Explainable Sparse Models: a Marriage between Machine Learning and Decision Theory

Patrice Perny
Sorbonne University
LIP6, team Decision
patrice.perny@lip6.fr

Nataliya Sokolovska
Sorbonne University
INSERM, team NutriOmics
nataliya.sokolovska@sorbonne-universite.fr

Context and Motivation Data integration or variables fusion is a process to combine multiple data sources or variables in a unified model, with the aim to construct a more accurate and reliable predictive models. In a number of real applications, there is an acute need for *explainable* simple (*sparse*) models, among them decision support, medicine, computer vision, remote sensing, e.g., smart cars operating in complex and dynamic environments using numerous sensors. The aim of this thesis is to propose new approaches based on non-additive integrals to construct explainable sparse models.

Traditional machine learning methods consider variables to be independent, even if they are highly correlated in a real task. However, in the last decade, an increasing attention was devoted to non-additive integrals, such as Choquet and Sugeno integrals, enabling to model interactions among variables and providing a fine control of synergies among them. They are more and more used in the context of supervised learning, i.e., where an algorithm has access to observations and their labels. The non-additive integrals are introduced into loss functions to learn reliable predictive models. The Choquet integral, e.g., was successfully used to extend the logistic regression [15], and the Sugeno integral was applied to ordinal aggregation problems [14]. One important specificity of these models is that the number of their weighting parameters exponentially grows with the number of variables. In a number of applications, especially when the model is used to make decisions that may impact human beings, the *interpretability* and the *explainability* of the model are required. There is a need to keep the models as simple and compact as possible.

Besides, decision theory proposes various models to support decision making in complex environments (involving multiple attributes or criteria, or multiple agents). The necessity of justifying decisions has motivated the elaboration of various formal models with different descriptive possibilities, kept as simple as possible to make it possible to explain recommendations (choice, ranking, scoring). In this area, non-additive integrals also are widely used for preference modelling and multicriteria decision support due to their high expressivity [3, 1, 6]. These models are parameterised by a set function named *capacity* assigning a weight to any subset of criteria. Some particular families of capacities are used to bound from above the number of parameters in these models (k-additive capacities) and keep the model as simple as possible. Some indices derived from capacities (e.g., Shapley values, interaction indices, Möbius transform) are used to provide some control of the model and facilitate its interpretability. These models must be adapted to data and could benefit from machine learning tools for their parameterization.

In this thesis, we would like to take the best of two worlds, machine learning and decision theory, both actively developed in Artificial Intelligence, to propose adaptive and interpretable evaluation models and contribute to produce reliable predictive models. Our main motivation is to explore non-additive capacities to construct compact interpretable models. In particular, we are interested to efficiently optimise an objective (loss) function which is based on a compact Choquet integral.

Models, Methods and Goals The existing machine learning approaches including the non-additive integrals are based on fitting the function to observational data based on the least squares error, or, if a robust model is needed, on the L_1 loss. Estimating the capacity vector is a computational challenge, even for moderate size applications. Such a high complexity comes from the exponential number of parameters (with respect to the number of features).

The current PhD project will focus on efficient learning of sparse capacities. We identified *three specific goals and the corresponding methods:*

1. Tackling a problem with a huge number of features, it would be natural to consider and to develop *methods of feature selection.* The problem is challenging, since the parameters of the model (the vector of capacities) are highly correlated, and redundant. The features can be naturally grouped, e.g., by data source. However, the obtained groups are highly overlapping, and there is a need for a specific feature selection approach. Although feature selection methods for overlapping groups exist, e.g., the overlapping group lasso [16], they cannot be applied directly to the non-additive integrals learning, since the monotonicity contraints are needed to be verified.

2. A more or less studied way to learn low-complexity non-additive integrals, is to learn *$k-$additive capacities* [17], for instance, learning 2-additive capacities. It was shown that capturing pairwise interactions between attributes can lead to an accurate model which can be efficiently optimised [17]. Very recently, a hierarchical approach to 2-additive capacities learning was proposed [13]. Together with the open question how to do the estimation efficiently, it is also important to find an optimal k.

3. The capacities are traditionally defined to be monotonic with respect to set inclusion. This is required for criteria aggregation to guarantee that preferences are consistent with Pareto dominance. However, monotonicity adds complexity to the learning of capacities, since the number of constraints that are to be satisfied is exponential (in the number of features). Yet, monotonicity is not always natural in a number of machine learning applications (the contribution of some variables may appear to be negative). The literature on the *non-monotonic capacities* is limited but some results on non-monotonic capacities in relation to machine learning methods already exist [18]. It is worth further analysing the learning of non-additive integrals without monotonicity.

Real Applications We aim to construct practical clinical models from real heterogeneous data. The scores developed in the context of this study will be validated by clinicians of the Pitié-Salpêtrière hospital. Our goal is to integrate bioclinical and environmental phenotyping together with personalized “omics” (metagenomics, metabolomics, transcriptomics, etc.) with the objective of developing new strategies for personalized medicine.

Requirements for the potential candidates. The candidate is expected to have a Master 2 in Computer Science (preferably in AI or Mathematics or Operations Research) or an equivalent engineering degree. A background in Machine Learning, optimization, and decision theory or any related field will be appreciated. An ideal candidate will propose, develop, and test numerically the developed methods. It is expected that the candidate provides some theoretical foundations for the methods and also implements them in R/Matlab/Python, and the final product will be publicly available.

References

[1] N. Benabbou, P. Perny, P. Viappiani. Incremental elicitation of Choquet capacities for multicriteria choice, ranking and sorting problems. *Artificial Intelligence Journal*, 246, 152-180, 2017.

[2] M. Clertant, N. Sokolovska, Y. Chevaleyre and B. Hanczar. Interpretable Cascade Classifiers with Abstention. *AISTATS*, 2019.

[3] M. Grabisch, C. Labreuche A decade of application of the Choquet and Sugeno integrals in multi-criteria decision aid. Annals of Operations Research, 175(1), 247-286, 2010.

[4] A. Kapoor, E. Horvitz. Breaking boundaries: active information acquisition across learning and diagnosis. *NIPS*, 2009.

[5] H. Lakkaraju, C. Rudin. Learning Cost Effective and Interpretable Treatment Regimes in the Form of Rule Lists. *AISTATS*, 2017.

[6] Hugo Martin, Patrice Perny. New Computational Models for the Choquet Integral. *ECAI*, 147-154, 2020.

[7] O. Sobrie, M.A. Lazouni, S. Mahmoudi, V. Mousseau, and M. Pirlot. A new decision support model for preanesthetic evaluation. *Computer Methods and Programs in Biomedicine*, 2016.

[8] N. Sokolovska, Y. Chevaleyre and J.-D. Zucker. A Provable Algorithm for Learning Interpretable Scoring Systems. *AISTATS*, 2018.

[9] N. Sokolovska, Y. Chevaleyre, J.-D. Zucker. Interpretable Score Learning by Fused Lasso and Integer Linear Programming. *DA2PL (From Multiple Criteria Decision Aid to Preference Learning)*, 2016.

[10] N. Sokolovska, Y. Chevaleyre, J.-D. Zucker. The Fused Lasso Penalty for Learning Interpretable Medical Scoring Systems. *IJCNN*, 2017.

[11] A.F. Tehrani, W. Cheng, K. Dembczyński, E. Hüllermeier. Learning monotone nonlinear models using the Choquet integral. Machine Learning, 89(1-2), 183-211, 2012.

[12] B. Ustun and C. Rudin. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, 2015

[13] R. Bresson, J. Cohen, E. Hüllermeier, C. Labreuche, M. Sebag. Neural Representation and Learning of Hierarchical 2-additive Choquet Integrals. *IJCAI*, 2020

[14] G. Beliakov, M. Glagolewski, S. James. Aggregation on ordinal scales with the Sugeno integral for biomedical applications Information Sciences, 2019

[15] A. F. Tehrani, W. Cheng, E. E. Hüllermeier. Choquistic Regression: Generalizing Logistic Regression using the Choquet Integral EUSFLAT, 2011

[16] L. Jacob, G. Obozinski, J.-P. Vert Group Lasso with Overlap and Group Lasso ICML, 2009

[17] E. Hüllermeier, A. F. Tehrani. Efficient Learning of Classifiers Based on the 2-Additive Choquet Integral Computational Intelligence in Intelligent Data Analysis, 2013

[18] T. C. Havens, D. T. Anderson. Machine Learning of Choquet Integral Regression with Respect to a Bounded Capacity (or Non-monotonic Fuzzy Measure) FUZZ-IEEE, 2019