Pitfalls of Local Explainability in Complex Black-Box Models

Antonio Maratea, Alessio Ferone

Department of Science and Technologies, University of Naples "Parthenope", Isola C4, Centro Direzionale, I-80143 Napoli, ITALY

Abstract

Post hoc models are becoming popular as additional tools to evaluate the results of black-box models and to provide explanations of the predictions they give. In this paper the main concerns that Local Induced models raise in the pointwise explanation of heavily overparametrized black-box models are discussed in depth, highlighting some vulnerabilities, some underrated issues and giving some warnings on the potentially negative effect on user trust of this explainability framework.

Keywords

XAI, Interpretable Machine Learning, Local Induced Models, User Trust

1. Introduction

The need for eXplainable Artificial Intelligence (XAI) has become an urgency due to the ease of implementation and stunning performances of overparametrized black–box models, especially Deep Neural Networks, that countless applications are testifying, often showing super-human abilities and rising new ethic concerns [1]. While there is an open debate on what an explanation should be, it can be reasonably assumed it requires a simple, transparent and understandable for humans metamodel, that reproduces to a certain degree the behavior of the underlying opaque oracle. Such metamodel can aim to either explain the predictions of the black-box model, or the motivations behind these predictions, requiring a different approach: the former more specifically focused with explaining a certain decision for a given input, the latter more generally targeted to the logic behind the model. A recent survey of explainable models can be found in [2].

In particular, given a black-box model, three problems can be considered [3]:

- the **model explanation** aims to open the black-box, that is to provide a global explanation of the model through a metamodel that is interpretable by the users.
- the **outcome explanation** aims to explain the correlation between the input data and the decision. Given a black-box and a a specific input instance, without explaining the whole underlying logic, it should provide an human-interpretable reason for the decision on that particular instance.

WILF'21: International Workshop on Fuzzy Logic, December 20–22, 2021, Vietri sul Mare, Italy
antonio.maratea@uniparthenope.it (A. Maratea); alessio.ferone@uniparthenope.it (A. Ferone)
0000-0001-7997-0613 (A. Maratea); 0000-0002-4883-0164 (A. Ferone)

- the **model inspection** problem aims to provide a representation to understand some specific properties of the black-box model or its predictions.

Model-agnostic are called the explainability models that are independent from the specific model to be explained. Recently, these methods have flourished and are seeing an increase in popularity due to their simplicity and fascinating "one fits all" promise. In the following the outcome explanation problem is tackled, focusing on pitfalls of current Local Induced *post hoc* models.

2. Global models

Given a problem $\mathscr{P}$ with domain D and a desired tolerance λ, let $C_{\mathscr{P}}(\lambda) : D \to \mathbb{N}$ be the complexity of the problem at hand, measured as the minimum number of free parameters or independent dimensions required to properly represent $\mathscr{P}$ (excluding the target variable): this is the so-called *Intrinsic Dimension* (ID) of the problem. $C(\lambda)$ is a monotonically decreasing function, that is $\lambda_1 \leq \lambda_2 \Rightarrow C(\lambda_1) \geq C(\lambda_2)$: the more fidelity is required from the model (lower tolerance λ), the more free parameters (or dimensions) are necessary and sufficient to reach it. Let E_m be the expressive power of a model m, measured as the number of free parameters or independent dimensions of the model, to guarantee the tolerance λ it must hold that:

$$E_m \geq C_{\mathscr{P}}(\lambda) \tag{1}$$

There are infinite models with the same expressive power coming from different families that can be used to model the problem $\mathscr{P}$, each with his trade-offs in terms of bias-variance, effectiveness-efficiency, exploration-exploitation, approximation-generalization, fidelity-interpretability, plasticity-stability and so on. The choice of the model and the discussion of its pros and cons is precisely the work of scientists and researchers, that hardly find consensus solutions, as the growing scientific literature demonstrates. Traditionally researchers have fought first to estimate the intrinsic dimension of a problem and then to find models that closely match this value, i.e. with $E_m = C_{\mathscr{P}}$, or reversely they have tried to find from the beginning the simplest, least expressive, model producing an acceptable value of λ. The reason for the second choice is that even if a well defined and transparent model actually exists and could be found, if the complexity of the problem is very high, the model with $E_m = C_{\mathscr{P}}$ could be still too complex to be understood and managed by humans. Going over tens of dimensions can easily produce models that are barely intelligible, and it can be said that to be useful for humans E_m must remain in the order of 10^2 or less. *Dimensionality reduction techniques* aims precisely to this, and on the positive side it must be said that many problems show a very low intrinsic dimension and even aggressive approximations produce good results.

The recent explosion of popularity of Deep Neural Networks (DNN) created a somewhat unprecedented situation: the temptation for prêt-à-porter accuracy has spread models with millions of parameters even when the intrinsic dimension of a problem is small: instead of looking for the model with the closest possible expressive power to the intrinsic dimension, or for the least expressive admissible representation, model overparametrization has become the norm. The extreme situation where the number of dimensions p exceeds even the number of

training samples n is no more a taboo. In this case it holds:

$$E_m >> C_{\mathscr{P}}(\lambda) \quad \text{and} \quad O_m = \frac{E_m}{C_{\mathscr{P}}(\lambda)} \tag{2}$$

Where $O_m >> 1$ is the overparametrization ratio for the chosen model. While DNN show stunning performances in terms of crude accuracy, they heavy overparametrize the problem, lack in transparency and generalization, are sensitive to adversarial attacks and ultimately do not provide a clear explanation of the motivation behind the results they give. It must be reminded that each new parameter/dimension increases exponentially the space of configurations to be explored.

3. Local models

A *Local Model*, called *lm*, is a model that approximates locally the model m on the problem $\mathscr{P}$. Local models make sense because even if a problem has complexity $C_{\mathscr{P}}(\lambda)$ on its full domain, there may be, and usually there are, subdomains where the complexity drops and simpler models are enough powerful to guarantee λ.

If d_i is a subdomain of the problem domain D, let $C_{\mathscr{P}_i}(\lambda) : d_i \rightarrow \mathbb{N}$ be the complexity of the considered problem in d_i, measured as the minimum number of free parameters or independent dimensions required to properly represent $\mathscr{P}$ locally: this is the so-called *Local Intrinsic Dimension* (LID) of the problem in d_i. This complexity is neither monotonic, nor constant along any direction, and again the expressive power of the chosen local model should dominate it. In formulae:

$$E_{lm_i} \geq C_{\mathscr{P}_i}(\lambda) \; \forall i \quad \text{and} \quad E_m \geq \max_i (E_{lm_i}) \tag{3}$$

Many local models are required to approximate the global one, and much like a piece-wise linear function can approximate any curve in the euclidean plane within a given tolerance if sufficiently small intervals are considered, any local model can approximate any global model to a given degree of tolerance λ if sufficiently small regions are considered. There is a positive correlation between the width of the regions (subdomains) and the minimum λ that can be guaranteed. In critical regions where the model shows an irregular pattern, locality should be increased and more local models, each valid for smaller subregions, are required.

4. Induced models for XAI

An *Induced Model*, called *im*, is a more interpretable, simpler and transparent model that approximates locally the model m. It follows from these requirements that:

$$E_{im} << E_m \quad \text{and} \quad E_{im} \leq E_{lm} \; \forall d_i \tag{4}$$

whereas there is not a clear relation between the expressive power of the induced model and the LID, with the risk of not being able to guarantee λ in d_i if $E_{im_i} < C_{\mathscr{P}_i}(\lambda)$ for some i. Hence *im* not only cannot properly model $\mathscr{P}$ on its full domain, but can even break the λ constraint locally,

and the more complex the original model is, the more likely the latter hypothesis becomes. To avoid loosing tolerance guarantees in the regions where $E_{im_i} < C_{\mathscr{P}_i}(\lambda)$, the positive correlation between the width of the regions and the minimum λ should be exploited, splitting them in one or more subregions where one or more further local models have to be considered. In the worst case a model with a very low expressive power could only approximate an hugely more complex model in regions so small that degenerate into a single point.

Further differences between local models *lm* and induced models *im* are that the former use the original features of the problem $\mathscr{P}$ and the original data in D for training, while the latter do not necessarily use the same features (they need more interpretability) and use as training data the predictions of model m. The induced model tries to model m, not $\mathscr{P}$, mimicking its behavior in an arbitrarily small region.

4.1. Interpretable feature spaces

Given a data matrix $X \in D$ of size $n \times (p+q)$, be $p+q$ the original features, q the subset of the original features that can be considered interpretable and be F a set of r disjoint ($D \cap F = \emptyset$) interpretable features measurable on the instances i. If according to the data analyst the interpretation requires features F, they should be measured on the raw data for each single instance before the analysis begin, so to produce a new data matrix $X \in E$ with domain $E = D \cup F$ of size $n \times (p+q+r)$. In theory features F could also be derived from the original features, but feature extraction techniques hinder precisely the interpretability of the original features, and build new features dependent from the original, so this option will not be considered in the following.

- Case 1: if the required interpretable features are disjoint from the original features, then $q = 0$ and the induced model should be built on the matrix $X \in F$ of size $n \times r$;
- Case 2: if the required interpretable features overlap with the original features, then the induced model should be built on the matrix $X \subset E$ of size $n \times (q+r)$;
- Case 3: if the required interpretable features are a subset of the original features, then the induced model should be built on the matrix $X \subset D$ of size $n \times q$.
- Case 4: if the required interpretable features coincide with the original features, then the induced model should be built on the matrix $X \in D$ of size $n \times (p+q)$.

Only in the last two cases the raw data should not be processed again, and when they are not available, only the last two cases are viable.

In case 1, let x_i be a single instance of the domain D to be explained, $\pi(x_i)$ its neighborhood in D and l_i the prediction that model m assigns to it, given a disjoint set of interpretable features F, f_i be the value that instance x_i has in the interpretable feature space and $\rho(f_i)$ its neighborhood in this feature space. As there is not an explicit mapping between D and F, and much less a continuous linear mapping, closeness is not preserved and π_{x_i} does not correspond to $\rho(f_i)$, in other terms the set of neighbors of instance i in D does not match the set of its neighbors in F. The local behavior of any model around f_i will be unrelated to the behavior of any model around x_i, as the set of neighbors is different.

In case 2, the set of actual neighbors is determined by the dominating components between the q original interpretable features and the r added interpretable features. In case 3, the problem

remains in its original domain but reduced in the interpretable feature space. In case 4, no simplification has been operated.

4.2. Local Induced models

One of the popular model-agnostic explainability models that uses a local induced model *im* is called LIME [4] and has spawned many variations [5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16]. What LIME and its variations produce is a pointwise explanation, that is an explanation that is valid for the prediction of a single instance i from model m.

Called N_i the set of neighbors of instance i in the interpretable feature space and $n_j \in N_i$ its generic neighbor, LIME trains a local model *im* such that the training data are the pairs $(n_j, l_j), n_j \in N_i$, where labels l_j are obtained from m. Labels are necessarily obtained passing to m the original and complete data vector x_j, but four cases should be distinguished:

- Case 1: if the required interpretable features are disjoint from the original features, there is not an unique, easy or efficient way to find the x_j corresponding to the n_j neighbors. This is doable only if the raw data are available and processed beforehand so that a new data matrix $X \subset E$ of size $n \times r$ is built on the same instances i represented in X.
- Case 2: if the required interpretable features overlap with the original features, then again raw data should be available and need preprocessing in order to have a new data matrix $X \subset E$ of size $n \times (p + r)$ is built on the same instances i represented in X
- Case 3: if the required interpretable features are a subset of the original features or coincide with them, then raw data are not required, they need no preprocessing and labels are obtained from the complete x_j.
- Case 4: if the required interpretable features coincide with the original features, then raw data are not required, they need no preprocessing and labels are obtained straightforwardly from the x_j.

In LIME the explanations are binary vectors that represent the presence/absence of interpretable features, the neighbors are found by random perturbation of x_i, and the considered family of models is sparse linear models of limited complexity. In case 3 and 4, LIME reduces to a pointwise feature selection technique, giving as explanations the best scoring input features that determined prediction l_j.

5. Trust and post-hoc local models

An human evaluator would not understand or accept a model that gives divergent explanations for very close instances in D, or different explanations for the same instance x_i; he would be fooled by plausibility instead of faithfulness, and would look for oversimplification even when it is not possible because the problem requires necessarily a complex model. The authors in [4] actually warn the readers about its unsuitability in difficult regions or with very hard problems, but that's exactly the realm of deep learning, where explainability is required the most.

A serious conundrum is the validation of an explanation in absence of a ground truth [17]. As there is not a "true" explanation to be targeted, the validation of the model can only be based

on the coherence of the explanation itself, on its local extension and on other common-sense related properties. A panel of human experts cannot be the solution and and cannot be invoked every time an explanation is needed. As the final purpose of explainability concerns gaining trust, the stability and a certain extended local validity of an induced model (not pointwise) is to be considered the most desirable property of an explanation. Called this *robustness* (there is no agreement on term definitions in the literature), is its lack that seriously hinders the trust of humans.

5.1. Some remarks on drawbacks of local explanations

First of all, the choice of the interpretable features is highly subjective: it depends on the context, the domain, the targeted experts, the purpose of the model and the time available to review the explanation. It is easy to imagine different explanations for the same instance x_i and model m choosing a different set of features F. In this new *Interpretable Feature Selection* (IFS) problem — where features are scored for interpretability instead of importance and where the criterium to measure a feature interpretability is both qualitative and subjective — the neglected risk is that the most appealing and credible features may be preferred to the most faithful and accurate ones. As human judgment is necessarily biased, a model designer will necessarily choose explainable features favoring his points of views and ultimately propagating his biases. Even worse, the risk that misleading explanation can be built by purpose by an attacker following the bias of an human evaluator to stole his trust in the model and pilot his decisions must be seriously taken into account [18].

Second, a point-wise explanation, only valid for a single instance x_i, does not generalize. An human evaluator may well find obnoxious and confusing that two arbitrarily close point in D may end up with completely different explanations in F. This can be due to x_i being actually close to the boundary of a class, or can be due to local variability of the induced model in a tangled area. Even if LIME takes into account locality in D weighting the instances n_j by the closeness of their corresponding x_j to x_i in D, it is $\rho(f_i)$ that is being explored, not $\pi(x_i)$. It must be stressed that the closeness relation is not kept changing the feature space and even if the model is locally faithful in $\rho(f_i)$, it is not a local model at all with respect to D, if the interpretable features are not a subset of the original ones. If they are, then it reduces to a pointwise feature selection. By consequence, uniformly sampling $\rho(f_i)$ not only does not guarantees an uniform coverage of $\pi(x_i)$ in general, but it can even create an imbalanced training set $(n_j, l_j), n \in N$, with all the nefarious consequences on training it implies. The variable density of data among regions can produce some model estimates based on very few actual neighbors and hence highly unreliable. The less dense is the region, the most unreliable is the explanation. The uncertainty of explanations, the inference of causal relationship among explanations and the dependence among interpretable features remain largely unaddressed issues [19, 20].

Third, the l_j come from model m, they are predicted values, and hence they are not 100% sure. This means that in regions where the model is less accurate the explanation for an instance may be plausible but referred to a wrongly labeled instance (explanation for wrongly labeled data); or the explanation may be wrong because the induced model is trained on wrongly labeled data surrounding a correctly labeled instance (wrong explanation); or both (wrong explanation for wrongly labeled data). The hardest is the instance to classify, the most unreliable (and sought

after) is its explanation.

Fourth, stability. Being based on random perturbations or sampling, LIME is not deterministic and the explanation generated running consecutively two or more times the algorithm are different. A minimum amount of variability is to be considered normal, but slightly more is enough to undermines the trust on the explanation by the human evaluator. The authors in [5] addressed this issue, but increasing the risk unreliable explanations in low density neighborhoods (only the original data are sampled). Authors in [16] utilizes a hypothesis testing framework to determine the number of synthetic data that guarantees stability of the resulting explanation and use synthetic data to obtain better explanations. In general the stability issue is the most discussed in the literature [5, 9, 20, 21, 22].

Fifth, spuriousness. The general risk in the construction of an induced model with significantly minor expressive power to figure out the behavior of a more complex model is that the explanations could be completely spurious and have no real connection to the model. Having by chance matching subclasses on the explainable features in training data, for examples all socks are red, and all trousers are blue, the explanation that the sea is a trouser because it is blue may appear perfectly plausible and there is no formal method to discard it. It is like trying to approximate a p-dimensional hyperplane with a bunch of lines in a different space, where the lines are drawn with sparse random sampling on predicted data, and calling this lines "explanations". Please see [23] for an insight.

The side effect of all these issues is loosing trust. Unfortunately the boundary regions are likely to contain the most difficult and interesting instances to explain, the ones where the explanations are needed the most. Their explanations come from induced models trained on the most uncertain and variable data predictions, with the most unstable, hardly generalizable, subjective and unreliable results.

6. Conclusions

Post hoc local models are becoming popular as additional tools to evaluate the results of black-box models and to provide explanations of the predictions they give. Notwithstanding some success in producing plausible explanations, the lack of expressive power and generalization ability of the induced local models, together with the pointwise, uncertain, unstable, unreliable and possibly unfaithful nature of the generated explanations, not neglecting the bias in the choice of the features to be considered as valid, engender serious concerns about the suitability of *post hoc* local models in case of deep or heavily overparametrized complex models. While a step forward would certainly be to use fuzzy variables and trying to learn non-local hierarchical cause-effect relationships, the key to understand at first glance if simplification is really obtainable cannot skip the evaluation of the Intrinsic Dimension of the problem.

References

[1] A. Maratea, A. Ferone, Deep neural networks and explainable machine learning, in: R. Fullér, S. Giove, F. Masulli (Eds.), Fuzzy Logic and Applications, Springer International Publishing, Cham, 2019, pp. 253–256.

[2] P. Linardatos, V. Papastefanopoulos, S. Kotsiantis, Explainable ai: A review of machine learning interpretability methods, Entropy 23 (2021) 18.
[3] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Comput. Surv. 51 (2018). URL: https://doi.org/10.1145/3236009. doi:10.1145/3236009.
[4] M. T. Ribeiro, S. Singh, C. Guestrin, "why should i trust you?": Explaining the predictions of any classifier, 2016. arXiv:1602.04938.
[5] M. R. Zafar, N. Khan, Deterministic local interpretable model-agnostic explanations for stable explainability, Machine Learning and Knowledge Extraction 3 (2021) 525–541. URL: https://www.mdpi.com/2504-4990/3/3/27.
[6] G. Visani, E. Bagli, F. Chesani, Optilime: Optimized LIME explanations for diagnostic computer algorithms, in: S. Conrad, I. Tiddi (Eds.), Proceedings of the CIKM 2020 Workshops co-located with 29th ACM International Conference on Information and Knowledge Management (CIKM 2020), Galway, Ireland, October 19-23, 2020, volume 2699 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: http://ceur-ws.org/Vol-2699/paper03.pdf.
[7] T. Botari, F. Hvilshøj, R. Izbicki, A. C. P. L. F. de Carvalho, Melime: Meaningful local explanation for machine learning models, CoRR abs/2009.05818 (2020). URL: https://arxiv.org/abs/2009.05818. arXiv:2009.05818.
[8] L. Hu, J. Chen, V. N. Nair, A. Sudjianto, Surrogate locally-interpretable models with supervised machine learning algorithms, arXiv preprint arXiv:2007.14528 (2020).
[9] S. Shi, X. Zhang, W. Fan, A modified perturbed sampling method for local interpretable model-agnostic explanation, CoRR abs/2002.07434 (2020). URL: https://arxiv.org/abs/2002.07434. arXiv:2002.07434.
[10] L. Hu, J. J. Chen, V. N. Nair, A. Sudjianto, Locally interpretable models and effects based on supervised partitioning (lime-sup), ArXiv abs/1806.00663 (2018).
[11] N. Gill, M. Kurka, W. Phan, Machine learning interpretability with h2o driverless ai, 2019.
[12] V. Haunschmid, E. Manilow, G. Widmer, audiolime: Listenable explanations using source separation, ArXiv abs/2008.00582 (2020).
[13] J. Rabold, H. Deininger, M. Siebers, U. Schmid, Enriching visual with verbal explanations for relational concepts - combining lime with aleph, in: PKDD/ECML Workshops, 2019.
[14] S. M. Shankaranarayana, D. Runje, Alime: Autoencoder based approach for local interpretability, in: H. Yin, D. Camacho, P. Tino, A. J. Tallón-Ballesteros, R. Menezes, R. Allmendinger (Eds.), Intelligent Data Engineering and Automated Learning – IDEAL 2019, Springer International Publishing, Cham, 2019, pp. 454–463.
[15] S. Mishra, B. L. Sturm, S. Dixon, Local interpretable model-agnostic explanations for music content analysis, in: ISMIR, 2017.
[16] Z. Zhou, G. Hooker, F. Wang, S-lime: Stabilized-lime for model explanation, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, KDD '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 2429–2438. URL: https://doi.org/10.1145/3447548.3467274. doi:10.1145/3447548.3467274.
[17] F. Yang, M. Du, X. Hu, Evaluating explanation without ground truth in interpretable machine learning, arXiv preprint arXiv:1907.06831 (2019).
[18] D. Slack, S. Hilgard, E. Jia, S. Singh, H. Lakkaraju, Fooling lime and shap: Adversarial attacks on post hoc explanation methods, in: Proceedings of the AAAI/ACM Conference

on AI, Ethics, and Society, AIES '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 180–186. URL: https://doi.org/10.1145/3375627.3375830. doi:10.1145/3375627.3375830.

[19] C. Molnar, G. Casalicchio, B. Bischl, Interpretable machine learning–a brief history, state-of-the-art and challenges, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2020, pp. 417–431.

[20] D. Slack, S. Hilgard, S. Singh, H. Lakkaraju, Reliable post hoc explanations: Modeling uncertainty in explainability, 2021. arXiv:2008.05030.

[21] A. H. A. Rahnama, H. Boström, A study of data and label shift in the lime framework, 2019. arXiv:1910.14421.

[22] S. Saito, E. Chua, N. Capel, R. Hu, Improving lime robustness with smarter locality sampling, 2021. arXiv:2006.12302.

[23] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (2019) 206–215. URL: https://doi.org/10.1038/s42256-019-0048-x. doi:10.1038/s42256-019-0048-x.