Project 1 : Searching for the Boson

Ducommun Colin, Elalamy Rayan, Van de Velde Anne-Sophie

October 29, 2018

1 Introduction

In this project, we were given a set of datas on which we had to detect the Higgs Boson. To do so, we were given a training set `train.csv` on which we performed our training, and a test set `test.csv` on which we performed our testing. We first had to clean and manipulate the data, and then we performed several methods in order to predict whether or not the Boson was there. The intuition would be that the logistic regression would give us the best ratio, because all the other methods implemented were prediction methods and not classification methods.

2 Models and Methods

2.1 Model based on the data

We first had a look on the features we were given. We saw that we had a lot of non-measured values, replaced with -999. Since in certain columns their number was way too significant to be able to remplace them with the median or the average, we chose to delete the columns where they appeared. To perform our first tests, we separated the features in two : a vector y containing -1 and 1 corresponding to the absence (resp. the presence) of the Boson, and a design matrix z containing the features.

Our approach consisted in separating the set training set in two : 80% of training and 20% of testing, to be able to test more simply our methods.

We then chose to add features by multiplying them together. This technic is quite classical in Machine Learning and adds some value to the results. Indeed our training was more performant once we added those new 'artificial' features. We then arrived at a new design matrix z.

We then decided to build a polynomial feature based on z, keeping in mind that it could cause some overfitting. Moreover, this treatment impact the interpretation of the features, and therefore the explanatory aspect of the data analysis. However the aim of this project is to perform the best performance, we thus decided to privilege the approach that gave us the best predictions.

By looking more closely at the features, we also remarked that one of them (jet_num) was just consisting in integers going from 0 to 3. We saw that the probability of the Boson being present was of 51% when this feature was 2, when this probability is only of 34% in general. We thus performed a so-called one-hot encoding, supposing that there were no hierarchy in jet_num. Unfortunately, the results were not good with this approach and we decided not to use it.

2.2 The methods

Our training consisted in the implementation of several methods: a linear regression using gradient descent; a linear regression using stochastic gradient descent; a least squares regression using normal equations; a ridge regression using normal equations; a logistic regression using gradient descend; a regularized logistic regression. Moreover as the majority of these methods are prediction methods, we use the ratio of predictions as the loss. We then compared these models and chose the appropriate one to find the Higgs Boson on test.csv . To see more precisely how to use each method, one should read the README file provided with our project.

It is important to notice that we tested ridge regression with several λ's, and then chose the one that gave us the best result.

3 Results

For the linear GD and the linear SGD, our prediction did not exceed the 73% of accuracy, while the least squares regression gave us an accuracy 80%. Logistic regression was not more performant than least squares. Overall, it is the ridge regression that gave us the best results with 81% of accuracy. We then tried again all the methods with a cross-validation, but the results were not better. We thus chose to implement the ridge regression.

4 Discussion

We could have separated the matrix z in four according to the categorical feature jet_num, and train on each of the four resulting matrices. The problem was on the merging of those 4 groups, and on the running time of such an approach. We thus decided to try with the one-hot encoding, which did not give us better results. Overall, we did not use this categorical feature.

Cross-validation did not give us better results, which is not so surprising as it is to prevent overfitting.

We could have done the average of all the weights of all the methods, but since ridge regression was more performant we decided to only use its weights.

One should notice that our approach of choosing the best λ in ridge regression could cause overfitting, and that we could have done an average over all the λ's but once again the results were better using the best one.

Another approach could be to look more closely at the physical meaning of the features. This could lead to a better treatment of the data and we tried to read some papers about the Higgs Boson, but our physical background was not deep enough to concentrate on this approach.

5 Summary

Even though it is not a categorical method, it is the ridge regression approach that gave us the best results. This is surprising, because we were expecting the logistic regression to be a better model for this classification problem.

Moreover, we made some choices in the treatment of the data that could have been different, and we are aware that it may lead to some overfitting.

Our main objective was to obtain the best result, and not to have the best explanatory model.