Higgs challenge

Matthias Kellner, Tim Kircher, Nearchos Potamitis
Machine Learning CS-433 : Project 1, EPFL

***Abstract*—This report describes our effort to tackle Cern's Higgs Challenge. The goal of the Higgs challenge is the identification of particle collisions for which the Higgs boson is present. Experimental data from the ATLAS Experiment in 2012 and machine learning algorithms are utilised to predict the presence of the Higgs boson.**

I. Introduction

After proton-proton collisions (*event*) in the Large Hadron Collider at CERN, a portion of the kinetic energy of the protons is converted into new particles. Most of these new particles are highly unstable and quickly decay into specific lighter particles (*channel*). The properties of these surviving particles are measured by the ATLAS detector. All events are filtered and a subset of relevant events is stored. In this challenge, we explore a dataset containing 4 different channels: Higgs boson (*signal*), Z boson, W boson and a pair of top quarks (*background*). The goal of the challenge is to identify if the Higgs boson is present, based on the experimental conditions. This reduces the classification problem into a binary classification; either the Higgs boson is present or the signal is to be considered noise.

II. Dataset Exploration

The provided dataset contains 30 features describing 250,000 events (training set).[1] The models are evaluated on a test set with 568,238 events. The ground truth labels of the test set are not provided and the goal is of the challenge is to accurately predict the ground truth labels of the test set. The signal to background ratio in the train set is 34.3 to 64.7. The type of all features is float except PRI_jet_num which is an integer and is equal to the number of jets detected (capped at 3). The features of the dataset are split in 2 categories. Firstly, there are the primitive features which are quantities that are measured directly from the detector. These quantities correspond to attributes of the final particles such as their transverse momentum and pseudorapidity. Then, there are the derived features which are calculated using the primitive ones and not measured directly, such as the estimated mass of the particle. The features of some of the data points are undefined in certain events. This is either due to experimental or logical constraints. Figure 1 shows the fraction of undefined values for each feature. If a data point is undefined, it is denoted with the value -999. The first feature, DER_mass_MMC, may be undefined due to experimental constraints. This is because the topology of certain events is too far from the expected topology, which prohibits the calculation of DER_mass_MMC for them. The logically constrained undefined data points are associated with the number of jet particles. Features, such as $DER\_mass\_jet\_jet$ (invariant mass for the 2 jets), are only defined if certain number jet particles are present. If the number of jet particles is below the number of required jet particles to define a data point of such a feature, the data point is denoted as undefined. Filling in values for the undefined data points is key for the performance of the model.

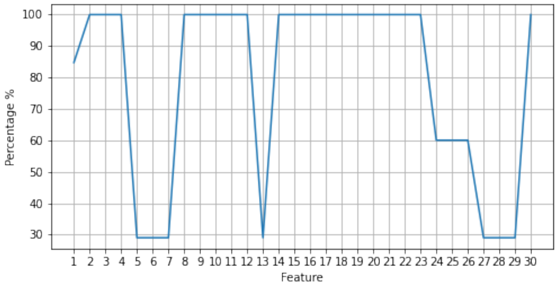

Figure 1. Missing values per feature

III. Methods

In total 6 machine learning algorithms are implemented. The provided training set, containing 250,000 events, is split into a training set and a validation set with a 80-20 ratio. Models are trained on the train set and validated on the validation set. Results presented in the results section correspond to the model performance on the validation set. The best models are then applied to the test set. Additionally, when considering ridge regression algorithm, 5-fold cross validation is performed to identify optimal regularization parameters. The regularization parameter is varied between [10^{-15},10^{-5}]. The model performance is defined by the categorical accuracy according to equation 1, where $N_{correct}$ denotes the number of correctly predicted events and N is the total number of events.

$$Accuracy = N_{correct}/N \cdot 100\% \tag{1}$$

A. Machine Learning Algorithms

The implemented and tested algorithms are gradient descent, stochastic gradient descent, closed-form least-squares, closed-form ridge-regression, logistic regression by gradient descent and regularized logistic regression by gradient descent. A maximum of 1000 iterations is utilised for gradient descent methods. The learning rate is set individually for gradient descent methods. For ridge regression, the optimal regularization parameter is identified by cross-validation. For regularised logistic regression a fixed regularization parameter is used.

B. Feature Engineering

During feature engineering several techniques were evaluated, choosing the approach that yields the highest categorical accuracy. First, filling in missing values is explored. Missing values are either filled with 0 or 1. Furthermore, missing values of the first feature are replaced with the median of the first feature. In this case, an additional column is added, where values are 0 if corresponding values of the first feature are undefined and 1 if they are defined. Also, logarithmic scaling of features is utilised for features with multi-scale behaviour (highest feature value $\geq$ 100). Min-max scaling is used to scale each feature to a range [0,1]. The features of the test set are scaled according to the transformation applied to the train set. Feature set expansion techniques are applied to enhance the model feature space. Polynomial feature expansion and feature interaction terms are investigated. Also, clipping outliers is analysed. In the latter, Data points that are below the bound of the 1.5 % or 92.5 % percentile are set equal to the corresponding bound.

IV. Results

First, baselines are created for the models, where missing values are replaced by 0 for all features except the first feature, where missing values are replaced by the median and then features are min-max scaled. Then, ridge regression models are further explored. The accuracy of the implemented model types on the test set and the utilised parameters are displayed in table I

Method	γ	λ	Val. Err.(%)
Gradient Descent	10^{-5}	-	70.55
Stochastic GD	10^{-5}	-	65.67
Least Squares	-	-	74.79
Ridge regression	-	$2.7 \cdot 10^{-14}$	74.78
Logistic regression	10^{-3}	-	65.67
Reg. Logistic regression	10^{-5}	10^{-5}	65.67

Table I
RESULTS FOR MACHINE LEARNING IMPLEMENTATIONS ON VALIDATION SET WITH MISSING VALUES REPLACED BY MEDIAN FOR FIRST FEATURE AND 0 FOR ALL OTHER FEATURES AND MIN-MAX SCALED FEATURES.

Ridge regression and least-squares show the best performance. When considering larger, expanded feature sets ridge regression is beneficial to prevent over-fitting [2]. Therefore ridge regression is further explored with expanded feature sets. Ridge regression models are trained on 3 different expanded feature sets containing polynomial features (Poly) of degree 2, binary interaction terms (Interaction) and polynomial features of degree 2 and binary interaction terms. In the third case, polynomial features are generated first and then binary interaction terms of all features, including the polynomial features, are computed. Then, treatment of outliers and logarithmic scaling of multi-scale features is considered (Add. Transf.). Table II shows the validation set accuracy for the described models. The checked columns correspond to the utilised tools for generating the feature expanded feature set. The standardization to a range of [0,1] is always performed as the last operation.

Poly	Interaction	Add. Transf.	λ	Val. Err.(%)
✓			$3.72 \cdot 10^{-7}$	77.46
	✓		$3.72 \cdot 10^{-7}$	79.57
✓	✓		$3.72 \cdot 10^{-7}$	81.52
✓	✓	✓	$5.17 \cdot 10^{-10}$	83.56

Table II
RESULTS FOR RIDGE REGRESSION WITH VARYING ALTERATIONS TO THE FEATURE SET AS INDICATED BY CHECK MARKS.

Table II shows the best result for ridge regression with an expanded feature set containing both polynomial features and feature interaction terms with logarithmic scaling of features with multi-scale behaviour and treatment of outliers. An accuracy of 83.56 % was achieved on the validation set. This model was then utilised to predict the labels on the test set and an accuracy of 83.69 % was achieved. Utilising both polynomials of degree 2 and interaction terms increases the number of features from 30 to 1895. Further increasing the number of features was limited by working memory.

V. Summary

In this project, the decay channel of the Higgs boson was classified based on a simulated dataset of collisions using machine learning methods. We verified our theoretical expectations with our results and achieved a categorical accuracy of 83.69 % on AIcrowd by implementing our final model. We realized that there are different ways to approach problems such as this but always exploring your data in advance is essential for good results. Moreover, we understood that good theoretical comprehension of the problem can be crucial when it comes to feature engineering. A detailed hyperparameter investigation of the gradient descent methods is required to achieve satisfactory results, and this could be considered for further studies. With regard to ridge regression, further improvement of the model accuracy can be made by looking into the utilised hyperparameters for the generation of the feature set utilised for the final model. The choice of cutoff values for removing outliers (Percentile 1.5 and 92.5) and logarithmic scaling (highest data point of feature $\geq$ 100) have not been optimised. Also, even larger feature sets with polynomials of order 3 could be considered.

References

[1] C. Adam-Bourdarios, G. Cowan, C. Germain, I. Guyon, B. Kégl, and D. Rousseau, "The Higgs boson machine learning challenge," in *Proceedings of the NIPS 2014 Workshop on High-energy Physics and Machine Learning*, ser. Proceedings of Machine Learning Research, G. Cowan, C. Germain, I. Guyon, B. Kégl, and D. Rousseau, Eds., vol. 42. Montreal, Canada: PMLR, 13 Dec 2015, pp. 19–55. [Online]. Available: https://proceedings.mlr.press/v42/cowa14.html

[2] D. W. Marquardt and R. D. Snee, "Ridge regression in practice," *The American Statistician*, vol. 29, no. 1, pp. 3–20, 1975. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/00031305.1975.10479105