CS433-Machine Learning Project 1

Amaury Combes - Vincenzo Bazzucchi - Alexis Montavon

***Abstract*—The Higgs Boson Kaggle challenge was put in place by physicists in CERN in order to analyze the massive data gathered during their research with the Large Hadron Collider. The idea was to use the best algorithms to predict if a particle collision event was a signal of the Higgs Boson. This challenge was actually one of the biggest ever on Kaggle and we reproduced it in our Machine Learning class at EPFL.**

I. Introduction

TODO: at the end

II. Model and Methods

TODO: explain step chosen (what works, failed, why, using tables)

A. Preprocessing

After diving into the dataset the first question that came to our attention was what to do with the undefined values *-999.0*. Three options came to mind, setting them to 0, to the average of every valid values in each feature or to the most frequent value in each feature. We opted for the second option as it seemed more coherent than the first one (although there wasn't any clear differences on the final accuracy result) and it turn out to be better than the last one when we cross validated our model. This is done by the *mean_spec* function in the *preprocessing.py* file.
We then decided to standardize the dataset to avoid too big variations in its values. We implemented a classic standardization function that subtract the mean of each feature from every value in the column and divide it by the standard derivation, see *stadardize* function in the *preprocessing.py* file.
As we saw in class, linear models are not very rich, so we used the polynomial augmentation technique used in the lab session. This is realized by the *polynomial_enhancement* function in the *preprocessing.py* file.
Finally, on the advice of different TAs and the article [1], we chose to train our model on each "categories" based on the numbers of jets (this is given by the column *PRI_jet_num*). This is done with the *category_iter* function in the *run.py* file.

B. Models

III. Results

TODO: I guess best results we got and the exact technics and parameters, give exact loss (mean of cross validation maybe)

IV. Summary

TODO: Retrace best option we used in short

References

[1] V. S. Bernard Ong, Nanda Rajarathinam and D. Khurana, "What it took to score the top 2 percent on the higgs boson machine learning challenge," 2016, https://blog.nycdatascience.com/student-works/machine-learning/top2p-higgs-boson-machine-learning/.