the original features.

3 Result

Our best result on our local test set, was obtained using the Ridge Regression algorithm.

Our best submission on AIcrowd was again with the Ridge Regression model and we obtained a score of $82,8\%$ of accuracy and a F1-score of $73,8\%$. Note that our actual best submission on AIcrowd was made with incorrect code and with our final code we never managed to surpass or reproduce this incorrect score that is visible on the leaderboard.

Discussion We notice that one of the feature, *PRI_jet_num* seemed very correlated with other features. To go further, one might select different models, do a different cleaning and choose different hyperparameters values according to the value of this feature.

4 Conclusion

We managed to predict, with different machine learning algorithms if a particular region of space that experienced an certain amount of events can be considered as a *signal* for a Higgs boson decay.

References

[1] Claire Adam-Bourdarios, Glen Cowan, Cecile Germain, Isabelle Guyon, Balazs Kegl, and David Rousseau. Learning to discover: the higgs boson machine learning challenge. *http://higgsml.lal.in2p3.fr/documentation*, 9, 2014.

[2] Jason Brownlee. How to remove outliers for machine learning. *https://machinelearningmastery.com/how-to-use-statistics-to-identify-outliers-in-data/*, 2018.

[3] Nicolas Flammarion and Martin Jaggi. Class project 1, ml higgs. *https://github.com/epfml/ML_course/blob/master/projects/project1/project1_description.pdf*, 2021.