

# Machine Learning - Higgs Boson Project

Lucien Michaël Iseli, Florian Maxime Charles Ravasi and Jules Eliot Gottraux  
*Master of Data Science, EPFL, Switzerland*

## I. INTRODUCTION

## II. PIPELINE

- 1) Copied 6 function we have to return from the labs
- 2) Changed the functions we copied from the labs such that they always assume that vectors are represented in  $(N,1)$
- 3) Plotted distributions of features to gain insight
- 4) Gradient descent on data to see what's going on
- 5) Try to remove features based on the plots, the ones that look like they don't add anything
- 6) Logistic regression, see if results make sense
- 7) Try polynomial expansion to have better accuracy
- 8) Try polynomial expansion with cross products to have better accuracy
- 9) Realize that the weird distributions of many features with extreme variance are most likely due to the fact that -999 values are unknown values. Seems weird that many of the values are around zero, and many of them are exactly -999
- 10) Normalize data without taking into account the -999 values, and set -999 values to 0. Such that they shouldn't affect the result as when 0 will be multiplied with the weight it'll be 0, i.e. not contribute.
- 11) Replotted the distributions of the features after normalizing and removing the -999 values
- 12) Notice that some plots look like uniform distribution or clearly don't give us insight on the result.  $=_i$  can remove them to have faster algorithm (NOT DONE YET)
- 13) When plotting the distributions of features without unknown values we noticed something very strange: all of the distributions are continuous except one! It only has 4 values
- 14) We thought that maybe this value is some sort of category, thus maybe we should treat them differently  $=_i$  we trained them separatly, we separate them in 4 categories based on that feature and train them separatly
- 15) Trained the model that way with linear regression (no expansion), obtained much better results
- 16) Tried logistic regression but results weren't as good
- 17) Tried to add feature expansion, square and sqrt (without cross products)