The Higgs boson machine learning challenge

Gianluca Mancini, Tullio Nutta
Department of Computer Science, EPFL Lausanne, Switzerland

***Abstract*—The vector of features of the decay signature of a boson is produced by colliding high speed protons at CERN; by anlyzing these features, it is possible to create models which predict the presence of a boson. This paper addresses the challenges of feature engineering aimed at discerning between the signal of a boson and background noise. Two feature engineering models were compared: feature engineering based on physical knowledge of the collisions, was proven to give better prediction compared to a Principal Component Analysis. Among a selection of basic regression methods, the Ridge Regression method resulted to provide the best results. The limitations of the feature engineering and of the regression methods used in this paper, however, suggest that a more complex and comprehensive analysis must be carried out to improve the models.**

I. Introduction

At CERN the particle accelerator smash protons bunch into one another at high speed: when they cross the ATLAS detector, sensors produce a vector of dimensions which is used to distinguish meaningful events from background noise. The bosons which originates from these collisions decay rapidly: hence the aim of the research is to find regions of the feature space in which there is a significant amounts of events which show the presence of a boson and create models to predict its presence for future analysis. Currently it is extremely important to improve the feature space selection to come up with weights functions which lead to better prediction of a boson [1].

The aim of this paper is selecting the meaningful feature space to predict the presence of a boson. Firstly the feature space was selected only considering the physical knowledge behind the experiment and with the help of linear and non-linear correlation coefficients. Secondly the feature space was selected carrying out a Principal Component Analysis. Both of the models aim at separating the important features to classify a boson from the complete set of attributes [2]. Eventually the results were compared with different regression methods in terms of stability, accuracy and MSE.

The methods section II describes the reasoning behind the feature selections based on the physical knowledge and the Principal Component Analysis stating which regression methods are used in the analysis. The choice of hyper-parameters and the optimization methods are presented in section III. Hence the results sections IV compares the outcome of the two models. A comparison of the results and a summary of the possible improvements and limitations regarding the regression methods used and the feature selection can be found in the discussion and conclusion section IV-C. It is extremely important to state that the methods and the analysis described in this paper have to be understood and they will be limited to the context of the challenge ”Learning to discover: the Higgs boson machine learning challenge” [1].

II. Feature Engineering Methods

The first step before applying any feature selection model is to inspect the data. Thirty attributes were provided divided in Primitive, raw quantities, and Derivative, derived from the raw quantities. Some data points could of certain features could at times have invalid values if their numerical value was -999.0. This could happen for a variety of reasons according to the specific feature; the difference in dealing with these values is at the base of the difference between the two methods provided.

A. Method 1: Feature selection based on Physics

Selecting the most important attributes from the ones given, is dependent on the physical meaning of each attribute. Firstly the physical background revealed that the most pivotal attribute of all. In fact from a physical point of view the value of a particular constant could directly deterministically and physically influence the value of another in a consistent manner. Hence the attribute space can be reduced by only considering the most pivotal quantity and eliminating the rest which would only constitute noise. Then the remaining feature space was further restricted by analyzing the Pearson Moment Correlation coefficient and the Spearman's correlation coefficient to identify respectively linear and non-linear relationships. Then a minimum threshold for meaningful correlation was set to be 0.7; eventually the attribute which shows the highest number of meaningful correlations was kept as representative of the other parameters which were discarded from the feature space. Finally the feature space was generated by polynomial expansion and degree optimization according to the regression method used as described in section III.

B. Method 2: Feature selection Principal Component Analysis

Principal Component Analysis is a powerful data representation method which captures the most variable data components of samples [3]. It allows to transform an attribute space in a feature space which encapsulates the greatest variation of the data. Consider the attribute space matrix to be X, $n \times m$ matrix. The covariance matrix Σ of the standardized matrix x can be found according to equation 1 [4].

$$\Sigma_{i,j} = corr(x_i, x_j) \tag{1}$$

Where $corr$ is the correlation coefficient between feature i and j. The eigenvalues in the diagonal matrix D and the eigenvectors matrix V are essential map the parameters space into a new feature space $V^{-1} \cdot \Sigma \cdot V = D$. On top of that they are useful to calculate the explained variance π as shown in equation 2 which highlights how much information can be attributed to each principal component [5].

$$\pi_j = \frac{\lambda_j}{\Sigma_{i=j}^{m} \lambda_i} \tag{2}$$

The eigenvectors with the highest value of π will be selected to compose the transformation matrix W. Using the cumulative variance information it was decided to include 95% of the variance and the number of principal components to include was determined based on that. Hence the new feature matrix T can be found as $T = x \cdot W$. Finally the feature space was generated by polynomial expansion and degree optimization according to the regression method used as described in section III. The challenge that it is currently being faced is a classification problem; as a consequence Regularized Logistic Regression and Ridge Regression have been selected as the most suited methods.

III. Feature Augmentation, Parameter Optimization, Cross Validation

Because of memory reasons a simplified version of a polynomial expansion is considered which doesn't include the cross terms of the features. A feature space matrix $n \times m$ is expanded to polynomial of degree p, becoming a $n \times (m \cdot p + 1)$ matrix. Other function such as trigonometric were considered for feature augmentation, but their optimization time was suited for the task.

Both the regression methods considered above involve the selection of hyper-parameters: λ, λ and γ must be chosen respectively for Ridge and Regularized Logistic regressions; on top of that the degree of the polynomial has also to be optimized. As a matter of fact an iterative method which finds the best combinations of hyper-parameter and degree has been set up.

In order to determine whether a method was underfitting or overfitting cross validation was implemented in the optimization algorithm with a folding of 4.

IV. Results

Firstly a baseline feature space was made by all attributes given without augmentation with polynomial expansion.

A. Feature Selection

Method 1 led to a selection of 16 features; in fact it was found that the value of PRI-jet-num was pivotal for a number of parameters as described in appendix B of the competition description [1]: when it attained the value of ¡1, a set of attributes were automatically invalid [1]. These attributes were discarded from the analysis for their numerical instability and dependency on PRI-jet-num. On top of that the estimated mass of the boson DER-mass-MMC could many times attain an invalid value if the topology of the events was unexpected [1]. Hence these values were substituted with the median value of column. Eventually it was found that 'DER-sum-pt', 'PRI-jet-num', 'PRI-met-sumet', 'PRI-jet-all' and 'DER-pt-h' had respectively with each other all a Pearson and Spearman correlation coefficient higher than 0.7. As a matter of fact all these features were deleted expect 'PRI-jet-num' which was the stronger correlated with all. Feature Space 1 (FS1) was created using the remaining features. Method 2 led to selecting 20 principal component features which explained 95% as shown in figure 1. Before doing so, however, for numerical reasons the invalid values were substituted with zeros. Hence the Feature Space 2 (FS2) was created using the principal components mapping.

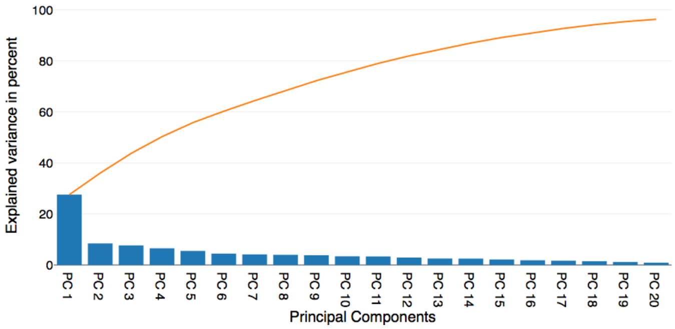

Fig. 1. Principal Components Analysis

B. Performance Matrix

A baseline test was made with the baseline feature space and performing least square regression. The performance of the Ridge regression method with different feature space are presented in table IV-B. All feature spaces were augmented with polynomial expansion. Regularized logistic regression which for a binary classification problem was expected to perform well; however due computational reason it wasn't possible to optimize the degree and the hyperparameter such that the regression would converge.

	MSE_ train	**Score Train**	**Score Test**	**Degree**	**Lambda**
Baseline	0.36	0.703	0.729	1	/
FS 1	0.29	0.805	0.806	11	4.6*10^-8
FS 2	0.31	0.776	0.789	13	10^{-6}

C. Discussion and Conclusions

As shown in table IV-B the feature space which produces the highest score is FS 1. All methods underfit since the MSE of the test was always higher than the MSE of the train set in the cross validation meaning that all estimation models had high variance. Both the feature selection method which performed better and PCA had some limitation: only linear relationships between features are considered and the potential multivariate nature of the data structure is not considered. On top of that the regression algorithm used are limited: neural network or support vector machine could potentially improve the results. Eventually in terms of feature engineering, a multiple regressions done according to categorical features PRI-jet-num, could have highlighted hidden relationships.

References

[1] C. . G. I. G. B. . K. . c. D. R. Claire Adam-Bourdariosa, Glen Cowanb, “Higgs challenge,” jul 2014.

[2] J. Brownlee, “Discover feature engineering, how to engineer features and how to get good at it,” https://machinelearningmastery.com/discover-feature-engineering-how-to-engineer-features-and-how-to-get-good-at-it, sep 2014.

[3] J. Y. Y. Xu, D. Zhang, “A feature extraction method for use with bimodal biometrics,” *Pattern Recognition*, vol. 43, no. Issue 3, pp. 1106–1115, mar 2010.

[4] V. M. Panaretos, *Statistics for Mathematicians, A Rigorous First Course*, ser. Springer: Compact Textbooks in Mathematics. Birkhuser, oct 2015.

[5] I. T. Jolliffe and J. Cadima, “Principal component analysis: a review and recent developments,” *Philos Trans A Math Phys Eng Sci*, apr 2016.