arXiv:1910.09358v1 [cs.LG] 21 Oct 2019

Making Bayesian Predictive Models Interpretable: A Decision Theoretic Approach

Homayun Afrabandpey[1], Tomi Peltola[1], Juho Piironen[2], Aki Vehtari[1], and Samuel Kaski[1]

[1]Helsinki Institute for Information Technology HIIT,
Department of Computer Science, Aalto University
[2]Curious AI
[1]firstname.lastname@aalto.fi, [2]juho@cai.fi

Abstract

A salient approach to interpretable machine learning is to restrict modeling to simple and hence understandable models. In the Bayesian framework, this can be pursued by restricting the model structure and prior to favor interpretable models. Fundamentally, however, interpretability is about users' preferences, not the data generation mechanism: it is more natural to formulate interpretability as a utility function. In this work, we propose an interpretability utility, which explicates the trade-off between explanation fidelity and interpretability in the Bayesian framework. The method consists of two steps. First, a reference model, possibly a black-box Bayesian predictive model compromising no accuracy, is constructed and fitted to the training data. Second, a proxy model from an interpretable model family that best mimics the predictive behaviour of the reference model is found by optimizing the interpretability utility function. The approach is model agnostic – neither the interpretable model nor the reference model are restricted to be from a certain class of models – and the optimization problem can be solved using standard tools in the chosen model family. Through experiments on real-word data sets using decision trees as interpretable models and Bayesian additive regression models as reference models, we show that for the same level of interpretability, our approach generates more accurate models than the earlier alternative of restricting the prior. We also propose a systematic way to measure stabilities of interpretabile models constructed by different interpretability approaches and show that our proposed approach generates more stable models.

1 Introduction and Background

Lack of interpretability remains a key barrier to the adoption of machine learning (ML) approaches in many applications. To bridge this gap, there is a growing interest among the machine learning community to ML interpretability methods, i.e. methods to make ML models understandable. Despite the large body of literature on *interpretable ML* (see Doshi-Velez and Kim [5], Du et al. [6], Murdoch et al. [20], and Weld and Bansal [29]), there has been little work on interpretability in the context of the Bayesian framework. Wang et al. [28] presented two probabilistic models for interpretable classification by constructing rule sets in the form of Disjunctive Normal Forms (DNFs). In this work, interpretability obtains by tweaking the prior

distributions to favor rule sets with a *smaller* number of *short* rules. This is achieved by allowing the decision maker to set the parameters of the prior distributions over the number and length of rules to encourage the model to have a desired size and shape. Letham et al. [18] obtained an interpretable classifier by using decision lists consisting of a series of *if ... then ...* statements. Interpretability factors are (i) number of rules in the list and (ii) size of the rules (number of statements in the left-hand side of rules). A prior distribution is defined over rule lists that favors decision lists with small number of short rules. Popkes et al. [23] proposed an interpretable Bayesian neural network architecture for clinical decision making tasks where interpretability is obtained by employing a sparsity inducing prior over feature weights. In a different but relevant scenario, Kim et al. [15] proposed an interactive approach with the goal to obtain from among a set of equally good clusterings, the one that best aligns with a user's preferences. User feedback affects the prior probability of prototypes being in a particular cluster (and therefore affects the clustering) either directly or indirectly, depending on the confidence level of the user feedback. In [27], the authors present a multi-value rule set for interpretable classification that allows multiple values in a condition and therefore induces more concise rules compared to the single-value rules. Same as the work of [28], interpretability is characterized by a prior distribution that favors *smaller* number of *short* rules.

In summary, a common practice for making ML models interpretable using the Bayesian framework is to fuse interpretability with the prior distribution such that in the inference, interpretable models become more favorable [11, 12, 30]. In the following, we call this approach *interpretabiliy prior*. We argue that this is not the best way of optimizing for interpretability for the following reasons:

1. Interpretability is about users' preferences, not about our assumptions about the data. The prior is meant for the latter. One should distinguish the data generation mechanism from the decision making process of interpretability optimization.

2. Optimizing interpretability naturally sacrifices some of the accuracy of the model. If interpretability is pursued by changing the prior, there is no reason why the trade-off between accuracy and interpretability would be optimal. This has been shown in a different but related scenario [22] where the authors showed that fitting a model using sparsity inducing priors that favor simpler models results in performance loss.

3. Formulating interpretability prior for certain classes of models such as neural networks could be difficult.

To solve these concerns, we propose to reserve the prior to assumptions on the data, and to include interpretability in the decision-making stage of how the model is used. This results in a two-step strategy to interpretability in the Bayesian framework. We first fit a highly accurate Bayesian predictive model, which we call reference model, to the training data without constraining it to be interpretable. In the second phase, we construct an interpretable proxy model that best describes locally and/or globally the behavior of the reference model. The proxy model is constructed by optimizing a utility function, referred to as *interpretability utility*, that consists of two terms: (I) a term to minimize the discrepancy of the proxy model from the reference model, and (II) a term to penalize the complexity of the model to make the proxy model as interpretable as possible. Term (I) corresponds to the *reference predictive model selection* idea in the Bayesian framework [26, Section 3.3]. The proposed approach is model-agnostic meaning that neither the reference model nor the interpretable proxy are constrained to a particular class of models. We also emphasize that the proposed approach is feasible for non-Bayesian models as well, which can be interpreted to produce point estimates of the parameters of the model instead of posterior

distributions.

Through experiments on real-world data sets using decision trees as interpretable proxies and Bayesian additive regression tree (BART) models [4] as reference models, we show that the proposed approach results in regression trees which are more accurate and more interpretable than the alternative of fitting an a-priori interpretable model to the data. We also show that this interpretability utility approach can construct more stable interpretable models.

1.1 Our Contributions:

Main contributions of the paper are:

- We present how Bayesian reference predictive model selection can be combined with interpretability utilities to produce more interpretable models in decision theoretically correct way. The proposed approach is model agnostic and can be used with different notions of interpretability.
- For the special case of classification and regression tree models [1] as interpretable models and BART model as the black-box Bayesian predictive model, we derive the formulation of the proposed approach and show that it outperforms the interpretability prior approach in accuracy, explicating the trade-off between explanation fidelity and interpretability.
- We propose a systematic approach to compare stability of interpretable models.

2 Motivation for Interpretability Utility

In this section, we discuss the motivation for formulating interpretability optimization in the Bayesian framework as a utility function. We also discuss how this formulation allows accounting for model uncertainty in the explanation. Both discussions are accompanied with illustrative examples.

2.1 Interpretability as a Decision Making Problem

Bayesian modelling allows encoding prior information into the prior probability distribution (similarly, one might use regularization in maximum likelihood based inference). This may tempt to changing the prior distribution to favour models that are easier for humans to understand, using some measure of interpretability. A simple example is using shrinkage priors in linear regression to find a smaller set of practically important covariates. We argue, however, based on the observation that interpretability is not an inherent characteristic of data generating processes. The approach can be misguiding and results in leaking user preferences about interpretability into the model of the data generation process.

We suggest separating the construction of a model for the data-generating process from construction of an interpretable proxy model. In a prediction task, the former corresponds to building a model that predicts as well as possible, without considering its interpretability. Interpretability is introduced in the second stage by building an interpretable proxy to explain the behaviour of the predictive model. We consider the second step as a decision making problem, where the task

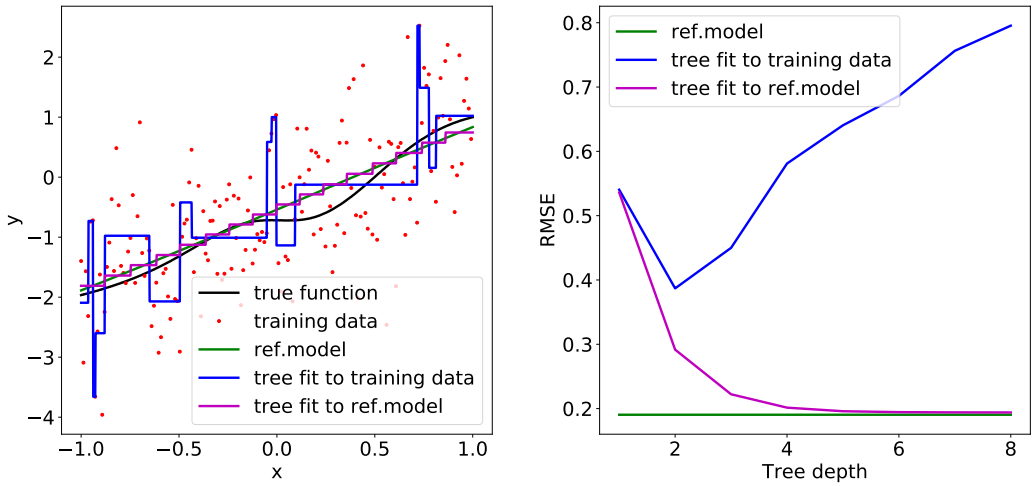


Figure 1: Illustrative example: The reference model (green) is a highly predictive non-interpretable model that approximates the true function (black) well. The interpretable model fitted to the reference model (magenta) approximates the reference model (and consequently the true function) well, while the interpretable model fitted to the training data (blue) fails to approximate the predictive behavior of the true function.

is to choose a proxy model that trades off between human interpretability and fidelity (w.r.t. the original model).

2.2 The Issue with Interpretability in the Prior

Let $\mathcal{M}$ denote the assumptions about the data generating process and $\mathcal{I}$ the preferences toward interpretability. Consider an observation model for data y, $p(y \mid \theta, \mathcal{M})$, and alternative prior distributions $p(\theta \,|\, \mathcal{M})$ and $p(\theta \,|\, \mathcal{M}, \mathcal{I})$. Here, θ can, for example, be continuous model parameters (e.g., weights in a regression or classification model) or it can index a set of alternative models (e.g., each configuration of θ could correspond to using some subset of input variables in a predictive model). Clearly, the posterior distributions $p(\theta \,|\, \mathcal{D}, \mathcal{M})$ and $p(\theta \,|\, \mathcal{D}, \mathcal{M}, \mathcal{I})$ (and their corresponding posterior predictive distribution) are in general different and the latter includes a bias towards interpretable models. In particular, when $\mathcal{I}$ does not correspond to prior information about the data generation process, there is no guarantee that $p(\theta \,|\, \mathcal{D}, \mathcal{M}, \mathcal{I})$ provides a reasonable quantification of our knowledge of θ given the observations $\mathcal{D}$ or that $p(\tilde{y} \,|\, \mathcal{D}, \mathcal{M}, \mathcal{I})$ provides good predictions. We will give an example of this below. In the special case where $\mathcal{I}$ does describe the data generation process, it can directly be included in $\mathcal{M}$.

For example, Lage et al [17] propose to find interpretable models in two steps: (1) fit a set of models to data and take ones that give high enough predictive accuracy, (2) build a prior over these models, based on an indirect measure of user interpretability (human interpretability score). It is not obvious that this leads to a good trade-off between accuracy and interpretability: in practice, it requires choosing the set of models for step 1 to contain interpretable models, mixing knowledge about the data generation process with preferences for interpretability.

2.2.1 Illustrative Example

We give an illustrative example, in a case where the assumptions in the interpretable model do not match with the data generating process, to demonstrate the difference between (1) fitting an interpretable model directly to the training data (the *interpretability prior* approach), and (2) the two-stage fitting process of first fitting a reference model and then approximating it with an interpretable model (the proposed *interpretability utility* approach). For simplicity of visualization, we use a one-dimensional smooth function as the data-generating process, with Gaussian noise added to observations (Figure 1:left, black curve and red dots). As an interpretable model, a regression tree is fitted with a fixed depth of 4 (Figure 1:left, blue). Being a piece-wise constant function, it doesn't correspond to true prior knowledge about the ground-truth function. A Gaussian process with the MLP kernel function is used as a reference model (Figure 1:left, magenta).

The regression tree (of depth 4) fitted directly to the data (blue line) overfits and doesn't give an accurate representation of the underlying data generation process (black line). The interpretability utililty approach, on the other hand, gives a clearly better representation of the smooth, increasing function, as the reference model (green line) captures the smoothness of the underlying data generation process and this is transferred to the regression tree (magenta line). The choice of the complexity of the interpretable model is also easier, because the tree can only "overfit" to the reference model, meaning that it only becomes a more accurate (but possibly less easy to interpret) representation of the reference model. In particular, the trade-off is between the interpretability and fidelity with regard to the reference model, but not the original training data, making the choice of complexity of the interpretable model significantly easier. Figure 1:right shows the root mean squared errors compared to the true underlying function as the tree depth is varied.

2.3 Interpreting Uncertainty

In many applications, such as treatment effectiveness prediction, knowing the uncertainty in the prediction is important. In Bayesian modelling, quantifying uncertainty is fundamental. Any explanation of the predictive model should also provide insight about the uncertainties and their sources. We demonstrate with an example that the proposed method can provide useful information about model uncertainty.

2.3.1 Practical Example

We provide an example of visualizing uncertainty with our proposed method in locally explaining Bayesian deep convolutional neural network predictions in the MNIST dataset of images of digits, with the task of classifying between 3s and 8s. We use the Bernoulli dropout method [9, 10], with a dropout probability of 0.2 and 20 Monte Carlo samples at test time, to approximate Bayesian neural network inference. Logistic regression is used as the interpretable model family[1].

Figure 2 shows visually the explanation model weights (mean, variance, and three samples out of the 20, explained using the linear model) for a digit, comparing the model in an early training

[1]The optimization of the interpretable model follows the general framework explained later, with logistic regression used as the interpretable model family instead of CART. No penalty for complexity was used here, since the logistic regression model weights are easy to visualize as pseudo-colored pixels

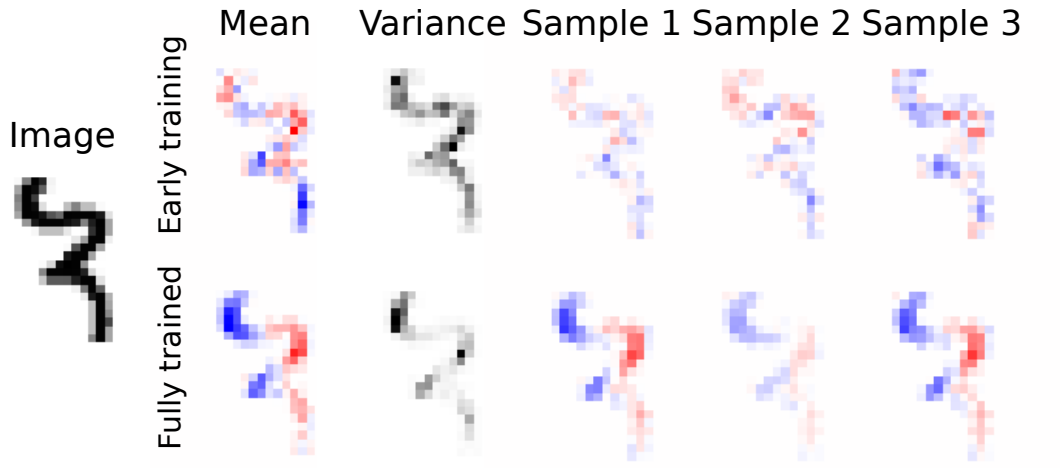


Figure 2: Mean explanation, explanation variance, and three sample explanations for a convolutional neural network 3-vs-8 MNIST-digit classifier early in the training and fully trained. Colored pixels show linear explanation model weights, with red being positive for 3 and blue for 8.

phase (upper row) and fully trained (lower row). The mean explanations show that the fully trained model has spatially smooth contributions to the class probability, while the model in early training is noisy. Moreover, being able to look at the explanations of individual posterior predictive samples shows, for example, that the model in early training has not yet been able to confidently assign the upper loop to either indicate a 3 or an 8 (samples 1 and 2 have reddish loop, while sample 3 has bluish). Indeed, the variance plot shows that the model variance spreads evenly over the digit. On the other hand, the fully trained model has little uncertainty about which parts of the digit indicate a 3 or an 8, with most model uncertainty being about the magnitude of the contributions.

3 Method: Interpretability Utility for Bayesian Predictive Models

Let $\mathcal{D} = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$ denote a training set of size N, where $\boldsymbol{x}_i = [x_{i1}, \cdots, x_{id}]^T$ is a d-dimensional feature vector and $y_i \in \mathbb{R}$ is the target variable. To construct an interpretable model using the Bayesian framework, we propose to separate the model construction and interpretability optimization. The idea is to first fit a highly predictive (reference) model $\mathcal{M}$ without concerning the interpretability of the fitted model. In the second phase, by optimizing a utility function which we call *interpretability utility*, we find an interpretable model that best explains the behavior of the reference model locally or globally.

Denote the likelihood of the reference model by $p(y|\boldsymbol{x}, \boldsymbol{\theta}, \mathcal{M})$ and the posterior distribution $p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M})$. For optimizing the interpretability, we introduce an interpretable model family with likelihood denoted by $p(y|\boldsymbol{x}, \boldsymbol{\eta}, \mathcal{T})$ belongs to a probabilistic model family $\mathcal{T}$ with parameters $\boldsymbol{\eta}$. The best interpretable model is the one that is the closest one to the reference model prediction wise, and at the same time easily interpretable. Assuming we want to locally interpret the

reference model, such explainable model can be found by optimizing the following utility function:

$$\hat{\boldsymbol{\eta}} = \arg\min_{\boldsymbol{\eta}} \int \pi_{\boldsymbol{x}}(\boldsymbol{z}) \mathrm{KL}\left[p(y \,|\, \boldsymbol{z}, \boldsymbol{\theta}, \mathcal{M}) \parallel p(y \,|\, \boldsymbol{z}, \boldsymbol{\eta}, \mathcal{T})\right] d\boldsymbol{z} + \Omega(\boldsymbol{\eta}) \tag{1}$$

where KL denotes the Kullback-Leibler divergence, Ω is the penalty function for the complexity of the interpretable model, and $\pi_{\boldsymbol{x}}(\boldsymbol{z})$ is a probability distribution defining the local neighborhood around $\boldsymbol{x}$, the data point which prediction is to be explained. The minimization of the KL divergence ensures that the interpretable model has similar predictive performance to the reference model while the complexity penalty guarantees the interpretability of the model.

We compute the expectation in Eq. 1 with Monte Carlo approximation by drawing $\{\boldsymbol{z}_s\}_{s=1}^{S}$ samples from $\pi_{\boldsymbol{x}}(\boldsymbol{z})$:

$$\hat{\boldsymbol{\eta}}^{(l)} = \arg\min_{\boldsymbol{\eta}} \frac{1}{S} \sum_{s=1}^{S} \mathrm{KL}\left[p(y \,|\, \boldsymbol{z}_s, \boldsymbol{\theta}^{(l)}, M) \parallel p(y \,|\, \boldsymbol{z}_s, \boldsymbol{\eta}, \mathcal{T})\right] + \Omega(\boldsymbol{\eta}), \tag{2}$$

for $l = 1, \ldots, L$ posterior draws from $p(\boldsymbol{\theta} \,|\, \mathcal{D}, \mathcal{M})$. Eq. 2 can be solved by first drawing a sample $\boldsymbol{\theta}^{(l)}$ from the posterior of the reference model, and then finding a sample $\boldsymbol{\eta}^{(l)}$ from the posterior of the interpretable model. It has been shown [22] that minimization of the KL-divergence in Eq. 2 is equivalent to maximizing the expected log-likelihood of the interpretable model over the likelihood obtained by a posterior draw from the reference model:

$$\arg\max_{\boldsymbol{\eta}} E_{y|\boldsymbol{z}_s, \boldsymbol{\theta}^{(l)}} \left[\log p(y \,|\, \boldsymbol{z}_s, \boldsymbol{\eta})\right]. \tag{3}$$

Using this equivalent form and by adding the complexity penalty term, the interpretability utility becomes:

$$\arg\max_{\boldsymbol{\eta}} E_{y|\boldsymbol{z}_s, \boldsymbol{\theta}^{(l)}} \left[\log p(y \,|\, \boldsymbol{z}_s, \boldsymbol{\eta})\right] - \Omega\left(\boldsymbol{\eta}\right). \tag{4}$$

Choice of the form of the complexity penalty term depends on the class of interpretable models; possible options are the number of leaf nodes for the class of decision trees, number of rules and/or size of the rules for the class of rule list models, number of non-zero weights for linear regression models, etc. Although the proposed approach is general and can be used for any family of interpretable models, in the following, we use the class of Classification and Regression Tree (CART) models [1] with the tree size (the number of leaf nodes) as the measure of interpretability. With this assumption, the interpretability prior is over the model space; it could also be defined over the parameter space of a particular model, such as tree shape parameters of Bayesian CART models [3].

A CART model describes $p(y \mid \boldsymbol{z}, \boldsymbol{\eta})$ with two main components $\boldsymbol{\eta} = (T, \boldsymbol{\phi})$: a binary tree T with b terminal nodes and a parameter $\boldsymbol{\phi} = (\phi_1, \phi_2, \cdots, \phi_b)$ that associates the parameter value ϕ_i with the ith terminal node. If $\boldsymbol{z}$ lies in the region corresponding to the ith terminal node, then $y \,|\, \boldsymbol{z}, \boldsymbol{\eta}$ has distribution $f(y|\phi_i)$, where f denotes a parametric probability distribution with parameter ϕ_i. For CART models, it is typically assumed that, conditionally on $\boldsymbol{\eta}$, values y within a terminal node are independently and identically distributed, and y values across terminal nodes are independent. In this case, the corresponding likelihood of the interpretable model for the lth draw from the posterior of $\boldsymbol{\theta}$ has the form

$$p(\boldsymbol{y} \,|\, \boldsymbol{Z}, \boldsymbol{\eta}^{(l)}) = \prod_{i=1}^{b} f(\boldsymbol{y}_i \,|\, \phi_i^{(l)}) = \prod_{i=1}^{b} \prod_{j=1}^{n_i} f(y_{ij} \,|\, \phi_i^{(l)}), \tag{5}$$

where $\boldsymbol{y}_i \equiv (y_{i1}, \cdots, y_{in_i})$ denotes the set of n_i observations assigned to the partition generated by the ith terminal node with parameter $\phi_i^{(l)}$, and $\boldsymbol{Z}$ is the matrix of all $\boldsymbol{z}_s$. For regression problems, assuming a mean-shift normal model for $f(y_{ij} \,|\, \phi_i^{(l)})$[2], we have the following likelihood for the interpretable model[3]:

$$\boldsymbol{y}_i \,|\, \phi_i^{(l)} \overset{\text{iid}}{\sim} N(\boldsymbol{y}_i \,|\, \mu_i^{(l)}, \sigma^{2(l)}), \;\; i = 1, \ldots, b. \tag{6}$$

With this formulation, the task of making a reference model M interpretable becomes finding a tree structure T with parameters $\boldsymbol{\phi}^{(l)} = (\boldsymbol{\mu}^{(l)} = \{\mu_i^{(l)}\}_{i=1}^b, \sigma^{2(l)})$ such that its predictive performance is as close as possible to M, while being as interpretable as possible as measured by the complexity term Ω.

With the normal likelihood defined for the terminal nodes, the log-likelihood of the ith partition generated by the ith terminal node is[4]

$$\mathcal{L}_i = -\frac{n_i}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{n_i}(y_{ij} - \mu_i)^2 \,,$$

and the log-likelihood of the tree for the S samples drawn from the neighborhood of $\boldsymbol{x}$ is

$$\mathcal{L} = -\frac{S}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{b}\sum_{j=1}^{n_i}(y_{ij} - \mu_i)^2 \,. \tag{7}$$

Projecting this into Eq. 4, the interpretability utility has the following form:

$$\begin{aligned}
&\arg\max_{\boldsymbol{\eta}} \; -\frac{S}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{b}\sum_{j=1}^{n_i} E_{y_{ij}|\boldsymbol{\theta}^{(l)}}\left[(y_{ij} - \mu_i)^2\right] \\
&- \Omega(T) \propto \\
&\arg\max_{\boldsymbol{\eta}} \; -\frac{S}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{b}\sum_{j=1}^{n_i}\left[\sigma_{ij}^2 + (\bar{y}_{ij} - \mu_i)^2\right] \\
&- \Omega(T),
\end{aligned} \tag{8}$$

where $\bar{y}_{ij}$ and σ_{ij}^2 are respectively the mean and variance of the reference model for the jth sample in the ith terminal node. $\Omega(T)$ is a function of the interpretability of the CART model. Here we set it to αb using α as a regularization parameter. The pseudocode of the proposed approach is shown in Algorithm 1.

When fitting a global interpretable model to the reference model, instead of drawing samples from $\pi_{\boldsymbol{x}}$, we will use training inputs $\{\boldsymbol{x}_n\}_{n=1}^N$ with their corresponding output computed by the reference model $\{y_n^{ref}\}_{n=1}^N$ as the target value.

The next subsection explains how to solve Eq. 8 for the CART model to obtain an interpretable model to interpret a reference model.

[2]For mean-variance shift model, each terminal node has its own σ_i^2 variable and the number of parameters is $2 \times b$.

[3]For classification problems, the likelihood follows a categorical distribution.

[4]For simplicity, for the rest of the paper we drop the index l from the parameters emphasizing that for each draw l of $p(\boldsymbol{\theta} \,|\, \mathcal{D}, \mathcal{M})$, a corresponding $\boldsymbol{\phi}^{(l)}$ will be computed

Algorithm 1: Decision theoretic approach for ML interpretability in the Bayesian framework

Input: training data $\mathcal{D} = \{(y_n, \boldsymbol{x}_n)\}_{n=1}^{N}$
Output: a decision tree explaining the prediction of a new sample $\boldsymbol{x}_{test}$

```
/* REFERENCE MODEL CONSTRUCTION                                    */
```
fit the Bayesian predictive model to the training data $\mathcal{D}$ without interpretability prior;
for *each sample* $\boldsymbol{x}_{test}$ *in the test set* **do**
  draw $\{\boldsymbol{z}_s\}_{s=1}^{S}$ from the neighborhood of $\boldsymbol{x}_{test}$ defined by $\pi_{\boldsymbol{x}}$;
  for *each draw* $\boldsymbol{z}_s$ **do**
    get the mean and variance of the Bayesian predictive distribution;
  end
  `/* INTERPRETABILITY OPTIMIZATION */`
  fit a CART model to $\{(\boldsymbol{z}_s, \bar{y}_s)\}_{s=1}^{S}$ by optimizing Eq. 8
end

3.1 Optimization Approach

We optimize Eq. 8 by using the backward fitting idea which involves first growing a large tree and then pruning it back to obtain a smaller tree with better generalization. For this goal, we use the formulation of maximum likelihood regression tree (MLRT) [25].

3.1.1 Growing a large tree

Given the training data[5], MLRT automatically decides on the splitting variable x_j and split point (a.k.a. pivot) c using a greedy search algorithm that aims to maximize the log-likelihood of the tree by splitting the data in the current node into two parts: the left child node satisfying $x_j \leq c$ and the right child node satisfying $x_j > c$. The procedure of growing the tree is as follows:

1. For each node i, determine the maximum likelihood estimate of its mean parameter μ_i given observations associated with the node, and then compute the variance parameter of the tree:

$$\hat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \bar{y}_{ij}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{b} \sum_{j=1}^{n_i} \left[\sigma_{ij}^2 + (\bar{y}_{ij} - \hat{\mu}_i)^2 \right]}{S}.$$

   The log-likelihood score of the node is then given, up to a constant, by $\mathcal{L}_i \propto -n_i \log(\hat{\sigma^2})$.

2. For each variable x_j, determine the amount of increase to the log-likelihood of the node i caused by a split r as

$$\Delta_{(r,x_j,i)} = \mathcal{L}_{i_R} + \mathcal{L}_{i_L} - \mathcal{L}_i,$$

   where $\mathcal{L}_{i_R}$ and $\mathcal{L}_{i_L}$ are the log-likelihood scores of the right and left child nodes of the parent node i generated by the split r on the variable x_j, respectively.

[5]Here, for local interpretation, training data refers to the S samples (with their corresponding predictions made by the reference model) taken from the neighborhood distribution to fit the explainable model.

3. For each variable x_j, select the best split r_j^* with largest increase to the log-likelihood.
4. Among the best splits, the one that causes the global maximum increase in the log-likelihood score will be selected as the global best split, r^*, for the current node, i.e. $r^* = \max_{r_j^*,\, j=1,\cdots,d} \Delta_{(r_j^*, x_j, i)}$.
5. Iterate steps 1 to 4 until reaching the stopping criteria.

In our implementation we used the minimum size of a terminal node (the number of samples lie in the region generated by the terminal node) as the stopping condition.

3.1.2 Pruning

We adopt the cost-complexity pruning using the following cost function:

$$C_\alpha(T) = S\log(\hat{\sigma}^2) + \alpha b, \tag{9}$$

where $\hat{\sigma}^2$ is the maximum likelihood estimate of the tree T. Pruning is done iteratively; in each iteration i, the internal node h that minimizes $\alpha = \frac{C(h)-C(T_i)}{|\text{leaves}(T_h)|-1}$ is selected for pruning where $C(h)$ refers to the cost of the decision tree with h as terminal node, $C(T_i)$ denote the cost of the full decision tree in iteration i, and T_h denotes the subtree with h as its root. The output of the above procedure is a sequence of decision trees and a sequence of α values. The best α and its corresponding subtree are selected using 5-fold cross-validation.

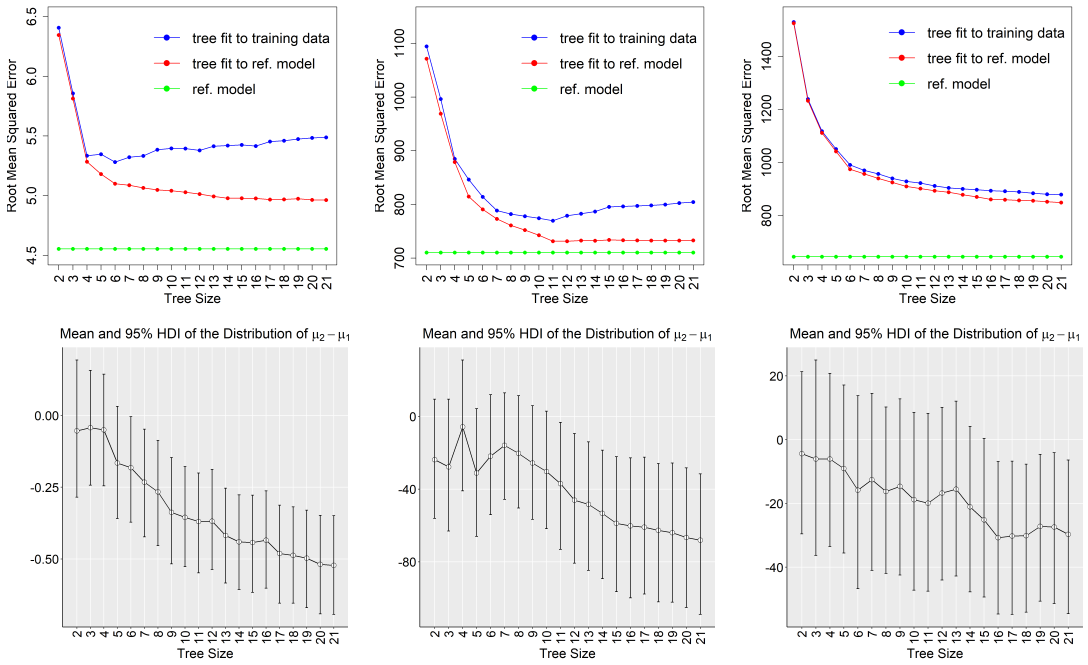
3.2 Connection With Local Interpretable Model-agnostic Explanation (LIME)

LIME [24] is a local interpretation approach that works by fitting a sparse linear model to the predictive model's response via sampling around the point being explained. Our proposed approach extends LIME to KL divergence based interpretation of Bayesian predictive models (although it can also be used for non-Bayesian probabilistic models as well) by combining the idea of LIME and projection predictive variable selection approach [22]. The approach is able to handle different types of predictions (continuous valued, class labels, counts,censored and truncated data, etc.) and explanations (local or global) as long as we can compute KL divergence between the predictive distributions of the original model and the explanation model. For a more detailed explanation of the connection check the preliminary work of [21].

4 Experiments

In this section, we compare the performance of the proposed method with the interpretability prior approach in three scenarios. First, we evaluate the ability of the two approaches to find a good trade-off between accuracy and interpretability when constructing global interpretations using CART models as the interpretable model family. Then, stability of models constructed by each approach is investigated. Finally, the approaches are compared in providing local explanation for each prediction. Our codes and data are available online at www.anonymous.com[6].

[6]The link will be available upon acceptance.



(a) Body Fat (b) Baseball Players (c) Bike Rental

Figure 3: *Top row*: Comparison of interpretability prior and interpretabiliy utility approach in trading off between accuracy and interpretability when using CART as explainable models and BART as reference model. *Bottom row*: Results of Bayesian t-test that shows the mean and 95% highest density interval of the distribution of difference of means.

4.1 Global Interpretation

4.1.1 Data

We test our proposed approach on three datasets: body fat [14], baseball players [13], and bike rental [8]. Each data set is divided into training and test set containing 75% and 25% of samples, respectively. As black-box reference model, we fit a Bayesian Additive Regression Tree (BART) model [4] to the training data using the BART package in R with all parameters set to the default values except the number of burn-in iterations (nskip) in the Markov Chain Monte Carlo (MCMC) sampling and the number of posterior draws (ndpost) which are set to 2000 and 4000, respectively. As the prediction of the BART model, we use the mean of the predictions of the posterior draws. CART models are used as the interpretable model family. The interpretability prior approach fits a CART model directly to the training data where the interpretability prior is on the fitted model, i.e. the CART model is simple to interpret. On the other hand, our approach fits the CART model to the reference model (the BART model), through interpretability utility optimization. The process was repeated for 50 runs with the dataset divided into training and test sets randomly.

Table 1: Bootstrap instability values in the form of mean ± std. Best values are bolded.

Interpretability	Body Fat	Baseball	Bike
Prior	0.61 ± 0.16	0.89 ± 0.06	$\mathbf{0.65 \pm 0.07}$
Utility	$\mathbf{0.58 \pm 0.14}$	$\mathbf{0.87 \pm 0.07}$	$\mathbf{0.65 \pm 0.07}$

4.1.2 Performance Analysis

Lundberg and Lee [19] suggested viewing an explanation model as a model itself. With this perspective, we quantitatively evaluate the explanation models as if they were models. In Figure 3, the top row compares the performance of the two approaches in trading off between accuracy and interpretability for different data sets. Using the number of leaf nodes in the CART models as the measure of interpretability, it can be seen that **the most accurate models with any level of complexity (interpretability) are obtained with our proposed approach**.

To test the significance of the differences in the results, we perform Bayes t-test [16]. The approach works by building up a complete distributional information for the means and standard deviations of each group (for each tree size, we have two groups of 50 RMSE values for the interpretability prior approach and interpretability utility approach) and constructing a probability distribution over their differences using MCMC estimation. From the distributions of the differences of the means, the mean and the 95% Highest Density Interval (HDI) (as the range were the actual difference of the two group is within 95% credibility) for each data sets is shown in the bottom row of Figure 3 where μ_1 refers to the mean values generated by the interpretability prior approach and μ_2 refers to the means of the distribution generated for the interpretability utility approach. When the 95% HDI does not include zero there is a credible difference between the two groups. As it is shown in the figure, for all three data sets, for highly interpretable models (highly inaccurate), the difference between the two approach is not significant (HDI contains zero). This is expected since by increasing the interpretability, the ability of the interpretable model to explain variability of the data or of the reference model decreases a lot and both approaches provide poor performance. However, by increasing the complexity to a reasonable level, we see that the differences of the two approaches become significant.

4.1.3 Stability Analysis

The goal of interpretable ML is to provide a comprehensive explanation of the prediction logic to the decision maker. However, perturbation in the data or new samples may affect the learned interpretable model and lead to a very different explanation. This instability can cause real problems for decision makers that need to take actions in critical situations. Therefore, it is important to evaluate stabilities of different interpretable ML approaches.

For this goal, using a bootstrapping procedure with 10 iterations, we compute pairwise dissimilarities of the interpretable models obtained using each approach and report the mean and standard deviation of the dissimilarity values as their instability measure (smaller is better). We used the dissimilarity measure proposed in [2]. Assuming we are given two regression trees T_1 and T_2, for each internal node t, the similarity of the trees at node t is computed by

$$S^t_{(1,2)} = I^t_{k=k'} \left(1 - \frac{|\,\delta^t_1 - \delta^t_2\,|}{\text{range}(X_k)}\right) \tag{10}$$

Table 2: Comparison of the local fidelity of LIME and Interpertability utility when being used to explain predictions of BART. Best values are bolded.

Dataset	LIME	Interpretability Utility
Boston housing	4.86	**2.94**
Automobile	0.014	**0.010**

where $I^t_{k=k'}$ is the indicator that determines whether the feature used to grow node t in T_1 is identical to the one used in T_2 ($I^t_{k=k'} = 1$) or not, δ^t_1 and δ^t_2 are pivots used to grow the node t in T_1 and T_2, respectively and range(X_k) is the range of values of feature k. Finally, the dissimilarity of the two decision trees is computed as $d(T_1, T_2) = 1 - \sum_{t \in \text{internal\_nodes}} q^t S^t_{(1,2)}$ where q^t are user specified weight value which we set to $1/b$ where b is the number of terminal nodes. The reported values are averaged over 45 values (10 bootstraping iterations result in $(10 \times 9)/2 = 45$ pairs of explainable models).

Table 1 compares the two approaches over the three data sets used in this subsection. The explanation models constructed using our proposed approach are more stable in two data sets and in one of them both approaches perform equally well (the differences in the Body fat and Baseball data sets are not statistically significant).

4.2 Local Interpretation

We next demonstrate the ability of the proposed approach in locally interpreting the predictions of a Bayesian predictive model. Same as before, we used BART model[7] as the black box model and CART as the interpretable model family. For the CART model, we set the maximum depth of the decision trees to 3 to obtain more interpretable local explanations. We compare with LIME[8] emphasizing that LIME is not an interpretability prior approach, however it is a commonly used baseline for local interpretation approaches. The comparison is done over 2 different data sets from the UCI repository [7]: Boston housing and automobile. Decision trees obtained by our approach to locally explain predictions of the BART model, used on average 2.03 features for the Boston housing data set and 2.4 for Automobile data set. Therefore, to have a fair comparison, we set the feature selection approach of LIME to ridge regression and select the 2 features with highest absolute weights to be used in the explanation[9]. We use the standard quantitative metric for local fidelity: $\mathbb{E}_x\left[\text{loss}\left(\text{interp}_x(\boldsymbol{x}), \text{pred}(\boldsymbol{x})\right)\right]$ where given a test data $\boldsymbol{x}$, $\text{interp}_x(x)$ refers to the prediction of the local interpretable model (fitted locally to the neighborhood of $\boldsymbol{x}$) for $\boldsymbol{x}$ and $\text{pred}(\boldsymbol{x})$ refers to the prediction of the black-box model for $\boldsymbol{x}$. We used locally weighted square loss as the loss function with $\pi_x = \mathcal{N}\left(\boldsymbol{x}, \sigma^2 \boldsymbol{I}\right)$ where $\sigma = 1$.

Each data set is divided into 90%/10% training/test split. For each test data, we draw 200 samples from the neighborhood distribution. Table 2 shows the results where our approach produces more accurate local explanation for both data sets. Figure 4 shows as an example, a decision tree constructed by our proposed approach to locally explain the prediction of the BART model for the particular test data shown in the figure from Boston housing data set. It

[7] In this experiment, we set the number of trees to 50 with nskip and ndpost set to 1000 and 2000 respectively, for faster run.

[8] We use the 'lime' package in R (https://cran.r-project.org/web/packages/lime/lime.pdf) for the implementation.

[9] MSEs of LIME with 3 features are respectively 2.48 and 0.006 for Boston housing and Automobile data set.

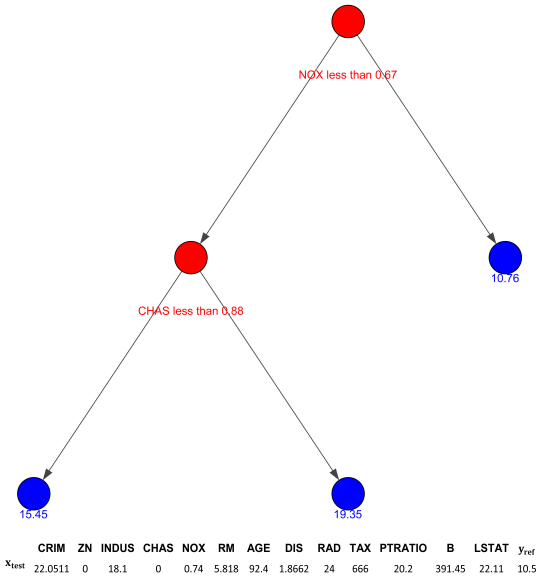


Figure 4: Example of a decision tree obtained by the interpretability utility approach to locally explain the prediction of the BART model (y_{ref} is the mean of the predictions of the 2000 posterior draws) for the particular test data x_{test}. Using only 2 features, our approach predicts the output 10.76. LIME with 2 features predicts the output to be 14.06, and with 3 features, LIME prediction is 13.18.

can be seen that using only two features, our proposed approach obtains good local fidelity while maintaining interpretability with a decision tree with only 3 leaf nodes.

5 Conclusion

We presented a novel approach to construct interpretable explanations in the Bayesian framework by formulating the task as optimizing a utility function instead of changing the priors. We first fit a Bayesian predictive model compromising no accuracy and then project the information in the predictive distribution of the model to an interpretable probabilistic model. This also allows accounting for model uncertainty in the explanations. We showed that the proposed approach outperforms the alternative approach of restricting the prior, in terms of accuracy, interpretability and stability.

6 Acknowledgments

This work was financially supported by the Academy of Finland (grants 294238, 319264 and 313195), by the Vilho, Yrjö and Kalle Väisälä Foundation of the Finnish Academy of Science and Letters, by the Foundation for Aalto University Science and Technology, and by the Finnish Foundation for Technology Promotion (Tekniikan Edistämissäätiö). We acknowledge the computational resources provided by the Aalto Science-IT Project.

References

[1] L Breiman, J Friedman, R Olshen, and C Stone. Classification and regression trees, 1999.

[2] Bénédicte Briand, Gilles R Ducharme, Vanessa Parache, and Catherine Mercat-Rommens. A similarity measure to assess the stability of classification trees. *Computational Statistics & Data Analysis*, 53(4):1208–1217, 2009.

[3] Hugh A Chipman, Edward I George, and Robert E McCulloch. Bayesian CART model search. *Journal of the American Statistical Association*, 93(443):935–948, 1998.

[4] Hugh A. Chipman, Edward I. George, and Robert E McCulloch. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

[5] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[6] Mengnan Du, Ninghao Liu, and Xia Hu. Techniques for interpretable machine learning. *arXiv preprint arXiv:1808.00033*, 2018.

[7] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

[8] Hadi Fanaee-T and Joao Gama. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, 2(2-3):113–127, 2014.

[9] Yarin Gal and Zoubin Ghahramani. Bayesian convolutional neural networks with Bernoulli approximate variational inference. In *4th International Conference on Learning Representations (ICLR) workshop track*, 2016.

[10] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 1050–1059, 2016.

[11] Jingyi Guo, Andrea Riebler, and Håvard Rue. Bayesian bivariate meta-analysis of diagnostic test studies with interpretable priors. *Statistics in medicine*, 36(19):3039–3058, 2017.

[12] Satoshi Hara and Kohei Hayashi. Making tree ensembles interpretable: A Bayesian model selection approach. In *International Conference on Artificial Intelligence and Statistics*, pages 77–85, 2018.

[13] David C Hoaglin and Paul F Velleman. A critical look at some analyses of major league baseball salaries. *The American Statistician*, 49(3):277–285, 1995.

[14] Roger W Johnson. Fitting percentage of body fat to simple body measurements. *Journal of Statistics Education*, 4(1), 1996.

[15] Been Kim, Elena Glassman, Brittney Johnson, and Julie Shah. ibcm: Interactive Bayesian case model empowering humans via intuitive interaction. *Technical Report: MIT-CSAIL-TR*, 2015.

[16] John K Kruschke. Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, 142(2):573, 2013.

[17] Isaac Lage, Andrew Slavin Ross, Been Kim, Samuel J Gershman, and Finale Doshi-Velez. Human-in-the-loop interpretability prior. *arXiv preprint arXiv:1805.11571*, 2018.

[18] Benjamin Letham, Cynthia Rudin, Tyler H McCormick, David Madigan, et al. Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, 9(3):1350–1371, 2015.

[19] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774, 2017.

[20] W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Interpretable machine learning: definitions, methods, and applications. *arXiv preprint arXiv:1901.04592*, 2019.

[21] Tomi Peltola. Local interpretable model-agnostic explanations of Bayesian predictive models via kullback-leibler projections. *arXiv preprint arXiv:1810.02678*, 2018.

[22] Juho Piironen, Markus Paasiniemi, and Aki Vehtari. Projective inference in high-dimensional problems: Prediction and feature selection. *arXiv preprint arXiv:1810.02406*, 2018.

[23] Anna-Lena Popkes, Hiske Overweg, Ari Ercole, Yingzhen Li, José Miguel Hernández-Lobato, Yordan Zaykov, and Cheng Zhang. Interpretable outcome prediction with sparse Bayesian neural networks in intensive care. *arXiv preprint arXiv:1905.02599*, 2019.

[24] 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*, pages 1135–1144. ACM, 2016.

[25] Xiaogang Su, Morgan Wang, and Juanjuan Fan. Maximum likelihood regression trees. *Journal of Computational and Graphical Statistics*, 13(3):586–598, 2004.

[26] Aki Vehtari and Janne Ojanen. A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142–228, 2012.

[27] Tong Wang. Multi-value rule sets for interpretable classification with feature-efficient representations. In *Advances in Neural Information Processing Systems*, pages 10835–10845, 2018.

[28] Tong Wang, Cynthia Rudin, Finale Doshi-Velez, Yimin Liu, Erica Klampfl, and Perry MacNeille. A Bayesian framework for learning rule sets for interpretable classification. *The Journal of Machine Learning Research*, 18(1):2357–2393, 2017.

[29] Daniel S Weld and Gagan Bansal. The challenge of crafting intelligible intelligence. *Communications of the ACM*, 62(6):70–79, 2019.

[30] Hongyu Yang, Cynthia Rudin, and Margo Seltzer. Scalable Bayesian rule lists. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3921–3930. JMLR. org, 2017.