AN INFORMATION-THEORETIC APPROACH TO EXPLAINABLE MACHINE LEARNING

Alexander Jung

Department of Computer Science, Aalto University, Finland; firstname.lastname(at)aalto.fi

ABSTRACT

A key obstacle to the successful deployment of machine learning (ML) methods to important application domains is the (lack of) explainability of predictions. Explainable ML is challenging since explanations must be tailored (personalized) to individual users with varying backgrounds. On one extreme, users can have received graduate level education in machine learning while on the other extreme, users might have no formal education in linear algebra. Linear regression with few features might be perfectly interpretable for the first group but must be considered a black-box for the latter. Using a simple probabilistic model for the predictions and user knowledge, we formalize explainable ML using information theory. Providing an explanation is then considered as the task of reducing the "surprise" incurred by a prediction. Moreover, the effect of an explanation is measured by the conditional mutual information between the explanation and prediction, given the user background.

1. INTRODUCTION

Machine learning (ML) methods allow to obtain predictions for certain quantities of interest based on statistical analysis of large amounts of historic data [2, 9, 12]. These methods are routinely used to power many services within our every day-life. A main application of ML methods is in recommender systems decide which job ads or which other user profiles could be interesting to us [13, 23]. Recent breakthroughs in ML, such as in image or text processing [7], holds the promise of boosting the level of automation in important industries [6, 19].

A key challenge for the successful and ethically sound deployment of ML methods to critical application domains is the (lack of) explainability of its predictions [8, 11, 15, 22]. One reason why explainable ML is difficult to obtain is that (good) explanations must be tailored to the knowledge of individual users ("explainee"). Thus, achieving explainable ML is easier for applications involving a homogenous group of users such as graduate students in a university program.

Large-scale applications, such as recommendation systems for video streaming providers typically involve users with very different backgrounds. Indeed, the user background can range from graduate studies in ML-related fields to users with now formal training in linear algebra. While linear models involving few hand-crafted features might be considered interpretable for the former group it might be considered a "black-box" for the latter group of users.

We study explainable AI within information theory by using a probabilistic model for the data and user background. Loosely speaking, we model the effect of providing an explanation for a prediction as a reduction of the "surprise" incurred by a prediction to the user. This qualitative interpretation of explaining a prediction leads naturally to measuring the quantitative effect of explanations via (conditional) mutual information between the explanation and the prediction, given the user background (see Section 2).

Our approach is different to existing work on explainable ML in the sense that we explicitly model the user knowledge. Existing methods for explainable ML can be roughly divided into two groups: (i) methods based on elementary models which are considered as intrinsically interpretable (ii) model-agnostic methods that probe the ML method as a black box.

The first category of explainable ML methods are obtained from models which are considered intrinsically interpretable. Such methods include linear models, decision trees and artificial neural networks [1, 8, 17]. Explaining the predictions obtained from such intrinsically interpretable models merely amounts to specifying the model parameters, such as the weights w_i of a linear predictor $h(\mathbf{x}) = \sum_i w_i x_i$, or the feature-wise thresholds used in decision trees [9].

Interpretable models allow to decompose its predictions into a combination of elementary properties of a data point. Defining elementary properties of a data point via the activations of a (deep) neural network renders those models also interpretable (see [17]).

Explainable models for sequential decision making have been studied in [14]. The authors of [14] obtain an explainable multi-armed bandit model by using the choice for the action space as the explanation (e.g., recommend only items that have already been purchased by the user). In contrast to [14], our approach uses a probabilistic model for the user background which is leveraged to compute explanations that are optimal in a precise (information-theoretic) sense.

A second category of explainable ML, referred to as model agnostic methods, is based on constructing explanations by probing a predictor as a black box [8, 21]. These methods aim at locally approximating black box models by simpler and interpretable models, such as linear models or shallow decision trees [21].

Our approach falls into this second category since it is also model agnostic. However, in contrast to existing approaches to model agnostic explainable ML, we do not use local approximations to explain a black box method. Instead, we use a probabilistic model for the model predictions and user knowledge.

Framing explainable ML within a probabilistic model allows to capture the act of explaining a prediction in an information-theoretic sense. Providing an explanation is then understood as providing the user additional information about the prediction delivered by the model.

Outline and Contribution. In Section 2, we propose a simple probabilistic model for the features, prediction and user summary of a data point. This probabilistic model allows to quantify the effect of explanations via the conditional mutual information between the explanation and the model prediction, given the user background. Section 3 then discuss how to maximize this conditional mutual information to obtain (information-theoretically) optimal explanations. A simple algorithm for computing optimal explanation given the model predictions and user summaries for i.i.d. samples is then presented in Section 4. The proposed algorithm allows to construct personalized explanations that are optimal in an information-theoretic sense.

2. PROBLEM SETUP

We consider a supervised ML problem involving data points with features $\mathbf{x} = \big(x_1, \ldots, x_n\big)^T \in \mathbb{R}^n$ and label $y \in \mathbb{R}$. Given some labelled training data

$$\big(\mathbf{x}^{(1)}, y^{(1)}\big), \big(\mathbf{x}^{(2)}, y^{(2)}\big), \ldots, \big(\mathbf{x}^{(m)}, y^{(m)}\big), \tag{1}$$

ML methods typically learn a predictor (map)

$$h(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R} : \mathbf{x} \mapsto \hat{y} = h(\mathbf{x}) \tag{2}$$

by requiring $\hat{y}^{(i)} \approx y^{(i)}$ [2,9,12].

After learning a predictor $\hat{y} = h(\mathbf{x})$, it is applied to new data points yielding the prediction $\hat{y} = h(\mathbf{x})$. In may application, the prediction $\hat{y}$ is then delivered to a human user. The user can be the subscriber of a streaming service [5], a dermatologist [4] or a city planner [24].

Each user has typically some conception or model for the relation between features $\mathbf{x}$ and label y of a data point. Based on the user background, she has some understanding of a data point with features $\mathbf{x}$.

Our approach to explainable ML is based on modelling the user understanding of a data point by some summary $u \in \mathbb{R}$. The summary is obtained by a stochastic) map from the features $\mathbf{x}$ of a data point. We will focus on summaries being obtained by a deterministic map

$$u(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R} : \mathbf{x} \mapsto u := u(\mathbf{x}). \tag{3}$$

However, our approach covers also stochastic maps characterized by a conditional probability distribution $p(u|\mathbf{x})$.

The (user-specific) quantity u represents the understanding of the specific properties of the data point given the user knowledge (modelling assumptions). We interpret u as a "summary" of the data point based on its features $\mathbf{x}$ and the intrinsic modelling assumptions of the user.

We formalize the act of explaining a prediction $\hat{y} = h(\mathbf{x})$ as presenting some additional quantity e to the user. This "explanation" e can be any quantity that helps the user to understand the prediction $\hat{y}$, given her understanding u of the data point.

For the sake of exposition, our focus will be on explanations obtained via a deterministic map

$$e(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R} : \mathbf{x} \mapsto e := e(\mathbf{x}), \tag{4}$$

from the features $\mathbf{x}$ of a data point. However, our approach can be generalized without difficulty to handle explanations obtained by a stochastic map. In the end, we only require the specification of the conditional probability distribution $p(e|\mathbf{x})$.

Explanations e can take very different forms. An explanation e could be a subset of features $\{x_i\}_{\mathcal{E} \subseteq \{1,\ldots,n\}}$ (see [20] and Section 3). More generally, explanations could be obtained from simple local statistics (averages) of features that are considered related, such as near-by pixels in an image. Alternatively, instead of (local statistics of) individual features, we could also use other data points as an explanation [14, 21].

To obtain explanations that are comprehensible and can be computed efficiently, we must typically restrict the space of possible explanations to a small subset of maps (10).[1] We denote by $\mathcal{F}$ the subset of maps (10) resulting in useful explanations.

We consider data points as independent and identically distributed (i.i.d.) realizations of a random variable with fixed underlying probability distribution $p(\mathbf{x}, y)$. Modelling the data point as random implies that the user summary u, prediction $\hat{y}$ and explanation e are also random variables. The joint distribution $p(u, \hat{y}, e, \mathbf{x}, y)$ conforms with the Bayesian network [18] depicted in Figure 2 since

$$p(u, \hat{y}, e, \mathbf{x}, y) = p(u|\mathbf{x}) \cdot p(e|\mathbf{x}) \cdot p(\hat{y}|\mathbf{x}) \cdot p(\mathbf{x}, y). \tag{5}$$

We measure the amount of additional information provided by an explanation e for a prediction $\hat{y}$ to some user u via the conditional mutual information [3, Ch. 2,8]

$$I(\hat{y}, e|u) := \mathrm{E}\bigg\{ \log \frac{p(\hat{y}, e|u)}{p(\hat{y}|u)p(e|u)} \bigg\}. \tag{6}$$

[1] This is conceptually similar to the restriction of the space of possible predictor functions in a ML method to a small subset of maps which is known as the hypothesis space.

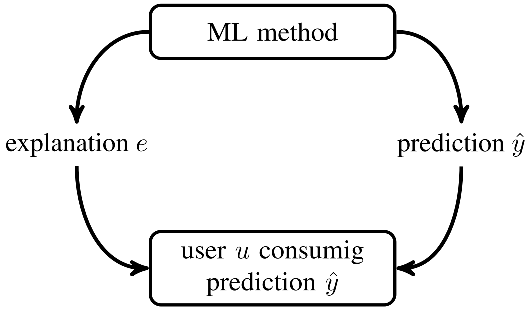

Fig. 1. An explanation e provides additional information $I(\hat{y}, e|u)$ to a user u about the prediction $\hat{y}$.

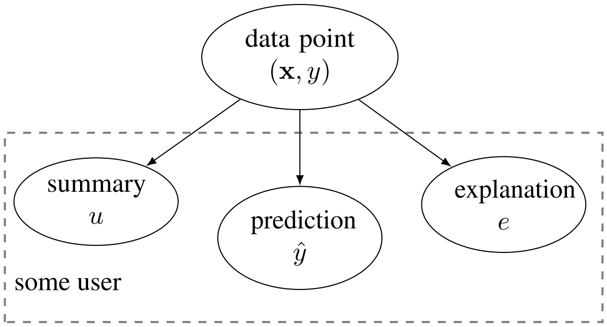

Fig. 2. A simple probabilistic model for explainable ML.

3. OPTIMAL EXPLANATIONS

Capturing the effect of an explanation using the probabilistic model (6) offers a principled approach to computing an optimal explanation e. We require an optimal explanation $e*$ to maximize the conditional mutual information (6) between the explanation e and the prediction $\hat{y}$ conditioned on the user summary u of the data point.

Formally, an optimal explanation e^* solves

$$I(e^*, \hat{y}|u) = \sup_{e \in \mathcal{F}} I(e, \hat{y}|u). \tag{7}$$

The choice for the subset $\mathcal{F}$ of valid explanations offers a trade off between complexity of explanations, additional information provided about the prediction and the computational cost of solving (7).

Let us illustrate the concept of optimal explanations (7) using a linear regression method. We model the features $\mathbf{x}$ as a realization of a multivariate normal random vector with zero mean and covariance matrix $\mathbf{C}_x$,

$$\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_x). \tag{8}$$

The predictor and the user summary are linear functions of the features,

$$\hat{y} := \mathbf{w}^T \mathbf{x}, \text{ and } u := \mathbf{v}^T \mathbf{x}. \tag{9}$$

A sensible construction of explanations is via subsets of individual features x_i that are considered most relevant for a user to understand the prediction $\hat{y}$ (see [17, Definition 2] and [16]).

In what follows, we restrict our attention to explanations of the form

$$e := \{x_i\}_{i \in \mathcal{E}} \text{ with some subset } \mathcal{E} \subseteq \{1, \ldots, n\}. \tag{10}$$

The complexity of an explanation e is measured by the number $|\mathcal{E}|$ of features that contribute to e. To keep explanations simple, we limit the number of features contributing to an explanation by a fixed (small) sparsity level,

$$|\mathcal{E}| \leq s (\ll n). \tag{11}$$

Modelling the feature vector $\mathbf{x}$ as Gaussian (8) implies that the prediction $\hat{y}$ and user summary u obtained from (9) is jointly Gaussian when conditioned on the explanation e (10). We can therefore rewrite problem (7) as

$$\begin{aligned}
&\sup_{\substack{\mathcal{E} \subseteq \{1,\ldots,n\} \\ |\mathcal{E}| \leq s}} I(e, \hat{y}|u) \\
&\quad = h(\hat{y}|u) - h(\hat{y}|e, u) \\
&\quad = (1/2) \log \mathbf{C}_{\hat{y}|u} - (1/2) \log \det \mathbf{C}_{\hat{y}|u,e} \\
&\quad = (1/2) \log \sigma^2_{\hat{y}|u} - (1/2) \log \sigma^2_{\hat{y}|u,e}.
\end{aligned} \tag{12}$$

Here, we used elementary properties of multivariate normal distributions [3, Ch. 8]. The last step in (12) follows from the fact that $\hat{y}$ is a scalar random variable.

The first component of the last expression in (12) does not depend on the choice $\mathcal{E}$ for the explanation e (see (10)). Therefore, the optimal choice $\mathcal{E}$ solves

$$\sup_{|\mathcal{E}| \leq s} -(1/2) \log \sigma^2_{\hat{y}|u,e}. \tag{13}$$

The maximization (13) is equivalent to

$$\inf_{|\mathcal{E}| \leq s} \sigma^2_{\hat{y}|u,e}. \tag{14}$$

In order to solve (14), we relate the conditional variance $\sigma^2_{\hat{y}|u,e}$ to a particular decomposition

$$\hat{y} = \alpha u + \sum_{i \in \mathcal{E}} \beta_i x_i + \varepsilon. \tag{15}$$

For an optimal choice of the coefficients α and β_i, the variance of the error term in (15) is given by $\sigma^2_{\hat{y}|u,e}$. Indeed,

$$\min_{\alpha, \beta_i \in \mathbb{R}} \mathrm{E}\big\{ \big(\hat{y} - \alpha u - \sum_{i \in \mathcal{E}} \beta_i x_i \big)^2 \big\} = \sigma^2_{\hat{y}|u,e}. \tag{16}$$

Inserting (16) into (14), an optimal choice $\mathcal{E}$ (of feature) for the explanation of prediction $\hat{y}$ to user u is obtained from

$$\inf_{|\mathcal{E}| \leq s} \min_{\alpha, \beta_i \in \mathbb{R}} \mathrm{E}\big\{ \big(\hat{y} - \alpha u - \sum_{i \in \mathcal{E}} \beta_i x_i \big)^2 \big\} \tag{17}$$

$$= \min_{\|\boldsymbol{\beta}\|_0 \leq s} \mathrm{E}\big\{ \big(\hat{y} - \alpha u - \boldsymbol{\beta}^T \mathbf{x} \big)^2 \big\}. \tag{18}$$

An optimal subset $\mathcal{E}_{\rm opt}$ of features defining the explanation e (10) is obtained from any solution $\boldsymbol{\beta}_{\rm opt}$ of (4) via

$$\mathcal{E}_{\rm opt} = \mathrm{supp}\, \boldsymbol{\beta}_{\rm opt}. \tag{19}$$

4. A SIMPLE XML ALGORITHM

Under a Gaussian model (8), Section 3 shows how to construct optimal explanations via the (support of the) solutions $\boldsymbol{\beta}_{\rm opt}$ of the sparse linear regression problem (4).

In order to obtain a practical algorithm for computing (approximately) optimal explanations (19), we need to approximate the expectation in with a an empirical average over i.i.d. samples $\big(\mathbf{x}^{(i)}, \hat{y}^{(i)}, u^{(i)}\big)$ of features, predictions and user summaries. This results in Algorithm 1.

Note that Algorithm 1 offer interaction with the user which has to provide samples $u^{(i)}$ of its summary for the data points with features $\mathbf{x}^{(i)}$.

```
Algorithm 1 XML Algorithm
Input: explanation sparsity $s$, i.i.d. samples $\big(\mathbf{x}^{(i)}, \hat{y}^{(i)}, u^{(i)}\big)$
    for $i = 1, \ldots, m$
compute $\widehat{\boldsymbol{\beta}}$ by solving
    $\widehat{\boldsymbol{\beta}} \in \arg\min_{\|\boldsymbol{\beta}\|_0 \leq s} \sum_{i=1}^{m} \big(\hat{y}^{(i)} - \alpha u^{(i)} - \boldsymbol{\beta}^T \mathbf{x}^{(i)}\big)^2 \qquad (20)$
Output: feature set $\widehat{\mathcal{E}} := \mathrm{supp}\widehat{\boldsymbol{\beta}}$
```

The sparse regression problem (20) becomes intractable for large feature length n. However, if the features are correlated weakly with each other and the user summary u, the solutions of (20) can be found by convex optimization. Indeed, for a wide range of settings, sparse regression (20) can be solved via a convex relaxation, known as the least absolute shrinkage and selection operator (Lasso) [10],

$$\widehat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^n} \sum_{i=1}^{m} \big(\hat{y}^{(i)} - \alpha u^{(i)} - \boldsymbol{\beta}^T \mathbf{x}^{(i)}\big)^2 + \lambda \|\boldsymbol{\beta}\|_1. \tag{21}$$

We have already a good understanding of choosing the Lasso parameter λ in (21) such that solutions of (21) coincide with solutions of (20) (see, e.g., [10]).

5. REFERENCES

[1] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):1–46, 07 2015.

[2] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[3] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, New Jersey, 2 edition, 2006.

[4] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542, 2017.

[5] C.A. Gomez-Uribe and N. Hunt. The netflix recommender system: Algorithms, business value, and innovation. *Association for Computing Machinery*, 6(4), January 2016.

[6] N.J. Goodall. Can you program ethics into a self-driving car? *IEEE Spectrum*, 53(6):28–58, June 2016.

[7] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[8] H. Hagras. Toward human-understandable, explainable ai. *Computer*, 51(9):28–36, Sep. 2018.

[9] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, NY, USA, 2001.

[10] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity. The Lasso and its Generalizations*. CRC Press, 2015.

[11] A. Holzinger. Explainable AI (ex-AI). *Informatik Spektrum*, 41:138–143, April 2018.

[12] A. Jung. Components of machine learning: Binding bits and flops. *arXiv preprint https://arxiv.org/pdf/1910.12387.pdf*, 2019.

[13] A. B. B. Martinez, J. J. P. Arias, A. F. Vilas, J. Garcia Duque, and M. Lopez Nores. What's on tv tonight? an efficient and effective personalized recommender system of tv programs. *IEEE Transactions on Consumer Electronics*, 55(1):286–294, 2009.

[14] J. McInerney, B. Lacker, S. Hansen, K. Higley, H. Bouchard, A. Gruson, and R. Mehrotra. Explore, exploit, and explain: personalizing explainable recommendations with bandits. In *Proceedings of the 12th ACM Conference on Recommender Systems*, 2018.

[15] B.D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi. The ethics of algorithms: Mapping the debate. *Big Data and Society*, 3(2), 2016.

[16] C. Molnar. *Interpretable Machine Learning - A Guide for Making Black Box Models Explainable*. [online] Available: https://christophm.github.io/interpretable-ml-book/., 2019.

[17] G. Montavon, W. Samek, and K. Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, 2018.

[18] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 1988.

[19] M. Di Renzo, M. Debbah, D.-T. Phan-Huy, A. Zappone, M.-S. Alouini, C. Yuen, V. Sciancalepore, G.C. Alexandropoulos, J. Hoydis, H. Gacanin, J. De Rosny, A. Bounceur, G. Lerosey, and M. Fink. Smart radio environments empowered by reconfigurable ai meta-surfaces: an idea whose time has come. *EURASIP Journal on Wireless Communications and Networking*, (1):1–20, 2019.

[20] M.T. Ribeiro, S. Singh, and C. Guestrin. "Why should i trust you?": Explaining the predictions of any classifier. In *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pages 1135–1144, Aug. 2016.

[21] M.T. Ribeiro, S. Singh, and C. Guestrin. Anchors: High-precision model-agnostic explanations. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[22] S. Wachter, B. Mittelstadt, and L. Floridi. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, 7(2):76–99, 2017.

[23] R. Wang, C. Chow, Y. Lyu, V.C.S. Lee, S. Kwong, Y. Li, and J. Zeng. Taxirec: Recommending road clusters to taxi drivers using ranking-based extreme learning machines. *IEEE Transactions on Knowledge and Data Engineering*, 30(3):585–598, 2018.

[24] X. Yang and Q. Wang. Crowd hybrid model for pedestrian dynamic prediction in a corridor. *IEEE Access*, 7, 2019.