Project 1

Monika Avila Marquez
EPFL

Abstract—

We develop a classification model that allows to determine whether the decay observed corresponds to a High Boson particle or to another one. We propose a boosted linear regression model and a logistic one. For the former, we estimate the vector of weights using least squares method which aims to minimize the MSE (Here descibe the boosted model!!). For the latter, we use the maximum likelihood method modified by adding a random vector. This modifications avoids convergence to local minima. Additionally, we only update the weights if we have an improvement of the loss function.

I. INTRODUCTION

The aim of this project is to develop a classification model for the decay processes observed. Such that we are able to predict if they correspond either to a High Boson particle or to another one. For this purpose, we propose to use a boosted linear regression model and a logistic one. For the former, we estimate the vector of weights using least squares method which aims to minimize the MSE (Here descibe the boosted model!!). For the latter, we use the maximum likelihood method. For the maximization of the likelihood we use Gradient Descent that we modify by adding a standard gaussian d-dimensional vector which allows to avoid to be converge to local minima. Additionally, we only update the weights if we have an improvement of the loss function. Results ..., Conclussions...

II. THE MODEL

In order to estimate the likelihood that an event signature is the result of a High boson process, we model the data using a boolean variable called $y_i, \forall i \in \{1, 2, ..., 2500\}$ as:

$$y_i = \begin{cases} 1, & \text{if} \quad prediction_i = b \\ 0, & \text{otherwise} \end{cases},$$

Since the expectation of a boleaan variable is equal to the probability that the variable is equal to 1, we can model the probability that the event is the result of a High Bosson process with a linear model:

$$p_w(y_i = 1|\mathbf{x}_i) = E(y_i|\mathbf{x}_i) = \mathbf{x}_i'w$$

Where $\mathbf{x}_i$ is a vector of D attributes considered in the regression and $w \in \mathbb{R}^D$ are the weights. However, as it is well known the biggest pitfall of this model is the fact that the estimated probability is out of the boundaries of the closed interval $(0, 1)$. In order to address the problem mentioned, we use the logistic regression. This model lies within the framework of a Generalized Linear Model. In this case, the probability of having a HB process is modeled as following:

$$p_w(y_i = 1|\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i'w)}{1 + exp(\mathbf{x}_i'w)}$$

For purposes of simplicity and because it is well known, we don't present the expression of the likelihood of the model.

III. METHODS

In the previous section, we have presented the models that we used. Now, we will describe the methods employed for the estimation of the vector of weights w. Since we have two different models, the estimation technique will be different. For the linear regression model, we use Ordinary Least Squares method aiming to minimize the MSE. For this, we used the normal equations and the closed from solution for the weights. For the logistic regression model, we aim to maximize the likelihood. For this we used Gradient Descent. We modified the gradient descent method by adding a d-dimensional gaussian random vector such that we avoid convergence to local minima.

$$w^{m+1} = w^m - \lambda \nabla l(w^m) + v^{m+1}$$

with $v^{m+1} \sim d - Gaussian(0, I)$. Where I is a DxD identity matrix and 0 a vector of zeros in $\mathbb{R}^D$. Moreover, we add the condition that we ill update the weight in each iteration only if the loss function is lower than in the previous step. While it is true that adding this condition increases the computational burden, the benefit is that we improve our prediction results since we reach better convergence.

IV. THE DATA

V. RESULTS

Show evidence to support your claims made in the introduction.

VI. DISCUSSION

Discuss the strengths and weaknesses of your approach, based on the results. Point out the implications of your novel idea on the application concerned.

VII. SUMMARY

Summarize your contributions in light of the new results.