Pattern classification and machine learning, Project I

Charlaix Ella, Rutagarama Maxime, Steinmann Raphaël
Department of Computer Science, EPF Lausanne, Switzerland

Abstract—**In this report we will analyze a dataset from the CERN which references different informations about particles and tells if a given particle is or not a Higgs Boson. We will fit different models to this dataset and find the best one in order to later be able to predict if a particle, about which we know some informations, is a Higgs Boson.**

I. Introduction

The dataset consists of 250000 observations that each list 30 values about a given particle and tells as a result if it is indeed a Higgs Boson with a binary output (-1 or 1). The values of the features can be missing, in which case it is set to -999. We will first of all proceed with the data pre-processing, and then apply our models to it. We will furthermore evaluate the efficiency of each of these models using the cross validation method and finally discuss about our results.

II. Models and Methods

Before considering models, the first thing we have to do is to analyze and modify the data in order to make it usable.

Data Pre-Processing We decide to consider a value as an outlier when it is too far from the interval between the first and the third quartiles. We replaced these outliers as well as the missing values by the mean of the other values of the column. In order to be able to compare the informations given by different dimensions of our data (which will be necessary when we will compare the variance of our data in the PCA procedure for example), we also standardize it.

To keep only the dimensions of our dataset that are relevant, we use the PCA method. This method consists in projecting our dataset onto a lower-dimensional space. For this purpose, we find the eigenvalues and the eigenvectors of our covariance matrix. After that, we keep only the ones that explained at least 0.95 of the variance After projecting our initial dataset onto our new dimensional space, we end up with 25 dimensions instead of our previous 30.

Now that our dataset has been cleaned, we can fit different models to it.

Our Models We will in the first place consider the following models seen in class:

- Linear regression using gradient descent,
- Linear regression using stochastic gradient descent,
- Least squares regression using normal equations,
- Ridge regression using normal equations,
- Logistic regression using gradient descent and
- Regularized logistic regression using gradient descent.

In addition to that, we will use polynomial basis to proceed with polynomial regression using some of the above models namely Least squares regression with normal equations, Ridge regression using normal equations and Logistic regression using gradient descent.

Model evaluation Once we know which models we will use to fit the data, we have to consider a way to evaluate and compare the efficiency of each of them. For that purpose, we split the data into two random subsets, the first one having 2/3 of the total size of the entire dataset and the second one 1/3. The first subset is the Training Set, on which we fit our model, and the second one is the Testing Set, which allows us to test our predictions.

Because we split the data in two, we have less information (both for the Training and the Testing set). To remedy this problem, we also implemented a cross validation method. The principle is the same but here we split the data into K set. The size of the Training Set will therefore be (K-1)/K the size of the entire data while the Testing set will be 1/K. The same procedure will be reiterated K times, with the Testing Set being another set for each iteration.

We use two values to evaluate the efficiency of our models: the percentage of right values obtained when making predictions on the Testing Set, and the AMS. AMS, or approximate median significance is a function described and derived in the Higgs Boson Challenge documentation, and the aim is to maximize its value.

III. Results

Models efficiency We show in the next table (Figure 1) the evaluations of each of the models we tested, with both fitting percentage and AMS as we discussed in previous section. Note that we finally did not perform PCA on the data since it was reducing the efficiency of our methods too significantly.

Method	Fitting %	AMS
Linear Regression, GD	72.5278%	625.3430
Linear Regression, SGD	72.5264%	625.3176
Least squares, normal eq.	77.3616%	698.7492
Ridge regression, normal eq.	77.3664%	698.8190
Logistic regression, GD	77.3440%	698.4929
Logistic regression, Newton	77.2896%	697.7004
Reg. logistic regression, GD	73.9312%	647.3215
Least Square poly. basis	79.5824%	730.5128
Ridge Regression poly.basis	81.4304%	756.1389

Fig. 1. Models efficiency evaluations

Degree and lambda selection for ridge regression We noted that our best results came from the ridge regression with a polynomial basis. We hence tried to optimize this model by finding the value of lambda and the degree that optimize the fitting percentage and AMS. We found the best result for degree 9 and lambda around 10^{-6}. Then we found more precisely the best lambda by plotting the fitting percentage against more values of lambda (Figure 2). The best result is the one we wrote in the previous table (Figure 1).

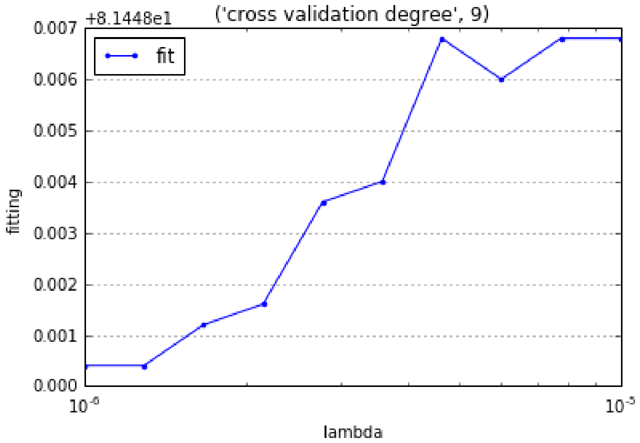

Fig. 2. Fitting percentage against lambda values for Ridge polynomial regression of degree 9

IV. Discussion

The submissions on kaggle were coherent with our performance estimations using data splitting and cross validation. Thanks from the determination of our models efficiency, we can argue that linear models are less accurate than our polynomial ones. This makes sense given that we give more degrees of freedom to our function to fit our data, which is quite complex. We are surprised that the logistic models did not perform very well since they are exactly made to predict binary outputs like the one which is given to us. Furthermore, we only used simple methods seen in class with some optimization and no package. Otherwise, neural networks seem to be the best methods for this problem.

V. Summary

The final objective of the project was to generate several models that could predict if a particle was a Higgs Boson or not, knowing some of the information of the later. The ridge regression model using a polynomial basis of degree 9 and a specific lambda, which was our best tested method, allowed us to make a 81.671% accurate prediction on kaggle.