Project 1

Monika Avila Marquez
EPFL

***Abstract*—**

We aim to predict whether an event is the result of a High Bosson process or not using a linear model and a logistic one. For the former, we estimate the vector of weights using least squares method which aims to minimize the MSE. In this case, we use the normal equations and not an iterative method. For the latter, we use the maximum likelihood method. In this case we use Gradient Descent for the maximization of the Likelihood.

I. Introduction

The aim of this project is to determine whether an event signature is the result of High bosson or other process. In order to ...

II. The Model

In order to estimate the likelihood that an event signature is the result of a High bosson 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:

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

Where $\mathbf{x}_i$ is the vector of attributes considered in the regression and w 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:

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

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.

IV. The Data

In this section we'll discuss about the data that we've used to test and compare the models that we've mentioned above. Our training dataset consists of in total 30 features and 1 label which indicates if it's from Higgs boson or from other event. The dataset itself is generated from simulation which mimics the actual particle collision events in which Higgs bosons (with fixed mass 125GeV) were produced and 3 other background processes.

The features in this dataset consist of both raw data which were measured by actual sensors and derived quantities computed from raw features. We can see description of first 5 features in table . And in figure we've checked separability of signal from background on sample 2 features.

Even thought we have considerably large variation of features ALMOST 70% of all events lacks at least one feature. We've simply set mean values of features instead of null variables, to minimize effect of missing data.

And finally, we've applied 2 discriminative transformations (square and sine) to each features and created pool of 90 features including original one. From here, we've selected best combination of features to use by forward selection method. In figure, we can see the relationship between MSE of trained model and number of features, using this algorithm.

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.