

Machine learning - Project 1 report

Group 21 – Volodymyr Loyko, Shiyue Nie, Thomas Sanchez

I. INTRODUCTION

This report will present the way we approached the problem of the classification of the Higgs Boson. We will first go through the transformation that we applied to the data, then move to the different methods we used to try to solve the problem and finally will speak in greater detail about our best submission.

II. FORMATTING THE DATA

The data we were given about the Higgs Boson contains 30 columns of measurements. Many of those columns contain NaNs, so it is needed to clean them before using them to predict the outcome of the experience on other data, in order to get unbiased results, as those NaNs, taking either the values -999 or 0 , are strong outliers. We will briefly describe all the techniques we used in order to format our data.

1. Counting the NaNs : First of all, we know that we will eventually get rid of the NaN values in our data, and we want to keep the information that they represent. Maybe having a NaN in a given column is correlated to the measurement of a Higgs Boson. What we did is add a column to our data which kept track of the number of NaN found in each row.

2. Sanitation of the Data : Now that the information about the NaN is stored in our data, we can safely remove them. We did the following for each column of our data which contained some invalid data. We first retrieved all the "valid" points in the column, i.e. those who did not have a NaN value, and then replaced the value of the unavailable points with the median of the valid points. Note that we used the median and not the mean because the median is robust to outliers. Indeed, imagine we had all the points in the region -1 to 1 , the NaNs to -999 and an outlier at a very large value, then taking the average of the valid data including this large outlier would largely influence the result, while taking the median would ignore this outlier.

3. Standardization of the Data : The second step in the formatting of our data was to standardize them. Having the gradient with the same magnitude in each direction yields a much faster convergence for the methods using gradient descent, like logistic regression. And it also turns out to be practical if we wish to construct a polynomial basis, as it will keep the large values of our data set bounded. It is important to note that in the case where a column has standard deviation 0 , i.e. all its values are the same, we then remove it from the data, as it contains no useful information.

4. Polynomial basis : The last step in the formatting of our data was to take a polynomial basis of degree D for our data. We did not add the cross terms of the form $x_1 x_2 x_3$ for polynomials of degree three, as this would have given far too much features for our model to fit. For a basis of degree D and an input with M features, our polynomial basis was of the form

$$[1 \ x_1 \cdots x_N \ x_1^2 \cdots x_N^2 \ x_1^D \cdots x_N^D] \text{ where } \mathbf{x}_i \in \mathbb{R}^M \quad (1)$$

5. Additional notice : When we turn to prediction, we will apply the same transformation to our testing set, but we will not recompute the median, mean and standard deviation, we will apply those that we had found for the training set, because we would otherwise end up with a biased result.

III. RESULTS

A. Methodology

The methodology we followed to obtain our results was an extensive use of the cross validation while varying all the other parameters (λ in the ridge regression as well as the polynomial degree) we had in a grid-search fashion. We would then choose the combination that yielded the minimal error, i.e. the smallest percentage of classification error. We did not iterate on γ for the iterative methods, as the γ will only change the speed of convergence of the method. We rather tried to pick on that seemed to work well and extended the number of maximal iterations, until convergence.

B. General results and comments

Let us now present quickly some results that we obtained with each method. Note that, as the Ridge Regression stood quickly out as the best for our purpose, we did not extensively try to use the other methods, hence the results shown here for all other methods could easily be improved, and they are just present here for comparison's sake.

Method used	GD	SGD	LS	RR	LR	RLR
Accuracy (%)	70.0	70.1	71.9	82.96	77.4	0.69
Degree	1	1	1	10 – 14	2	3
γ	2^{-6}	2^{-12}			10^{-8}	10^{-6}
max_iters	2000	2000			2000	2000
λ				0.7 – 54		7.74

Table I

BEST ERROR IN THE MEASUREMENTS FOR RESPECTIVELY GD (GRADIENT DESCENT), SGD (STOCHASTIC GRADIENT DESCENT), LS (LEAST SQUARES), RR (RIDGE REGRESSION), LR (LOGISTIC REGRESSION), RLR (REGULARIZED LOGISTIC REGRESSION). THE DEGREE FIELD CORRESPONDS TO THE DEGREE OF THE POLYNOMIAL BASIS WE USED FOR THE PREDICTION, THE γ IS THE PARAMETER FOR EVERY METHOD IMPLYING SOME KIND OF GRADIENT DESCENT, MAX_ITERS, λ IS THE PARAMETER FOR EVERY METHOD THAT IMPLIED PENALIZATION OF THE LARGE WEIGHTS. NOTE THAT THE RESULTS FOR RIDGE REGRESSION SEEM QUITE VAGUE, AND IT IS BECAUSE THEY WERE PERFORMED ON A DIFFERENTLY MODIFIED DATASET. THEY WILL BE GIVEN EXTENSIVELY IN THE TABLE II

First of all, let us talk about our biggest surprise in the results. We would have expected both Logistic Regression (LR) and Regularized Logistic Regression (RLR) to do extremely well in this problem, as this is actually a classification problem : we must find whether the output is a Boson. However, we were not able to get very successful results with them, due to the fact that the convergence was very slow and that the picking of the parameter γ

was quite complicated. Indeed, choosing it too large would cause the cost function to oscillate after a few hundreds of iterations, and picking it too small would cause the convergence to take tens of thousands of iterations. We did not try to dynamically adapt it, rather focusing on getting better results using the Ridge Regression. This convergence was slow even when trying to start with initial weights that were the result of a previous iteration with another method (e.g. the best w from Ridge Regression, ...). Another problem linked to the fact that those methods are very slow is that they make cross-validation extremely painful, as the algorithm must compute thousands of slow steps for each method each time. However, we believe that if we had a more time and a deeper understanding of Regularized Logistic Regression, we could have got very good results with it, as it seems instinctively to be the most suited method for this problem : it is a classification method which allows for high degree polynomial basis, which allows to fit very non linear behaviors while not overfitting the data.

We did not use the Gradient Descent and Stochastic Gradient descent very much, as they did not in this case present any advantage over the Least Squares and the Ridge Regression : they are iterative, depending on γ , while in the least squares, we only solve a linear system of equations. However, they are slightly faster than the logistic regression methods, as they do not require the evaluation of an exponential for each entry at each iteration. Another reason for not using it was that they tend to overfit the data if we start considering high degree polynomial basis as we typically did, as they do not penalize large weights.

The Least Squares has the advantage of being quick to compute, and giving a good solution immediately (the results it gives are good even with the raw data !). Its only problem is that it also overfits the data for bases with large polynomial bases, as it does not penalize weights, like (stochastic) gradient descent and logistic regression.

C. Best results : Ridge Regression

1) *Special formatting of our data:* Due to lack of time, we did not test this additional formatting of data with the other models. We only selected the model that was doing best already, and it was the Ridge Regression. What we did is split the dataset into four distinct datasets, using the JET_split attribute to do so, as it ranges from 0 to 3 with only integer entries. It appeared then that some columns were filled with NaNs, and we then discarded them from the concerned datasets. The number of features then varied from a split to the other. We then split the testing data in the same fashion, and perform the prediction on each split dataset, before merging all the results we get. We also allowed each of the split dataset to be cross-validated on different parameters, in order to find relatively quickly the best value for each of them.

2) *Details of the results:* The detail of the results we obtained is summarised in the table II below. We see that the prediction error is overall higher than in what we have in the table I. This could be because we have a smaller dataset each time that we train on, and so the result is slightly less precise than when we simultaneously train on the whole dataset. Note that the splits vary a lot in the number of elements in them, and that the splits 0 and 1 have way less features than others : it is due to the fact that some columns

end up being filled only with $NaNs$ and are hence excluded as they do not bring any information.

	Split 0	Split 1	Split 2	Split 3
Size of the split	99913	77544	50379	22164
Number of features	19	23	30	30
Cross Validation Error (%)	19.5	24.2	20.6	20.5
Degree	10	12	14	14
λ	0.695	2.976	54.556	26.367

Table II
EXPANDED RESULTS FOR THE BEST SUBMISSION WITH RIDGE REGRESSION ON A SPLIT DATASET. THE RESULTS ARE THE ONES WITH OBTAINED WITH A CROSS-VALIDATION WITH SEVERAL POLYNOMIAL DEGREES AND λ RANGING FROM 10^{-2} TO 10^2

Having this split dataset allows us a lot more flexibility in picking the weights and matches together particles that have the same behavior, hence the fact that it is our best result. We want also to show an example of cross-validation, on the figure 1 below.

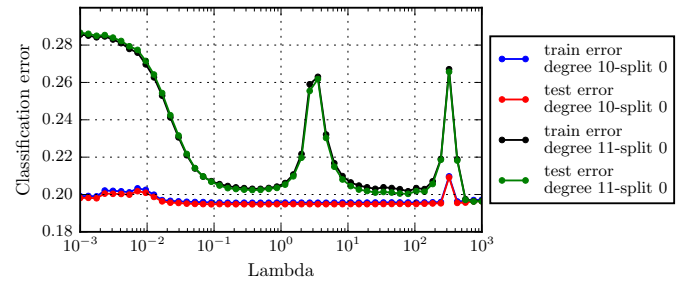


Figure 1. Cross-Validation for the split 0 for polynomials of degree 10,11 and λ ranging from 10^{-3} to 10^3 .

We have here in our figure a behaviour that could seem strange, because the cross-validation plot has peaks, and is flat otherwise. It is due to the fact we do not use a cost function here, but that we rather compute the classification error (in percentage of error). We know that a small cost function implies a small classification error, but we also know that the converse is not true : a small classification error will not imply a small cost function. Hence, the graph that we have here can seem quite unpredictable. We choose to use the classification error over to usual cost function because it allowed for searching for a wider range of solutions : some that would not minimize the cost function, but would actually classify our data quite well.

IV. CONCLUSION

We were able to approach the problem of finding the Higgs Boson from with quite different Machine Learning methods, but were very surprised in the result. While we expected the (Regularized) Logistic Regression to perform well, it presented too many fallbacks for us to use it reliably : the convergence was very slow, the result not extremely accurate. We used then the Ridge regression, which isn't iterative, and got our best results with it.