Explanation Scores in Data Management

Leopoldo Bertossi[*]

SKEMA Business School Canada
Montreal, Canada
leopoldo.bertossi@skema.edu

1 Introduction

In data management, artificial intelligence (AI), and machine learning (ML) in particular, one wants *explanations* for certain results. For example, for query answers in databases (DBs). In ML, one wants explanations for automated classification results. Explanations that are based on *numerical scores* assigned to elements of a model that may contribute to an outcome have become popular. These *attribution scores* attempt to capture the quantitative degree of relevance of a tuple to a query answer; or a of a feature value to the label assigned to an entity.

In this presentation, we will survey some of the recent advances on the definition, use and computation of score-based explanations for query answering in DBs, and some extensions for ML. Special emphasis is placed on the use of counterfactual reasoning for score specification and computation. This presentation is heavily influenced by our recent research.

2 Explanation Scores

Different scores have been proposed in the literature. Among them we find the *responsibility score* as found in *actual causality* [7, 6], where the notion of *counterfactual intervention* is fundamental. In data management, responsibility, in the form of a *Resp*-score has been used to quantify the strength of a tuple as a cause for a query result [9, 3].

Database repairs are common when dealing with inconsistent DBs [2]. Connections between repairs and actual causality in DBs has been useful to obtain complexity and algorithm results for responsibility [3]. On the basis of database repairs, a measure (or global score) to quantify the degree of inconsistency of a DB has also been introduced.

The *Resp* score has to be generalized to deal with non-binary features in ML [4], which could also be used to define a fine-grained responsibility in DBs at the attribute level. The *causal-effect score* has also been defined and applied to explain query answers in DBs [10].

The Shapley value of *coalition game theory* can be used to define attribution scores in DBs [8, 5]. Since *several tuples together*, much like players in a coalition game, are necessary to produce a query result, some may contribute more than others to a *game function* represented by the query result.

The Shapley value has also been used to define explanation scores to feature values in ML-based classification. Since its computation is intractable in general, tractable classes of models have been identified [1].

References

[1] Arenas, M., Barcelo, P., Bertossi, L. and Monet, M. On the Complexity of SHAP-Score-Based Explanations: Tractability via Knowledge Compilation and Non-Approximability Results. *Journal of Machine Learning Research*, 2023, 24(63):1-58.

[2] Bertossi. L. *Database Repairing and Consistent Query Answering*. Synthesis Lectures in Data Management. Morgan & Claypool, 2011.

[3] Bertossi, L. and Salimi, B. From Causes for Database Queries to Repairs and Model-Based Diagnosis and Back. *Th. Comp. Sys.*, 2017, 61(1):191-232.

[4] Bertossi, L., Li, J., Schleich, M., Suciu, D. and Vagena, Z. Causality-Based Explanation of Classification Outcomes. Proc. 'Data Management for End-To-End Machine Learning', DEEM WS at SIGMOD 2020.

[5] Bertossi, L., Kimelfeld, B., Livshits, E. and Monet, M. The Shapley Value in Database Management. *ACM Sigmod Record*, 2023, 52(2):6-17.

[6] Chockler, H. and Halpern, J. Y. *Resp*onsibility and Blame: A Structural-Model Approach. *Journal of Artificial Intelligence Research*, 2004, 22:93-115.

[7] Halpern, J. and Pearl, J. Causes and Explanations: A Structural-Model Approach. Part I: Causes. *The British J. Phil. of Sci.*, 2005, 56(4):843-887.

[8] Livshits, E., Bertossi, L., Kimelfeld, B. and Sebag, M. The Shapley Value of Tuples in Query Answering. *Log. Methods Comput. Sci.*, 2021, 17(3).

[9] Meliou, A., Gatterbauer, W., Moore, K. F. and Suciu, D. The Complexity of Causality and Responsibility for Query Answers and Non-Answers. Proc. VLDB, 2010, pp. 34-41.

[10] Salimi, B., Bertossi, L., Suciu, D. and Van den Broeck, G. Quantifying Causal Effects on Query Answering in Databases. Proc. TaPP, 2016.

[*] Prof. Emeritus, Carleton University, Ottawa, Canada