Machine Learning - Project 1

Yann Mentha, Maxime Epars, Gianni Giusto
Machine Learning course (CS-433), EPF Lausanne, Switzerland
October 28, 2019

***Abstract*—As part of the CS-433 Machine Learning course at EPFL, this project aimed to apply the methods and concepts seen in class to a real-world dataset. As such, the computational process that was implemented to prove the existence of the Higgs particle at CERN was recreated and a robust model was built, which provided a classification accuracy of 82.9%.**

I. Introduction

The Higgs boson is a non-stable elementary particle decays rapidly in other particles when produced. To observe and report its existence, scientists at CERN measured multiple physical variables that characterize the decay signature of the Higgs boson. The goal here was to learn and test a model from this generated data to segregate signals that are relative to background noise or indeed produced by the Higgs particle. To develop a robust binary classifier, the prediction accuracy and robustness of various analytical methods were assessed following a standard machine learning workflow.

II. Models and Methods

The dataset is divided into a training set and a testing set composed of 250'000 and 568'238 entries, respectively. The training set is paired with labels where each sample was associated to a category (-1 for background noise and 1 for the presence of a Higgs Boson). The model construction was based on the training set, while the testing set was only used as a final estimator of the model performance.

A. Exploratory data analysis

The dataset is composed of 30 features, from which all but one are continuous (floating point). The remaining regressor, called `PRI_jet_num` is categorical, taking integer values of 0, 1, 2, or 3. As reported by the CERN[1], this variable represents the number of jets (*i.e.* shower of hadrons). By inspecting the dataset documentation, it was observed that many features are linked somehow to this categorical attribute. Finally, many features contain values of -999, which correspond to meaningless variables or values that cannot be computed (as explained in the dataset documentation).

B. Feature processing and engineering

Before feeding the model with the input data, a pipeline was designed to clean the dataset and extract more meaningful information.

1) Categorical data: From the physical interpretation of `PRI_jet_num`, it was hypothesized that this feature was likely to have an impact on the distributions of the other variables and therefore the data should be split into subsets based on their value of this categorical feature. Consequently, models were built independently on each subset, to improve the global accuracy of the classifier. Note that classes 2 and 3 (`PRI_jet_num` values of 2 and 3, respectively) were deliberately grouped in one subset (Subset 2) mainly to have balanced classes. Indeed, these 2 classes had lower number of samples compared to the other ones and it could have induced a model weakening due to lack of data.

2) Feature selection: Features with constant values (*i.e.* with null-variance), were discarded as they could not be used for discrimination. Moreover, principale component analysis (PCA) was also implemented in order to reduce the complexity of the problem by keeping only the most informative dimensions and hence reduce overfitting. However, the dimensionality reduction did not meet the expectations as it did not provide any more performance gain, both in terms of F1 score and accuracy. Therefore, PCA was not retained in the final model for simplicity purpose.

3) Missing values and standardization: To tackle the -999 values issue, the mean and standard deviation were calculated for each feature without accounting for these aberrant values and each feature was standardized individually to produce a zero-mean and approximately unitary variance dataset. This allowed to subtract any influence of the range differences between variables on the model construction and thus to prevent a bias. Finally, -999 values were set to 0, neutralizing their influence on the model training.

4) Feature expansion: To account for non-linear relations between the variables and the labels, feature expansion was performed with a polynomial basis up to an optimized degree.

C. Models

Different models were implemented to perform the classification task, namely linear regression, ridge regression, logistic regression and regularized logistic regression. For the methods without closed-form solution, the optimization was achieved using the canonical stochastic gradient descent (SGD), which is an unbiased estimator and is less time-demanding than full gradient descent. Grid search was also

[1]ATLAS collaboration (2014). Dataset from the ATLAS Higgs Boson Machine Learning Challenge 2014. CERN Open Data Portal. DOI:10.7483/OPENDATA.ATLAS.ZBP2.M5T8

used to look for best hyperparameters such as λ for ridge and regularized logistic regression and for the degree of the polynomial basis expansion.

D. Performance estimation

Hyperparameters estimation and the model performance were assessed using a k-fold cross-validation for each subset. Both the accuracy and the F1-score as a metric were used. The cutoff value allowing classification was systematically optimized in the k-fold and in the submission data as well, according to the used metric (F1 score or accuracy) following a grid search approach.

III. Results

Following the data cleaning step, Figure 1 displays features for which the categorical value (`PRI_jet_num`) has the biggest impact on their respective distribution hence emphasizing our hypothesis to build different models for each class.

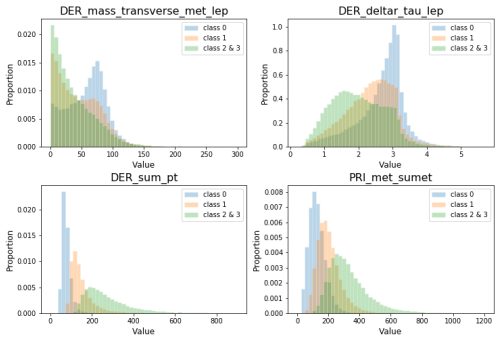

Figure 1. Distribution of 4 specific features for the different classes of the categorical feature `PRI_jet_num`. Classes 0, 1 and 2 & 3 are represented in blue, orange and green respectively. The variations in the distribution of the different classes emphasize our hypothesis that building different models for each class may help discriminating between noise and Higgs Boson.

Following hyperparameters optimization and model testing, ridge regression turned out to be the most efficient model in the classification task, exhibiting the highest accuracy, as it is represented in Table I.

Table I
MODELS PERFORMANCE BASED ON CLASSIFICATION ACCURACY.

Model	Subset 0	Subset 1	Subset 2	Test
Reg. logistic reg.	0.81	0.69	0.70	**0.761**
Ridge reg.	0.84	0.80	0.83	**0.829**

On Figure 2 is displayed the grid search that was performed with ridge regression to optimize the polynomial degree and λ. Subsets 1 and 2 exhibit quite similar maps, while subset 0 differs from the rest. The optimal parameters for each subset are summed up in in Table II.

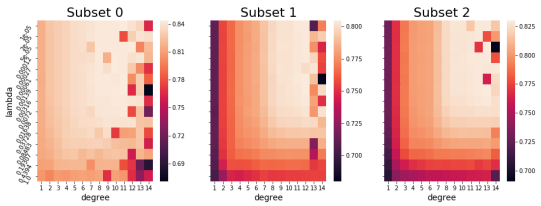

Figure 2. Accuracies for different combinations of polynomial degree expansion and λ parameters for the ridge regression.

Table II
BEST HYPERPARAMTERS FROM A 5-FOLD CROSS-VALIDATION WITH RIDGE REGRESSION.

Parameters	Subset 0	Subset 1	Subset 2
λ	5.18e-5	0.14e-2	0.14e-2
Degree	12	12	13

IV. Discussion

It is important to note that ridge regression was selected over other models solely based on its accuracy to segregate background noise from Higgs boson associated signal. Indeed, its high representational power and its closed-form solution allowed to obtain satisfying accuracy and F1-score in a time-effective way. However, as it is not originally developped for classification, the loss function of ridge regression was not inline with the classification performance. With larger time resources, a regularized logistic regression would probably have been more suited for this task, as the logistic function loss output is bounded between 0 and 1, and can thus be considered as a probability estimate. Nevertheless, this method is based on gradient descent and the convergence is very time-consuming and prone to multiple-parameter optimization.

In further research, the model could be enriched with a more exhaustive augmentation of the features. For instance, logarithmic and sinusoidal expansion of the features could be implemented and evaluated, potentially unveiling new relations between the features and the labels.

Secondly, feature selection could also be performed based on pair-wise correlations of features. As it is reported by CERN that variables containing the prefix `DER_` are derived from the ones starting with `PRI_`, there is a high probability that some features are highly correlated to others. Nevertheless, it is believed that this would not generate significantly different results as PCA did not improve the performance of the model after discarding the constant features, suggesting that the remaining features are quite independent in terms of information they provide.