Explainability Queries for ML Models and its Connections with Data Management Problems

Pablo Barceló ✉
Universidad Católica de Chile, Macul, Chile

Abstract

In this talk I will present two recent examples of my research on explainability problems over machine learning (ML) models. In rough terms, these explainability problems deal with specific queries one poses over a ML model in order to obtain meaningful justifications for their results. Both of the examples I will present deal with "local" and "post-hoc" explainability queries. Here "local" means that we intend to explain the output of the ML model for a particular input, while "post-hoc" refers to the fact that the explanation is obtained after the model is trained. In the process I will also establish connections with problems studied in data management. This with the intention of suggesting new possibilities for cross-fertilization between the area and ML.

The first example I will present refers to computing explanations with scores based on Shapley values, in particular with the recently proposed, and already influential, SHAP-score. This score provides a measure of how different features in the input contribute to the output of the ML model. We provide a detailed analysis of the complexity of this problem for different classes of Boolean circuits. In particular, we show that the problem of computing SHAP-scores is tractable as long as the circuit is deterministic and decomposable, but becomes computationally hard if any of these restrictions is lifted. The tractability part of this result provides a generalization of a recent result stating that, for Boolean hierarchical conjunctive queries, the Shapley-value of the contribution of a tuple in the database to the final result can be computed in polynomial time.

The second example I will present refers to the comparison of different ML models in terms of important families of (local and post-hoc) explainability queries. For the models, I will consider multilayer perceptrons and binary decision diagrams. The main object of study will be the computational complexity of the aforementioned queries over such models. The obtained results will show an interesting theoretical counterpart to wisdom's claims on interpretability. This work also suggests the need for developing query languages that support the process of retrieving explanations from ML models, and also for obtaining general tractability results for such languages over specific classes of models.

2012 ACM Subject Classification Theory of computation → Models of learning

Keywords and phrases ML models, Explainability, Shapley values, decision trees, OBDDs, deterministic and decomposable Boolean circuits

Digital Object Identifier 10.4230/LIPIcs.ICDT.2021.1

Category Invited Talk

Short Bio. Pablo Barceló is a Full Professor at Pontificia Universidad Católica de Chile, where he also acts as Director of the Institute for Mathematical and Computational Engineering. He is the author of more than 80 technical papers, has chaired ICDT 2019, will be chairing ACM PODS 2022, and is currently a member of the editorial committee of Logical Methods in Computer Science. From 2011 to 2014 he was the editor of the database theory column of SIGMOD Record. His areas of interest are database theory, logic in computer science, and the emerging relationship between these areas and machine learning.