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. Standardization

DO WE REALLY NEED IT ????

In order to work on comparable data, and not to overweight some of them, a standardization must be applied on them.

B. 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)

C. Features selection

WE MUST TEST IT AND CHOSE IF WE KEEP

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 **??**)

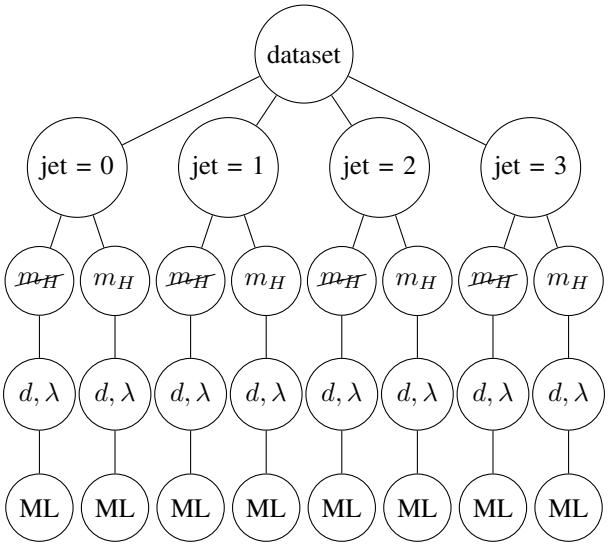

Figure 1. Overall view of modelling pipeline. Split of the dataset into eight different cathegories, search of the parameters and calculation of weight by Ridge Regression.

D. Polynomial building

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.

E. Selection of Ridge Regression parameters

We apply a 2-steps grid-search method on each of the eight data subset for two parameters: **polynomial degree** $d \in [0, 15]$ and **regularization parameter** $\lambda \in [1^{-10}, 1]$. Each pair of parameter is test through a 4-fold cross-validation and scored by RMS error. Figure 2 shows the result. Once the best parameters are chosen for each subsets, we zoomed on them and apply again a grid-search on the zoomed interval, to gain precision. Table I sums up the chosen parameters.

III. Results

This model gives us an accuracy of 83.08% when tested by 4-fold cross-validation. However, the accuracy against

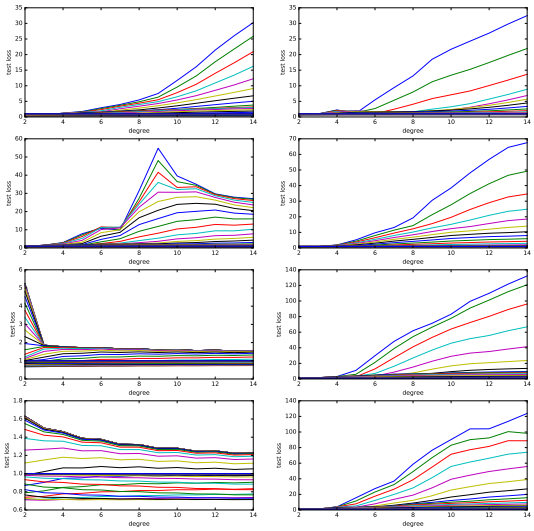

Figure 2. Grid search for parameter selection d and λ on the eight data subsets.

Subset	lambda λ	degree d
jet0 no mass	1.5998e-05	2
jet0 mass	3.2374e-05	2
jet1 no mass	0.1526	13
jet1 mass	3.2374e-05	2
jet2 no mass	0.0184	2
jet2 mass	7.9060e-06	3
jet3 no mass	0.6250	3
jet3 mass	0.0091	3

Table I
SELECTED PARAMETERS OF RIDGE REGRESSION

test set provided on Kaggle.com only scored 77.748%. The details of the accuracy of the model over each data subsets are listed on table II.

Subset	Percentage of Higgs	Accuracy of the model
jet0 no mass	0.0596	0.9437
jet0 mass	0.3243	0.7866
jet1 no mass	0.0932	0.9067
jet1 mass	0.3858	0.7236
jet2 no mass	0.1581	0.8699
jet2 mass	0.5327	0.7767
jet3 no mass	0.0704	0.9295
jet3 mass	0.3203	0.7102

Table II
DETAILS OF MODEL ACCURACY

IV. DISCUSSION

The fact that the prediction is better on the local test than in Kaggle.com could point some overfitting on our training method.

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.