Project 1 on Machine Learning team Yoor

Sergei Volodin[1], Baran Nama[1], and Omar Mehio[1]

[1]EPFL
{*sergei.volodin,baran.nama,omar.mehio*}*@epfl.ch*

***Abstract*—A classification dataset from the Large Hadron Collider simulations is being studied. First, the data is thoroughly explored using visual aids. After that, several basic Machine Learning methods are applied on preprocessed data. Results are evaluated using cross-validation. Model overview is given for each considered algorithm and the best model is chosen.**

I. Introduction

The Higgs boson is a famous elementary particle which was first predicted in 1960s and then discovered in 2012 [1]. Its famousness is due to two facts, first being the collosal amount of effort put into construction of the Large Hadron Collider and conducting the ATLAS experiment and the second being the fact that the Higgs boson is considered to be connected to the fact that particles have a mass.

The ATLAS experiment consists of protons colliding at near-relative speed. After collision, the resulting particles sometimes contain the Higgs boson. Itself, it is not detectable by LHC. However, it is possible to detect the particles that it is decaying to.

The data being studied comprises of 250 000 objects (train) each having 30 features. Each object represents a collision of a stream of protons. The data was not obtained during the ATLAS experiment, but rather from the simulation [2]. Features represent properties of detected particles. It is required to determine if the particles represent the Higgs boson.

The following paper claims that it is possible to use simple methods, such as Linear and Logistic Regressions to classify the data. In the following sections, the data is thoroughly studied and then the model is chosen based on reasoning and cross-validation.

1) What is the data (Simulation from LHC, details from physics)
2) What are we trying to do? Get the best classification score
3) Overview of data, diagrams of features, feature selection, feature augmentation
4) Methods and their choice (Linear regression, logistic regression, ridge regression) because of simplicity

II. Models and Methods

1) Least squares. Problem: missing data, overfit
2) Mean imputation. Problem: overfit, meaninglessness for some features
3) Feature binarization, add new feature 'feature missing', add squares for features for mass
4) Ridge regression using k-fold. Problem: low accuracy (?)
5) Logistic regression
6) Nearest neighbors?

III. Results

Shows that accuracy is good enough meaning that model selection was good

IV. Discussion

State that we can improve the accuracy by using non-linear classifiers?

V. Summary

We have shown that it is possible to detect the Higgs boson using linear methods and feature augmentation.

References

[1] https://en.wikipedia.org/wiki/Higgs_boson

[2] REPLACE ME