Master Internship Subject:

Automatic Neighborhood Design for Localized Model-interpretation

Luis Galárraga, IRISA - LACODAM, INRIA, luis.galarraga@inria.fr
Romaric Gaudel, CREST, ENSAI, romaric.gaudel@ensai.fr

November 5, 2018

Location: IRISA, team WIDE, Rennes ; or Ensai, Bruz (near Rennes)

Keywords: Machine Learning, Interpretable Artificial Intelligence

1 Context and Objective

The application of Machine Learning in real life raises more prominently the need for models that are "understandable by a human being". The requirement for a comprehensible model typically arises when it comes to correct errors of the model or to support its predictions on specific examples.

A model is too complex to be readable by a human being because the function learned by this model is itself complex. Thus, it is unreasonable to seek for a simple explanation of the model as a whole. On the other hand, when we look more locally (at the neighborhood of an example), the model is more likely to be simple. This hypothesis is at the root of approaches like LIME [1] and Anchors [2] which respectively look for a local linear approximation of the model, and a neighborhood of the target example on which the value predicted by the model is constant.

Both approaches currently lack a process to derive the appropriate neighborhood. On the one hand LIME lets the user search for himself the concept of neighborhood that suits him best (without any tool defining "better"). On the other hand, Anchor builds a neighborhood in which the function learned is constant, but in the meanwhile it loses information on the conditions that would bring out of this constant area.

As part of the internship we will study alternatives and/or additions to LIME and Anchors to automatically adapt the neighborhood, while maintaining information about attributes that must be modified so that an example is classified in a different category.

2 Planning for Research Activity

We will first focus on binary classification models (for each entry, the model has to choose a label among a set of two labels) on $\mathbb{R}^d$. We will extend the neighborhood selected by Anchors and characterize the border met.

We may also explore new strategies to build the discretization of continuous features at the root of Anchors.

Extensions will look either at multi-class settings, or at relevant ways to extend the neighborhood in the context of categorical features.

3 Practical Information and Skills

The intern will have to manipulate linear classifiers, first order logic rules, and decision trees. The candidates should not be repelled by such objects.

References

[1] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144, New York, NY, USA, 2016. ACM.

[2] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 1527–1535. AAAI Press, 2018.