Deriving Local Internal Logic for Black Box Models

Deriving Local Internal Logic for Black Box Models

Eliana Pastor
Supervised by Elena Baralis

Politecnico di Torino, Italy
{eliana.pastor, elena.baralis}@polito.it

Abstract. Despite the widespread use, machine learning methods produce black box models. It is hard to understand how features influence the model prediction. We propose a novel explanation method that explains the predictions of any classifier by analyzing the prediction change obtained by omitting relevant subsets of attribute values. The local internal logic is captured by learning a local model in the neighborhood of the prediction to explain. The explanations provided by our method are effective in detecting associations among attributes and class label.

Keywords: Interpretability · Prediction Explanation · Local model.

1 Introduction

Machine learning algorithms are widely applied in every aspect of our society. Their growing popularity and their widespread use have made it increasingly important to understand why a classification model take a particular decision. The call for more explainable predictions comes also for institution. The European Union approved the GDPR, a regulation for ensuring personal data protection. It states that individuals have the right to receive "meaningful information about the logic involved" in case of automated decision-making. For some authors, this requirement legally mandates a "right to explanation" [2].

We propose a novel model-agnostic explanation method that explains the predictions made on single instances by any classifier. The explanation highlights the internal logic of the model in a neighborhood of the prediction. It is based on the knowledge of the local behavior of the model, captured by an interpretable local model.

2 Related Work

Many algorithms have been proposed for improving the interpretability of already existing classification models. Model-dependent solutions are proposed for

SEBD 2018, June 24-27, 2018, Castellaneta Marina, Italy.

handling only specific models. Model-agnostic solutions instead treat the machine learning model as a black box. Our research is focused on these approaches for their general applicability and the advantages derived from it. These methods in fact are applicable to any classification methods without making any assumption on their internal logic. Thus, the comparison among different techniques in terms of model interpretability is possible.

While some approaches try to explain the original model globally, others propose a general method for explaining individual predictions, i.e. why particular decisions are made. Ribeiro et al. [4] introduce a model-agnostic method for explaining individual prediction by learning an interpretable and linear model in the locality of the prediction to be explained. However, the linear approximation may not be faithful if the model is highly non-linear even in the locality of the prediction [4]. The locality is captured by randomly perturbed samples around the instance. Thus, non-existing configurations of attribute values can also be generated. Hence, the local model cannot be considered fully trustworthy.

Several works study how a prediction changes if parts of the input components are omitted. Lemaire et al. [3] and Robnik-Šikonja and Kononenko [5] consider how each attribute value is relevant for the prediction for tabular data, by omitting one attribute value at a time. Štrumbelj et al. study also the omission of more attribute values together, thus also addressing the attribute interaction [6]. The information of how attributes interact with the others is summarized in one single contribution for each attribute value. Hence, the information of interaction relevance is lost. Moreover, they compute the omission effect for the power set of the attributes. Hence, the method is affected by an exponential time complexity. We propose a novel solution that highlight not only the influence of each attribute value for a particular prediction but also of relevant attribute interactions. Moreover, we overcome the problem of exponential time complexity exploiting local properties of the original model to be explained.

3 Method

We propose a novel method applicable to explain individual predictions of any classification method. Given the particular prediction that we want to explain, we omit one or more attribute values at a time and we measure how the prediction changes. The relevance of the change is estimated as a difference of prediction probabilities with respect to a particular target class. The greater is this difference, the more the omitted attribute values are relevant for the prediction.

With respect to existing approaches, we are interested in understanding not only how each single attribute value is significant for the prediction but also how it interacts with the others. An attribute value can determine the prediction alone or only if it is in conjunction with others. In the latter case, we need to omit more attributes at the time for observing how the prediction changes. Omitting sets of attribute values allows also to deal with the disjunction case. The disjunction condition occurs when more than one configurations of attributes values determines the prediction. We can observe a change of the prediction probability only if we considered the omission of the attribute values together.

The feature values that in conjunction or disjunction are relevant for a prediction are highlighted by a local interpretable model. The local model is an associative classifier learned in the neighborhood of the prediction that we want to explain. The local rules, being understandable, provide preliminary insights of why a decision is made by the considered model. Each rule is in AND form. Thus, it gives the information of what attributes together determine the prediction. Moreover, a prediction may be determined by more rules. We can deal with the disjunction condition considering jointly the omission of the subsets highlighted by the local rules. Only the relevant attribute subsets provided by the local model are considered, instead of the complete power set of all attribute combinations. Hence, our approach overcomes the exponential time complexity.

4 Preliminary Results

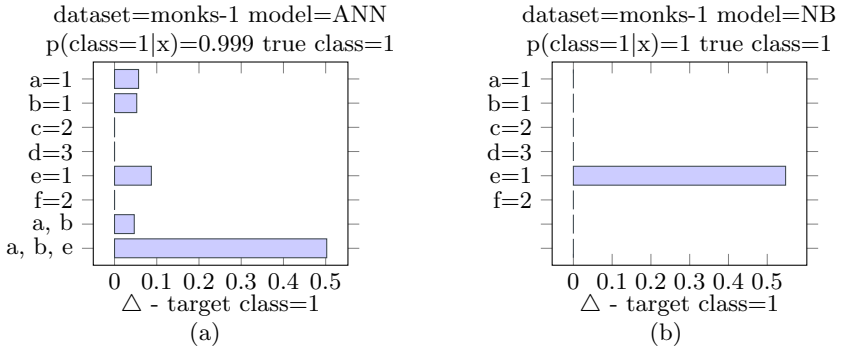

Fig. 1: Explanation of (a) the neural network and (b) the Naive Bayes prediction of a particular instance of the *monks-1* data set.

In this section, preliminary outcomes of our novel explanation method are presented. The *Monk1* data set is an artificial data set composed by 6 discrete attributes *a,b,c,d,e,f* and the class label can take value 1 or 0 [1]. Being artificial, the relation between the attributes and the class value is known. The class is 1 if *a=b* or if *e=1*, 0 otherwise. We train a MLP ANN using the *Monk1* data set. Let x = (a=1, b=1, c=2, d=3, e=1, f=2) be the instance that we want to explain. We know that the "true class" is 1 because *e=1* and *a=b*. The ANN correctly predicts the class label as 1. To estimate the relevant subsets of feature values, we train the associative classifier in the locality of instance x. The local model returns the following association rules: $\{e = 1\} \rightarrow class = 1,\ \{a = 1,\ b = 1\} \rightarrow class = 1.$ Hence, if *e=1* the instance is assigned to the class 1 or if a and b are both equal to 1. These relations should indeed determine the class. Thus, the local behavior captures the true explanation. Once that the relevant subsets are determined, the prediction differences are computed. The estimation is made for each attribute value, for the relevant subset {*a=1,b=1*} and for the OR of the relevant rules, thus for {*a=1,b=1,e=1*}. The results are shown in Figure 1a. The terms *e=1*, *a=1* and *b=1* have alone a positive importance in the determination of the class. It is interesting to notice that {*a=1,b=1*} together have not a great prediction

difference but it is comparable to the ones when they are considered alone. If a and b are removed together, the class label does not change. The prediction is in fact still 1 because of *e=1*. Same considerations can be made for *e=1*. Removing the rules in OR, the prediction probability drastically changes. Only if considered together, we can observe and quantify how these attributes interact.

If we explain the prediction for the same instance, made by another model, we may obtain a different result. The explanation should capture how the model behaves in the locality of the instance. Different models work differently. This difference may be on the predicted class label, but also on the feature values that drive the prediction. Consider the explanation of the same instance x and still built with respect to class 1, but classified by the Naive Bayes classifier (NB). The local model returns a single relevant rule: $\{e = 1\} \rightarrow class = 1$.
The results are shown in Figure 1b. The NB classifier assigns correctly the instance x to class 1, but only because *e=1*. The local model and the explanation highlight that the Naive Bayes classifier has not learned the association that if *a=b* then *class=1*. Because of its assumption of independence between features, it is not able to learn the importance that a and b have together. Hence, the local model and the explanation in this case successfully reflect the model behavior.

5 Conclusions and Future Work

Preliminary tests show that our technique is able to capture the diverse internal logic of classification techniques. Differently than existing approaches, the importance of relevant subsets of feature values to the prediction is computed. Thus, our method provides to end users the information of what attributes together determine the prediction and the quantification of their influence.

As future work we plan to (*i*) formalize our approach proposing formal definitions of the prediction change estimation, (*ii*) evaluate the effect of the neighborhood in the local rules and the resulting explanations and (iii) apply the proposed method to real-world data sets, validating explanations through the assistance of domain experts.

Acknowledgements This work is partially funded by SmartData@PoliTO.

References

1. Dheeru, D., Karra Taniskidou, E.: UCI machine learning repository (2017)
2. Goodman, B., Flaxman, S.: European union regulations on algorithmic decision-making and a "right to explanation". arXiv preprint arXiv:1606.08813 (2016)
3. Lemaire, V., Feraud, R., Voisine, N.: Contact personalization using a score understanding method. In: 2008 IEEE Int. Joint Conf. on Neural Networks. pp. 649–654
4. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: Proc. of the 22Nd ACM SIGKDD Int. Conf. on KDD. pp. 1135–1144. KDD '16, ACM, New York, NY, USA (2016)
5. Robnik-Šikonja, M., Kononenko, I.: Explaining classifications for individual instances. IEEE Trans. Knowl. Data Eng. **20**(5), 589–600 (2008)
6. Štrumbelj, E., Kononenko, I., Robnik Šikonja, M.: Explaining instance classifications with interactions of subsets of feature values. DKE **68**(10) (2009)