Project 1: Machine Learning

Computer science department. EPFL. October 2021.

Sierra, Sandra
sandra.sierravega@epfl.ch

Gómez, Belén
belen.gomezgarcia@epfl.ch

Mozo, Rafael
rafael.mozoarias@epfl.ch

***Abstract*—The Higgs boson is an elementary particle in the Standard Model of physics which explains why other particles have mass. The goal of the Higgs Boson Machine Learning Challenge is to estimate the likelihood that a given event's signature was the result of a Higgs Boson (signal) or some other process/particle (background). In this project, we have faced the problem using different regression and clasification methods, achieving the best result using Ridge Regression with a categorical accuracy of 0.828 on AIcrowd.**

I. Introduction

The aim of this project is to find a model that predicts whether a Higgs boson signal or background decay after a simulated particle collision event. Given a training data with several experiments and its outcome, we will manipulate the dataset to deal with missing values, remove related features, apply different methods of both regression and classification and estimate how well our method is doing using cross-classification. The final goal will be to find the best model with the lowest loss and that reaches the highest categorical accuracy according to AIcrowd.

II. Exploratory data analysis

After importing the training set, we observe that there is an amount of 250.000 events, 30 features and only two prediction values, 'b' for background and 's' for signal. Hence, the problem is a Binary Classification with $Y \in \{b, s\}$ and a proportion of 65.73% and 34.27% respectively.

Let's now deal with missing values. According to the documentation, the value for the mass is -999.0 when the topology of the event was too far from the expected one. We can see that there are 38114 missing values for this feature (i.e. DER_mass_MMC). To replace this values, we are going to use the median of the rest of the values for the feature. Other option will be to use the mean, but we will stick with the first option as it is more robust when we have outliers.

Regarding the other features, the missing values depend on the number of jets of the event (i.e. PRI_jet_num).

- If it had no jets, a specific set S of the features presents missing values.
- If it had 1 jet, a specific subset $S' \subset S$ of the features presents missing values.
- If it had either 2 or 3 jets, there are no missing values.

In order to have this into account, we will create 3 different masks so we can create 3 different models that fit better.

III. Feature processing

We have used a correlation table to check if there are any relationships between the features. There are some features with a correlation higher than 0.9. This value is enough to consider that there is a close correlation between them, so they don't give us new information. Hence, we will remove them.

IV. Methods and visualization

As it was required in the project description, we have implemented six regression methods in order to fit our model: Gradient Descent, Stochastic Gradient Descent, Least Squares, Ridge Regression, Logistic Regression and Regularized Logistic Regression.

A. Training

The first step is to import the training and the testing sets. We remove the related features and split the data into the three events using the masks we have created. Now, it is time for training, we will use all the methods with the following parameters.

$max\_iters = 50$ and $w_{initial} = np.zeros(tX.shape[1])$

- Least Squares
- Gradient Descent: $\gamma = 10^{-10}$
- Stochastic Gradient Descent: $\gamma = 10^{-10}$
- Ridge Regression: $\lambda = 10^{-5}, degree = 4$
- Logistic Regression: $\gamma = 10^{-30}$
- Regularized Logistic Regression: $\gamma = 10^{-10}, \lambda = 10^{-10}$

V. Estimating the generalization error

In order to estimate how well our methods are doing, we use cross-validation to compute the accuracy of the model. After running the cross-validation with all the models we observe that, as we have found out before checking the losses, ridge regression is still the best method, with an average accuracy of 0.803486.

VI. Results

Analyzing Table I, we conclude that Ridge regression is the best model, obtaining the lowest loss and the highest accuracy.

Model	Loss	Mean accuracy	Standar deviation
Least squares	1.018563	0.662481	0.074804
Gradient Descent	1.499854	0.687682	0.042842
Stochastic Gradient Descent	1.500395	0.662499	0.074803
Ridge Regression	0.865104	0.803486	0.023132
Logistic regression	1.499999	0.646651	0.078735
Regularized Logistic Regression	1.499999	0.646651	0.078735

Fig. 1. Loss and accuracy of the different models

A. Discussion

As the problem is a Binary Classification, we expected that the best result would be obtained using Logistic Regression. However, the results show that Ridge Regression obtains a better approximation. Even fixing γ and augmenting the number of interactions the result is still worse.

B. Submission

To make the final submission we apply Ridge Regression with parameters $\lambda = 10^{-5}, degree = 4$ to the split and processed testing set, obtaining a categorical accuracy of 0.828 in AIcrowd.

VII. Summary

In this project, we have performed all the regression methods taught in the Machine Learning course. Beginning with a first approach to data processing and following with different analysis using all the methods. We finally compare the accuracy of each model and conclude saying that Ridge Regression is the best method to model the Higgs Boson Machine Learning Challenge.