Project 1: Higgs Boson Challenge

Kaan Okumuş, Mihaela Diana Zanoaga, Roxane Burri
Department of Computer Science, EPFL, Switzerland

***Abstract*—In this report it is shown the implementation of several Machine Learning methods used to solve a binary classification model. The dataset used was obtained from CERN and the aim of the challenge is to predict if a particle is Higgs boson or not. The focus will be both on handling data, which is significant for generating powerful models, and on the investigation and analysis of different regression models, in order to find the one which gives the more accurate solution.**

I. Introduction

Higgs Boson Challenge goal is to predict if a collision event is a Higgs boson (signal), an elementary particle discovered at CERN, or some other particle (backgound). The Higgs boson decays very rapidly into other particles, so it's impossible to observe it directly, but it's necessary to measure different decay signature of the event.
Data pre-processing are in the next section, those decay-related features and then, with this cleaned data and different Machine Learning models are implemented with this data. The third section is to show how models performed, and explaining the way of selection of the best one. Finally, it will be a report of all results and a discussion about all the elaboration process.

II. Models and Methods

In order to solve this task, it performed two main steps: first explored data analysis and after implement several Machine Learning algorithms to find the best parameters to optimize the prediction.

A. Exploratory Data Analysis

It is firstly focused on data pre-processing by examining the train data-set which was composed of 250000 observations and 30 features. It is looked for missing values, categorical features and correlated features. Furthermore, it is attempted to standardize data and to apply polynomial expansion. [3]

1) Handling missing values: the first step was to check for missing values in the data-set. In order to manage them, it is firstly decided to delete columns with more than 70% of missing values and to impute median at place of -999 in those columns with missing value ratio less than 70%. However, deleting columns may result in the loss of the useful data. As the effects of the results, only the imputation by median is applied. It is preferred median to mean and mod since it is much more robust in regards to outliers and all the features with missing values are numerical data [4].

2) Handling categorical features: The column PRI_jet_num is a categorical feature characterized by the presence of 4 unique values. Therefore, it is better to handle them with the "One hot encoding technique". That means that for each level of the categorical column a new variable is created and mapped with a binary variable 0 and 1. Therefore, in total, 4 new columns of 0 and 1 are added, and the original categorical feature is deleted [1].

3) Standardize data: Data is standardized, since many machine learning algorithms are very sensitive to feature scaling.

4) Polynomial expansion: An off-term (constant column of 1 values) is added. And for each column, the feature expansion is computed to achieve more complex model until a certain degree that it is tuned maximizing the accuracy through grid search and cross-validation technique [5], [2].

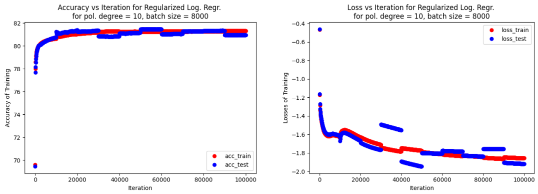
B. Models

After having analysed and cleaned data, several machine learning models are started to be implemented. In order to have generalization error and good estimates we applied cross-validation method with 10 folds. And therefore we tuned parameters through grid search.
Firstly, the least square solution is implemented using normal equation. Its hyper-parameter is the degree of the polynomial extension and grid search algorithm is designed to tune it.
In order to solve any over-fitting problem, ridge regression is then performed from its closed-from solution. Its regularization parameter is also obtained using grid search method.
Both least square and ridge regression cannot demonstrate how the learning proceeded, but they provide converged solution. In order to ensure the learning is done well, gradient descent algorithm is applied. However, due to the computational cost, achieving optimal solution with reasonable computational power is very challenging. Thus, mini-batch stochastic gradient descent is implemented. Learning rate, maximum number of iterations and batch-size are tuned by trial and error method with respect to accuracy/loss graphics and execution time. This is because implementation of grid search is highly costing.
As linear regression is generally not suitable to create decision boundaries on classification problems, logistic regression, much robust method, is computed using stochastic gradient descent. Trial and error method is

implemented to tune hyper-parameters. Moreover, in case of over-fitting issues, the regularization technique is added to the logistic regression to increase the accuracy.

III. Results

For each of these models, the best parameters are chosen for getting the optimal model, searching to avoid over-fit. Calculating the train and test accuracy regularized logistic regression with SGD seems to be the best model for our classification problem. Indeed, even if least-square estimation method with normal equation and best tuned degree parameter 9, gives higher accuracy, it may reasonably over-fitting. For this we computed ridge regression with a regularization parameter of 1e-10. The model although gives less accuracy even if low variance and converging of loss may think it's now not over-fitting. On the other hand, also linear regression using gradient descent, with tuned values for learning rate and degree of 4×10^{-2} and 9 respectively, performed not as well as logistic regression. Moreover, the algorithms were computer both with batch size 1 and batch-size of 8000, which it the one giving the best accuracy values with the same execution time compared to the ones with different batch-sizes. Finally, logistic regression, mini-batch stochastic gradient descent implemented with the tuned polynomial degree 9, batch-size 8000, number of iteration 100000 and learning rate tuned to 1×10^{-1} is still giving lower values as regularized one and it may also overfit. Regularized logistic regression, with larger polynomial degree 13 with batch-size 8000, 100000 iteration, tuned learning rate of 1×10^{-4} and regularization parameter tuned to 1×10^{-4} seems therefore to give, through cross-validation estimation, the best accuracy and with high confidence of not over-fitting. The learning process in regularized logistic regression can be observed from training and testing accuracy and loss values through iterations as seen from Fig. 1



(a) Train and test accuracy vs iteration

(b) Train and test loss vs iteration

Figure 1: Performance metrics through learning in regularized logistic regression

IV. Discussion

It can be analyzed that in the data pre-processing part, number of degree affects the robustness of the model. As the polynomial degree increases, the model becomes more complex, which may result in over-fitting. Ridge regression gives better results with larger degree than Least Square, as it can be seen from Tab.I. This is because it prevents over-fitting for complex models and gives a more powerful solution. On the other hand, by using linear regression with gradient descent, it is observed that the speed of the learning process is very low. While, stochastic gradient descent gives very high variance and reduced accuracy even if it computes much faster. Therefore, mini-batch stochastic gradient descent with 8000 batch-size is preferred since can give result closer to the least square or ridge regression methods even if these ones better as they are the converged solution of linear regression. Applying logistic regression with mini-batch stochastic gradient descent gives better results compared to that of the linear regression, probably because the problem is a binary classification. In order to apply logistic regression with larger number of degree, it is implemented with regularization technique, which results in slightly better model and gives us the best performance as it can be seen from Tab. I.

Table I: Compare model performance

Model	Train Acc.	Test Acc.	Train Loss	Test Loss
Least Square GD	77.31	77.32	0.31	0.32
Least Square miniSGD	78.60	78.70	1.09	0.77
Least Square	81.57	81.57	0.276	2.39e14
Ridge Regr.	81.02	81.02	0.2836	0.2836
Log. Regr.SGD	81.43	81.33	-1.73	-1.71
Reg. Log. Regr.SGD	81.66	81.34	-1.92	-1.9

V. Summary

In this binary classification problem, it is seen that in order to implement machine learning models, it is compulsory to apply data pre-processing technique such as handling missing values, categorical data, standardization. Feature augmentation makes the model more complex, which may improve the performance or cause over-fitting problems. With the regularization technique, the system can perform better with more complex model by avoiding over-fitting. Therefore, rigde regression and regularized logistic regression are the best models. Moreover, for binary classification, logistic regression is more suitable because of its loss function. However, for some classification problems as in this experiment, the decision boundary may be close to be linear. Therefore, the accuracy for ridge and logistic regression is very close. However, for large number of iteration, the one can achieve the best model with regularized logistic regression.

References

[1] Alice, C. *Feature Engineering for Machine Learning & Principles and Techniques for Data Scienties*, 2 ed. O'Reilly, 2018.

[2] Dalton, J., Dietz, L., and Allan, J. Entity query feature expansion using knowledge base links. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval* (2014), pp. 365–374.

[3] García, S., Luengo, J., and Herrera, F. *Data preprocessing in data mining*, vol. 72. Springer, 2015.

[4] Max, K. *Feature Engineering and Selection: A Practical Approach for Predictive Models*. CRC Press, 2020.

[5] Peckov, A. *A machine learning approach to polynomial regression*. PhD thesis, PhD thesis, Jozef Stefan International Postgraduate School, Ljubljana, 2012.