Machine Learning (2020) 109:1855–1876
https://doi.org/10.1007/s10994-020-05901-8

# A decision-theoretic approach for model interpretability in Bayesian framework

**Homayun Afrabandpey[1] · Tomi Peltola[1] · Juho Piironen[3] · Aki Vehtari[1] · Samuel Kaski[1,2]**

Received: 10 January 2020 / Revised: 3 July 2020 / Accepted: 11 August 2020 /
Published online: 4 September 2020


**Abstract**
A salient approach to interpretable machine learning is to restrict modeling to simple 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 which does not compromise accuracy, is 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 a certain class of models—and the optimization problem can be solved using standard tools. 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 alternative of restricting the prior. We also propose a systematic way to measure stability of interpretabile models constructed by different interpretability approaches and show that our proposed approach generates more stable models.

**Keywords** Interpretable machine learning · Bayesian predictive models

## 1 Introduction and background

Accurate machine learning (ML) models are usually complex and opaque, even to the modelers who built them (Lipton 2018). This lack of interpretability remains a key barrier to the adoption of ML models in some applications including health care and economy.

Editors: Ira Assent, Carlotta Domeniconi, Aristides Gionis, Eyke Hüllermeier.

✉ Homayun Afrabandpey
homayun.afrabandpey@aalto.fi

Extended author information available on the last page of the article

To bridge this gap, there is growing interest among the ML community to interpretability methods.

Such methods can be divided into (1) interpretable model construction, and (2) post-hoc interpretation. The former aims at constructing models that are understandable. Post-hoc interpretation approaches can be categorized further into (1) model-level interpretation (a.k.a. global interpretation), and (2) prediction-level interpretation (a.k.a. local interpretation) (Du et al. 2018). Model-level interpretation aims at making existing black-box models interpretable. Prediction-level interpretation aims at explaining each individual prediction made by the model (Doshi-Velez and Kim 2017). In this paper, we focus mostly on post-hoc interpretation.

Prior research on the construction of interpretable models has mainly focused on restricting modeling to simple and easy-to-understand models. Examples of such models include sparse linear models (Ustun and Rudin 2016), generalized additive models (Lou et al. 2012), decision sets (Lakkaraju et al. 2016), and rule lists (Jung et al. 2017). In the Bayesian framework, this approach maps to defining model structure and prior distributions that favor interpretable models (Letham et al. 2015; Wang et al. 2017; Popkes et al. 2019; Wang 2018). We call this approach *interpretability prior*. Letham et al. (2015) established an interpretability prior approach for classification by use of decision lists. Interpretability measures used to define the priors were (1) the number of rules in the list and (2) the size of the rules (number of statements in the left-hand side of rules). A prior distribution was defined over rule lists to favor decision lists with a small number of short rules. Wang et al. (2017) developed two probabilistic models for interpretable classification by constructing rule sets in the form of Disjunctive Normal Forms (DNFs). In this work, interpretability is achieved similar to Letham et al. (2015), using prior distributions which favor rule sets with a smaller number of short rules. In Wang (2018), the authors extended (Wang et al. 2017) by presenting a multi-value rule set for interpretable classification, which allows multiple values per condition and thereby induces more concise rules compared to single-value rules. As in Wang et al. (2017), interpretability is characterized by a prior distribution that favors a smaller number of short rules. Popkes et al. (2019) built up an interpretable Bayesian neural network for clinical decision-making tasks, where interpretability is attained by employing a sparsity-inducing prior over feature weights. For more examples, see Kim et al. (2015), Hara and Hayashi (2018), Yang et al. (2017), Guo et al. (2017).

A common practice in model-level interpretability is to use simple models as interpretable surrogates to highly predictive black-box models (Craven and Shavlik 1996; Zhou and Hooker 2016; Bastani et al. 2018; Lakkaraju et al. 2019; Kuttichira et al. 2019). Craven and Shavlik (1996) were among the first to adopt this approach for explaining neural networks. They used decision trees as surrogates and trained them to approximate predictions of a neural network. In Zhou and Hooker (2016), the authors presented an approach to approximate the predictive behavior of a random forest by use of a single decision tree. With the same objective as Zhou and Hooker (2016), Bastani et al. (2018) developed an approach to interpret random forests using simple decision trees as surrogates. They employed active learning to construct more accurate decision trees with help from a human. Lakkaraju et al. (2019) established an approach to interpret black-box classifiers by highlighting the behavior of the black-box model in subspaces characterized by features of user interest. In Kuttichira et al. (2019), the authors used decision trees to extract rules to describe the decision-making behavior of black-box models. For more examples of this approach, see Breiman and Shang (1996), Meinshausen (2010), Wu et al. (2018), Deng (2019). The common characteristic of these

approaches is that they seek an optimal trade-off between interpretability of the surrogate model and its faithfulness to the black-box model. To the best of our knowledge, there is no Bayesian counterpart for this approach in the interpretability literature.

We argue that an interpretability prior is not the best way to optimize interpretability in the Bayesian framework 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, which in this case includes optimization of interpretability.
2. Optimizing interpretability may sacrifice some of the accuracy of the model. If interpretability is pursued by revising the prior, there is no reason why the trade-off between accuracy and interpretability would be optimal. This has been shown for a different but related scenario in Piironen et al. (2018) where the authors showed that fitting a model using sparsity-inducing priors that favor simpler models results in performance loss.
3. Formulating an interpretability prior for certain classes of models such as neural networks could be difficult.

To address these concerns, we develop a general principle for interpretability in the Bayesian framework, formalizing the idea of approximating black-box models with interpretable surrogates. The approach can be used to both constructing, from scratch, interpretable Bayesian predictive models, or to interpreting existing black-box Bayesian predictive models. The approach consists of two steps: first, a highly accurate Bayesian predictive model, called a reference model, is fitted to the training data without compromising the accuracy. In the second step, an interpretable surrogate model is constructed which best describes locally or globally the behavior of the reference model. The proxy model is obtained by optimizing a utility function, referred to as *interpretability utility*, which consists of two terms: (1) a term to minimize the discrepancy of the proxy model from the reference model, and (2) a term to penalize the complexity of the model to make the proxy model as interpretable as possible. Term (1) corresponds to selection of reference predictive model in the Bayesian framework (Vehtari and Ojanen 2012, Section 3.3).

The proposed approach can be used both for constructing interpretable Bayesian predictive models and to generate post-hoc interpretation for black-box Bayesian predictive models. When using the approach for post-hoc interpretability, it can be used to generate both global or local interpretation. The approach is model-agnostic, meaning that neither the reference model nor the interpretable proxy are constrained to a particular model family. However, when using the approach to construct interpretable Bayesian predictive models, the surrogate model should be from the family of Bayesian predictive 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. Table 1 compares the characteristics of the proposed approach with some of the related works from literature.

We demonstrate with experiments on real-world data sets that the proposed approach generates more accurate and more stable interpretable models than the alternative of fitting an a priori interpretable model to the data, i.e., using the interpretability prior approach. For the experiments in this paper, decision trees and logistic regression were used as interpretable proxies, and Bayesian additive regression tree (BART) models

**Table 1** Characteristics of different intereptation approaches

| Approach | References | Domain | Interp. model | Black-box model | Task | Bayesian |
|---|---|---|---|---|---|---|
| Trepan | Craven and Shavlik (1996) | G | DT | NN | C | ✗ |
| – | Bastani et al. (2018) | G | DT | TE | C | ✗ |
| BATrees | Breiman and Shang (1996) | G | DT | TE | C/R | ✗ |
| inTrees | Deng (2019) | G | DR | TE | C | ✗ |
| – | Kuttichira et al. (2019) | G | M/A | M/A | C/R | ✗ |
| Node harvest | Meinshausen (2010) | G | TE | TE | R | ✗ |
| Our approach | – | G/L | M/A | M/A | C/R | ✓ |

*G* global, *L* local, *DT* Decision Tree, *DR* Decision Rules, *M/A* Model Agnostic, *TE* Tree Ensemble, *NN* Neural Network, *C* Classification, *R* Regression

(Chipman et al. 2010), Bayesian neural networks, and Gaussian Processes (GP) were used as reference models.

### 1.1 Our contributions

Main contributions of this paper are:

- We propose a principle for interpretable Bayesian predictive modeling. It combines a reference model with interpretability utility to produce more interpretable models in a decision-theoretically justified 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 (CART) (Breiman et al. 1984) as interpretable models and BART as the black-box Bayesian predictive model, we show that the proposed approach outperforms the earlier interpretability prior approach in accuracy, explicating the trade-off between explanation fidelity and interpretability. Further, through experiments with different reference models, i.e., GP and BART, we demonstrate that the predictive power of the reference model positively affects the accuracy of the interpretable model. We also demonstrate that our proposed approach can find a better trade-off between accuracy and interpretability when compared to its non-Bayesian counterparts, i.e., BATrees (Breiman and Shang 1996) and node harvest (Meinshausen 2010).
- We propose a systematic approach to compare stability of interpretable models and show that the proposed method produces more stable models.

## 2 Motivation

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 to account for model uncertainty in the explanation. Both discussions are accompanied with illustrative examples.

### 2.1 Interpretability as a decision-making problem

Bayesian modeling allows encoding prior information into the prior probability distribution (similarly, one might use regularization in maximum likelihood based inference). This might be tempting to change the prior distribution to favor models that are easier for humans to understand, as has been done in earlier works, using some measure of interpretability. A simple example is to use shrinkage priors in linear regression to find a smaller set of practically important covariates. However, we argue that based on the observation, interpretability is not an inherent characteristic of data generation processes. The approach can be misleading and results in leaking user preferences about interpretability into the model of the data generation process.

We suggest to separate 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 accurately as possible, without restricting it to be interpretable. Interpretability is introduced in the second stage by building an interpretable proxy to explain the behavior of the predictive model. We consider the second step as a decision-making problem, where the task 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 $p(\theta \mid \mathcal{M})$ and $p(\theta \mid \mathcal{M}, \mathcal{I})$. Here, $\theta$ 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 $\theta$ could correspond to using some subset of input variables in a predictive model). Clearly, the posterior distributions $p(\theta \mid \mathcal{D}, \mathcal{M})$ and $p(\theta \mid \mathcal{D}, \mathcal{M}, \mathcal{I})$ (and their corresponding posterior predictive distributions) 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 $p(\theta \mid \mathcal{D}, \mathcal{M}, \mathcal{I})$ provides a reasonable quantification of our knowledge of $\theta$ given the observations $\mathcal{D}$, or that, $p(\tilde{y} \mid \mathcal{D}, \mathcal{M}, \mathcal{I})$ 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 $\mathcal{M}$.

Lage et al. (2018) 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). In practice, the process requires the set of models for step 1 to contain interpretable models, which means that there is still the possibility of leaking user preferences for interpretability into the knowledge about the data generation process. This may lead to an unreasonable trade-off between accuracy and interpretability.

#### 2.2.1 Illustrative example

We give an example to illustrate the effect of adding interpretability constraints to the prior distribution when these constraints do not match data generating process. For simplicity, we define a single interpretability constraint which is over the structure of the model:

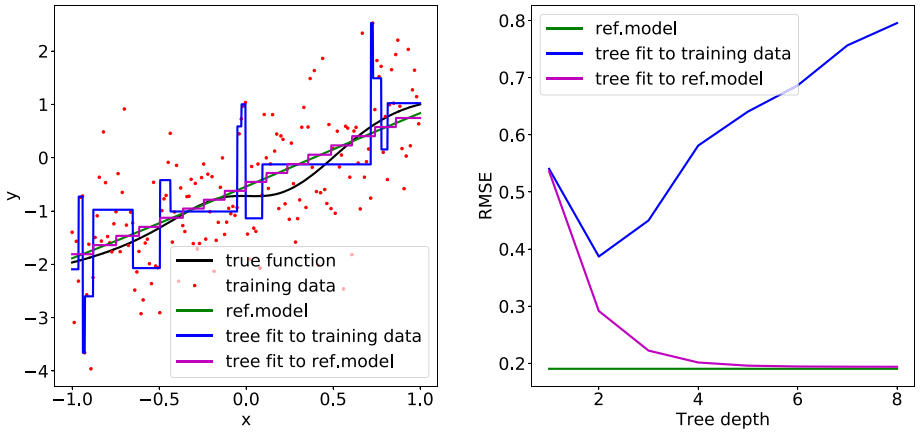



**Fig. 1** Left: The reference model (green) is a highly predictive non-interpretable model that approximates the true function (black) well. The interpretable model fitted to the reference model (magenta) approximates the reference model (and consequently the true function) well, while the interpretable model fitted to the training data (blue) fails to approximate the predictive behavior of the true function. Right: Root Mean Squared Errors (RMSE) compared to the true underlying function as the tree depth is varied. By increasing the complexity of the interpretable model (decreasing its interpretability), predictive performance of the reference model and its corresponding interpretable model converge; the interpretable model overfits to the reference model (Color figure online)

regression tree with a fixed depth of 4. The interpretability prior approach corresponds to fitting an interpretable model with the above constraint directly to the training data. In the alternative approach, first a reference model is fitted to the data, and then the reference model is approximated with a proxy model that satisfies the interpretability constraint, using the interpretability utility introduced in Sect. 3. For simplicity of visualization, we use a one-dimensional smooth function as the data-generating process, with Gaussian noise added to observations (Fig. 1: left, black curve and red dots). Regression tree is a piece-wise constant function which does not correspond to the true prior knowledge about the ground-truth function, i.e. being a 1D smooth function. A Gaussian process with the MLP kernel function is used as a reference model for the two-stage approach (Fig. 1: left, magenta).

The regression tree of depth 4 fitted directly to the data (blue line) overfits and does not give an accurate representation of the underlying data generation process (black line). The two-stage approach, on the other hand, gives a clearly better representation of the smooth, increasing function. This is because 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 becomes a more accurate (but possibly less easy to interpret) representation of the reference model as shown in Fig. 1: right.

### 2.3 Interpreting uncertainty

In many applications, such as medical treatment effectiveness prediction (Sundin et al. 2018), knowing the uncertainty in the prediction is important. Any explanation of the

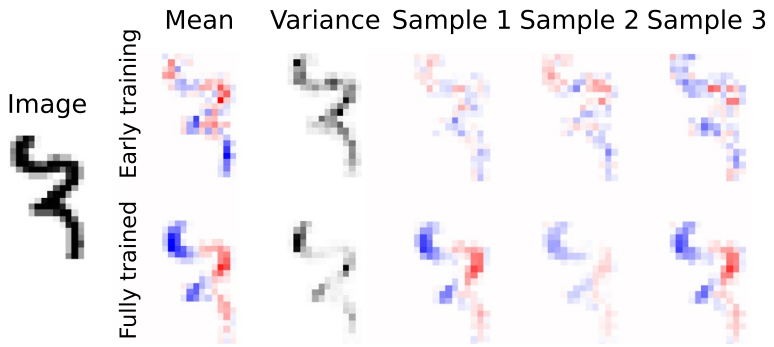



**Fig. 2** Mean explanation, explanation variance, and three sample explanations for a convolutional neural network 3-vs-8 MNIST-digit classifier early in the training and fully trained. Colored pixels show linear explanation model weights, with red being positive for 3 and blue for 8 (Color figure online)

predictive model should also provide insight about the uncertainties and their sources. The posterior predictive distribution of the reference model contains both the aleatoric (predictive uncertainty given the model parameter, i.e., noise in the output) and the epistemic uncertainty (uncertainty about model parameters). We can capture both of these into our interpretable model, since it is fitted to match the reference posterior predictive distribution. The former is captured by conditioning the interpretable model on a posterior draw from the reference model, while the latter is captured by fitting the interpretable model on multiple posterior draws. Details will be given later in Sect. 3. Here, we demonstrate with an example that the proposed method can provide useful information about model uncertainty.

### 2.3.1 Practical example

We demonstrate uncertainty interpretation in locally explaining a prediction of a Bayesian deep convolutional neural network in the MNIST dataset of images of digits (LeCun et al. 1998). The reference model is classifying between digits 3 and 8. We use the Bernoulli dropout method (Gal and Ghahramani 2016a, b), with a dropout probability of 0.2 and 20 Monte Carlo samples at test time, to approximate Bayesian neural network inference (the posterior predictive distribution). Logistic regression is used as the interpretable model family.[1]

Since we are classifying images, we can conveniently visualize the explanation model. Figure 2 shows visually the logistic regression weights for a digit, comparing the reference model in an early training 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 illustrates the epistemic uncertainty. For example, the reference model in early training has not yet been able

[1] The optimization of the interpretable model follows the general framework explained in Sect. 3, 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.

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

Here we first explain the procedure to obtain interpretability utility for regression tasks. The case of classification models is similar and is explained in Sect. 3.3.

### 3.1 Regression models

Let $\mathcal{D} = \left\{\left(\boldsymbol{x}_n, y_n\right)\right\}_{n=1}^{N}$ denote a training set of size $N$, where $\boldsymbol{x}_i = \left[x_{i1}, \ldots, x_{id}\right]^T$ is a $d$-dimensional feature vector and $y_i \in \mathbb{R}$ is the target variable. Assume that a highly predictive (reference) model $\mathcal{M}$ is fitted to the training data without concerning interpretability constraints. Denote the likelihood of the reference model by $p(y \mid \boldsymbol{x}, \boldsymbol{\theta}, \mathcal{M})$ and the posterior distribution $p(\boldsymbol{\theta} \mid \mathcal{D}, \mathcal{M})$. Posterior predictive distribution of the reference model obtains as $p(\tilde{y} \mid \mathcal{D}) = \int_\theta p(\tilde{y} \mid \theta) p(\theta \mid \mathcal{D}) d\theta$. Our goal is to find an interpretable model that best explains the behavior of the reference model locally or globally. We introduce an interpretable model family $\mathcal{T}$ with likelihood $p(y \mid \boldsymbol{x}, \boldsymbol{\eta}, \mathcal{T})$ and posterior $p(\boldsymbol{\eta} \mid \mathcal{D}, \mathcal{T})$, belongs to a probabilistic model family with parameters $\boldsymbol{\eta}$. The best interpretable model is the one closest to the reference model prediction–wise, and at the same time easily interpretable. To measure the closeness of the predictive behavior of the interpretable model to the reference model, we compute the Kullback-Leibler (KL) divergence between their posterior predictive distribution. Assuming we want to locally interpret the reference model, and following simplifications of Piironen et al. (2018) for computing the KL divergence of posterior predictive distributions, the best interpretable model can be found by optimizing the following utility function:

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

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

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

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

for $l = 1, \ldots, L$ posterior draws from $p(\boldsymbol{\theta} \mid \mathcal{D}, \mathcal{M})$. Equation 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 that minimizes the objective function. It has been shown in Piironen et al. (2018) 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:

$$\arg\max_{\boldsymbol{\eta}} \frac{1}{S}\sum_{s=1}^{S} E_{\tilde{y}_s|z_s,\boldsymbol{\theta}^{(l)}}\left[\log p\left(\tilde{y}_s \mid z_s, \boldsymbol{\eta}\right)\right]. \tag{3}$$

Using this equivalent form and by adding the complexity penalty term, the interpretability utility obtains as

$$\arg\max_{\boldsymbol{\eta}} \frac{1}{S}\sum_{s=1}^{S} E_{\tilde{y}_s|z_s,\boldsymbol{\theta}^{(l)}}\left[\log p\left(\tilde{y}_s \mid z_s, \boldsymbol{\eta}\right)\right] - \Omega(\boldsymbol{\eta}). \tag{4}$$

The complexity penalty term should be chosen to match the resulting model; possible options are the number of leaf nodes for decision trees, number of rules and/or size of the rules for 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 CART models with tree size (the number of leaf nodes) as the measure of interpretability. With this assumption, similar to the illustrative example in Sect. 2.2.1, the interpretability constraint is defined 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 (Chipman et al. 1998). The interpretability prior approach corresponds to fitting a CART model to the training data, i.e. samples drawn from the neighborhood distribution of $\boldsymbol{x}$.

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 vector $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_b)$ that associates the parameter value $\phi_i$ with the $i$th terminal node. If $\boldsymbol{z}$ lies in the region corresponding to the $i$th terminal node, then $y \mid \boldsymbol{z}, \boldsymbol{\eta}$ has distribution $f(y \mid \phi_i)$, where $f$ denotes a parametric probability distribution with parameter $\phi_i$. 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 $l$th draw from the posterior of $\boldsymbol{\theta}$ has the form

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

where $\boldsymbol{y}_i \equiv (y_{i1}, \ldots, y_{in_i})$ denotes the set of the $n_i$ observations assigned to the partition generated by the $i$th terminal node with parameter $\phi_i^{(l)}$, and $\boldsymbol{Z}$ is the matrix of all the $z_s$. For regression problems, assuming a mean-shift normal model for each terminal node $i$,[2] the likelihood of the interpretable model is defined as

$$f\left(\boldsymbol{y} \mid \boldsymbol{\phi}^{(l)}\right) = \prod_{i=1}^{b}\prod_{j=1}^{n_i} N\left(y_{ij} \mid \mu_i^{(l)}, \sigma^{2(l)}\right), \tag{6}$$

where $\boldsymbol{\phi}^{(l)} = (\boldsymbol{\mu}^{(l)} = \{\mu_i^{(l)}\}_{i=1}^{b}, \sigma^{2(l)})$. With this formulation, the task of finding an interpretable proxy to the reference model $M$ is reformed to find a tree structure $T$ with parameters

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

$\phi^{(l)}$ such that its predictive performance is as close as possible to $M$, while being as interpretable as possible. Interpretability is measured by the complexity term $\Omega$.

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

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

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

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

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

**Algorithm 1:** Decision–theoretic approach for local interpretability in the Bayesian framework

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

```
/* REFERENCE MODEL CONSTRUCTION                                   */
```
fit the Bayesian predictive model to $\mathcal{D}$ without concerning interpretability constraints;
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 $\bar{y}_s$ and variance $\sigma_s^2$ 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

When fitting a global interpretable model, instead of drawing samples from $\pi_{\boldsymbol{x}}$, we 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 CART models.

### 3.2 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) (Su et al. 2004).

### 3.2.1 Growing a large tree

Given the training data,[3] 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 $\mu_i$ given observations associated with the node, and then compute the variance parameter of the tree given $\{\mu_i\}_{i=1}^{b}$:

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

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

   The log-likelihood score of the node is then computed, 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 in 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.
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. $r^* = \max_{r_j^*,\, j=1,\ldots,d} \Delta_{\left(r_j^*, x_j, i\right)}$.
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.

### 3.2.2 Pruning

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

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

Pruning is done iteratively; in each iteration $i$, the internal node $h$ that minimizes $\alpha = \frac{\left(C(h) - C(T_i)\right)}{\left(|\text{ leaves } (T_h)| - 1\right)}$ 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 $T_h$

[3] 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.

denotes the subtree with $h$ as its root. The output is a sequence of decision trees and a sequence of $\alpha$ values. The best $\alpha$ and its corresponding subtree are selected using 5-fold cross-validation.

### 3.3 Classification models

For classification problems, assuming the CART models as the interpretable model family, the form of the interpretability utility is the same as Equation 4 except that the likelihood of the interpretable model follows a multinomial distribution with the following log-likelihood:

$$\mathcal{L} = \sum_{i=1}^{b} \sum_{j=1}^{n_i} \sum_{k=1}^{K} I\left(y_{ij} \in C_k\right) \log p_{ik} \quad \text{s.t.} \quad p_{ik} \geq 0, \quad \sum_{k=1}^{K} p_{ik} = 1 \tag{10}$$

where $I(y_{jk} \in C_k)$ is the indicator function determining wheter or not the $j$th sample of the $i$th node belongs to the $k$th category assuming that there are in total $K$ categories. The $p_{ik}$ denote the probability of the occurrence of the $k$th category in the $i$th terminal node and the set of parameters are $\phi = \{\boldsymbol{p}_i = \left(p_{i1}, \ldots, p_{ik}\right)\}_{i=1}^{b}$. Therefore, the final form of the interpretability utility for Bayesian classification models is

$$\arg\max_{\boldsymbol{\eta}} \frac{1}{S} \sum_{i=1}^{b} \sum_{k=1}^{K} n_k \log p_{ik} + \Omega(T) \quad \text{s.t.} \quad p_{ik} \geq 0, \quad \sum_{k=1}^{K} p_{ik} = 1 \tag{11}$$

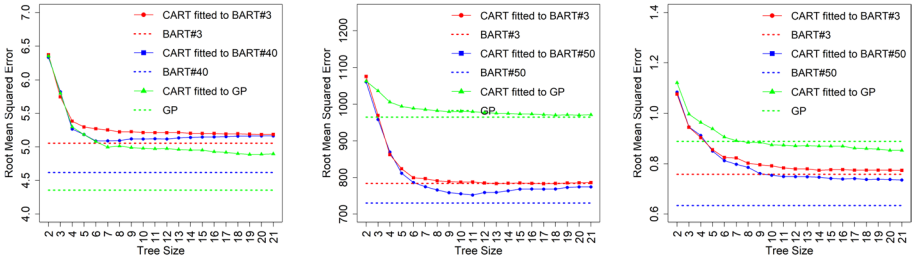
where $\boldsymbol{\eta} = (T, \phi)$ and $n_k = \sum_{j=1}^{n_i} I\left(y_{jk} \in C_k\right)$. The optimization approach is again similar to the process explained in Sect. 3.2 with the difference that the maximum likelihood estimate of the parameters of each node $i$ obtains as $\hat{p}_{ik} = \frac{n_k}{n_i}$. Finally, the log-likelihood score of each node $i$ is determined by $\mathcal{L}_i = \sum_{k=1}^{K} n_k \log \hat{p}_{ik}$.

### 3.4 Connection with local interpretable model-agnostic explanation (LIME)

LIME (Ribeiro et al. 2016) is a prediction-level interpretation approach that fits a sparse linear model to the black-box model's prediction via drawing samples from the neighborhood of the data point to be 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). This is achieved by combining the idea of LIME with the idea of projection predictive variable selection (Piironen et al. 2018). The approach is able to handle different types of predictions (continuous valued, class labels, counts, censored and truncated data, etc.) and interpretations (model-level or prediction-level) 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 Peltola (2018).

## 4 Experiments

We demonstrate the efficacy of the proposed approach through experiments on several real-world data sets. Section 4.1 discusses the experiments related to global intepretation. We first investigate the effect of reference models with different predictive powers




**(a)** Body fat **(b)** Baseball **(c)** Auto risk

**Fig. 3** Effect of reference models with different predictive performance on the performance of the interpretable models fitted to them. More accurate reference models result in interpretable models with higher predictive performance. The values of "ntree" used for the BART models are shown by "#"

on the performance of the final interpretable model. Secondly, we compare our approach with the interpretability prior alternative, of fitting directly an interpretable model to the data, in terms of their capability to trade off between accuracy and interpretability. We also compare the performance of our approach with non-Bayesian counterparts introduced in Sect. 1. Further, we investigate the stability of our approach and the interpretability prior approach. Section 4.2 examines local interpretation, where we compare our approach with LIME. Our codes and data are available online at https://github.com/homayunafra/Decision_Theoretic_Approach_for_Interpretability_in_Bayesian_Framework.

## 4.1 Global interpretation

### 4.1.1 Data

In our experiments, we use the following data sets: body fat (Johnson 1996), baseball players (Hoaglin and Velleman 1995), auto risk (Kibler et al. 1989), bike rental (Fanaee-T and Gama 2014), auto mpg (Quinlan 1993), red wine quality (Cortez et al. 2009), and Boston housing (Harrison Jr and Rubinfeld 1978).

Each data set is divided into training and test set containing 75% and 25% of samples, respectively.

### 4.1.2 Effect of reference model

The purpose of this test is to evaluate how the predictive power of the reference model affects the performance of the interpretable model when it is used to globally explain the reference model. Three data sets are adopted for this test: body fat, baseball players, and auto risk. Furthermore, three reference models with different predictive powers are adopted: two BART models, and a Gaussian process (GP).

For the BART models, we used the BART package in R with two different values for the "ntree" (number of trees) parameter. For one model, "ntree" is set to the value that gives the highest predictive performance on the validation set (blue dotted line in Fig. 3), while for another one, this parameter is set to 3, a low value, which gives poor predictive

performance (red-dotted line in Fig. 3). The rest of the parameters are set to their default values except "nskip" and "ndpost", which are set to 2000 and 4000, respectively. For the BART models, mean of the predictions of the posterior draws is used as their output. For the GP (green-dotted line in Fig. 3), "Matern52" is used as the kernel with variance and length scales obtained by cross-validation over a small grid of values.[4]

CART models are used as the interpretable model family. The size of the tree, i.e., the total number of leaf nodes, is used as the measure of interpretability (Bastani et al. 2018; Hara and Hayashi 2018).

Figure 3 demonstrates the results, which are averaged over 50 runs. The difference in the predictive performance of the interpretable models fitted to different reference models suggests that using more accurate reference models (BART in Baseball and Auto risk data sets, and GP in Body fat data set) can generate more accurate interpretable models as well. This is expected since by the performance of the interpretable model converges to the performance of the reference model; therefore the interpetable model will be more accurate when fitted to a more accurate reference model. The gap between the predictive performance of the interpretable models and their corresponding reference models is due to the limited predictive capability of the interpretable model. For some tasks, this gap can be made narrower by increasing the complexity of the interpretable model, while for others, a different family of interpretable models may be needed.

Finally, in Fig. 3c, the performance of the interpretable model fitted to the GP reference model is better than the reference model itself, for some complexities. This may be because of different extrapolation behavior of CART and GP. In the high-dimensional space, the test data may be outside the support of the training data; thus, extrapolation behavior matters. Simpler models can make more conservative extrapolations which may be helpful in this case.

#### 4.1.3 Interpretability prior versus interpretability utility

In this subsection, we compare our approach with the interpretability prior approach, in terms of the capability of the methods to trade off between accuracy and interpretabilityr. BART is used as the reference model, and CART is used as the interpretable model family. The interpretability prior approach fits a CART model directly to the training data where the prior assumption is that CART models are simple to interpret. On the other hand, our approach fits the CART model to the reference model, by optimization of the interpretability utility.

Figure 4 demonstrates the results using all the data sets introduced in Sect. 4.1.1. The results are averaged over 50 runs. It can be seen that the most accurate models with any level of complexity (interpretability) are obtained with our proposed approach.[5]

[4] This may not be the best setting for the GP. We did not attempt to optimize that since our objective is not to compare the performance of GP with BART, but instead to compare the performance of the interpretable models fitted to them.

[5] The single exception happened in auto risk data set with tree size of 3.

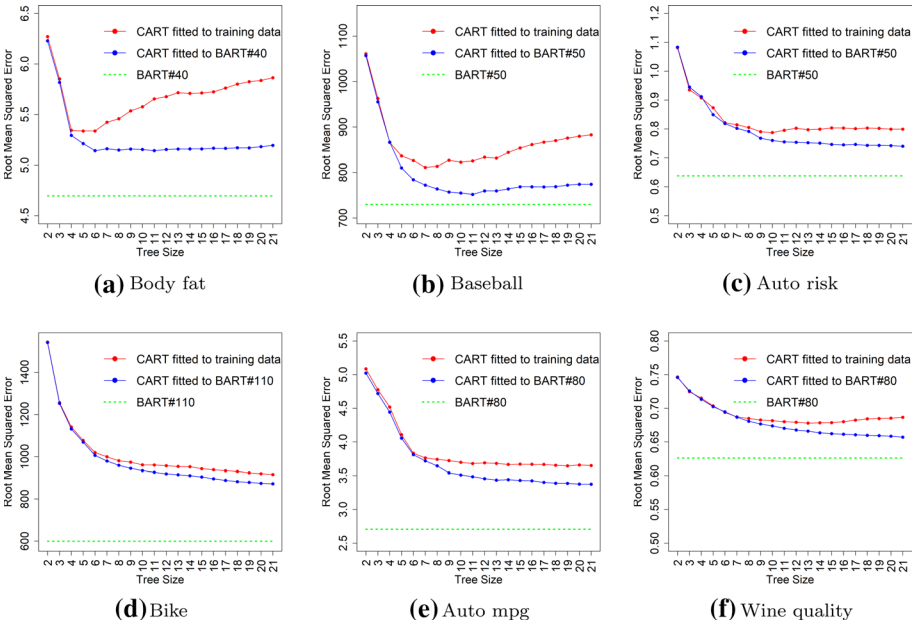



**Fig. 4** Comparison of interpretability prior (red) and interpretability utility (blue) approach in trading off between accuracy and interpretability when using CART as explainable models and BART as reference model. The values of "ntree" used for the BART models are shown by "#" (Color figure online)

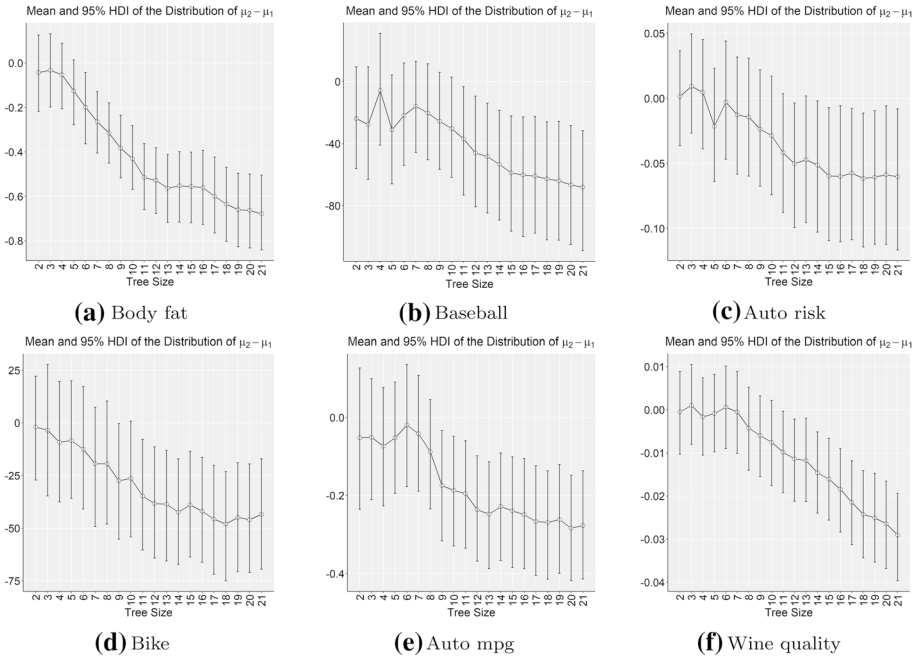



**Fig. 5** Results of Bayesian *t*-test that shows the mean and 95% highest density interval of the distribution of difference of means. $\mu_1$ and $\mu_2$ refer to the means of the distributions obtained for the interpretability prior and interpretability utility approach, respectively

**Table 2** Comparison of our approach with two non-Bayesian counterparts: node harvest and BATrees

| | RMSE | | | | Size | |
|---|---|---|---|---|---|---|
| | NodeHarvest | BATrees | Our approach | | Node Harvest | BATrees |
| | | | Size = 10 | Size = 15 | | |
| Body fat | **5.14 ± 0.37** * | 5.26 ± 0.42 (4.84 ± 0.38) | 5.15 ± 0.36 | 5.16 ± 0.37 | 34.4 ± 6.5 | 19.5 ± 5.7 |
| Baseball | 783.1 ± 72.5 * | 1020.1 ± 69.9 (755.9 ± 67.8) | **755 ± 81.1** | 768.6 ± 92.5 | 35.5 ± 5.4 | 15.3 ± 6.5 |
| Auto risk | 0.78 ± 0.06 * | 0.79 ± 0.1 (0.64 ± 0.09) | 0.76 ± 0.1 | **0.75 ± 0.1** | 37.2 ± 12.4 | 20.4 ± 4.8 |
| Bike | 913.3 ± 67.2 * | 907.4 ± 71.5 (681.8 ± 60.7) | 934.4 ± 68.9 | **904.2 ± 60.6** | 47.6 ± 9.1 | 33.5 ± 6.3 |
| Auto mpg | 3.47 ± 0.33 * | **3.39 ± 0.27** (2.83 ± 0.34) | 3.51 ± 0.31 | 3.43 ± 0.33 | 58.8 ± 8.1 | 28 ± 4.4 |
| Wine quality | 0.67 ± 0.03 * | **0.66 ± 0.02** (0.6 ± 0.02) | 0.67 ± 0.02 | **0.66 ± 0.02** | 51.3 ± 7 | 43 ± 13.2 |

The RMSE values are shown in terms of mean ± SD. Sizes are shown in terms of mean ± sd of number of leaf nodes for BATrees and number of nodes with non-zero coefficients for node harvest. For BATrees, the predictive performance of its reference model is shown in the parantheses. For node harvest, it was not possible to obtain the performance of the reference model since the R package provides no means for that. Best RMSE values are bolded

To test the significance of the differences in the results, we performed the Bayes *t*-test (Kruschke 2013). The approach works by building up a complete distributional information for the mean and standard deviation of each group[6] and constructing a probability distribution over their differences using MCMC estimation. From this distribution, the mean credible value as the best guess of the actual difference and the 95% Highest Density Interval (HDI) as the range were the actual difference is with 95% credibility are shown in Fig. 5. When the 95% HDI does not include zero, there is a credible difference between the two groups. As shown in the figure, for all data sets and for highly interpretable models (highly inaccurate), the difference between the two approaches 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, and both approaches provide almost equally poor performance. However, by increasing the complexity (equivalently decreasing interpretability) to a reasonable level, we see that the differences of the two approaches become significant for all data sets.

Finally, we further compared the performance of our proposed approach with two non-Bayesian counterparts, i.e., BATrees (Breiman and Shang 1996) and node harvest (Meinshausen 2010). BATrees employs a single decision tree that best mimics the predictive behavior of a tree ensemble. Random forest is used as the reference model for BATrees. Node harvest simplifies a tree ensemble, i.e., random forest, by use of the shallow parts of

[6] For each tree size, there are two groups of 50 RMSE values: one for the interpretability prior approach, and one interpretability utility approach.

**Table 3** Bootstrap instability values in the form of mean ± std

| Interpretability | Body fat | Baseball | Auto risk | Bike | Auto mpg | Wine quality |
|---|---|---|---|---|---|---|
| Prior | $0.71 \pm 0.11$ | $0.84 \pm 0.08$ | $\mathbf{0.79 \pm 0.19}$ | $\mathbf{0.68 \pm 0.09}$ | $0.70 \pm 0.14$ | $0.74 \pm 0.11$ |
| Utility | $\mathbf{0.62 \pm 0.19}$ | $\mathbf{0.83 \pm 0.07}$ | $0.81 \pm 0.16$ | $\mathbf{0.68 \pm 0.09}$ | $\mathbf{0.64 \pm 0.14}$ | $\mathbf{0.70 \pm 0.13}$ |

Best values are bolded

the trees. We chose these approaches with random forest as their black-box model for the comparison for two reasons:

1. to the best of our knowledge, there is no approach particularly established for explaining Bayesian tree ensemble models, i.e., BART. The approach of Hara and Hayashi (2018) can be modified for this objective; however, it requires inputs in terms of rules extracted from the tree ensemble, which calls for considerable of extra work.
2. BART can be considered as a Bayesian interpretation of random forest, and it has been revealed with some synthetic and real-data experiment that they have similar predictive performances (Hernández et al. 2018).

For node harvest, we used the R implementation with default setting. For BATrees, the Python implementation in Hara and Hayashi (2018) is used with the depth of BATrees chosen from $\{3, 4, 5, 6\}$ using 5-fold cross validation. The measure of complexity for node harvest is the total number of nodes with non-zero coefficients.

Table 2 demonstrates the results. The results are averaged over 50 runs with the same seed value used for the experiments in Fig. 4. The table shows that our proposed approach attained much better trade-off between accuracy and interpretability compared to BATrees and node harvest. For 4 data sets, our approach provides higher accuracies even with smaller sizes. For the rest, still our approach provides comparable predictive performance with a complexity of about half of the complexities of node harvest and BATrees. The differences between the bolded RMSE values with the rest of the RMSEs in Baseball, Auto risk and Wine quality data sets are significant using the Bayes $t$-test, while for other data sets the differences are not significant. According to the table, node harvest tends to generate more complex surrogate models. This is expected since in node harvest, the surrogate model is still an ensemble of shallow trees.

#### 4.1.4 Stability analysis

The goal of interpretable ML is to provide a comprehensive explanation of the predictive behavior of the black-box model to the decision maker. However, perturbation in the data or adding new samples may affect the learned interpretable model and lead to a very different explanation. This instability can cause problems for decision makers. Thereby, it is important to evaluate the stability of different interpretable ML approaches. For this objective, we propose the following procedure for stability analysis of interpretable ML approach.

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 Briand et al. (2009). Assuming we are given two

**Table 4** Comparison of the local fidelity of LIME and Interpertability utility when being used to explain predictions of BART

| Dataset | LIME | Interpretability Utility |
|---|---|---|
| Boston housing | 4.86 | **2.94** |
| Auto risk | 0.014 | **0.010** |

Best values are bolded

regression trees $T_1$ and $T_2$, for each internal node $t$, the similarity of the trees at node $t$ is computed by

$$S^t_{(1,2)} = I^t_{k=k'}\left(1 - \frac{\mid \delta^t_1 - \delta^t_2 \mid}{\text{range}\,(X_k)}\right) \tag{12}$$

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, $\delta^t_1$ and $\delta^t_2$ 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 $d\big(T_1, T_2\big) = 1 - \sum_{t\in \text{ internal \_ nodes}} q^t S^t_{(1,2)}$ 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 3 compares the two approaches over the data sets introduced in Sect. 4.1.1. The interpretability utility approach generated on average more stable models for most data sets; however, drawing a general conclusion is not possible because except body fat, for the rest of the data sets, the differences are not significant according to the Bayes $t$-test.

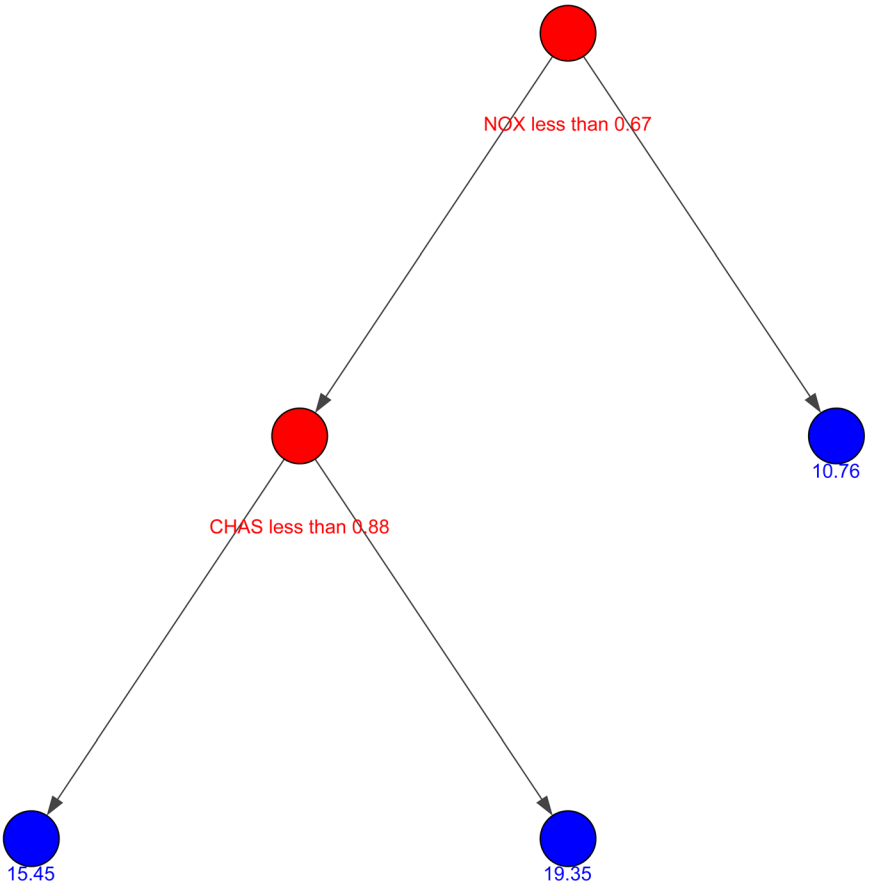
### 4.2 Local interpretation

We next demonstrate the ability of the proposed approach in locally interpreting the predictions of a Bayesian predictive model. BART[7] is used as the black box model and CART is used 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] which is a commonly used baseline for local interpretation approaches. Decision trees obtained by our approach to locally explain predictions of the BART model, used on average 2.03 and 2.4 features for the Boston housing and the auto risk data sets, respectively. Therefore, to maximize comparability, we set the feature selection approach of LIME to ridge regression and select the 2 features with the highest absolute weights to be used in the explanation.[9] We use the standard quantitative metric for local fidelity: $\mathbb{E}_x\big[\text{ loss }\big(\text{ interp}_x(\boldsymbol{x}),\ \text{pred}\,(\boldsymbol{x})\big)\big]$ 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 pred $(\boldsymbol{x})$

[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 Auto risk data sets.




| | CRIM | ZN | INDUS | CHAS | NOX | RM | AGE | DIS | RAD | TAX | PTRATIO | B | LSTAT | $y_{ref}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $x_{test}$ | 22.0511 | 0 | 18.1 | 0 | 0.74 | 5.818 | 92.4 | 1.8662 | 24 | 666 | 20.2 | 391.45 | 22.11 | 10.5 |

**Fig. 6** 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

refers to the prediction of the black-box model for $\boldsymbol{x}$. We used locally weighted square loss as the loss function with $\pi_x = \mathcal{N}(\boldsymbol{x}, \sigma^2 \boldsymbol{I})$ 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 4 shows that our approach produces more accurate local explanation for both data sets. Figure 6 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 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. This is obtained by first fitting a Bayesian predictive model which does not compromise accuracy, termed as a reference model, to the training data, and then project the information in the predictive distribution of the reference model to an interpretable model. The approach is model agnostic, implying that neither the reference model nor the interpretable model is restricted to a certain model. In the current implementation, the interpretable model, i.e., CART, is not a Bayesian predictive model; however, it is straightforward to extend the formulation to the case where a Bayesian predictive model, e.g., Bayesian CART (Denison et al. 1998), is used as the interpretable model. This remains for future. The approach also allows accounting for model uncertainty in the explanations. Through experiments, we demonstrated that the proposed approach outperforms the alternative approach of restricting the prior, in terms of accuracy, interpretability and stability. Furthermore, we showed that the proposed approach performs comparable to non-Bayesian counterparts such as BATrees and node harvest even when they have higher complexities (equivalently less interpretability).

**Acknowledgements** This work was supported by the Academy of Finland (Flagship programme: Finnish Center for Artificial Intelligence FCAI, 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.

**Funding** Open access funding provided by Aalto University.



## References

Bastani, H., Bastani, O., & Kim, C. (2018). Interpreting predictive models for human-in-the-loop analytics. arXiv preprint arXiv:1705.08504 (pp. 1–45).

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Breiman, L., & Shang, N. (1996). *Born again trees*. Technical report, University of California, Berkeley, Berkeley, CA (Vol. 1, p. 2).

Briand, B., Ducharme, G. R., Parache, V., & Mercat-Rommens, C. (2009). A similarity measure to assess the stability of classification trees. *Computational Statistics & Data Analysis*, *53*(4), 1208–1217.

Chipman, H. A., George, E. I., & McCulloch, R. E. (1998). Bayesian CART model search. *Journal of the American Statistical Association*, *93*(443), 935–948.

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, *4*(1), 266–298.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, *47*(4), 547–553.

Craven, M., & Shavlik, J. W. (1996). Extracting tree-structured representations of trained networks. In *Advances in neural information processing systems* (pp. 24–30).
Deng, H. (2019). Interpreting tree ensembles with intrees. *International Journal of Data Science and Analytics*, *7*(4), 277–287.
Denison, D. G. T., Mallick, B. K., & Smith, A. F. M. (1998). A Bayesian CART algorithm. *Biometrika*, *85*(2), 363–377.
Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.
Du, M., Liu, N., & Hu, X. (2018). Techniques for interpretable machine learning. arXiv preprint arXiv :1808.00033.
Fanaee-T, H., & Gama, J. (2014). Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, *2*(2–3), 113–127.
Gal, Y., & Ghahramani, Z. (2016a). Bayesian convolutional neural networks with Bernoulli approximate variational inference. In *4th international conference on learning representations (ICLR) workshop track*.
Gal, Y., & Ghahramani, Z. (2016b). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd international conference on machine learning* (pp. 1050–1059).
Guo, J., Riebler, A., & Rue, H. (2017). Bayesian bivariate meta-analysis of diagnostic test studies with interpretable priors. *Statistics in Medicine*, *36*(19), 3039–3058.
Hara, S., & Hayashi, K. (2018). Making tree ensembles interpretable: A Bayesian model selection approach. In *International conference on artificial intelligence and statistics* (pp. 77–85).
Harrison, D, Jr., & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, *5*(1), 81–102.
Hernández, B., Raftery, A. E., Pennington, S. R., & Parnell, A. C. (2018). Bayesian additive regression trees using Bayesian model averaging. *Statistics and Computing*, *28*(4), 869–890.
Hoaglin, D. C., & Velleman, P. F. (1995). A critical look at some analyses of major league baseball salaries. *The American Statistician*, *49*(3), 277–285.
Johnson, R. W. (1996). Fitting percentage of body fat to simple body measurements. *Journal of Statistics Education*. https://doi.org/10.1080/10691898.1996.11910505.
Jung, J., Concannon, C., Shroff, R., Goel, S., & Goldstein, D. G. (2017). Simple rules for complex decisions. arXiv preprint arXiv:1702.04690.
Kibler, D., Aha, D. W., & Albert, M. K. (1989). Instance-based prediction of real-valued attributes. *Computational Intelligence*, *5*(2), 51–57.
Kim, B., Glassman, E., Johnson, B., & Shah, J.. (2015). *ibcm: Interactive Bayesian case model empowering humans via intuitive interaction*. Technical report: MIT-CSAIL-TR.
Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573.
Kuttichira, D. P., Gupta, S., Li, C., Rana, S., & Venkatesh, S. (2019). Explaining black-box models using interpretable surrogates. In *Pacific Rim international conference on artificial intelligence* (pp. 3–15). Springer.
Lage, I., Ross, A. S., Kim, B., Gershman, S. J, & Doshi-Velez, F. (2018). Human-in-the-loop interpretability prior. arXiv preprint arXiv:1805.11571.
Lakkaraju, H., Bach, S. H, & Leskovec, J. (2016). Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, (pp. 1675–1684).
Lakkaraju, H., Kamar, E., Caruana, R., & Leskovec, J. (2019). Faithful and customizable explanations of black box models. In *Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society* (pp. 131–138).
LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., et al. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.
Letham, B., Rudin, C., McCormick, T. H., Madigan, D., et al. (2015). Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, *9*(3), 1350–1371.
Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, *61*(10), 36–43.
Lou, Y., Caruana, R., & Gehrke, J. (2012). Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 150–158).
Meinshausen, N. (2010). Node harvest. *The Annals of Applied Statistics*, *4*(4), 2049–2072.

Peltola, T. (2018). Local interpretable model-agnostic explanations of Bayesian predictive models via Kullback–Leibler projections. arXiv preprint arXiv:1810.02678.
Piironen, J., Paasiniemi, M., & Vehtari, A. (2018). Projective inference in high-dimensional problems: Prediction and feature selection. arXiv preprint arXiv:1810.02406.
Popkes, A.-L., Overweg, H., Ercole, A., Li, Y., Hernández-Lobato, J. M., Zaykov, Y., & Zhang, C. (2019). Interpretable outcome prediction with sparse Bayesian neural networks in intensive care. arXiv preprint arXiv:1905.02599.
Quinlan, J. R. (1993). Combining instance-based and model-based learning. In *Proceedings of the tenth international conference on machine learning* (pp. 236–243).
Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). 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* (pp. 1135–1144). ACM.
Su, X., Wang, M., & Fan, J. (2004). Maximum likelihood regression trees. *Journal of Computational and Graphical Statistics*, *13*(3), 586–598.
Sundin, I., Peltola, T., Micallef, L., Afrabandpey, H., Soare, M., Majumder, M. M., et al. (2018). Improving genomics-based predictions for precision medicine through active elicitation of expert knowledge. *Bioinformatics*, *34*(13), i395–i403.
Ustun, B., & Rudin, C. (2016). Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, *102*(3), 349–391.
Vehtari, A., & Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, *6*, 142–228.
Wang, T. (2018). Multi-value rule sets for interpretable classification with feature-efficient representations. In *Advances in neural information processing systems* (pp. 10835–10845).
Wang, T., Rudin, C., Doshi-Velez, F., Liu, Y., Klampfl, E., & MacNeille, P. (2017). A Bayesian framework for learning rule sets for interpretable classification. *The Journal of Machine Learning Research*, *18*(1), 2357–2393.
Wu, M., Hughes, M. C, Parbhoo, S., Zazzi, M., Roth, V., Doshi-Velez, F. (2018). Beyond sparsity: Tree regularization of deep models for interpretability. In *Thirty-second AAAI conference on artificial intelligence*.
Yang, H., Rudin, C., & Seltzer, M. (2017). Scalable Bayesian rule lists. In *Proceedings of the 34th international conference on machine learning* (Vol. 70, pp. 3921–3930). JMLR.org.
Zhou, Y., & Hooker, G. (2016). Interpreting models via single tree approximation. arXiv preprint arXiv:1610.09036.



## Affiliations

**Homayun Afrabandpey[1] · Tomi Peltola[1] · Juho Piironen[3] · Aki Vehtari[1] · Samuel Kaski[1,2]**

Tomi Peltola
tomi.peltola@tmpl.fi

Juho Piironen
juho.t.piironen@gmail.com

Aki Vehtari
aki.vehtari@aalto.fi

Samuel Kaski
samuel.kaski@aalto.fi

[1] Department of Computer Science, Helsinki Institute for Information Technology HIIT, Aalto University, Espoo, Finland

[2] University of Manchester, Manchester, UK

[3] Curious AI, Helsinki, Finland