Project 1 on Machine Learning team Yoor

Sergei Volodin[1], Baran Nama[1], and Name here 3[1]

[1]EPFL
{*sergei.volodin,baran.nama,email3*}*@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

Claim: it is possible to use simple methods for this dataset

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.