Project 1 - To the pursuit of the Higgs Boson

Léa Bommottet, Nicolas Brunner, Karine Perrard
EPFL, Switzerland

I. Introduction

Sexy introduction about ML and the problem, we choose to have with 4 different model in function of the data, and all the different models???

II. Models and methods

A. Choice of Models

First of all, we have to select the core of our model. In this project, there is 6 possible models on which we can base.

Linear regression using gradient descent
    Add description + formulas
Linear regression using stochastic gradient descent
    Add description + formulas
Least squares regression using normal equations
    Add description + formulas
Ridge regression using normal equations
    Add description + formulas
Logistic regression using gradient descent or SGD
    Add description + formulas
Regularized logistic regression using gradient descent or SGD
    Add description + formulas

The reason we based our model on ??? is because we have the least global error on the dataset. Fig.1

Furthermore, in function of the feature named "fit", some parameters are unavailable. Thanks to this important information, we don't train only one model, but 4, since "fit" can be 0,1,2,3. This idea reduce our global error, when compared to an single model without this separation. Fig.2

B. Choice of parameters

Blablabla Cross-correlation by K-fold:

As seen in Fig.3, the parameters that reduces most the train error are lambda = ??? and gamma = ???.

C. Choice of features

We can improve our model by adding new features, which are function of other parameters. We want more features that have a Normal distribution.

Natural logarithm
    Add description + picture where a parameter is nice
Square root
    Add description + picture where a parameter is nice
Threshold
    Add description + picture where a parameter is nice
Nothing max
    Add description + picture where a parameter is nice
Nothing norm
    Add description + picture where a parameter is nice
Distance
    When we plot 2 parameters in a graph, we can find some correlation between (Fig.3). To reduces this property to a normal distribution, we compute the Manhattan distance. + picture where a parameter is nice
Nth power
    Add description + picture where a parameter is nice

III. Results

On compare entre model simple et le 4 jets.

On montre les graph features, pourquoi certaines sont sqrt, del, cube.

On compare avec et sans ces features en plus.

On montre notre erreur global, on parle de kaggle?

IV. Discussion

Separating our model into 4 different ones was a great idea, reducing considerably our global error (Fig.1) in the same way that including more parameters (Fig.2). Those added criterion improve the quality of our model, adjusting from ??? to ???.

V. Summary

0.18% of error is kinda ok with the simplicity of the core models. If we want to improve it more, we would need to find more useful parameters to boost our data.

Acknowledgements

On est tous trop beau: et c'est pas nos référence en dessous si jamais

References

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

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

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

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

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

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

[7] 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.

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