Project 1 - Machine Learning - Spot the Boson

Team KLR - Loïc Busson, Razvan Mocan, Kamran Nejad-Sattary

26 October 2020

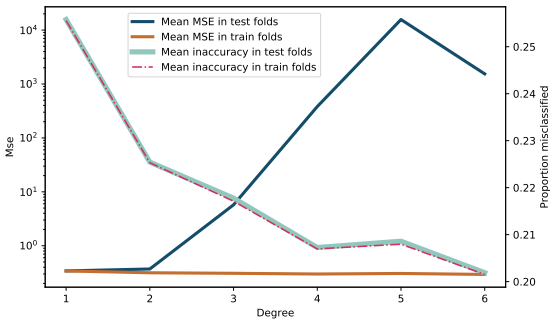
Introduction

Attempting to recreate the discovery process of the Higgs particle, this project provided a training and testing dataset of 250'000 and 568'238 observations respectively, including missing values distinguished by the value -999. The training set's class was labelled either -1 or 1, coining this problem a classification one, a question of whether a statistical method could be devised to spot whether an event's signature resulted from a Higgs boson signal, or merely some background particles.

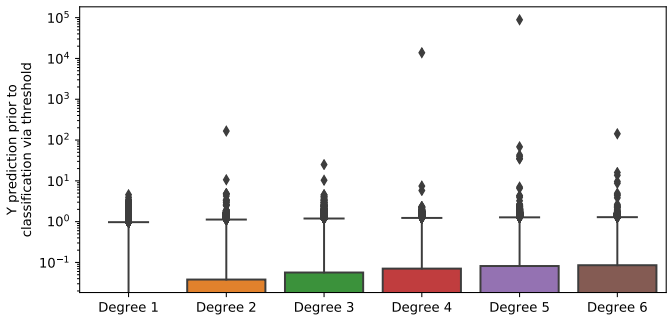
Methodology

Statistical approaches considered and implemented to varying success to accomplish this were the linear regression, with or without regularization, logistic regression variants, and a first attempt at a neural network. However, pressing on the fundamentals taught, emphasis was particularly put on exploiting finer details (cross-validation, comparing loss methods, and attempting to best tune parameters) more so than employing more difficult statistical models - leading to the decision to use Ridge Regression (despite the problem being of classification nature) offering a normal equation solution to reduce computation necessities.

In a first part, as indeed the data included missing numerous values, these were filled using the median (though later compared with other filling methods) of each feature, rather than removing them which would otherwise lead to removing three quarters of the data off the bat. Similarly, in attempt to also rescale the data, minmax normalization was exploited to linearly transform the data into a range of 0 and 1. The minimum, maximum and the median were stored to later transform the testing set as naturally they had to be scaled and adjusted with the same parameters to benefit from the trained model.



(a) Comparing MSE and Misclassification Error

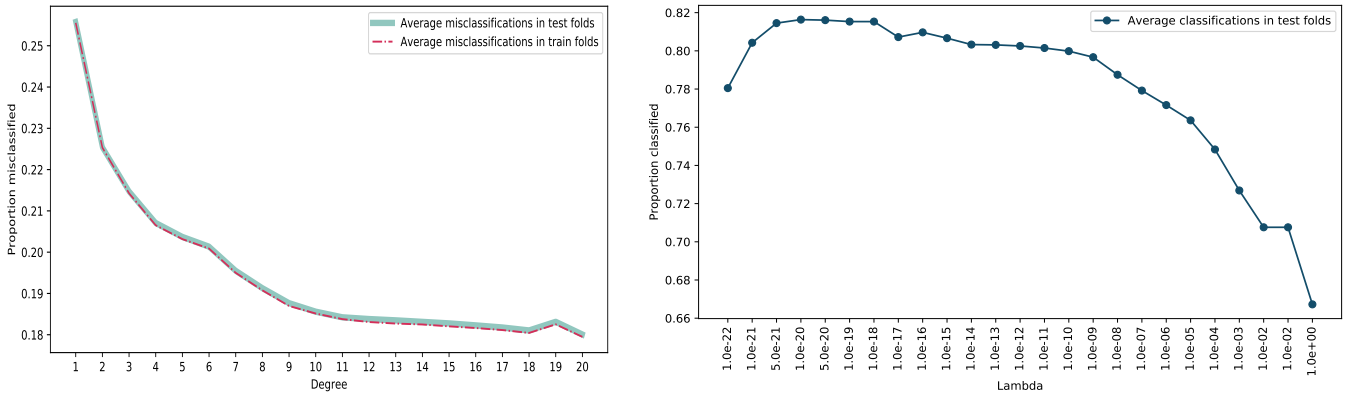


(b) Illustrating Distant Predicted Regressand Values

Figure 1: Choosing between MSE and misclassification loss for hyperparameter tuning

Moving on to our L_2-Regularized Linear Regression, a first decision had to be made as to which metric would be used to determine the best parameters, be it the λ weighting attributing importance to the regularization, or even the degree of polynomial augmentation employed to add complexity to the model. While Least Squares does minimize the Mean Squared Error given these parameters, picking hyperparameters according to another criterion is still feasible. Indeed, intuitively the concept of distance is not a one-to-one match with the classification problem at hand, since so long as the class is correctly predicted, the regressand being more or less distant may not be entirely deterministic as to the strength of the model. Therefore the best degree was

decided by testing a variety of λ and degrees, and selecting whichever presented the lowest misclassifications on the testing folds during cross validation. Cross validation served the multi-purpose of avoiding overfitting on the training set, while retrieving as much training as possible from the data but limiting bias. In addition, cross validation provided for an estimate of performance of the model on data which it had not directly seen.



(a) Determining the best degree

(b) Determining the best regularization λ

Figure 2: Tuning the expansion degree and λ

Results & Discussion

In figure 2 are displayed the results of this tuning, portraying diminishing marginal increases in accuracy above the degree of 10 - note that for each degree, λ's between 1e-16 and 1e2 were tried, with the best accuracy of the given degree reported, most degrees preferring (maximising test fold accuracy) a λ as small as 1e-14. Due to computation time, these were performed on only 5-fold cross validation. Definitely, given the increased accuracy by augmenting the degree of the data, other transformative methods may yield better tuning. Having devised a plan of how to determine whether the tuning seems better or not, it is left to trial.

For finer tuning after honing in on degree 13, upon decision that above this degree marginal returns were so minimal that it would be nothing better than trial and error with the final scoring system and putting Occam's Razor to use, λ was more granularly tuned. Figure (2b) illustrates this λ tuning for 10-fold cross validation. Figure (2b) sets forth the proposal of $\hat{\lambda} = 1e - 20$. Applying the weights determined by Ridge regression on our test set, with $\hat{\lambda} = 1e - 20$, degree 13 expansion, missing values filled by median values of the train data, and minmax normalized, the final estimate of accuracy was 81.9% which is almost identical to that of the internally trained data. Using the mean or standardizing instead of the median or minmax resulted each in an approximate half to a full percent decrease in accuracy. Similarly, using Lasso to feature-select prior to Ridge did not yield better results, as displayed in the notebook selecting the best λ, reinforcing the result of the small λ originally found with Ridge. In addition, increasing the folds of cross validation did not result in drastically different results. The variance differences between 5 and 10-fold cross validation for high degrees are portrayed in the Ridge Tuning notebook and reinforces why upon tuning cross validation is beneficial.

Conclusion

All in all, a plan was devised to best tune the Ridge Regression and make use of the so-far taught concepts, in attempt to best classify whether event's signatures resulted from the Higgs Boson particle or background processes. Final results led to a preference of selecting the polynomial expansion degree, the λ weighting of regularization, the replacement method for missing values, as well as normalization techniques based on whichever yielded the best accuracy on test folds during cross validation, which itself served the purpose of estimating the model's expected performance on unseen data while maximizing training opportunity for the model. The final and best result was 81.9% correctly classified, and an F1-score of 0.72 using this approach. Attempts at both a logistic regression, and an L_2 regularized one, as well a Neural Network were made, with scored accuracy of 80.0% for the non-tuned Neural network.