About Interpreting and Explaining Machine Learning and Statistical Models

Gilbert Saporta[1]

[1] CEDRIC-CNAM, 292 rue Saint Martin, 75003 Paris, France
(E-mail: gilbert.saporta@cnam.fr)

Keywords: Machine learning, Interpretability, Model-agnostic, Importance measures, Post-processing

The use of black-box models for decisions affecting citizens is a hot topic of debate. Some authors like Rudin [5] are in favour of stopping the use of machine learning models and going back to models which are interpretable by design.

We will focus in this communication on the statistical aspects, leaving aside ethics, despite its importance. The dilemma between explaining and predicting has been addressed by Breiman [1], Saporta [6] and Shmueli [5] among others.

First of all, it seems to us necessary to distinguish between explicability: how does the model work, is the algorithm auditable? and interpretability: what are the important variables and values that may change the decision?

A first approach to make black-boxes interpretable is by performing some post-processing. The idea is to plug in an interpretable model to its outputs, *eg* try to get close predictions by a decision tree or a linear model. One example is given in Liberati *et al.* [3] where a non-linear SVM is approximately reconstructed by a linear classifier, with a small loss of efficiency.
A second approach is to derive variable importance measures by some kind of sensitivity analysis: following Breiman, this can be done by measuring the decrease in prediction accuracy when the values of a predictor are randomly permuted. Molnar [4] gives many examples of such “model-agnostic” interpretation methods.

The concept of an interpretable model deserves to be discussed. In addition to logic or rule based models like decision trees, it is often considered that linear models are easily interpretable. Grömping [2] and Wallard [8] has shown that it is not the case: there exist more than 10 metrics for measuring variable importance in linear regression.

6th SMTDA Conference Proceedings, 2-5 June 2020, Barcelona, Spain

All of the above approaches are still based on correlations and can only provide an imperfect answer to the question: what would be the response if one or more predictors were changed intentionally or unintentionally? Interpretable models should be causal models.

References

1. L. Breiman, Statistical Modeling: The Two Cultures, Statistical Science, 16, 3, 199–231, 2001.
2. U. Grömping, Variable importance in regression models. WIREs Computational Statistics, 7, 137-152, 2015.
3. C. Liberati, F. Camillo, G. Saporta, Advances in credit scoring: combining performance and interpretation in kernel discriminant analysis. Advances in Data Analysis and Classification, Springer Verlag, 11, 1, 121-138, 2017.
4. C. Molnar, Interpretable Machine Learning, A Guide for Making Black Box Models Explainable, https://christophm.github.io/interpretable-ml-book, 2019
5. C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1, 206–215, 2019.
6. G. Saporta, Models for Understanding versus Models for Prediction, In P.Brito, ed., Compstat Proceedings, Physica Verlag, 315-322, 2008.
7. G. Shmueli, To explain or to predict? Statistical Science, 25, 289–310, 2010.
8. H. Wallard Using Explained Variance Allocation to analyse Importance of Predictors, 16th ASMDA conference proceedings, 1043-1054, 2015.