Machine Learning: Project I

João Luis Soares Lopes, Rémy Pétremand, Halima Hannah Schede

***Abstract*—The objective of the project is to identify and describe a model that is capable of predicting whether or not a given decay signature originated from a Higgs boson. The data used for the project is original and provided by CERN. Different regression models were tested and hyperparameter optimization was performed. The optimal learning algorithm was found to be regularized logistic regression and the model results in an accuracy of 83.909% on Kaggle.**

I. Introduction

In physics, the Higgs Boson is a product of quantum excitation of the Higgs field. The objective of this project is to discriminate between different decay signatures from particles which may or may not have been Higgs bosons.

The data set is original data provided by CERN and was generated by the ATLAS experiment. It is comprised of thirty features, describing experimental variables such as momenta of the hadronic tau and invariant masses of other quantum particles. There are 250000 samples with corresponding labels which indicate whether a given event was a ”signal” (i.e. the given event was produced by a Higgs boson) or ”background”.

To obtain the optimal vector of weights for the classification assignment, it is necessary to test different models (such as linear and logistic regression) under different hyperparameters. Hyperparameter optimization includes choosing values for the following variables: the learning rate, γ, the degree of polynomial expansion and the λ parameter in regularized models.

After selecting a model along with its hyperparameters based on their accuracy and variance scores, the model was applied to the entire data set. The retrieved weight vector w was then used to test samples in order to predict which events originated from Higgs bosons.

II. Methods

The steps involved in selecting an optimized model include data pre-processing, determining the best learning algorithm, selecting the model hyperparameters and fitting the model to the dataset to retrieve the optimal weight vector, w.

A. *Data pre-processing*

The data set consists of a number of features with high rates of missing values set to -999. Some of these features are undefined if a sample has less than two jets, which are elementary particles that emerge from high-energy particle collisions. To classify the different types of samples more accurately, samples were placed into one of three groups: those which lack jets, those with only one jet and finally those with more than one jet. For each group, undefined features were removed. This implies that three models were fitted using an identical model-building procedure. In order to classify incoming test samples, the group was first inferred with respect to the jets, and then the appropriate model was applied for prediction.

Different filtering methods were applied to remove missing values after jet separation. The methods either comprised outlier detection and replacement of missing values by zero or the mean or median of the features. The outliers were defined to be samples that are outside the confidence interval: $[\mu{+}3\sigma; \mu{-}3\sigma]$. The methods were tested to understand their effects on the learning algorithms. After applying the filtering methods, the data set was standardized and principal component analysis (PCA) was implemented.

B. *Splitting data and cross-validation procedure*

One must evaluate the variance and bias of the potential models and check the error on an unseen test set.

In order to have an unbiased error, the data was split into a training and test set. The test set had 20% of the samples of the data set and was standardized using means and variances acquired from the training set. The test set was transformed into a new space using the eigenvectors calculated from the training set.

To retrieve the variance and accuracy of the various models, cross-validation was performed on the training set by splitting it into six subsets. It should be noted that for every training subset within cross-validation, the test subset is standardized and transformed into a new space for PCA as described above.

After determining these statistical moments, a model is fit on the whole training set to retrieve w, as well as the standardization parameters (means and variance of training columns), which are used to predict unclassified data.

C. *Learning algorithms*

Linear regression and logistic regression, both of which were also tested with regularization hyperparameters, were implemented to determine which type of learning algorithm was most capable of predicting Higgs boson-related events.

Linear regression was tested by gradient descent and direct least squares calculations. Logistic regression was calculated via Newton’s method which uses both the gradient and Hessian matrix. A grid search of resolution 5×5 with ranges [0.05, 0.08] for γ and $[1 \times 10^{-2}, 1 \times 10^{2}]$ for λ optimized these values for logistic regression.

D. *Feature set expansion and dataset transformations*

Three methods were implemented to better understand the effects of feature expansion techniques - polynomial basis expansion, interacting terms and logarithmic transformations. The first method refers to bringing the original features to an exponent, the second multiplies features among one another and the last consists of computing the logarithm of features. This methodology drastically increased the number of features which motivated the use of dimensionality reduction via PCA.

Lastly, a column of ones is added to the data set. This is required as the first value of the w vector is a bias term.

III. Results

THE figures show prediction accuracy of the assessed models. The prediction accuracy of w retrieved from the training set to use on the test set is consistently higher than cross-validation averages as it used significantly more data than the cross-validation folds.

A. *Pre-processing and partitioning samples based on jets*

Depending on how missing values were handled, the model's prediction accuracy varied. Replacement by mean values of the columns may make it difficult for learning algorithms to capture important differences between samples. The fact that removing outliers is more effective than removing features and samples that have higher rates of unknown values may be because there is important information in such samples.

Fig. 1 demonstrates that partitioning the samples based on jets before learning the model improves prediction accuracy to as high as 84.5% depending on the jet number. *Fig. 2* (logistic regression) shows a prediction accuracy of 76.8% without partitioning using the same model-building methodology.

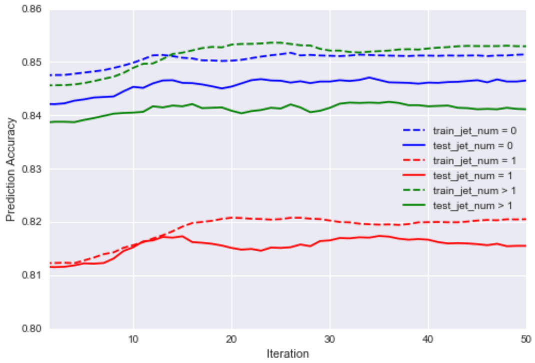

Fig. 1. **Effects of partitioning samples into groups based on jet number** Blue: samples with no jets. Red: samples with one jet. Green: samples with two or more jets. Solid lines are test set accuracy predictions while dottes lines represent training set accuracy predictions.

B. *Performance of learning algorithms*

Applying the linear regression model on the test set provided an accuracy of 76.3% by direct least squares. Since least squares linear regression can be solved directly for an optimal w, the solution has a higher prediction accuracy than gradient descent algorithms. Ridge regression is slightly beneficial. Logistic regression performs best with an accuracy of 76.8% likely because it is appropriate for binary inputs. The variance of the test set for each algorithm was on the order of 1×10^{-5}, implying w was not overfit. Hyperparameters are summarized in *Tab. 1* for the three sample groups. Notably the learning rate was very sensitive, if it was slightly too large w could diverge.

Sample Group	γ	λ
No Jets	0.08	0.01
1 Jet	0.0575	10
2+ Jets	0.065	10

Table 1: Optimized hyperparameters for each sample group according to jets. γ: learning rate. λ: regularizing parameter.

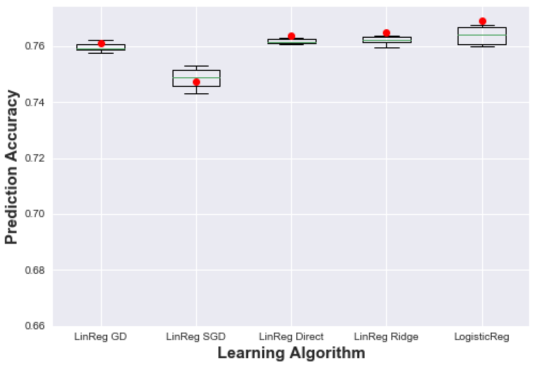

Fig. 2. **Comparison of Learning Algorithms** Boxes correspond to values retrieved from cross-validation procedure within training set. LinReg: Linear Regression; GD: Gradient Descent; SGD: Stochastic Gradient Descent. Red dots refer to prediction accuracy of w retrieved from training set on test set

C. *Feature set expansion and transformations*

Feature set expansion was a crucial step, and is necessary to reach a prediction accuracy greater than 76.8% in the test set. Applying basis expansion to a linear or logistic regression model increased accuracy of the test set gradually until the eighth degree where the model began to fail. The most beneficial effect is for the fourth degree, which is selected for our final model.

Logarithmic transformation of the data in regression problems is a common technique in the field. It may have increased prediction accuracy of the model because a linear model was used on features that are potentially exponentially distributed. Logging them will normally distribute them, and linear models work at best when data is normally distributed.

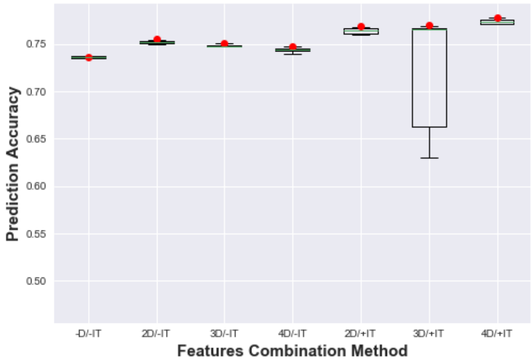

Fig. 3. **Effects of Feature Set Expansions and Polynomial Models** D: degree calculated from original features. IT: interacting terms. Boxes correspond to values retrieved from cross-validation within training set. Red dots refer to prediction accuracy of w retrieved from training set on test set

IV. Conclusion

TO conclude, we observed that logistic regression was the most appropriate learning algorithm for the given data. Throughout the project, we saw that feature expansion has the most powerful effect on prediction, followed by pre-processing methods. We also report that accuracy increases when models are tailored to samples based on jet group. After following our methodology, we produced an accuracy on Kaggle of 83.909%.