PCML CS-433: Higgs Challenge Project

Joachim Muth, SCIPER 214757, joachim.muth@epfl.ch
Junze Bao, SCIPER 266983, junze.bao@epfl.ch
Chanhee Hwang, SCIPER 260872, chanhee.hwang@epfl.ch

School of Computer and Communication Sciences, EPF Lausanne, Switzerland

***Abstract*—...**

I. Introduction

The ATLAS experiment consists of collisions between protons. Particles created by these collisions are detected by sensors, producing a sparse vector of about a hundred thousand dimensions. Analyzing these data, ATLAS team try to estimate if the detected particles comes from a Higgs boson decay.

The *Higgs boson machine learning challenge* consists of a large dataset of particles decay detections labeled as Higgs or background. The dataset is composed of thirty features and is already cleaned from a lot of well-known background effect well-known by the ATLAS team. Also, in order to balance the great number of background events compared to Higgs events, the size of both dataset is balanced. [1]

II. Models and Methods

A. Split of the data

In order to chose which features to keep in the machine learning modelling we did, and to deal with the great number of *NaN* in the data set, we analyzed their distrubution.

It show that it's possible to cathegorize the events into four different sets based on the number of jets (*int* $\in [0, 3]$). There is a physics reason behind it, since some measures make no sens for some jet numbers. [1].

Once this split proceeded, subsets share all the same defined features, except the estimated mass m_H of the Higgs boson candidate. Once again, these subsets are split into two subsets to obtain, finally, eight different subsets to model (see figure 1)

B. Features selection

As our eight datasets are well splitted, we can just discard the features only containing *NaN* values. The remaining features are well completed. This does not mean that they are all interesting. The histogram of their frequencies show candidates of possibly useless features which have almost no variance (see figure 2)

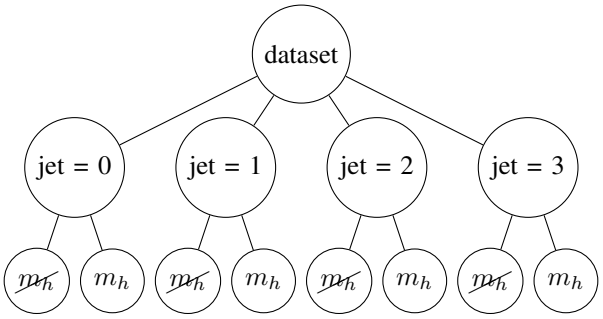

Figure 1. Split of the dataset into eight different cathegories

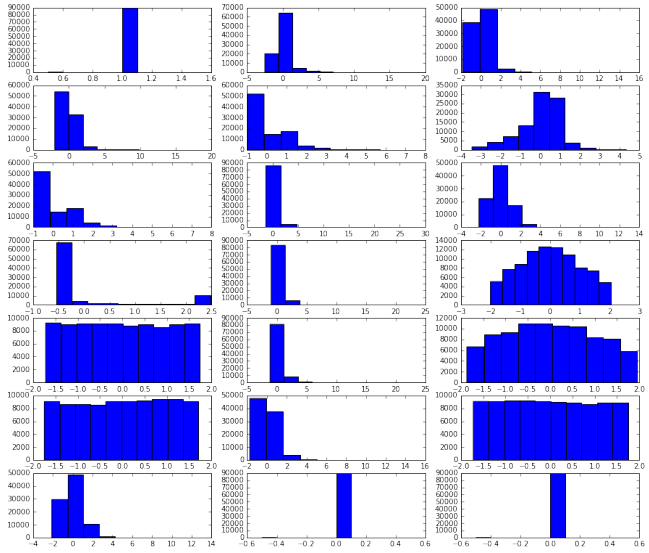

Figure 2. Frequencies histograms of feature values of subset jet0 m_h

C. Polynoms build

We construct a polynom matrix of degree d with our features matrix. The degree is chosen by 4-fold cross validation and is different for each of our eight datasets.

D. Logistic Regression

As we face a *classification* problem, we choose to use a *Logistic Regression* algorithm which shows good result as binary classifier. [2] This choice leads to two problems to problems to take care:

1) Rescale of the y's: As LR must be used for result values 0 or 1, we rescale $yin-1,1$ to $y \in 0,1$.

2) Deal with overflow: Double value can handle number until $\sim 10^{304}$. This limit is reached with $exp(x)$ for $x \simeq 700$. To handle this problem, we use Taylor series of first order in both exponential and sigma functions. [3]

For $s >> 1 \rightarrow 1 + exp(s) \simeq exp(s)$ then

$$log(1 + exp(s)) \simeq log(exp(s)) = s.$$

$$exp(s)./(1 + exp(s)) \simeq 1$$

For $0 < s << 1 \rightarrow exp(s) \simeq 1 + s$ and then

$$log(1 + exp(s)) \simeq log(2 + s) \simeq log(2) + s/2$$

$$exp(s)./(1 + exp(s)) \simeq 1/2 + s/4.$$

E. Selection of best parameter

We apply a grid-search method on each of the eight data subset for two parameters: **polynomial degree** $d \in [0, 15]$ and **regularization parameter** $\lambda \in [1^{-10}, 1]$. For each pair of parameter is test through a 4-fold cross-validation and scored by RMS error.

III. Results

IV. Discussion

V. Summary

...

References

[1] C. Adam-Bourdarios, G. Cowan, C. Germain, I. Guyon, B. Kgl, and D. Rousseau, "Learning to discover: the higgs boson machine learning challenge," 2014.

[2] M. E. Khan, "Logistic regression," 2015, machine Learning Course - CS-433.

[3] A. Donda, "Avoiding numerical overflow when calculating the value and gradient of the logistic loss function," 2013, stackoverflow.com.