Higgs Boson - Machine Learning Project 1

Shadi Naguib, Léo Alvarez, Daniel Gönczy
Department of Computer Science, EPF Lausanne, Switzerland

***Abstract*—The Higgs Boson discovery, announced at the Large Hadron Collider at CERN in March 2013, was the center of a new machine learning challenge. This particle can result from the collision of two accelerated protons. Physicist were able to identify it by measuring the decay signature of a collision event. The challenge was to identify those signals among a noisy background. This report relates and details our methodology and results while trying to recreate the "discovery" of the Higgs Boson using Machine Learning methods, based on real data. By comparing different models and hyperparameters, we achieved 82.9% of accuracy using the Ridge regression method.**

I. Introduction

The aim of this Machine Learning project is to predict if a vector composed of primitive sensors values and derived features is related to the signature decay of a Higgs Boson or if it is simply due to the background noise ([1]). In order to do so, we are given a dataset of $250'000$ events composed of 30 columns of features. Each event is classified as either a background noise or a Higgs boson signal. In order to classify a signal using only its features, we investigate several machine learning regression models and compare their classification performances first using K-Fold cross validation on our labeled dataset, and then by uploading our predictions on an unlabeled dataset of $568'238$ events on AICrowd platform.

II. Models and Methods

In order to achieve the best accuracy, we started by understanding and cleaning the data. We proceeded to test out multiple machine learning techniques to achieve the best results.

A. Data Pre-Processing

Very quickly we found out that events had a lot of undefined features, represented by a value of -999 ([2]). These values were replaced by the mean and the median of the feature, and it seemed as if the results achieved were better by using the median.

We also noticed that the PRI-Jet feature was categorical, which means that it can only take discrete values of 0, 1, 2 and 3. The events with PRI-jet 0 and 1 have a lot of undefined features. Therefore, the strategy here was to divide the events into four datasets depending on which PRI-jet they correspond to and to train a different model with specific hyperparameters on each one of these datasets. Finally, each model aimed at finding the corresponding optimal hyperparameters. Of course, every dataset was first standardized by setting each feature to have a mean of 0 and a standard deviation of 1.

B. Regression models

We have written the following functions and tested the corresponding models to make predictions whether the events of our dataset correspond to the decay signature of a Higgs Boson:

- Linear Regression with Gradient Descent
- Linear Regression with Stochastic Gradient Descent
- Least squares regression (normal equations)
- Ridge Regression
- Logistic Regression
- Regularized Logistic Regression

C. Feature expansion

We have found that enlarging the feature space by adding interaction terms between the features gave us significantly better results. For each feature, we have also added columns containing the polynomials of this feature up to a degree M, as explained in the equation below :

$$X = (1|X^K), \text{ K=k for } k \in (0, M) \tag{1}$$

The optimal degree is found by performing grid search and averaging the accuracies found by 5-fold cross validation.

D. Research protocol

For every hyperparameter combination and each regression model, the accuracy is estimated using a 5-fold cross validation on the training set. This means the data is split into 5 random groups: each group is taken out as the test group, and the model is fitted taking the other groups as the training set. It is then evaluated on the extracted test group. The accuracy scores are then averaged to compute a more accurate performance of our model. By doing so, we avoid overfitting our model over a particular training set. Once the optimal hyperparameter combination is found, the weights are trained on the whole training dataset, and the resulting model is used to make predictions on the unlabeled testing set. A random seed of 6 is used to ensure reproducibility of our results.

III. Results

In the table below, the results corresponding to the chosen hyperparameters and the accuracy for each tested classification method are presented for PRI-Jet 0.

Indeed, it is clear that the Ridge Regression performs better than other methods, with an accuracy of 82.7% when predicting if the signature decay of the particle collision correspond to a Higgs Boson signal for PRI jet 0 dataset.

Method	λ	γ	K	maxiter	Accuracy
Linear Regression (GD)	-	0.1	12	500	0.743
Linear Regression (SGD)	-	0.01	12	500	0.731
Least squares	-	-	12	-	0.765
Ridge regression	1e-4	-	9	-	0.848
Logistic regression	-	0.4	1	1000	0.809
Reg. Logistic regression	0.5	0.5	1	1000	0.813

TABLE I
HYPERPARAMETERS OF DIFFERENT REGRESSION MODELS (PRI-JET 0)

A. Feature selection using PCA

We tried reducing the dimension of the dataset by using PCA ([3] and [4]) and the results obtained were not better than without reduction of the feature dimension space. For the PRI-Jet 0 dataset with Ridge Regression, we obtained an accuracy of 82.6% using the first principal components contained in 95% of the variance. Compared to 84.8% with the original features, the PCA technique is not useful in our case.

B. Regression Hyperparameters

As explained in the previous sections, polynomial terms were added to every feature up to a degree M. The table below summarizes the accuracies obtained on the validation set in function of the chosen degree M for jet 0. The same table can be built similarly for jets 1, 2 and 3.

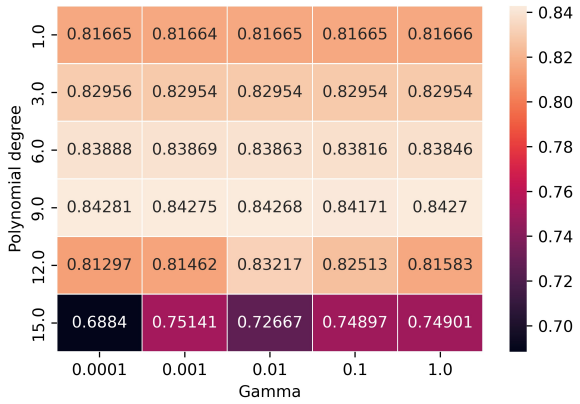

Fig. 1. Hyperparameters for Jet 0 with Ridge Regression model

The best score found by cross validation is using a 12 degree polynomial expansion, giving us a result of 82.7% accuracy on the whole dataset (PRI jet 0, 1, 2 and 3). Thus, the selected Ridge Regression model and its optimal parameters for the AICrowd submission are presented in the table below:

Jet	γ	Degree	Accuracy
0	1e-4	9	0.848
1	1e-2	12	0.802
2	1e-3	12	0.832
3	1	12	0.834

TABLE II
HYPERPARMAMETERS FOR EVERY JET OF THE FINAL RIDGE REGRESSION MODEL

IV. DISCUSSION

The model with the worst performance is the least squares with Stochastic Gradient Descent with a cross validation accuracy of 71.3% while selecting the best tuned hyperparameters. This is significantly worse than the second worse model (least squares with Gradient Descent) that has a total accuracy of 73.2%. The ridge regression is clearly better than all other methods.

Moreover, the polynomial expansion and the accuracy of the model aren't linearly dependent, meaning a higher polynomial degree doesn't necessarily lead to a higher accuracy. This is due to an over-fitting of the model on the training data when we use too high degrees.

V. SUMMARY

Starting from an initial accuracy of 72% obtained by using cross validation after cleaning our training data and fitting a Ridge regression model on it, we managed to increase this accuracy up to 82.9% by extending the feature space using polynomial expansion and tuning our hyperparameters using 5-fold cross validation. Several other dimension reduction techniques have been tried (PCA) but the accuracy was not improved. In addition to the python files, all our code can be found in our project notebook in ([5]).

REFERENCES

[1] W. Khan, "Extracting signals of higgs boson from background noise using deep neural networks," 2020, https://arxiv.org/pdf/2010.08201.pdf.

[2] D. Rousseau, "The higgs boson machine learning challenge," 2014, university Paris-Sud, France.

[3] "What are the pros and cons of the pca?" *i2Tutorials*, 2019. [Online]. Available: https://www.i2tutorials.com/what-are-the-pros-and-cons-of-the-pca/

[4] W. Navarrete, "Principal component analysis with numpy," *Towards Data Science*, 2020. [Online]. Available: https://towardsdatascience.com/pca-with-numpy-58917c1d0391

[5] "Github project1," 2021. [Online]. Available: https://github.com/shadinaguib/higgs-boson-prediction