CS433-Machine Learning Project 1

Ahmed Nour Achich - Ahmed Amine Ghariani
Department of Computer Science, EPFL, Switzerland

***Abstract*—Machine learning has grown in importance, providing methods and approaches to solve a wide range of issues in a variety of scientific domains. We explore and evaluate various supervised learning algorithms and how they work with a data set from CERN to predict the presence of the Higgs Boson**

I. Introduction

The Higgs Boson is a fundamental particle that has sparked a lot of interest in the physics world. Collision experiments, such as the Atlas experiment at CERN, have generated massive amounts of data that can be utilized to discover Higgs Boson signals. We're working on a binary classification problem in this project. Our goal is to predict particle collision events at CERN as either Higgs Boson signals or mere background noise. To obtain the results described in the next sections, we used approaches from exploratory data analysis, feature processing, hyper-parameter estimation through cross validation, visualization, and implementation of six basic machine learning algorithms.

II. Models and Methods

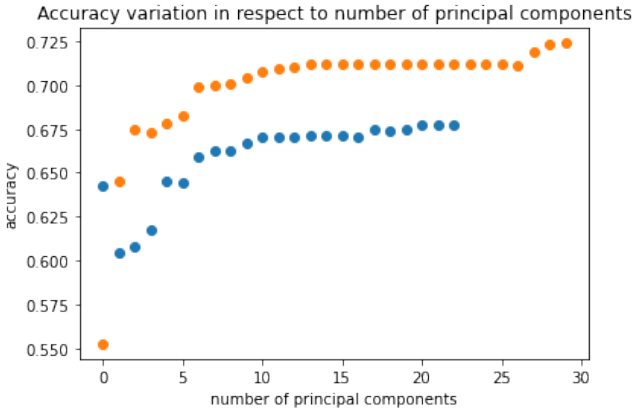
A. Machine Learning methods implementation

We began by implementing the six methods provided in the lectures, and then ran each one for the first time to see how the algorithms performed on the raw dataset.

B. Exploratory data analysis

In order to construct a better model for this classification problem, we attempted to better understand the data.

- There are 250.000 points in the data set, each with 30 features one categorical and the rest continuous. Raw quantities measured by the experiment's detectors and derived variables estimated by CERN's physicists using the raw quantities are separated into two groups.
- 11 of the 30 features had a lot of missing values or values that couldn't be computed. -999.0 is used to denote these values.
- All of those features are floating point numbers except for the number of jets, labeled PRI jet num, which can take 4 integer values in 0,1,2,3.
- The number of jet pseudo particles appearing in the detector is represented by PRI jet num.We noticed that the jet feature had a significant impact on the distribution of missing values. In order to avoid the Simpson's Paradox, we opted to segment our dataset based on the jet num value. We investigated partitioning our data-set into four parts (one for each jet num number), however the data-sets when jet num was two and three were substantially smaller, and overall performance was poorer than when we partitioned into three parts. As a result, tX0 denotes all data points with jet num = 0, tX1 denotes all data points with jet num = 1, and tX2 denotes all data points with jet num = 2 or 3.



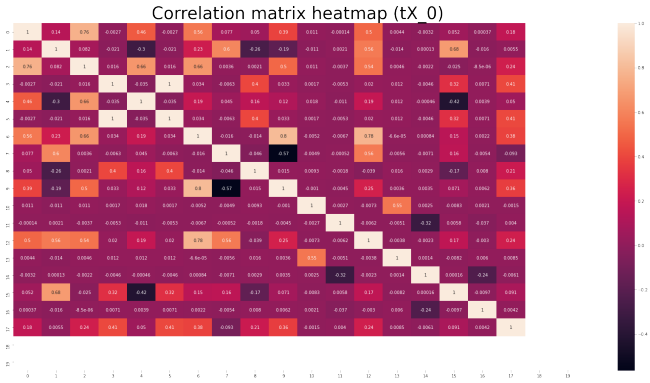


Figure 1. correlation heatmap for $tX0$

After removing the missing features from each data set, we computed the correlation matrix of the remaining features and attempted to reduce dimensionnality. We excluded features that were highly correlated with all others (where the correlation coefficient's absolute value was greater than 0.8). Surprisingly, this did not improve accuracy, and we

decided to not use it.We also attempted a PCA dimension reduction but as shown in the figure it did not impact well on the accuracy.

C. Feature Engineering

The problem of predicting the presence of a Higgs Boson is far too difficult for a linear model based solely on our current features. In order to fit more complex hypotheses, we introduced non linearity by increasing the feature space .

D. Cross Validation

We performed a 5-fold cross validation to find the best lambda,learning rate gamma and degree for our polynomial expansion .When performing ridge regression, we concluded that the optimal degree was 7 or 4 depending on the subset and obtained the best lambda. It's worth noting that the accuracy cost function is utilized instead of MSE.

III. Summary and Discussion

Given that we were working with a binary classification task, the algorithms' prediction findings were quite surprising. We expected logistic regression (or a regularized version of it) to outperform all other methods, but least squares and ridge regression produced the best results.