PhD Subject: Explicable Machine Learning via Classification Trees and Discrete Optimization

The research institute LIPN

The LIPN (Laboratoire d'Informatique de Paris-Nord) was created in 1985, and has been associated with CNRS since January 1992. It eventually became a CNRS UMR in January 2001. The LIPN is the computer science laboratory of University Sorbonne Paris Nord (USPN). The laboratory comprises 5 teams: A3, AOC, CALIN, LoVe, RCLN. It has overall more than 150 members including full-time researchers, associate and full professors, technical and administrative staff, as well as PhD students and postdocs. The scientific activities display many collaborations, may they be national, international, in particular with the industry. LIPN members are extremely involved on the national scene, such as in national boards and in institutional cooperation projects, as well on the international scene. Members are widely present in editorial boards, programme committees, expert pools of different countries, international schools, and conduct collaborative research with colleagues worldwide.

The research team AOC

The AOC team mainly focuses on two large areas: "Graphs and Polyhedra" and "Mathematical Programming". The first studies optimization problems in graphs with a focus on polyhedral, approximation and complexity theory. One core topic is the use of geometric and algebraic tools from polyhedral theory to derive new insights towards the essential properties of the underlying combinatorial problem. Such problems can rarely be solved by enumeration, and their study often requires unveiling structural properties of the sets of solutions. The second theme deals with the design and analysis, both theoretical and empirical, of mathematical programming approaches and algorithms. Both exact and heuristic approaches for solving hard combinatorial optimization problems are developed. Mixed integer linear programming and mixed integer non linear programming are among our areas of expertise. We are interested also in other subjects, such as bilevel optimization, robustness and, last but not least, the connection between optimization and machine learning. The strengths of the team are the quantity and quality of publications together with open source software development. The team actively participated, and coordinated, in national projects and has several international collaborations. The international expertise of the team is also assessed by its ability to attract international researchers, including as visiting scholars.

Research topic

Modern deep neural networks are powerful but with the drawback of lack of explainability. Therefore, many researchers are moving on to a different way of thinking. The "easiest" one is represented by random classification trees or forests [9]. In 2017, Bertsimas and Duhn formulate the problem of determining the optimal classification tree (OCT), that is the one minimizing the misclassification error, as a mixed integer linear problem (MILP) [1]. Since then, different models have been proposed [3, 6, 7, 8]. Experimental results show better classification with OCT with respect to heuristic algorithms for determining classification trees like CART [5] or C4.5 [10], but still do not achieve better performances than random forests. Moreover, addressing these formulations by the use of state-of-the-art solvers, like gurobi or cplex, allows to solve the problem for middle-size datasets. On the other hand, these solvers get stuck for large and realistic datasets.

Recently, a MILP formulation has been proposed for the generalization of optimal classification forests (OCF) [2]. It consists in duplicating the formulation of Bertisimas and Duhn [1] k times and in adding

constraints to classify each example by using the majority of the k decisions given by the trees. Experiments show that optimal classification forests allows to obtain predictions with 3 trees of size at most 2 as accurately as by performing Random Forest [4] with 500 trees. However, these models do not scale: indeed, in addition to the complexity inherent to each tree formulation, symmetries among trees appear. This provides a lot of space for improvement. The aim of this PhD project is to define models and algorithms for optimal classification forests and variants for large datasets.

The PhD student will start by performing a deep literature review and attacking the following research questions:

- The formulation of [2] uses a specific MILP formulation for each tree. As many have been proposed since then, are some of them more suitable for being used to formulate each tree in OCF?
- Since in the different formulations trees are of small size (generally with a maximum depth of 3 or 4), and accuracy increases with the number of trees in the forests, we ask the following question: is it possible to decompose the problem into two parts? More precisely, how can we find the best classification with a given set of trees and generating dynamically trees and then, starting from this decomposition, how can we define an efficient algorithm?
- In some domains like medicine, predicting false positive or false negative has clearly not the same impact. Hence, classification trees and forests should forbid one of these two misclassifications. How can we handle this in the different formulations?

Advisors

The prospective PhD student will be supervised by three researchers: Roberto Wolfler Calvo (`wolfler@lipn.fr`), Silvia Di Gregorio (`digregorio@lipn.fr`), and Mathieu Lacroix (`lacroix@lipn.fr`). If you have any questions or would like to hear more about the project and the position, please feel free to contact any of them.

References

[1] Bertsimas, D., and Dunn, J. Optimal classification trees. *Machine Learning 106*, 7 (July 2017), 1039–1082.

[2] Blanco, V., Japón, A., Puerto, J., and Zhang, P. A Mathematical Programming Approach to Optimal Classification Forests, Apr. 2023. arXiv:2211.10502 [cs, math].

[3] Blanquero, R., Carrizosa, E., Molero-Río, C., and Romero Morales, D. Optimal randomized classification trees. *Computers & Operations Research 132* (Aug. 2021), 105281.

[4] Breiman, L. Random forests. *Machine Learning 45*, 1 (2001), 5–32.

[5] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. *Classification and Regression Trees*. Wadsworth, 1984.

[6] Firat, M., Crognier, G., Gabor, A. F., Hurkens, C., and Zhang, Y. Column generation based heuristic for learning classification trees. *Computers & Operations Research 116* (Apr. 2020), 104866.

[7] Hu, X., Rudin, C., and Seltzer, M. Optimal Sparse Decision Trees. In *Advances in Neural Information Processing Systems* (2019), H. Wallach, H. Larochelle, A. Beygelzimer, F. d. Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, Curran Associates, Inc.

[8] Lin, J., Zhong, C., Hu, D., Rudin, C., and Seltzer, M. Generalized and scalable optimal sparse decision trees. In *Proceedings of the 37th International Conference on Machine Learning* (13–18 Jul 2020), H. D. III and A. Singh, Eds., vol. 119 of *Proceedings of Machine Learning Research*, PMLR, pp. 6150–6160.

[9] Rokach, L., and Maimon, O. *Data Mining with Decision Trees*, 2nd ed. WORLD SCIENTIFIC, 2014.

[10] Salzberg, S. L. C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Machine Learning 16*, 3 (Sept. 1994), 235–240.