Pilot Study
—
Inside the Black Box: How to Explain Individual Predictions of a Machine Learning Model

Marc Beillevaire
marcbei@kth.se

February 20, 2017

Abstract

This is the pilot study for my master thesis for KTH Institute of technology, conducted at Dataiku, a datascience company in Paris. This documents aims at describing work related to the topic of this thesis: how to explain individual outputs of predictive models based on machine learning algorithms. This topic has already been studied by several research teams using many different approaches.

1 Background and objectives

Today, most of machine learning algorithms look like black boxes from the user perspective: it is not easy - if not impossible - to understand how a given prediction is made by, let's say, a Random Forest with several hundreds of trees. Some tools already exist to understand models but they are mostly global insights, like the variable importance in a decision tree.

Yet in many cases having a local insight, related to a particular prediction, would be really helpful to users. An insurance company would like to know why a particular client has been detected as a fraudster, or a manufacturer doing predictive maintenance would like to know not only which device is likely to break soon but also how this device's lifetime could be increased practically.

There are several articles trying to formalize in a high level fashion what explanations are or should be – [6] and [10]. Another one focuses on the General Data Protection Regulation recently approved by the EU parliament – [8]

These articles does not provide technical solutions to the problem of finding good algorithms to extract explanations, but are essential to understand the concepts and what is at stake in interpretable explanations.

2 Related work

As this problem is about explaining models, there are two main solutions to it:

- Model dependant explanations
- Model agnostic explanations

2.1 Model dependant explanations

The first one implies understand how a model is build, making use of its particular type, and extract explanations from it. This is usually faster to compute, but explanations of this kind are not generalizable to other types of models.

An explaining algorithm for tree-based methods – individual decision trees, random forests and gradient boosted trees – has been developed by Saba in [13]. When examining the prediction for a given point, his algorithm takes the same path in the tree as the original point. At each node, the tree splits part of the feature space into two smaller and disjoint subspaces. It is then possible to compute the difference of proportion of the predicted class in the bigger space and the smaller space where the point ended. This difference means that at a given node, the criterion changed the predicted class probability by the value of the difference computed before. The criterion is based on a single feature, therefore this feature's influence is the difference of proportion in the bigger space and the smaller space. To get the influence of all features, Saba's algorithm considers all the split in the prediction's path in the tree.

This way of exploring trees through the path of the original datapoint makes this algorithm efficient: its complexity is proportional to the height of the tree (as well as linear in the number of trees if the algorithm is an ensemble).

There are also specific methods for linear models in a quite old book [1] by Achen. Linear models are very used in areas where only causal relationships are desired, such as economics. The linearity makes the explanation very straightforward as decomposing the influence of each feature is very easy.

2.2 Model independent explanations

Although the previous methods are computationally efficient, they still rely on specific models. A model-independent algorithm would be preferable for at least two reasons: its ability to explain any type of classifiers and regressors, and the possibility to compare two explanations from two different model even from different types.

Several algorithms has been developed, most notably: [12], [2], [9], [5], [14], [7], [3].

Some of them tries to estimate the local output distribution from the model like: *Lime* [12] and *explanation vectors* [2]. If the model is a classifier that outputs probabilities, then the output distribution is a probability distribution in the feature space. *Lime* algorithm by Ribeiro, Singh, and Guestrin [12] aims at modeling this local distribution by a simpler model, namely a linear model with few variables that is therefore easy to interpret. Baehrens, Schroeter, and Harmeling's algorithm in [2] mimic the local distribution using Gaussian kernel estimation, and then computes the local gradient of this estimated distribution.

Thus, both these methods outputs what happens when the datapoint shifts a bit in the feature space, which is very convenient when someone wants to know how to optimize a model output by changing the variables.

Other articles focus on changing the value of a given attribute in the input vector and measure how the model behave when this attribute is modified. This is how explanations are computed in "Explaining Classifications For Individual Instances" [9]: for each attribute A_i they compute a probability difference:

$$\text{predDiff}_i(x) = f(x) - f(x \setminus A_i)$$

as well as the information difference:

$$\text{infDiff}_i(x) = logp(y|x) - logp(y|x \setminus A_i)$$

This methods is somewhat similar to partial dependence plots, and to another article [7] where the authors develop a similar visualization tool called Individual Conditional Expectation plots. The principle is similar to computing a partial dependence plot, but instead of averaging it over all the examples, the authors decide to keep one curve for each instance. If the estimator is the function $f : (x_1, x_2, ..., x_n) \mapsto f(x_1, x_2, ...x_n)$, then for each example i and feature j, a curve is drawn with the following shape:

$$x \mapsto f(x_1, ..., x_{j-1}, x, x_{j+1}, ..., x_n)$$

This lets the user having far more insights on the model behaviour on individual instances, without being longer to compute than a partial dependence plot.

Finally, another interesting approach in extracting explanation uses game theory and the *Shapley value* [14]. The authors consider explanations as a set of feature influence, where each influence is the Shapley value of the feature. The Shapley value could be described as the influence of a feature regarding potential coalitions formed with other features. For example in the Council of the European Union, each country get a number of votes proportional to its population, and bigger countries have a larger power to make strong coalitions, while smaller ones have less. The Shapley value reflects this power. In the case of a machine learning model: features that are able to form strong "coalitions" to change the output of the model gain a large Shapley value, and should explain more importantly the outcome.

2.3 Other articles

Some other articles proposes solutions for specific applications such as:

- Textual data: “Explaining data-driven document classifications” [11]
- Health care data: “Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission” [4], but this one is more about how to build intelligible models than understand complicated ones.
- Breast cancer data: “Explanation and reliability of prediction models: the case of breast cancer recurrence”, by authors of [14], [3] and [9].

References

[1] Christopher H. Achen. *Interpreting and using regression.* SAGE, Oct. 1982.

[2] D. Baehrens, T. Schroeter, and S. Harmeling. “How to Explain Individual Classification Decisions”. In: *Journal of Machine Learning Research* 11 (2010), pp. 1803–1831.

[3] Z. Bosnić and I. Kononenko. “Estimation of individual prediction reliability using the local sensitivity analysis”. In: *Journal of Applied Intelligence* 29.3 (Aug. 2007), pp. 187–203. DOI: 10.1007/s10489-007-0084-9.

[4] R. Caruana et al. “Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission”. In: KDD ’15. 2015, pp. 1721–1730.

[5] I. De Falco et al. “An evolutionary approach for automatically extracting intelligible classification rules”. In: *Journal of Knowledge and Information Systems* 7.2 (Feb. 2015), pp. 179–201. DOI: 10.1007/s10115-003-0143-4.

[6] Mary T. Dzindoleta et al. “The role of trust in automation reliance”. In: *International Journal of Human-Computer Studies* 58 (2003), pp. 697–718.

[7] Alex Goldstein et al. “Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation”. In: *Journal of Computational and Graphical Statistics* 24.1 (2015), pp. 44–65. DOI: 10.1080/10618600.2014.907095.

[8] B. Goodman and S. Flaxman. “European Union regulations on algorithmic decision-making and a “right to explanation””. 2016. URL: https://arxiv.org/pdf/1606.08813v3.pdf.

[9] I. Konenko and Robnik-Šikonja. “Explaining Classifications For Individual Instances”. In: *IEEE Transactions on Knowledge & Data Engineering* 20 (May 2008), pp. 589–600.

[10] Zachary Chase Lipton. “The Mythos of Model Interpretability”. In: *CoRR* abs/1606.03490 (2016). URL: http://arxiv.org/abs/1606.03490.

[11] D. Martens and F. Provost. “Explaining data-driven document classifications”. In: *MIS Quarterly* 38 (Mar. 2014), pp. 73–100.

[12] M. T. Ribeiro, S. Singh, and C. Guestrin. “Why Should I Trust You? Explaining the Predictions of Any Classifier”. In: KDD ’16. 2016. URL: http://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf.

[13] A. Saba. “Interpreting random forests”. In: *Diving into data* (2015). URL: `http://blog.datadive.net/interpreting-random-forests/`.

[14] E. Strumbelj and I Kononenko. “An Efficient Explanation of Individual Classifications using Game Theory”. In: *Journal of Machine Learning Research* 11 (2010), pp. 1–18.

[15] E. Štrumbelj et al. “Explanation and reliability of prediction models: the case of breast cancer recurrence”. In: *Journal of Knowledge and Information Systems* 7.2 (Aug. 2010), pp. 305–324.