Towards Explainable Optimisation Criteria

Indrė Žliobaitė

Towards Explainable Optimisation Criteria

Indrė Žliobaitė*
University of Helsinki, Finland

June 4, 2024

Abstract

Explainability has been at the forefront of machine learning research in recent years. Despite large volumes of research already conducted, a consensus on what should be explained and in what contexts is still lacking. Perhaps a generic consensus is not even possible. Our position is to bring forward explanations of the machine learning process rather than, or in addition to, explaining machine-learned outcomes. In most practical tasks for machine learning, evaluation criteria often evolve during the modelling process. Many model variants are tested before the final model is selected if it is ever final. Explainability research should pay closer attention to explaining the optimisation criteria used for model fitting, evaluation, and selection in more realistic ways.

1 About once in a lifetime

Nearly twenty years have passed since the famous Netflix Prize [1] was handed over to the winning team BellKor's Pragmatic Chaos. The team blended a set of already well-performing advanced predictors into a boosted ensemble [2]. The evaluation criteria for the Prize was elegant, easy to explain and easy to measure. As the race concluded and the publicity receded, the company was left with a dilemma regarding the acquired solution. We now know that the winning solution was never deployed for multiple reasons [3]. The needs have changed, the market has evolved, making the solution scale computationally was too prohibitive. The company concluded that "the additional accuracy gains that [were] measured did not seem to justify the engineering effort needed to bring them into a production environment" [3].

*indre.zliobaite@helsinki.fi

In the early 2010s, Kaggle came to the arena of machine learning competitions [4]. Primarily, they provided a platform for data owners to host competitions. More importantly, they provided consultancy for the hosts on how to formulate concrete evaluation criteria for the purpose of competitions. The evaluation would typically focus on a single measure to match the nature of competitions; after all, outcomes must be precisely measurable to rank the participants [5].

We do not know how many of the winning solutions of Kaggle competitions actually were deployed. Undoubtedly, some of the solutions did. Still, it seems that the most precious outcome of those competitions for the hosts has been access to the people – the winners or near winners open for hire.

Our position is that we, the community, need to strive for explainable optimisation criteria that would be different from the Netflix Prize or Kaggle contests. We do not advocate for abandoning the contests. Yet, the real criteria that make sense for an organisation are rarely mathematically elegant or conceptually concise. In practice, machine learning solutions often follow chains of reasoning, and multiple afflictions happen while testing many solutions without really knowing the ground truth, despite nominal ground truth being available. If there were an easily accessible way to know the real ground truth, there would be no pressing need for an advanced machine learning solution, would there?

The following example comes from a presentation viewed at a leading machine learning conference a few years ago. After the presentation, a member of the audience asked:
"How long did it take you to formulate this into a machine learning problem, prepare the data and then solve it?"
"9-12 months", answered the presenter, "the work included an advisor guiding two students".
"And how often do we run into this task?", asked the audience member.
"About once during a lifetime", the presenter replied.
"So what would it take to solve this by hand, not using machine learning?", the audience member continued.
"About 2-3 months for one person", the presenter replied.

Of course, we have the teaching duty, and teaching machine learning in the way of Kaggle-style contests works well. After all, many standard machine learning research papers are like mini contests themselves. A study would present a proposed solution and compare it to a baseline and perhaps several advanced competitive solutions. The proposed solution is expected to win, unlike in the contests, not on a single but on many datasets, as many as possible. Keeping in mind that there is no "free lunch" [6] reviewers

would still grin at solutions that do not simultaneously excel across multiple datasets.

Indeed, we would like machine learning research to generalise. We admire neat theory, even more so when it is spiced up with the real world flavours. Building theory is undoubtedly essential, but so is keeping in touch with reality. Hence, there is a lot to explain on how we do things and why we do things in machine learning, especially machine learning research, in order for the users, the regulators, and the general public [7] to understand better what it is to be expected from machine-learned solutions.

Explainability of optimisation criteria in our position thus encompasses multiple aspects. Explainability of optimisation criteria is not only about being mathematically convenient. It is also about being practically meaningful, articulated, and perhaps repeatable? We are not sure about repeatability, though. Would repeatability (as opposed to replicability and in relation to reproducibility) imply that given the same task two independent machine learning design teams would be expected to arrive to the same solutions without reusing the previous code? There is a lot to think about here.

2 No magic in the method

Explainability research (XAI) rose to mainstream along with the popularity of deep learning. The general focus of XAI is to communicate how machine-learned systems reason [8–12]. The majority of XAI research focuses on ways of explaining how systems make predictions when the models have already been trained and are ready to be used [13,14]. Explanations focus on either how a given model works or how predictions for a particular individual are made with this model. Popular ways of explaining include reasoning about feature importance [15], making simpler surrogate models [16], interpreting model parameters or comparing to known reference data points prototypes.

These ways of explaining dissect models in retrospect. Sometimes, such explanations sound nearly apologetic. It is clear that before attempting to explain to others we, model makers, first and foremost need to explain the modelling process to ourselves. And the first question to answer is whether the computational task we are trying to solve makes sense in reality? Do we need this model? Is it worth the effort? How much human labor and other resources go into it? [17]. Or would one better solve the task by hand, like in the conference dialogue example? Any claim of real world relevance should be accompanied by an answer to these questions, and if our answer is negative, let us call our exercise a toy example, a synthetic experiment, a

benchmark for a competition.

There is no magic in the method. Some madness, perhaps. Before even starting to explain the method, we need to explain first why we chose to make a machine-learned model over other ways of decision making.

If the task is real, the next question is whether we are after automating something that humans can easily solve or already know, but machine learning is expected to do it cheaper or faster? Or are we after discovering new knowledge, something that is not known and perhaps not even knowable at the time of using machine-learned models? The former, perhaps, does not even need explanations, while explaining the latter is notoriously difficult. Explaining, in this case, effectively requires solving the task manually in addition to solving it via machine-learning.

And if we can really solve the knowledge discovery task manually, do we still need to solve it via machine learning? Suppose we do. Explaining how we solve it is challenging when there are many choices to be made along the way unless we are in a setting like chess, where the rules of the game are clear and the world is fully observable [18].

If we need to explain, in a broad sense, why we chose this way of making a model over many alternatives, one way is to look at the differences. Differences are often easier to explain than absolutes. Machine-learned solutions can differ in terms of the input data they use (e.g., different features), they can differ in the model architectures and the shapes of the decision boundaries that they can capture (e.g. linear vs. non-linear), they can use different optimisation criteria for fitting the model (e.g., ordinary least squares vs. regularised regression), or they can post-process the outputs in different ways (e.g. preventing the system from outputting negative predictions for the amount of precipitation).

After all, the users of the knowledge-discovery type of models would be interested in differences. They would ask: "So you have predictions from all those model variants; why do they differ?" To explain this, one will inevitably need to explain the modelling process.

Questions to explain will include questions:

- How do we come up with and select the final model?
- Is the selection based on accuracy of the model fit? Simplicity? Robustness (e.g. to outliers)?
- Have we selected to optimise for computational costs in model fitting and model predictions?

The challenge is that this way of explaining is difficult to standardise, let alone quantify. What is not easy to standardise is not easy to research, not easy to pack into the expected formats of research papers, not easy to produce PhD theses on. Neither is it easy to repeat. Are we bound to contest-style research to dominate even the field of XAI?

3 Discussion potential

We have no easy solution to propose. Points of departure for a discussion could include questions like:

- Should we modify or expand research evaluation criteria?
- Should we add evaluation of using a machine-learned vs. manual predictions?
- Should we add human labor as a dimension?
- Should we ask to document and list all the variants ever tested? Should we ask to list the sequence of testing? (until there is a better way)
- Can optimisation criteria ever be fully objective? Can explanations be objective? [19]
- Can we test objectivity?
- Would another model developer, given the same task come up with the same optimisation criteria? If so, would it be a matter of convention[1], mathematical convenience?

A true interdisciplinary collaboration between sciences, or art and science, is about understanding each other's processes. Making machine learning models is an interdisciplinary collaboration with society. Thus, we need to understand the processes of society, and we need to communicate our processes, the processes of machine learning developers, and researchers. We need to explain how we do modelling in general and how we can solve a predictive task at hand in particular. We cannot realistically explain machine-learned outcomes without explaining the machine learning process.

[1] An example of a reviewer's comment: "Although I agree with the authors' rationale for not wanting to use PGLS, in this day and age it is simply not acceptable to NOT do the analyses using PGLS."

References

[1] J. Bennett and S. Lanning. The Netflix Prize. In *Proceedings of KDD Cup and Workshop*, 2007.

[2] Y. Koren. The BellKor Solution to the Netflix Grand Prize, 2009.

[3] X. Amatriain and J. Basilico. Netflix recommendations: Beyond the 5 stars (part 1), 2012.

[4] J. Carpenter. May the best analyst win. *Science*, 331(6018):698–699, 2011.

[5] A. Blum and M. Hardt. The ladder: A reliable leaderboard for machine learning competitions. In *Proc. of the 32nd Int. Conf. on Machine Learning*, pages 1006–1014, 2015.

[6] D. Wolpert and W. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1:67–82, 1997.

[7] M. Langer, D. Oster, T. Speith, H. Hermanns, L. Kästner, E. Schmidt, A. Sesing, and K. Baum. What do we want from Explainable Artificial Intelligence (XAI)? – a stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, 296:103473, 2021.

[8] D. Gunning and D. Aha. DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine*, 40(2):44–58, 2019.

[9] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020.

[10] A. Adadi and M. Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018.

[11] E. Tjoa and C. Guan. A survey on explainable artificial ntelligence (XAI): Towards medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11):4793–4813, 2021.

[12] A. Das and P. Rad. Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. *ArXiV*, 2020.

[13] C. Molnar. *Interpretable Machine Learning*. LeanPub, 2019.

[14] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5), 2018.

[15] S. Lundberg and S. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4768–4777, 2017.

[16] M. Ribeiro, S. Singh, and C. Guestrin. "Why should i trust you?": Explaining the predictions of any classifier. In *Proc. of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, KDD, pages 1135–1144, 2016.

[17] K. Crawford. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press, 2021.

[18] R. McIlroy-Young, S. Sen, J. Kleinberg, and A. Anderson. Aligning superhuman AI with human behavior: Chess as a model system. In *Proc. of the 26th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, KDD, pages 1677–1687, 2020.

[19] L. Daston and P. Galison. *Objectivity*. Princeton University Press, 2007.