Project 1 : Searching for the Boson

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

October 27, 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 different 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, remplaced 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 there were some -999. 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.

By looking more closely at the features, we also remarked that one of them was just consisting in integers going from 0 to 3. By looking more closely, we saw that the probability of the Boson being present was of 51% when this feature was 1, when this probability is only of 34% in general. We thus separated the features in 4 groups : one for each value between 0 and 3, and performed our methods on each of those groups.

2.2 The methods

Our training consisted in the implementation of several methods :

i) a linear regression using gradient descent, called least_squares_GD ,

ii) a linear regression using stochastic gradient descent, called lest_squares_SGD ,

iii) a least squares regression using normal equations, called least_squares ,

iv) a ridge regression using normal equations, called ridge_regression ,

v) a logistic regression using gradient descend, called logistic_regresssion ,

vi) a regularized logistic regression, called reg_logisitic_regression .

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.

3 Results

For the linear GD and the linear SGD, our prediction did not exceed the 65% of accuracy, while the least squares regression gave us an accuracy 80%.

4 Discussion

5 Summary