Rigorous Explanations for Machine Learning Models*

Extended Abstract

Joao Marques-Silva

Faculty of Science, University of Lisbon, Portugal
jpms@ciencias.ulisboa.pt

The recent successes of Machine Learning (ML) motivate an ever growing range of applications. In some settings, e.g. in safety critical applications, one is often expected to explain the predictions made by ML models. For example, this is the case when such predictions are to be assessed by a human decision maker, or used for later diagnosis in the case of failure. Some ML models are naturally amenable to interpretation. This is the case with logic models, including decision trees, lists and sets. In such cases, the models represent the explanations explicitly, and so the goal is to synthesize models such that the resulting explanations are as succinct as possible [8,16,3,12,7]. However, in many settings the most successful ML models are *not* naturally interpretable and, from the perspective of a human decision maker, operate as black-boxes. Concrete examples include (Deep) Neural Networks ((D)NNs), Support Vector Machines (SVMs), Bayesian Network Classifiers (BNCs), model ensembles, among many others. Approaches for explaining non-interpretable ML models are most often heuristic [13,4,11,14,10,17,1,2,9,5], in that explanations are computed by only exploiting information that is *local* to a given instance. Alternatively, some recent works focused on devising rigorous approaches for computing explanations. One such example is a compilation-based approach for BNCs [15]. This recent work also established a natural relationship between explanations and prime implicants of the classification function, concretely prime implicant explanations. A different approach [6] is to bypass the need for compilation, and relate explanations with abduction. The special setting of ML predictions enables relating abduction with the computation of prime implicants. Furthermore, and instead of exploiting compilation, this approach develops dedicated algorithms for computing explanations. More importantly, these two works [15,6] enable the computation of explanations which hold *globally*, in clear contrast with existing approaches for computing local explanations. More formally, given some logic-based representation $\mathcal{M}$ of a target ML model, a concrete instance $\mathcal{I}$, and a prediction π for that instance, a (global) explanation $\mathcal{E} \subseteq \mathcal{I}$ is such that $\mathcal{E} \models (\mathcal{M} \rightarrow \pi)$. This problem formulation enables the computation of both cardinality-minimal and subset-minimal explanations, provided a oracle (i.e. a reasoner) for the decision problem $(\mathcal{M} \rightarrow \pi)$ exists. This talk provides an overview of these recent approaches for computing global explanations. The current focus is on explaining NNs models, but the approach can conceptually be applied to any other setting where the ML model accepts a logic-based

* Work supported by FCT grants ABSOLV (PTDC/CCI-COM/28986/2017) and FaultLocker (PTDC/CCI-COM/29300/2017). Joint work with Alexey Ignatiev and Nina Narodytska.

representation. Moreover, the talk summarizes existing experimental results and highlights ongoing research work.

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I.J., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: NeurIPS. pp. 9525–9536 (2018)
2. Alvarez-Melis, D., Jaakkola, T.S.: Towards robust interpretability with self-explaining neural networks. In: NeurIPS. pp. 7786–7795 (2018)
3. Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., Rudin, C.: Learning certifiably optimal rule lists for categorical data. Journal of Machine Learning Research **18**, 234:1–234:78 (2018)
4. Frosst, N., Hinton, G.E.: Distilling a neural network into a soft decision tree. In: CExAIIA (2017)
5. Ibrahim, M., Louie, M., Modarres, C., Paisley, J.: Global explanations of neural networks. In: AIES (2019)
6. Ignatiev, A., Narodytska, N., Marques-Silva, J.: Abduction-based explanations for machine learning models. In: AAAI (2019)
7. Ignatiev, A., Pereira, F., Narodytska, N., Marques-Silva, J.: A SAT-based approach to learn explainable decision sets. In: IJCAR. pp. 627–645 (2018)
8. Lakkaraju, H., Bach, S.H., Leskovec, J.: Interpretable decision sets: A joint framework for description and prediction. In: KDD. pp. 1675–1684 (2016)
9. Lakkaraju, H., Kamar, E., Caruana, R., Leskovec, J.: Faithful and customizable explanations of black box models. In: AIES (2019)
10. Li, O., Liu, H., Chen, C., Rudin, C.: Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In: AAAI. pp. 3530–3537 (2018)
11. Montavon, G., Samek, W., Müller, K.: Methods for interpreting and understanding deep neural networks. Digital Signal Processing **73**, 1–15 (2018)
12. Narodytska, N., Ignatiev, A., Pereira, F., Marques-Silva, J.: Learning optimal decision trees with SAT. In: IJCAI. pp. 1362–1368 (2018)
13. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should I trust you?": Explaining the predictions of any classifier. In: KDD. pp. 1135–1144 (2016)
14. Ribeiro, M.T., Singh, S., Guestrin, C.: Anchors: High-precision model-agnostic explanations. In: AAAI (2018)
15. Shih, A., Choi, A., Darwiche, A.: A symbolic approach to explaining bayesian network classifiers. In: IJCAI. pp. 5103–5111 (2018)
16. Wang, T., Rudin, C., Doshi-Velez, F., Liu, Y., Klampfl, E., MacNeille, P.: A bayesian framework for learning rule sets for interpretable classification. Journal of Machine Learning Research **18**, 70:1–70:37 (2017)
17. Wu, M., Hughes, M.C., Parbhoo, S., Zazzi, M., Roth, V., Doshi-Velez, F.: Beyond sparsity: Tree regularization of deep models for interpretability. In: AAAI. pp. 1670–1678 (2018)