ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240462

Actively Learning from Machine Learning Models with Queries and Counterexamples (Extended Abstract)

Ana Ozaki [a,*]

ORCID (Ana Ozaki): https://orcid.org/0000-0002-3889-6207

Abstract. We consider the exact and probably approximately correct (PAC) learning frameworks from computational learning theory and discuss opportunities and challenges for applying notions developed within these frameworks to extract information from black-box machine learning models, in particular, from language models. We discuss recent works that consider algorithms designed for the exact and PAC frameworks to extract information in the format of automata, Horn theories, and ontologies from machine learning models and possible applications of these approaches for understanding the models, studying biases, and knowledge acquisition.

1 Introduction

In active learning [3], the learner can formulate questions to a teacher, in particular, ask for the classification of an input, in this way, actively navigating the search space of possible hypotheses. To formalize the learning procedure, we consider Angluin's exact learning framework [2]. In this framework, a learner interacts with a teacher using queries in a predefined format. The learning task is seen as an identification task, that is, there is a space of possible hypothesis and the goal is to identify a "correct" one. For example, consider a patient that goes to the doctor with some infection and the task of the doctor is to identify which kind of infection is attacking the patient. In this scenario, the doctor plays the role of the learner and tries to learn the medical condition of the patient, who plays the role of the teacher. The most studied types of queries in Angluin's framework are *membership* and *equivalence* queries.

Membership queries embody the basis of active learning: the learner chooses an input, also called *example*, and asks the teacher to classify it into 'yes' (positive) or 'no' (negative) [3]. In the scenario described earlier, the doctor may ask the patient whether the patient had fever, which the patient replies with 'yes' or 'no'. Equivalence queries can be more challenging for the teacher. In this case, the learner formulates a hypothesis and asks the teacher whether it is correct and, if not, it asks for a *counterexample*, that is, an example that illustrates the hypothesis is not correct. In our toy scenario, this would not make sense as it would mean the doctor would ask the patient whether she/he has a particular medical condition X and ask for a counterexample if X is not correct (but finding X is what the patient was trying to do). Fortunately, equivalence queries can be *simulated* by asking batches of membership queries. The learning outcome may not be fully correct but, under certain conditions, one can give a probabilistic guarantee based on Valiant's probably approximately correct (PAC) framework [17]. The exact and PAC frameworks have been extensively investigated in the literature and several authors have proposed algorithms for learning classical concept classes in computer science within these frameworks such as automata, decision trees, Horn theories, and ontologies. An emerging line of research is to study these algorithms and their applicability for extracting information from black-box machine learning models. Can we extract automata from neural networks [20]? Can we extract a Horn theory [5] or an ontology from a language model [14]?

* Corresponding Author. Email: anaoz@uio.no.

2 Motivation

There are multiple reasons for why considering black-box machine learning models as teachers and learning from them can be interesting. The first is that, when investigating a black-box model, one may not know which datasets were used to train the model (this is indeed the case for many machine learning models used in practice, such as proprietary language models), which hinders the possibility of training a new model with the same data, for analysis and comparison. Secondly, even when the data is available, training a new model may be overly expensive. Moreover, it is known that modern architectures may not be calibrated [11], meaning that what is extracted from a machine learning model may not be the same as if learning from the data, so querying the model could help to detect and understand a possible biased behaviour of the model. The fourth reason is that even when the data used to train a model is available, not too large, and the model is calibrated, by actively querying the model, one can take advantage of its generalization ability and get a classification for instances not originally present in the data.

The active learning approach can be used to verify whether the behaviour of a model is as expected, by covering a space of inputs/outputs in a systematic way. This could be useful to create *adversarial inputs* [20] and to explore *out of distribution* inputs/outputs where usually models tend to fail [10]. Interestingly, one can see the active learning approach as some kind of *knowledge compilation*, where the relevant part of the information in a black-box model is extracted in a desired format, possibly easier to compute, e.g. a Horn theory, or easier to interpret and explain, e.g. a decision tree.

Several authors have proposed strategies to extract decision trees from black-box machine learning models [13, 18, 6, 16, 8, 15, 7], as an attempt understand/explain the models or to create a surrogate model that is explainable. While one cannot take the explanation of a formalism extracted from a black-box model, such as a decision tree, as an explanation of the *internal behaviour* of the model, an extracted decision tree with its explanations may offer useful information about the overall behaviour of the model, since it is an approximation of

it. However, current works on extracting decision trees from black-box models tend to focus on empirical evidence (by indicating that the explanations are *plausible*) instead of theoretical guarantees on the fidelity of the extracted trees. In the next section, we discuss in more details works based on algorithms designed for Angluin's exact learning framework, with a strong theoretical foundation, where the teacher is a machine learning model.

3 Automata, Horn Theories, Ontologies

The most well-known algorithm designed within Anlguin's exact learning framework, called L^*, creates an automaton by posing membership and equivalence queries to a teacher [1]. Automata are interesting because they represent how an instance can go from one state to the other. This algorithm has been applied in different kinds of context in the literature and, more recently, it has been applied to extract automata from recurrent neural networks (RNNs) [19]. One of the challenges in applying algorithms designed within Angluin's framework with membership and equivalence queries is how to answer equivalence queries, which as we mentioned before, in practice, need to be simulated. Another challenge is that the L^* algorithm assumes the teacher is trying to teach a regular language. While this can be a theoretical assumption explored by the algorithm and its correctness proof, in practice, this may not be the case when you consider that answers are coming from a neural network. Indeed the authors report that this happened in their experiments with RNNs [19].

Horn expressions have also been extensively studied, in particular, as a class of concepts that can be efficiently learned within Angluin's framework with membership and equivalence queries [4]. This class is interesting because Horn expressions naturally represent 'if'-'then' statements and reasoning can be performed in polynomial time. Extracting Horn expressions from BERT-based language models has been recently explored [5]. The authors address the challenge that the answers of the machine learning model may not consistent with a Horn theory by analysing the theoretical problem without the assumption that the target expression to be learned is a Horn theory. They also need to simulate equivalence queries and deal with the fact that the input/output of an algorithm designed within Angluin's framework does not match with the format of the input/output of BERT-based models. The experiments are used to validate results in the literature that indicate occupational gender biases in these models but never used Horn expressions for this purpose.

Actively learning ontologies has been investigated for various fragments of lightweight ontologies [12], though, without using language models as teachers. On this direction, there have been recent efforts in extracting ontologies from language models [9, 14]. It has been established that there is statistical evidence of correlation between what is expressed in ontologies created by ontology engineers and answers given by language models [14]. Although this may not be a surprise at all to users with experience in language models such as ChatGPT, the approach allows to systematically investigate these models, provide a quantitative overview of the results for ontologies in various domains of knowledge, and pave the way for knowledge acquision in the format of ontologies from language models.

References

[1] D. Angluin. Learning regular sets from queries and counterexamples. *Inf. Comput.*, 75(2):87–106, 1987. doi: 10.1016/0890-5401(87)90052-6. URL https://doi.org/10.1016/0890-5401(87)90052-6.

[2] D. Angluin. Queries and concept learning. *Machine Learning*, 2(4): 319–342, 1988. ISSN 0885-6125.

[3] D. Angluin. Computational learning theory: Survey and selected bibliography. In S. R. Kosaraju, M. Fellows, A. Wigderson, and J. A. Ellis, editors, *Proceedings of the 24th Annual ACM Symposium on Theory of Computing*, pages 351–369. ACM, 1992. doi: 10.1145/129712.129746.

[4] D. Angluin, M. Frazier, and L. Pitt. Learning conjunctions of horn clauses. *Mach. Learn.*, 9:147–164, 1992. doi: 10.1007/BF00992675.

[5] S. Blum, R. Koudijs, A. Ozaki, and S. Touileb. Learning horn envelopes via queries from language models. *International Journal of Approximate Reasoning*, page 109026, 2023. ISSN 0888-613X. doi: https://doi.org/10.1016/j.ijar.2023.109026. URL https://www.sciencedirect.com/science/article/pii/S0888613X23001573.

[6] O. Boz. Extracting decision trees from trained neural networks. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-26, 2002, Edmonton, Alberta, Canada*, pages 456–461. ACM, 2002. doi: 10.1145/775047.775113. URL https://doi.org/10.1145/775047.775113.

[7] M. W. Craven and J. W. Shavlik. Extracting tree-structured representations of trained networks. In D. S. Touretzky, M. Mozer, and M. E. Hasselmo, editors, *NeurIPS*, pages 24–30. MIT Press, 1995. URL http://papers.nips.cc/paper/1152-extracting-tree-structured-representations-of-trained-networks.

[8] D. Dancey, D. McLean, and Z. Bandar. Decision tree extraction from trained neural networks. In V. Barr and Z. Markov, editors, *Proceedings of the 17th Int. Florida Artificial Intelligence Research Society Conference, Miami Beach, Florida, USA*, pages 515–519. AAAI Press, 2004. URL http://www.aaai.org/Library/FLAIRS/2004/flairs04-089.php.

[9] M. Funk, S. Hosemann, J. C. Jung, and C. Lutz. Towards ontology construction with language models. In S. Razniewski, J. Kalo, S. Singhania, and J. Z. Pan, editors, *KBC-LM and LM-KBC*, volume 3577 of *CEUR Workshop Proceedings*, 2023.

[10] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z.

[11] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *ICML*. JMLR.org, 2017.

[12] B. Konev, C. Lutz, A. Ozaki, and F. Wolter. Exact learning of lightweight description logic ontologies. *J. Mach. Learn. Res.*, 18: 201:1–201:63, 2017.

[13] R. Krishnan, G. Sivakumar, and P. Bhattacharya. Extracting decision trees from trained neural networks. *Pattern Recognit.*, 32(12):1999–2009, 1999. doi: 10.1016/S0031-3203(98)00181-2.

[14] R. S. Matteo Magnini, Ana Ozaki. Actively learning ontologies from llms: First results (extended abstract). In J. C. J. Laura Giordano and A. Ozaki, editors, *DL*, volume 3739 of *CEUR Workshop Proceedings*, 2024. URL https://ceur-ws.org/Vol-3739/abstract-18.pdf.

[15] G. Nanfack, P. Temple, and B. Frénay. Global explanations with decision rules: a co-learning approach. In C. P. de Campos, M. H. Maathuis, and E. Quaeghebeur, editors, *UAI*, volume 161 of *Proceedings of Machine Learning Research*, pages 589–599. AUAI Press, 2021. URL https://proceedings.mlr.press/v161/nanfack21a.html.

[16] G. P. J. Schmitz, C. Aldrich, and F. S. Gouws. ANN-DT: an algorithm for extraction of decision trees from artificial neural networks. *IEEE Trans. Neural Networks*, 10(6):1392–1401, 1999. doi: 10.1109/72.809084. URL https://doi.org/10.1109/72.809084.

[17] L. G. Valiant. A theory of the learnable. In R. A. DeMillo, editor, *Proceedings of the 16th Annual ACM Symposium on Theory of Computing, April 30 - May 2, 1984, Washington, DC, USA*, pages 436–445. ACM, 1984. doi: 10.1145/800057.808710. URL https://doi.org/10.1145/800057.808710.

[18] N. Vasilev, Z. Mincheva, and V. Nikolov. Decision tree extraction using trained neural network. In C. Klein and M. Helfert, editors, *Proceedings of the 9th International Conference on Smart Cities and Green ICT Systems, SMARTGREENS 2020, Prague, Czech Republic, May 2-4, 2020*, pages 194–200. SCITEPRESS, 2020. doi: 10.5220/0009351801940200. URL https://doi.org/10.5220/0009351801940200.

[19] G. Weiss, Y. Goldberg, and E. Yahav. Extracting automata from recurrent neural networks using queries and counterexamples. In J. G. Dy and A. Krause, editors, *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 5244–5253. PMLR, 2018. URL http://proceedings.mlr.press/v80/weiss18a.html.

[20] G. Weiss, Y. Goldberg, and E. Yahav. Extracting automata from recurrent neural networks using queries and counterexamples (extended version). *Mach. Learn.*, 113(5):2877–2919, 2024. doi: 10.1007/S10994-022-06163-2. URL https://doi.org/10.1007/s10994-022-06163-2.