Using Machine Learning to Discover the Higgs Boson

Leonhard X. Driever[a], Fridtjof Storm Flaate[b] and Hedda C. Soland[c]
[a] Computational Science and Engineering, [b] Computer Science, [c] Material Science and Engineering
[a,b,c] École Polytechnique Fédérale de Lausanne, Switzerland

Abstract

The aim of this project was to distinguish whether a data point represents Higgs boson signal (s) or background noise (b), a famous and challenging classification problem. After thorough data examination and processing, six well-known machine learning algorithms were investigated, using cross-validation to optimize the relevant hyperparameters. With an accuracy of 74.1%, ridge regression was found to be the best algorithm for this problem. Applying the method to subsets of the data based on the jet number allowed increasing the overall prediction accuracy to 74.8%.

1. Introduction

When the Higgs boson was finally detected by the Large Hadron Collider (LHC) in 2012, machine learning served as an important tool for the engineers at CERN. Distinguishing Higgs boson signal from background noise is a classification problem that can be solved with the use of machine learning algorithms. Using a train and test data set from the ATLAS experiment, this report investigates such algorithms and optimizes the best one.

2. Models and Methods

This section outlines how the available data was handled, which machine learning models where examined, and how the best model was determined and further improved.

2.1. Exploratory Data Analysis and Feature Processing

The train and test set consist of 32 columns that were split into identifiers, features tX, and predictions y, where y was either b or s . To represent this binary classification problem numerically, b entries where assigned a value of 0, s entries a value of 1. Out of the 30 features in tX, ten were undefined for jet numbers (PRI_jet_num) of zero and one [1]. These 10 features, as well as the feature indicating the jet number, were removed from the data to avoid training the models with faulty values. As explained in 2.4 for one model the data was divided into four sub-groups based on the jet number.

In addition to the undefined features correlated to jet number, the feature DER_mass_MMC was undefined for 15% of the data points. A simple approach would have been to replace these missing values with a constant, such as the mean of the feature. However, it was decided that this method was insufficiently accurate. Instead, the correlation matrix for all features was examined and it was found that there is a very strong correlation of 0.91 between DER_mass_MMC and the feature DER_mass_vis. Thus, it was decided that linear regression is suitable for predicting the missing DER_mass_MMC values using the line fitted to the DER_mass_vis data. The strong correlation and the fitted line are visualized in Figure 1.

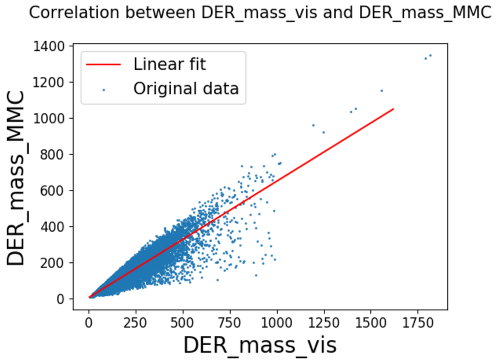

Figure 1: Correlation between features DER_mass_vis and DER_mass_MMC for train data.

2.2. Machine Learning Models

In total six different models were investigated. The first four belong to the class of linear regression, with the variants being to either solve using the normal equations, using gradient descent, or using stochastic gradient descent. The fourth method of this class is ridge regression solved using the normal equations. The final two methods are both logistic regression methods solved using stochastic gradient descent, where the second of the two penalizes large weights through regularization.

2.3. Selection of Best Model

To select the best one, all models were evaluated using their performance on the test set (which was measured using the EPFL AIcrowd web interface). This performance indicates which fraction of data points the models are able to classify correctly. The first step was to, where applicable, optimize the model hyperparameters. This was done using three-fold cross-validation on the training set, picking those hyperparameter values that minimize the cross-validation testing error. The models where then trained on the training set using the optimal hyperparameters, and the model

[1] A complete explanation of each parameter can be seen in the CERN Open Data Portal [1].

achieving the highest fraction of correct predictions was deemed as the best one.

2.4. *Further Improvement of Model Performance*

With the best model identified, further strategies to improve model performance where investigated. Firstly, feature processing was used to investigate if polynomial expansion allows for better prediction accuracy. Cross-validation was used to assess different order polynomial expansions and to simultaneously re-optimize the model hyperparameter for each version of the processed features. Secondly, during exploratory data analysis, patterns dependent on jet number were discovered. Thus, the data was split into four subsets based on the jet number. The model was trained for each set separately and predictions where made by applying the different trained versions to the corresponding subsets of the testing data.

3. Results

Table 1 summarizes the optimized hyper parameter values and the prediction accuracies achieved for the different models.

Table 1: Hyper parameters and accuracy for models.

Model	**Hyperparam.**	**Accuracy**
Least Squares	NA	73.4%
Least Squares GD	$\gamma = 1.34 \cdot 10^{-5}$	51.7%
Least Squares SGD	$\gamma = 6.31 \cdot 10^{-11}$	63.9%
Ridge Regression	$\lambda = 4.64 \cdot 10^{-5}$	74.1%
Logistic Regression	$\gamma = 5.62 \cdot 10^{-6}$	69.1%
Regularized Logistic Regression	$\gamma = 5.30 \cdot 10^{-8}$ $\lambda = 50$	63.7%

As clearly follows from Table 1, ridge regression proved to be the best model for this problem. Thus, the methods discussed in subsection 2.4 were applied to this model. Using cross-validation it was found that, out of the different examined polynomial expansions (not including the unmodified features), the best option was to simply extend the existing features by a feature of value one, allowing for a model offset ω_0. However, this extended model was found to lower the prediction accuracy to 73.4%. In contrast, applying ridge regression with the original features and the same λ to the subsets based on the jet numbers increased the prediction accuracy to 74.8%.

4. Discussion

This section examines the results presented in section 3 and comments on how the different models performed relative to one another. The success of the performance improving methods for the best model is also assessed.

4.1. *Performance of Different Models*

One surprising outcome of the investigation is the superior performance of the least squares models over the logistic regression models. The opposite was expected as logistic regression models are more directly designed for binary classification problems. However, the results are supported by the theorem that a good regressor also implies a good classifier. Furthermore, the performance of the logistic regression models may improve if more computational power is available, allowing for a greater number of iterations during training.

Considering the different least squares models, the model solved using the normal equations is expectedly superior to those trained with GD and SGD, as it represents the best result that can be achieved with pure least squares. The value of regularization is demonstrated by the ridge regression being able to further enhance the performance of the pure least squares technique.

4.2. *Performance-Improving Methods*

Surprisingly, adding an offset term to the features used for ridge regression worsened the model performance. Examining the model weights found through training, it becomes clear that the performance loss is due to the effect that the inclusion of the offset weight has on the other weights. More successful was the strategy to train and apply the model separately for the data points corresponding to the different jet numbers. Further increasing the model performance by 0.7%, the method highlights the value of examining and understanding the nature of the data at hand.

5. Summary

In this report, different machine learning models have been investigated for predicting Higgs boson detection from a data set. Although this model is not specifically designed for classification problems, ridge regression proved to provide the most accurate predictions. By examining the meaning of the data at hand and training and applying the model separately for different jet numbers, it was possible to further increase the prediction accuracy to 74.8%. Based on the success of this improvement method, it is recommended that scientists with a deeper understanding of the data at hand are involved in developing similar methods that make use of the physical significance of the data. This will allow finding new ways in which the data can be used to more efficiently train the examined machine learning models.

References

[1] collaboration, ATLAS *et al.*: *Dataset from the atlas higgs boson machine learning challenge 2014*. CERN Open Data Portal, 2014.