Machine Learning - Project 1

Marion Chabrier, Valentin Margraf, Octavianus Sinaga
Department of Computer Science, EPFL Lausanne, Switzerland

***Abstract*—The goal of this project is to apply Machine Learning techniques on data from CERN generated by smashing protons into one another and measuring the decay signature of the possibly resulted Higgs boson. With this decay signature as input our model predicts whether it was result of a Higgs boson or something else (noise). We use binary classification methods to tackle this problem.**

I. Introduction

First we preprocess the data i.e. standardize it and get rid of missing values and outliers. Then we implement the six different methods: least_squares, least_squares_GD, least_squares_SGD, ridge_regression, logistic_regression, regularized_logistic_regression. We use each method to learn a model on the training data and see how well they perform. For each model we additionally vary the hyperparameters to optimize the performance. Finally we compare their performances on the test data from CERN.

II. Data Preprocessing

standardize (test and train data)
replace by 0 values with = -999 (test and train data)
delete outliers (train data)

III. Models

For each model we run 4-fold validation on our training data to tune our hyperparameters in order to optimize our model. The hyperparameters in this case are the *degree* for all the models and the constant *lambda_* for the Ridge Regression and the Regularized Logistic Regression.
Figure 1 shows how the choice of the *degree* affects the *RMSE* in the case of Least Squares. We see that for degree = 11 we get our best result, whereas for higher degrees the model will overfit. Lower degrees instead give a bigger *RMSE*, hence the model underfits.

In Figure 2 one can see that *lambda_* ridge works with lambda = 0 because then its like least squares!!!! gives a better fit than. The higher the value the more we penalize large weights i.e. the more we favor small weights.

After having optimized the hyperparameters for each model we want to see how the different models perform on the test data from CERN. We therefore submit each prediction on *AICrowd* and see what result it gives us. In table 2 they can be easily compared.

Least Squares performs best ending up with an accuracy of 0.821. For the *degree* we chose the value 11. Ridgre regression performs good as well with an accurate choice of *degree* 10 and *lambda_* 0.01. It gives an accuracy of 0.815.

All the other methods perform not as good as two mentioned above. Maybe this is because...´

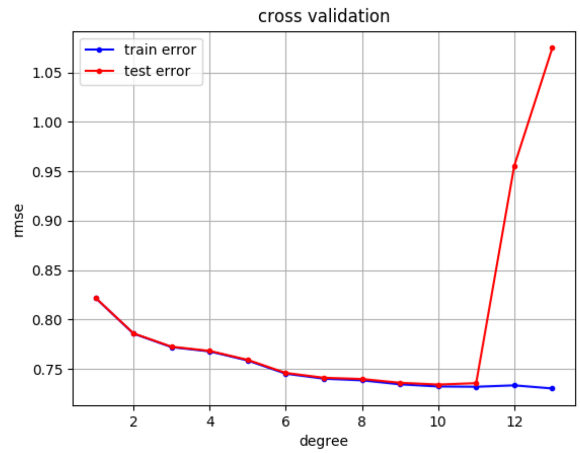

Figure 1. Rmse for different degrees using least_squares.

Methods	lambda_	d
Least Square	-	11
Least Square GD	-	10
Least Square SGD	-	10
Ridge Regression	0.01	10
Logistic Regression	-	10
Reg. Logistic Regression	0.01	10

Table I
Hyperparameters giving the best performance of our models calculated through cross validation.

IV. Results and Discussion

As mentioned before, least_squares perform best concerning accuracy and F1-score. We did not take computational

Methods	Accuracy	F1-Score	lambda_	d
Least Square	0.821	0.723	-	11
Least Square GD	0.566	0.012	-	10
Least Square SGD	0.391	0.394	-	10
Ridge Regression	0.815	0.71	0.01	10
Logistic Regression	0.673	0.12	-	10
Reg. Logistic Regression	0.673	0.12	0.01	10

Table II
Performance of our models submitted on AICrowd.

cost in account in order to compare the methods. Is it actually surprising, that Least Squares Gradient performs that poor because in theory it would converge to the same optimum as Least Squares. Possible causes for this may be that we did not choose a good *gamma* for the stepsize or we did not do enough iterations.

V. Summary

References

[1] S. P. Jones, "How to write a great research paper," 2008, microsoft Research Cambridge.

[2] M. Schwab, M. Karrenbach, and J. Claerbout, "Making scientific computations reproducible," *Computing in Science and Engg.*, vol. 2, no. 6, pp. 61–67, 2000.

[3] J. B. Buckheit and D. L. Donoho, "Wavelab and reproducible research," Stanford University, Tech. Rep., 2009.

[4] R. Gentleman, "Reproducible research: A bioinformatics case study," *Statistical Applications in Genetics and Molecular Biology*, vol. 4, no. 1, 2005. [Online]. Available: http://www.bepress.com/sagmb/vol4/iss1/art2

[5] Editorial, "Scientific writing 101," *Nature Structural & Molecular Biology*, vol. 17, p. 139, 2010.

[6] G. Anderson, "How to write a paper in scientific journal style and format," 2004, http://abacus.bates.edu/ ganderso/biology/resources/writing/HTWtoc.html.

[7] R. H. Kallet, "How to write the methods section of a research paper," *Respiratory Care*, vol. 49, no. 10, pp. 1229–1232, 2004.

[8] A. Hunt and D. Thomas, *The Pragmatic Programmer*. Addison Wesley, 1999.

[9] J. Spolsky, *Joel on Software: And on Diverse & Occasionally Related Matters That Will Prove of Interest etc..: And on Diverse and Occasionally Related Matters ... or Ill-Luck, Work with Them in Some Capacity*. APRESS, 2004.