Evaluation of binary classification methods to detect Higgs Bosons

Lawrence Brun, Florian Genilloud, Dario Müller
Department of Computer Science, EPF Lausanne, Switzerland

***Abstract*—In order to detect the presence of Higgs Bosons from feature vectors representing the decay signature of a collision event, different binary classification techniques from machine learning were carried out to determine the best method with the highest accuracy of prediction. The findings showed that the ridge regression method yielded the best model, with correct predictions of 80.7% and an F1-score of 0.697.**

I. Introduction

Machine Learning offers a wide range of methods to develop models for various prediction applications. However, it is often difficult to gauge the accuracy of each approach when compared to one another. To assess the performance of various widely-used methods, they were applied to a classical problem from physics: detecting the past presence of Higgs bosons based on their decay patterns. The methods used include:

- Linear regression using gradient descent
- Linear regression using stochastic gradient descent (SGD)
- Least squares regression using normal equations
- Ridge regression using normal equations
- Logistic regression using gradient descent (GD)
- Regularized logistic regression using GD

II. Models and Methods

A. Preprocessing

Before implementing and training the different models, preprocessing steps had to be carried out. To this extent, 7 columns containing one single value of -999 were removed since no model contribution is to be expected from a constant feature. At certain positions, missing values were imputed by the feature's median value, accounting for eventual outliers. Lastly, these outliers were removed all together by applying a function replacing all values outside of three times the standard deviation range with the respective feature's median. Since the ranges of different feature's values differed significantly, all the values were standardized by centering them and reducing their range. This was achieved by subtracting the feature's median from all of its values and dividing them by their standard deviation, respectively.

B. Feature Expansion

In order to account for interactions between different features, diverse feature expansion methods were used.

- Polynomial expansion using different powers to describe a feature
- Cross-term expansion combining different features to account for feature correlation incidents
- Log-term expansion to take into account logarithmic influences of features on the label values.

To make sure the models also perform well on unknown data, k-fold cross-validation was applied [1]. As a trade-off between meaningfulness and computation time, k was chosen as 5.

III. Results

A comparison juxtaposing the methods' results is presented in the following Figure 1.

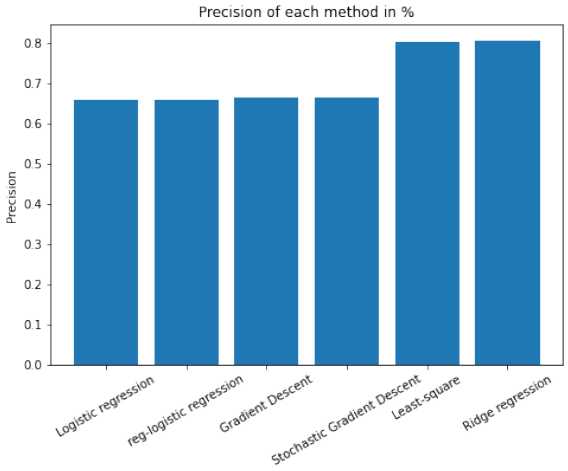

Fig. 1: Comparison of the different methods evaluated with their best feature expansion

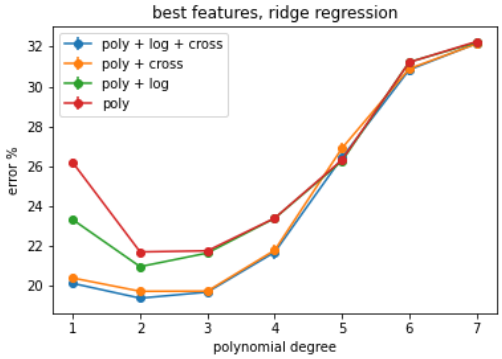

Fig. 2: Comparison of different feature expansion methods with regards to their performance in the Ridge Regression model

With a precision of 0.6578 each, both "normal" and Regularized Logistic Regression are outperformed by all other methods. Gradient Descent (GD) and Stochastic Gradient Descent (SGD) perform equally well with a precision of 0.6644. However, all of the previously methods performed worse than the remaining ones: Least Squares Regression and the Ridge Regression which are only different due to an introduction of a weight penalization parameter in the case of Ridge Regression. The latter allows to increase the precision from 0.803 to 0.807, corresponding to an increase by about 0.4%. The Ridge Regression method can therefore be considered as the most accurate method for this application. With an optimal lambda of around 10^{-8}, it in fact resembles a simple Least Square Regression since the weight penalization is quite low.

Considering the feature expansion performance for the most accurate model - the Ridge Regression - depicted in Figure 2, it becomes evident that the best results are achieved by considering polynomial feature expansion of degree 2 while also deploying cross-term expansion. Adding a log-term expansion slightly reduces the loss and consequently improves the overall model accuracy.

IV. Discussion

Both SGD and GD yielding similar errors implies that both approaches reached a local minimum since the SGD is considered as robust towards settling on local minima. Hence, the only difference between the two is the shorter training time when using the stochastic gradient descent [2]. The Ridge Regression being the optimal method to be used can be explained by considering the small penalization lambda which implies that it actually resembles a simple least square regression since the weight penalization of 10^{-8} is quite low. This is confirmed by considering higher lambdas which yield an exponentially higher loss, caused by subsequent underfitting.

V. Summary

To conclude, it can be stated that for applications using extensive datasets like the one treated in this study, a method based on a Least Square Regression should be implemented. Furthermore, the benefit of additionally using a Ridge Regression is quite limited since over-penalization by introducing large weight parameters (lambdas) inherently leads to a model underfitting the data. This is can be seen when considering Figure 3 in the Annex.

VI. Annex

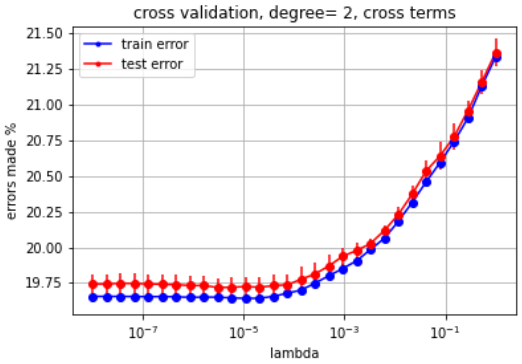

Fig. 3: Losses of a Ridge Regression with polynomial feature expansion of degree 2

References

[1] Prashant Gupta. *Cross-Validation in Machine Learning*. URL: https://towardsdatascience.com/cross-validation-in-machine-learning-72924a69872f (visited on 10/24/2020).

[2] Aerin Kim. *Difference between Batch Gradient Descent and Stochastic Gradient Descent*. URL: https://towardsdatascience.com/difference-between-batch-gradient-descent-and-stochastic-gradient-descent-1187f1291aa1 (visited on 10/24/2020).