Binary classification for Higgs bosons

Simon Canales
EE master's student

Claudio Loureiro
EE master's student

Jordan Willemin
EE master's student

***Abstract*—The discovery of the Higgs boson was one of the most outstanding physics discovery of the last decades. The proof of its existence was made possible thanks to CERN particle accelerator. Basically, physicists collide protons into one another, which generates smaller particles (hopefully a Higgs boson). This paper presents a way to determine if a collision event's signature is the result of a Higgs boson or background. To do this, binary classification is made on feature vectors corresponding to collision events' signatures.**

I. Introduction

When two protons collide into each other, smaller particles are generated as by-products of the collisions. Higgs boson's lifetime is very short, which makes it impossible to observe it directly. However, scientists are able to measure its "decay signature". Decay signatures of collisions involving Higgs boson and those of collisions not involving Higgs boson look similar, but using machine learning techniques we will try to determine if a given decay signature is that of a Higgs boson or not.

This paper proposes a way of designing a classifier in order to determine if an unlabeled feature vector corresponds to a Higgs boson or not. Each vector is composed of 30 features corresponding to the decay signature of a collision event. To train our model, 250,000 labeled feature vectors of the CERN dataset are used. Different models and model parameters are compared and cross-validation is used to determine which classifier is the most efficient.

First, the dataset is explored and pre-processing techniques are applied on it in order clean the data. Then, a way of computing the model parameters is presented. At last, classification of the test dataset is made.

II. Models and methods

Before putting our data through machine learning algorithms, let's have a closer look at it. Our data is composed of a training and testing set. The training set contains 250,000 samples each composed with 30 features. In addition for each sample we have its correct label. The testing set is composed of 568,237 samples with also 30 features each. The label for each sample denote with the event corresponds to an Higgs boson or if it is likely to be another type of event.

A. Observing and cleaning data

A quick look at the training dataset tells that we have 65% of one category of data and other 35% of the other category. Therefore without doing any machine learning algorithm, one could already achieve a 65% accuracy. Looking more deeply into the features, we denote two types of features: continuous type of data and categorical type of data. Statistics can be made out of this categorical feature and compute the proportion of "Higgs bosons event" for each category. The distribution is summarized in the following table.

	Higgs Boson	Other event
Category 1	25%	75%
Category 2	36%	64%
Category 3	51%	49%
Category 4	30%	70%

Another thing which is easily identifiable is the fact that some of the data is missing. Indeed, for most of the features, some data has an arbitrary value of -999 (for the last feature vector, corrupted data takes an arbitrary value of 0 instead of -999). Figure 1 shows the training dataset for two features. Vectors corresponding to "Higgs boson" event are in yellow. Missing data is clearly identifiable for each feature.

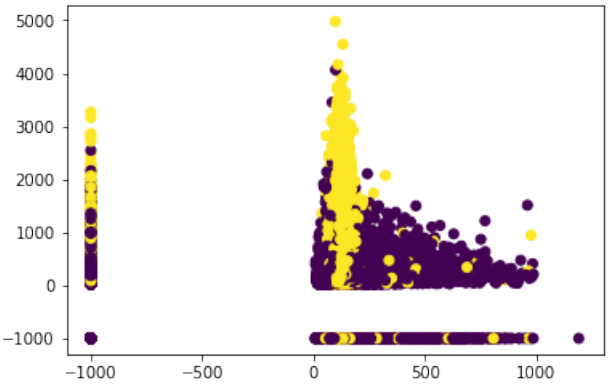

Figure 1. Training dataset with missing data (Higgs boson events are in yellow, unknown units)

In order to achieve better separability of the dataset, the missing data was given the mean value of the feature it is corresponding to. The dataset was then standardized so that each feature has the same "weight", no matter its unit. The cleaned, standardized dataset for two features can be seen in Figure 2. Visualizing these features shows a good separability. Two features, however, contained a lot of missing data and assigning a mean value to them resulted in feature vectors which would not correspond to the actual underlying

distribution of the feature. These features were considered not relevant for the classification task and removing them from our dataset resulted in faster computation and better accuracy.

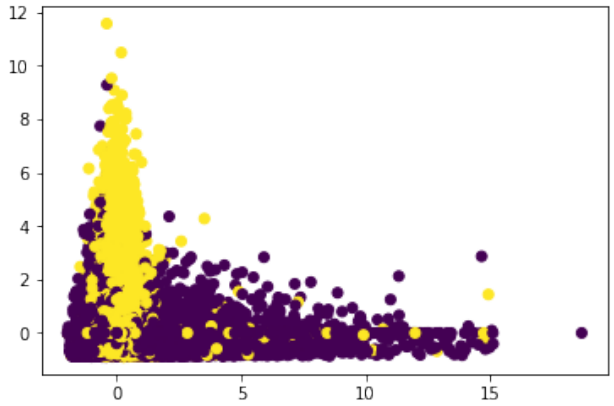

Figure 2. Cleaned and standardized training dataset (Higgs boson are in yellow, no units)

B. Model selection

In this section, methods to improve the classification accuracy are presented. All models are tested with cross-validation using a quarter of the training dataset to train the classification model and the remaining three quarters of the training dataset to test it. This process is iterated 4 times so that each quarter is used to train the model once, and accuracy results are averaged in order to give the actual accuracy of our model.

The different tested methods are listed bellow:

- Linear regression using gradient descent
- Linear regression using stochastic gradient descent (SGD)
- Least squares regression using normal equations
- Ridge regression using normal equations
- Logistic regression using gradient descent or SGD
- Regularized logistic regression using gradient descent or SGD

In general, for well-chosen initial conditions, the Ridge regression using normal equations was the method giving the highest accuracy.

In order to make the dataset more separable and avoid overfitting, we added polynomial basis functions $\phi(x_n) := [1, x_n, x_n^2, x_n^3, ..., x_n^M]$ of degree M for the input data x_n as new features. Our results tends to prove that increasing the degree of the polynomial basis results does not improve accuracy for the least squares regression using gradient descent. Our guess is that since there is no regularization term for this method, adding features resulted in overfitting. This observation could be generalized for all our methods that do not use regularization.

For methods that uses a regularization term, the accuracy tends to improve as the degree of the polynomial basis increases. Figure 3 shows the accuracy of the Ridge regression model, which uses a regularizer of the form $\Omega(\mathbf{w}) = \lambda\|\mathbf{w}\|_2^2$ to penalize large $\mathbf{w}$'s, for different values of the regularization term λ and different degrees of the polynomial basis M.

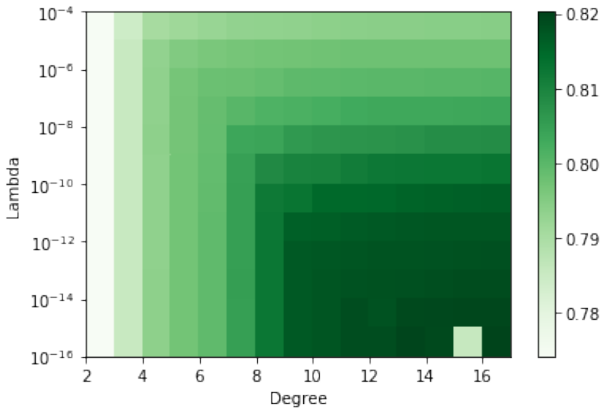

Figure 3. Accuracy of the Ridge regression for different polynomial basis degrees and values of lambda. Dark green means higher accuracy, scale on the right.

The best accuracy is given for $\lambda = 10^{-15}$ and for a polynomial degree of 16.

Finally, for logistic regressions, the value -1 has to be allocated in order of the 0 labels in order to improve their efficiency. However these methods does not prove to be better than our ridge regression.

III. Results

We can summarize the best accuracy obtained for each method with tuned parameters.

- Linear regression using gradient descent : 74.8%
- Linear regression using SGD : 71%
- Least squares : 78.5%
- Ridge regression : 81.9%
- Logistic regression : 69.1%
- Regularized logistic regression : 69.2%

IV. Conclusion

In this paper, a binary classifier was designed and tested on the CERN decay signature feature vectors dataset, in order to determine if a given decay signature is that of an event involving a Higgs boson or not. To have a more separable dataset, some pre-processing steps were proposed.

In particular, some features which were judged unnecessary were removed, and others were added using polynomial basis functions for the input data. Different regression methods were tested through cross-validation, and it was concluded that the best accuracy (approximately 82%) is achieved using Ridge regression.