Has my supervised learning model really learnt which are the most important features of the data? The case of financial applications

Argimiro Arratia

Universitat Politècnica de Catalunya, Spain.
E-mail address: argimiro@cs.upc.edu
URL: http://www.cs.upc.edu/~argimiro

Explaining machine learning models predictions is of crucial importance: to understand if we can trust the trained model and make changes consequently, or in many industrial scenarios explaining why a model makes a certain prediction is more important than the prediction itself (e.g. in the banking business when denying a loan or a mortgage to a client, the lender must provide an explanation to the borrower).

A new (and necessary) trend in machine learning applications to the financial industry is to develop criteria to assess if a trained model actually learnt from the data provided, for the specific goalof predictions interpretation. In supervised learning the goodness of a model is commonly measured with some *error-based metrics* that only considers the actual output y and the predicted output $\hat{y}$. The main limitation of this measure is that one does not receive any information about the relations between the output and the features. This causes a lack of explainability of the results that we would like to overcome by introducing a novel method based on Shapley values [1, 2], and developed in the thesis [3]. Instead of measuring the goodness of the fit of a model in terms of an error function (i.e. a function of y and $\hat{y}$, $f(y, \hat{y})$), we propose a performance metric that takes into account also the relations between the inputs and the outputs (a function also of the features: $f(x, y, \hat{y})$). This allows to understand what actually the model have learnt and to start a first diagnostic analysis of the model.

We illustrate the power of our model predictions explanation method with real examples on models for loan default risk analysis and deep neural networks for financial time series forecasting.

Acknowledgments: Research supported by grant TIN2017-89244-R from MINECO (Ministerio de Economía, Industria y Competitividad) and the recognition 2017SGR-856 (MACDA) from AGAUR (Generalitat de Catalunya).

References

[1] Shapley, L. S. (1952). *A value for n-person games* (No. RAND-P-295). Rand Corp Santa Monica CA.

[2] Lundberg, S. M., & Lee, S. I. (2017). *A unified approach to interpreting model predictions*. In: Advances in neural information processing systems, 4765-4774.

[3] Noci, A. (2020) *Explaining Machine Learning models predictions: theory and empirical analysis*. Master Thesis in Quantitative Finance, Mathematical Engineering at Politecnico di Milano, Italy (directors: A. Arratia, L. Belanche)