Machine Learning (2022) 111:205–242
https://doi.org/10.1007/s10994-021-06121-4

# InfoGram and admissible machine learning

**Subhadeep Mukhopadhyay**[1]

Received: 3 December 2020 / Revised: 27 October 2021 / Accepted: 28 October 2021 /
Published online: 10 January 2022


**Abstract**
We have entered a new era of machine learning (ML), where the most accurate algorithm with superior predictive power may not even be deployable, unless it is *admissible* under the regulatory constraints. This has led to great interest in developing fair, transparent and trustworthy ML methods. The purpose of this article is to introduce a new information-theoretic learning framework (admissible machine learning) and algorithmic risk-management tools (InfoGram, L-features, ALFA-testing) that can guide an analyst to *redesign* off-the-shelf ML methods to be regulatory compliant, while maintaining good prediction accuracy. We have illustrated our approach using several real-data examples from financial sectors, biomedical research, marketing campaigns, and the criminal justice system.

**Keywords** Admissible machine learning · InfoGram · L-Features · Information-theory · ALFA-testing · Algorithmic risk management · Fairness · Interpretability · COREml · FINEml

## 1 Category: fairness, explainability, and algorithm bias

Machine learning (ML) methods are rapidly becoming an essential part of automated decision-making systems that directly affect human lives. While substantial progress has been made toward developing more powerful computational algorithms, the widespread adoption of these technologies still faces several barriers—the biggest one being ensuring adherence to regulatory requirements, without compromising too much accuracy. Naturally, the question arises: how to systematically go about building such regulatory-compliant fair and trustworthy algorithms? This paper offers new statistical principles and

Editors: João Gama, Alípio Jorge, Salvador García

✉ Subhadeep Mukhopadhyay
deep@unitedstatalgo.com

[1] United Analytics and Computational Intelligence Inc and H20.ai, Mountain View, CA, USA

information-theoretic graphical exploratory tools that engineers can use to "detect, mitigate, and remediate" off-the-shelf ML-algorithms, thereby making them *admissible* under appropriate laws and regulatory scrutiny.[1]

## 2 Introduction

First-generation "prediction-only" machine learning technology has served the tech and eCommerce industry pretty well. However, ML is now rapidly expanding beyond its traditional domains into highly regulated or safety-critical areas—such as healthcare, criminal justice systems, transportation, financial markets, and national security—where achieving high predictive-accuracy is often as important as ensuring regulatory compliance and transparency in order to ensure the trustworthiness. We thus focus on developing *admissible machine learning* technology that can balance fairness, interpretability, and accuracy in the best manner possible. How to systematically go about building such algorithms in a fast and scalable manner? This article introduces some new statistical learning theory and information-theoretic graphical exploratory tools to address this question.

**Going beyond "Pure" prediction algorithms**: Predictive accuracy is not the be-all and end-all for judging the 'quality' of a machine learning model. Here is a dazzling example: Researchers at the Icahn School of Medicine at Mount Sinai in New York City found that (Zech et al. 2018; Reardon 2019) a deep-learning algorithm, which showed more than 90% accuracy on the x-rays produced at Mount Sinai, performed poorly when tested on data from other institutions. Later it was found that "the algorithm was also factoring in the odds of a positive finding based on how common pneumonia was at each institution—not something they expected or wanted." This sort of unreliable and inconsistent performance can be clearly dangerous. As a result of these safety concerns, despite lots of hype and hysteria around AI in imaging, only about 30% of radiologists are currently using machine learning (ML) for their everyday clinical practices (Allen et al. 2021). To apply machine learning appropriately and safely– especially when human life is at stake–we have to think beyond predictive accuracy. The deployed algorithm needs to be comprehensible (by end-users like doctors, judges, regulators, researchers, etc.) in order to make sure it has learned *relevant and admissible features* from the data, which is meaningful in light of investigators' domain knowledge. The fact of the matter is, an algorithm that is solely focused on *what* is learned, without reasoning *how* it learned what it has learned, is not intelligent enough. We next expand on this issue using two real data applications.

**Admissible ML for industry**: Consider the UCI Credit Card data (discussed in more details in Sec 3.2.3), collected in October 2005, from an important Taiwan-based bank. We have records of $n = 30,000$ cardholders. The data composed of a response variable $Y$ denoting: default payment status (Yes = 1, No = 0), along with $p = 23$ predictor variables (e.g., gender, education, age, history of past payment, etc.). The goal is to accurately predict the probability of default given the profile of a particular customer.

On the surface, this seems to be a straightforward classification problem for which we have a large inventory of powerful algorithms. Yeh and Lien (2009) performed an exhaustive comparison between six machine learning methods (logistic regression, K-nearest neighbor, neural net, etc.) and finally selected the neural network model, which attained

[1] This article is written for the Special Issue on 'Foundations of Data Science'

83% accuracy on a 80-20 train-test split of the data. However, traditionally build ML models are not deployable, unless it is *admissible* under the financial regulatory constraints[2] (Wall 2018), which demand that (i) the method should not discriminate people on the basis of protective features[3], here based on `gender` and `age`; and (ii) The method should be simpler to interpret and transparent (compared to those big neural-nets or ensemble models like random forest and gradient boosting).

To improve fairness, one may remove the sensitive variables and go back to business as usual by fitting the model on the rest of the features–known as 'fairness through unawareness.' Obviously this is not going to work because there will be some proxy attributes (e.g, zip code or profession) that share some degree of correlation (information-sharing) with race, gender, or age. These proxy variables can then lead to the same unfair results. It is not clear how to define and detect those proxy variables to mitigate hidden biases in the data. In fact, on a recent review by Chouldechova and Roth (2020) on algorithmic fairness, the authors forthrightly stated

> *But despite the volume and velocity of published work, our understanding of the fundamental questions related to fairness and machine learning remain in its infancy.*

Currently, there exists no systematic method to directly construct an admissible algorithm that can mitigate bias. To quote a real practitioner of a reputed AI-industry: "I ran 40,000 different random forest models with different features and hyper-parameters to search a fair model." This ad-hoc and inefficient strategy could be a significant barrier for an efficient large-scale implementation of admissible AI technologies. Figure 1 shows a fair and shallow tree classifier with four decision nodes, which attains 82.65% accuracy; this was built in a completely automated manner without any hand-crafted manual tuning. Section 2 will introduce the required theory and methods behind our procedure. Nevertheless, this simple and transparent anatomy of the final model makes it easy to convey *which* are the key drivers of the model: variables `Pay_0` and `Pay_2`[4] are the most important indicators to default. These variables have two key characteristics: they are highly predictive and at the same time safe to use in the sense that they share very little predictive information with the sensitive attributes age and gender, and for that reason, we call them *admissible* features. The model also convey *how* the key variables impacting credit risk: the simple decision tree shown in Fig. 1 is fairly self-explanatory, and its clarity facilitates an easy explanation of the predictions.

**Admissible ML for science**: Legal requirement is not the only reason why we want to build admissible ML. In scientific investigations, it is important to know whether the deployed algorithm helps researchers to better understand the phenomena by refining their "mental model." Consider, for example, the prostate cancer data where we have $p = 6033$ gene expression measurements from 52 tumor and 50 normal specimens. Figure 2 shows a 95% accurate classification model for prostate data with only two "core" driver genes! This compact model is admissible in the sense that it confers the following benefits: (i) it identifies a two-gene signature (composed of gene-1627 and gene-2327) as the top factor associated with prostate cancer. They are *jointly* overexpressed in the tumor samples but interestingly they have very little marginal information (not individually differentially expressed,

[2] The Equal Credit Opportunity Act (ECOA) is a major federal financial regulation law enacted in 1974.

[3] `https://en.wikipedia.org/wiki/Protected_group`.

[4] `Pay_0` and `Pay_2` denote the repayment status of the last two months (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, and so on).

as shown in Fig. 6). Accordingly, traditional linear-model-based analysis will fail to detect this gene-pair as a key biomarker. (ii) The simple decision tree model in Fig. 2 provides a mechanistic understanding and justification as to why the algorithm thinks a patient has prostate cancer or not. (iii) Finally, it provides the needed guidance on what to do next by having a control over the system. In particular, a cancer biologist can choose between different diagnosis and treatment plans with the goal to regulate those two oncogenes.

**Goals and organization**: The primary goal of this paper is to introduce some new fundamental concepts and tools to lay the foundation of *admissible machine learning* that are efficient (enjoy good predictive accuracy), fair (prevent discrimination against minority groups), and interpretable (provide mechanistic understanding) to the best possible extent.

Our statistical learning framework is grounded in the foundational concepts of information theory. The required statistical formalism (nonparametric estimation and inference methods) and information-theoretic principles (entropy, conditional entropy, relative entropy, and conditional mutual information) are introduced in Sect. 2. A new nonparametric estimation technique for conditional mutual information (CMI) is proposed that scales to large datasets by leveraging the power of machine learning. For statistical inference, we have devised a new model-based bootstrap strategy. The method was applied to the problem of conditional independence testing and integrative genomics (breast cancer multi-omics data from Cancer Genome Atlas). Based on this theoretical foundation, in Sect. 3, we laid out the basic elements of admissible machine learning. Section 3.1 focuses on algorithmic interpretability: how can we efficiently search and design self-explanatory algorithmic models by balancing accuracy and robustness to the best possible extent? Can we do it in a completely model-agnostic manner? Key concepts and tools introduced in this section are: Core features, infogram, L-features, net-predictive information, and COREml. The procedure was applied to several real datasets, including high-dimensional microarray gene expression datasets (prostate cancer and SRBCT data), MONK's problems, and Wisconsin breast cancer data. Section 3.2 focuses on algorithmic fairness, which tackles the challenging problem of designing admissible ML algorithms that are *simultaneously* efficient, interpretable, and equitable. There are several key techniques introduced in this section: admissible feature selection, ALFA-testing, graphical risk assessment tool, and FINEml. We illustrate the proposed methods using examples from criminal justice system (ProPublica's COMPAS recidivism data), financial service industry (Adult income data, Taiwan credit card data), and marketing ad campaign. We conclude the paper in Sect. 4 by reviewing the challenges and opportunities of next-generation *admissible* ML technologies.

## 3 Information-theoretic principles and methods

The foundation of admissible machine learning relies on information-theoretic principles and nonparametric methods. The key theoretical ideas and results are presented in this section to develop a deeper understanding of the conceptual basis of our new framework.

### 3.1 Notation

Let $Y$ be the response variable taking values $\{1, \ldots, k\}$, $\mathbf{X} = (X_1, \ldots, X_p)$ denotes a $p$-dimensional feature matrix, and $\mathbf{S} = (S_1, \ldots, S_q)$ is additional set of $q$ covariates (e.g., collection of sensitive attributes like race, gender, age, etc.). A variable is called `mixed` when it can

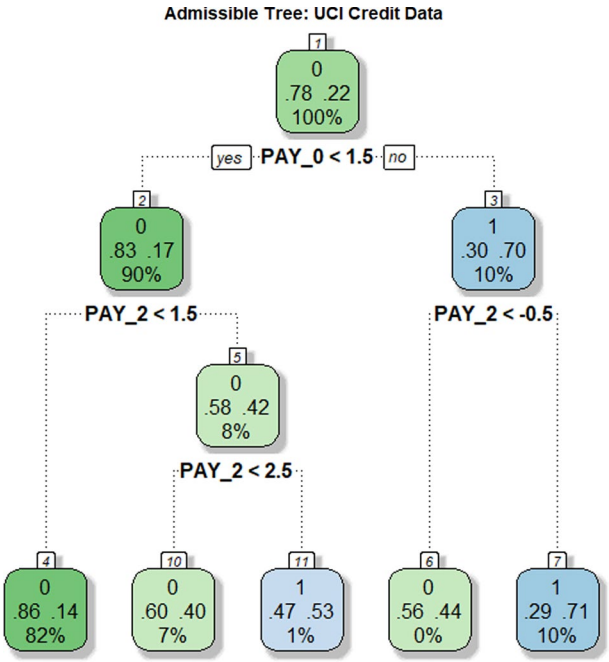



**Fig. 1** A shallow admissible tree classifier for the UCI credit card data with four decision nodes, which is as accurate as the most complex state-of-the-art ML model

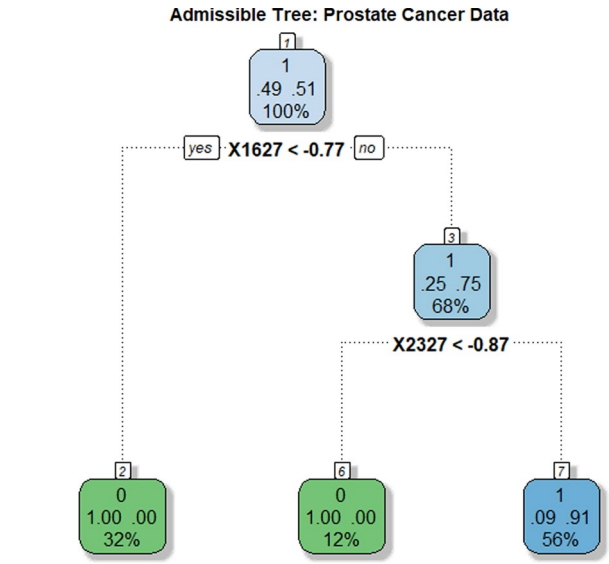



**Fig. 2** A two-gene admissible tree classifier for prostate cancer data with $p = 6033$ gene expression measurements on 50 control and 52 cancer patients

take either discrete, continuous, or even categorical values, i.e., completely unrestricted data-types. Throughout, we will allow both $\mathbf{X}$ and $\mathbf{S}$ to be `mixed`. We write $Y \perp\!\!\!\perp \mathbf{X}$ to denote the independence of $Y$ and $\mathbf{X}$. While, the conditional independence of $Y$ and $\mathbf{X}$ given $\mathbf{S}$ is denoted by $Y \perp\!\!\!\perp \mathbf{X} \mid \mathbf{S}$. For a continuous random variable, $f$ and $F$ denote the probability density and distribution function, respectively. For a discrete random variable the probability mass function will be denoted by $p$ with proper subscript.

### 3.2 Conditional mutual information

Our theory starts with an information-theoretic view of conditional dependence. Under conditional independence:

$$Y \perp\!\!\!\perp \mathbf{X} \mid \mathbf{S}$$

the following decomposition holds for all $y, \mathbf{x}, \mathbf{s}$

$$f_{Y,\mathbf{X}|\mathbf{S}}(y, \mathbf{x}|\mathbf{s}) = f_{Y|\mathbf{S}}(y|\mathbf{s}) f_{\mathbf{X}|\mathbf{S}}(\mathbf{x}|\mathbf{s}).$$

More than testing independence, often the real interest lies in *quantifying* the conditional dependence: the average deviation of the ratio

$$\frac{f_{Y,\mathbf{X}|\mathbf{S}}(y, \mathbf{x}|\mathbf{s})}{f_{Y|\mathbf{S}}(y|\mathbf{s}) f_{\mathbf{X}|\mathbf{S}}(\mathbf{x}|\mathbf{s})}, \tag{2.1}$$

which can be measured by conditional mutual information (Wyner 1978).

**Definition 1** Conditional mutual information (CMI) between $Y$ and $\mathbf{X}$ given $\mathbf{S}$ is defined as:

$$\mathrm{MI}(Y, \mathbf{X} \mid \mathbf{S}) = \iiint_{y,\mathbf{x},\mathbf{s}} \log \left( \frac{f_{Y,\mathbf{X}|\mathbf{S}}(y, \mathbf{x}|\mathbf{s})}{f_{Y|\mathbf{S}}(y|\mathbf{s}) f_{\mathbf{X}|\mathbf{S}}(\mathbf{x}|\mathbf{s})} \right) f_{Y,\mathbf{X},\mathbf{S}}(y, \mathbf{x}, \mathbf{s}) \, \mathrm{d}y \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{s}. \tag{2.2}$$

**Two Important Properties**. (P1) One of the striking features of CMI is that it captures multivariate non-linear conditional dependencies between the variables in a completely nonparametric manner. (P2) CMI possesses the necessary and sufficient condition as a measure of conditional independence, in the sense that

$$\mathrm{MI}(Y, \mathbf{X}|\mathbf{S}) = 0 \text{ if and only if } Y \perp\!\!\!\perp \mathbf{X} \mid \mathbf{S}. \tag{2.3}$$

Conditional independence relation can be described using graphical model (also known as Markov network), as shown in Fig. 3 below.

### 3.3 Net-predictive information

One of the major significances of CMI as a measure of conditional dependence comes from its interpretation in terms of additional 'information gain' on $Y$ learned through $\mathbf{X}$ when we already know $\mathbf{S}$. In other words, CMI measures the Net-Predictive Information (NPI) of $\mathbf{X}$ —the *exclusive* information content of $\mathbf{X}$ for $Y$ beyond what is already subsumed by $\mathbf{S}$. To formally arrive at this interpretation, we have to look at CMI from a different angle, by expressing it in terms of conditional entropy. Entropy is a fundamental information-theoretic uncertainty measure. For a random variable $Z$, entropy $H(Z)$ is defined as $-\mathbb{E}_Z[\log f_Z]$.

**Definition 2** The conditional entropy $H(Y|\mathbf{S})$ is defined as the expected entropy of $Y|\mathbf{S} = \mathbf{s}$

$$H(Y \mid \mathbf{S}) = \int_{\mathbf{s}} H(Y \mid \mathbf{S} = \mathbf{s}) \, \mathrm{d}F_{\mathbf{s}}, \tag{2.4}$$

which measures how much uncertainty remains in $Y$ after knowing $\mathbf{S}$, on average.

**Theorem 1** *For $Y$ discrete and $(\mathbf{X}, \mathbf{S})$* `mixed` *multidimensional random vectors, $\mathrm{MI}(Y, \mathbf{X}|\mathbf{S})$ can be expressed as the difference between two conditional-entropy statistics*:

$$\mathrm{MI}(Y, \mathbf{X} \mid \mathbf{S}) \;=\; H(Y \mid \mathbf{S}) \;-\; H(Y \mid \mathbf{S}, \mathbf{X}). \tag{2.5}$$

The proof involves some standard algebraic manipulations, and is given in Appendix A.1.

***Remark 1*** (Uncertainty Reduction) The alternative way of defining CMI through eq. (2.5) allows us to interpret it from a new angle: Conditional mutual information $\mathrm{MI}(Y, \mathbf{X}|\mathbf{S})$ measures the *net impact* of $\mathbf{X}$ in reducing the uncertainty of $Y$, given $\mathbf{S}$. This new perspective will prove to be vital for our subsequent discussions. Note that, if $H(Y|\mathbf{S}, \mathbf{X}) = H(Y|\mathbf{S})$, then $\mathbf{X}$ carries no *net*-predictive information about $Y$.

### 3.4 Nonparametric estimation algorithm

The basic formula (2.2) of conditional mutual information (CMI) that we have presented in the earlier section, is, unfortunately, not readily applicable for two reasons. First, the practical side: in the current form, (2.2) requires estimation of $f_{Y,\mathbf{X}|\mathbf{S}}$ and $f_{\mathbf{X}|\mathbf{S}}$, which could be a herculean task, especially when $\mathbf{X} = (X_1, \ldots, X_p)$ and $\mathbf{S} = (S_1, \ldots, S_q)$ are large-dimensional. Second, the theoretical side: since the triplet $(Y, \mathbf{X}, \mathbf{S})$ is mixed (not all discrete or continuous random vectors) the expression (2.2) is not even a valid representation. The necessary reformulation is given in the next theorem.

**Theorem 2** *Let $Y$ be a discrete random variable taking values $1, \ldots, k$, and $(\mathbf{X}, \mathbf{S})$ be a* `mixed` *pair of random vectors. Then the conditional mutual information can be rewritten as*

$$\mathrm{MI}(Y, \mathbf{X} \mid \mathbf{S}) \;=\; \mathbf{E}_{\mathbf{X},\mathbf{S}}\Big[\mathrm{KL}\big(p_{Y|\mathbf{X},\mathbf{S}} \parallel p_{Y|\mathbf{S}}\big)\Big], \tag{2.6}$$

*where Kullback-Leibler (KL) divergence from $p_{Y|\mathbf{X}=\mathbf{x},\mathbf{S}=\mathbf{s}}$ to $p_{Y|\mathbf{S}=\mathbf{s}}$ is defined as*

$$\mathrm{KL}\big(p_{Y|\mathbf{X},\mathbf{S}} \parallel p_{Y|\mathbf{S}}\big) \;=\; \sum_{y} p_{Y|\mathbf{X},\mathbf{S}}(y|\mathbf{x}, \mathbf{s}) \log\left(\frac{p_{Y|\mathbf{X},\mathbf{S}}(y|\mathbf{x}, \mathbf{s})}{p_{Y|\mathbf{S}}(y|\mathbf{s})}\right). \tag{2.7}$$

To prove it, first rewrite the dependence-ratio (2.1) solely in terms of conditional distribution of $Y$ as follows:

$$\frac{\Pr(Y = y \mid \mathbf{X} = \mathbf{x}, \mathbf{S} = \mathbf{s})}{\Pr(Y = y \mid \mathbf{S} = \mathbf{s})} \;=\; \frac{p_{Y|\mathbf{X},\mathbf{S}}(y|\mathbf{x}, \mathbf{s})}{p_{Y|\mathbf{S}}(y|\mathbf{s})}$$

Next, substitute this into (2.2) and express it as

$$\mathrm{MI}(Y, \mathbf{X} \mid \mathbf{S}) \;=\; \iint_{\mathbf{x},\mathbf{s}} \left[\sum_{y} p_{Y|\mathbf{X},\mathbf{S}}(y|\mathbf{x}, \mathbf{s}) \log\left(\frac{p_{Y|\mathbf{X},\mathbf{S}}(y|\mathbf{x}, \mathbf{s})}{p_{Y|\mathbf{S}}(y|\mathbf{s})}\right)\right] \mathrm{d}F_{\mathbf{X},\mathbf{S}}$$

Replace the part inside the square brackets by (2.7) to finish the proof. □

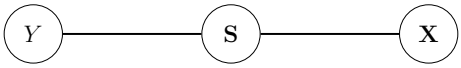



**Fig. 3** Representing conditional independence graphically, where each node is a random variable (or random vector). The edge between $Y$ and $\mathbf{X}$ passes through the $\mathbf{S}$

***Remark 2*** CMI measures how much information is shared only between $\mathbf{X}$ and $Y$ that is not contained in $\mathbf{S}$. Theorem 2 makes this interpretation explicit.

**Estimator**: Goal is to develop a practical nonparametric algorithm for estimating CMI from $n$ i.i.d samples $\{\mathbf{x}_i, y_i, \mathbf{s}_i\}_{i=1}^n$ that works for large$(n, p, q)$ settings. Theorem 2 immediately leads to the following estimator of (2.6):

$$\widehat{\mathrm{MI}}(Y, \mathbf{X} \mid \mathbf{S}) \;=\; \frac{1}{n} \sum_{i=1}^{n} \log \frac{\widehat{\Pr}(Y = y_i | \mathbf{x}_i, \mathbf{s}_i)}{\widehat{\Pr}(Y = y_i | \mathbf{s}_i)}. \tag{2.8}$$

**Algorithm 1**: *Conditional mutual information estimation*: the proposed ML-powered nonparametric estimation method consists of three simple steps:

`Step 1`. Choose a machine learning classifier (e.g., support vector machines, random forest, gradient boosted trees, deep neural network, etc.), and call it $\texttt{ML}_0$.

`Step 2`. Train the following two models:

$$\begin{aligned} \texttt{ML.train}_{y|\mathbf{x},\mathbf{s}} &\leftarrow \texttt{ML}_0\big(Y \sim [\mathbf{X}, \mathbf{S}]\big) \\ \texttt{ML.train}_{y|\mathbf{s}} &\leftarrow \texttt{ML}_0\big(Y \sim \mathbf{S}\big) \end{aligned}$$

`Step 3`. Extract the conditional probability estimates $\widehat{\Pr}(Y = y_i|\mathbf{x}_i, \mathbf{s}_i)$ from $\texttt{ML.train}_{y|\mathbf{x},\mathbf{s}}$, and $\widehat{\Pr}(Y = y_i|\mathbf{s}_i)$ from $\texttt{ML}_0\big(Y \sim \mathbf{S}\big)$, for $i = 1, \ldots, n$.

`Step 4`. Return $\widehat{\mathrm{MI}}(Y, \mathbf{X} \mid \mathbf{S})$ by applying formula (2.8).

***Remark 3*** We will be using the gradient boosting machine (`gbm`) of Friedman (2001) in our numerical examples (obviously, one can use other methods), whose convergence behavior is well-studied in literature (Breiman et al. 2004; Zhang 2004), where it was definitively shown that under some very general conditions, the empirical risk (probability of misclassification) of the gbm classifier approaches the optimal Bayes risk. This Bayes risk consistency property surely carries over to our conditional probability estimates in (2.8), which justifies the good empirical performance of our method in real datasets.

***Remark 4*** Taking the base of the log in (2.8) to be 2, we get the measure in the unit of *bits*. If the log is taken to be the natural $\log_e$, then it is in nats unit. We will use $\log_2$ in all our computation.

The proposed style of nonparametric estimation provides some important practical benefits:

- Flexibility: Unlike traditional conditional independence testing procedures (Candes et al. 2018; Berrett et al. 2019), our approach requires neither the knowledge of the exact parametric form of high-dimensional $F_{X_1,\ldots,X_p}$ nor the knowledge of the conditional distribution of $\mathbf{X} \mid \mathbf{S}$, which are generally *unknown* in practice.

- Applicability: (i) Data-type: The method can be safely used for *mixed* $\mathbf{X}$ and $\mathbf{S}$ (any combination of discrete, continuous, or even categorical variables). (ii) Data-dimension: The method is applicable to *high-dimensional* $\mathbf{X} = (X_1, \ldots, X_p)$ and $\mathbf{S} = (S_1, \ldots, S_q)$.
- Scalability: Unlike traditional nonparametric methods (such as kernel density or $k$-nearest neighbor-based methods), our procedure is scalable for *big datasets* with large$(n, p, q)$.

### 3.5 Model-based bootstrap

One can even perform statistical inference for our ML-powered conditional-mutual-information statistic. In order to test $H_0 : Y \perp\!\!\!\perp \mathbf{X} \mid \mathbf{S}$, obtain bootstrap-based $p$-value by noting that under the null $\Pr(Y = y|\mathbf{X} = \mathbf{x}, \mathbf{S} = \mathbf{s})$ reduces to $\Pr(Y = y|\mathbf{S} = \mathbf{s})$.

**Algorithm 2**: *Model-based Bootstrap*: The inference scheme proceeds as follows:

`Step 1.` Let

$$\widehat{p}_{i|\mathbf{s}} \;=\; \widehat{\Pr}(Y_i = 1|\mathbf{S} = \mathbf{s}_i), \quad \text{for } i = 1, \ldots, n$$

as extracted from (already estimated) the model $\texttt{ML.train}_{y|\mathbf{s}}$ (step 2 of Algorithm 1).

`Step 2.` Generate the null $Y^*_{n\times 1} = (Y^*_1, \ldots, Y^*_n)$ by

$$Y^*_i \;\leftarrow\; \text{Bernoulli}(\widehat{p}_{i|\mathbf{s}}), \quad \text{for } i = 1, \ldots, n$$

`Step 3.` Compute $\widehat{\text{MI}}(Y^*, \mathbf{X} \mid \mathbf{S})$ using the Algorithm 1.

`Step 4.` Repeat the process $B$ times (say, $B = 500$); compute the bootstrap null distribution, and return the $p$-value.

***Remark 5*** A parametric version of this inference was proposed by Rosenbaum (1984) in the context of observational causal study. His scheme resamples $Y$ by estimating $\Pr(Y = 1|\mathbf{S})$ using a logistic regression model. The procedure was called conditional permutation test.

### 3.6 A few examples

***Example 1*** Model: $X \sim$ Bernoulli(0.5); $S \sim$ Bernoulli(0.5); $Y = X$ when $S = 0$ and $1 - X$ when $S = 1$. In this case, it is easy to see that the true $\text{MI}(Y, X|S) = 1$. We simulated $n = 500$ i.i.d $(x_i, y_i, s_i)$ from this model and computed our estimate using (2.8). We repeated the process 50 times to access the variability of the estimate. Our estimate is:

$$\widehat{\text{MI}}(Y, X \mid S) \;=\; 0.994 \pm 0.00234.$$

with (avg.) $p$-value being almost zero. We repeated the same experiment by making $Y \sim$ Bernoulli(0.5) (i.e., now true $\text{MI}(Y, X|S) = 0$), which yields

$$\widehat{\text{MI}}(Y, X \mid S) \;=\; 0.0022 \pm 0.0017.$$

with (avg.) pvalue being 0.820.

***Example 2*** *Integrative Genomics*. The wide availability of multi-omics data has revolutionized the field of biology. It is a general consensus among practitioners that combining individual omics data sets (mRNA, microRNA, CNV and DNA methylation, etc.) leads

to improved prediction. However, before undertaking such analysis, it is probably worthwhile to check what is the additional information we gain from a combined analysis compared to a single-platform one. To illustrate this point, we use a Breast cancer multi-omics data that is a part of The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov/). It contain the expression of three-kinds of omics data sets: miRNA, mRNA, and proteomics from three kinds of breast cancer samples ($n = 150$): Basal, Her2, and LumA. $\mathbf{X}_1$ is $150 \times 184$ matrix of miRNA, $\mathbf{X}_2$ is $150 \times 200$ matrix of mRNA, and $\mathbf{X}_3$ is $150 \times 142$ matrix of proteomics.

$$\begin{aligned} \mathrm{MI}(Y, \mathbf{X}_2 \mid \mathbf{X}_1) &= 0.013; \ p\text{-value} = 0.356 \\ \mathrm{MI}(Y, \mathbf{X}_3 \mid \mathbf{X}_1) &= 0.0186; \ p\text{-value} = 0.235 \\ \mathrm{MI}\big(Y, \{\mathbf{X}_2, \mathbf{X}_3\} \mid \mathbf{X}_1\big) &= 0.0192; \ p\text{-value} = 0.501. \end{aligned}$$

It shows: neither mRNA or proteonomics add any substantial information beyond what is already captured by miRNAs.

## 4 Elements of admissible machine learning

How to design admissible machine learning algorithms with enhanced efficiency, interpretability, and equity?[5] A systematic pipeline for developing such admissible ML models is laid out in this section, which is grounded in the earlier information-theoretic concepts and nonparametric modeling ideas.

### 4.1 COREml: algorithmic interpretability

#### 4.1.1 From predictive features to core features

One of the first tasks of any predictive modeling is to identify the key drivers that are affecting the response $Y$. Here we will discuss a new information-theoretic graphical tool to quickly spot the "core" decision-making variables, which are going to be vital in building interpretable models. One of the advantages of this method is that it works even in the presence of correlated features, as the following example illustrates; also see Appendix A.7.

***Example 3*** *Correlated features.* $Y \sim \text{Bernoulli}(\pi(\mathbf{x}))$ where $\pi(\mathbf{x}) = 1/(1 + e^{-\mathcal{M}(\mathbf{x})})$ and

$$\mathcal{M}(\mathbf{x}) \ = \ 3\sin(X_1) - 2X_2. \tag{3.1}$$

$X_1, \ldots X_{p-1}$ be i.i.d $\mathcal{N}(0, 1)$ random variables, and

$$X_p = 2X_1 - X_2 + \epsilon, \ \text{where } \epsilon \sim \mathcal{N}(0, 2), \tag{3.2}$$

[5] However, the general premise of admissible ML is extremely broad and flexible, and will continue to evolve with the regulatory requirements to ensure rapid development of trustworthy algorithmic methods.

which means $X_p$ has no additional predictive value beyond what is already captured by the core variables $X_1$ and $X_2$. Another way of saying this is that $X_p$ is *redundant*—the conditional mutual information between $Y$ and $X_p$ given $\{X_1, X_2\}$ is zero:

$$\mathrm{MI}\big(Y, X_p \mid \{X_1, X_2\}\big) = 0.$$

The top of Fig. 4 graphically depicts this. The following nomenclature will be useful for discussing our method:

$$\begin{aligned} \text{CoreSet} &= \{X_1, X_2\} \\ \text{Imitator} &= \{X_p\} \\ \text{Probes} &= \{X_3, \ldots, X_{p-1}\}. \end{aligned}$$

Note that the imitator $X_p$ is *highly predictive* for $Y$ due to its association with the core variables. We have simulated $n = 500$ samples with $p = 50$. For each feature we compute,

$$R_j \;=\; \text{overall relevance score of jth predictor, } j = 1, \ldots, p. \tag{3.3}$$

The bottom-left corner of Fig. 4 shows the relative importance scores (scaled between 0 and 1) for the top seven features using `gbm` algorithm[6], which correctly finds $\{X_1, X_2, X_{50}\}$ as the important predictors. However, it is important to recognise that this modus operandi—irrespective of the ML algorithm—can not distinguish the 'fake imitator' $X_{50}$ from the real ones $X_1$ and $X_2$. To enable refined characterization of the variables, we have to 'add more dimension' to the classical machine learning feature importance tools.

### 4.1.2 InfoGram and L-features

We introduce a tool for identification of core admissible features based on the concept of net-predictive information (NPI) of a feature $X_j$.

**Definition 3** The net-predictive (conditional) information of $X_j$ given all the rest of the variables $\mathbf{X}_{-j} = \{X_1, \ldots, X_p\} \backslash \{X_j\}$ is defined in terms of conditional mutual information:

$$C_j \;=\; \mathrm{MI}(Y, X_j \mid \mathbf{X}_{-j}), \quad \text{for } j = 1, \ldots, p. \tag{3.4}$$

For easy interpretation, we standardize $C_j$ by $\frac{C_j}{\max_j C_j}$ and convert it between 0 and 1. Infogram, which is the acronym for information diagram, is a scatter plot of $\{(R_j, C_j)\}_{j=1}^{p}$ over the unit square $[0, 1]^2$; see the bottom-right corner of Fig. 4.

**L-Features**. The highlighted L-shaped area contains features that are either irrelevant or redundant. For example, notice the position of $X_{50}$ in the plot, indicating that it is highly predictive but contains no new complementary information for the response. Clearly, there could be an opposite scenario: a variable carries valuable net individual information for $Y$, despite being moderately relevant (not ranked among the top few); see Sec. 3.1.4.

[6] based on whether a particular variable was selected to split on during learning a tree, and how much it improves the Gini impurity or information gain.

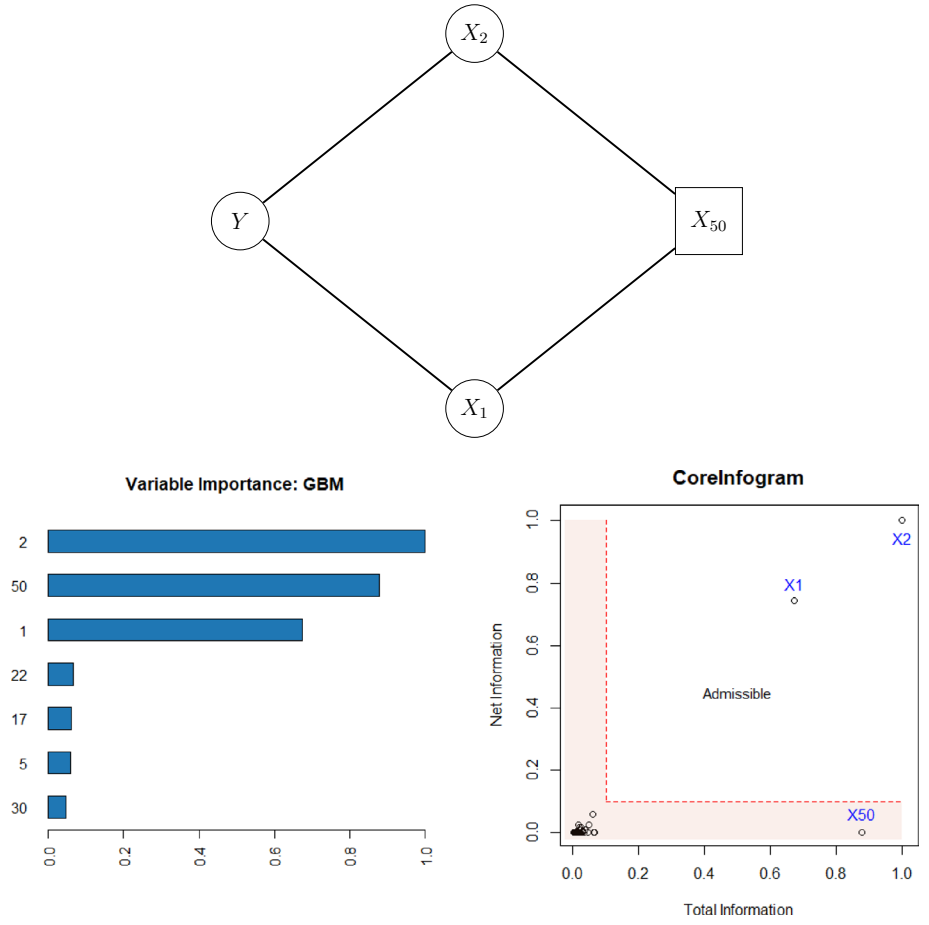



**Fig. 4** *Top*: The graphical representation of example 3 is shown. *Bottom-left*: The gbm-feature importance score for top seven features; rest are almost zero thus not shown. *Bottom-right*: infogram identifies the core variables $\{X_1, X_2\}$ from the $X_{50}$. The L-shaped area with 0.1 width is highlighted in red; it contains inadmissible variables with either low relevance or high redundancy

***Remark 6*** (Predictive Features vs. CoreSet) Recall that in Example 3, the irrelevant feature $X_{50}$ is strongly correlated with the relevant ones $X_1$ and $X_2$ through (3.2), thus violate the so-called "irrepresentable condition"–for more details see the bibliographic notes section of Hastie et al. (2015, p. 311). In this scenario (which may easily arise in practice), it is hard to recover the "important" variables using traditional variable selection methods. The bottom line is: identifying `CoreSet` is a much more difficult undertaking than merely selecting the most predictive ones. The goal of infogram is to facilitate this process of discovering the key variables that are driving the outcome.

***Remark 7*** (CoreML) Two additional comments before diving into a real data examples. First, machine learning models based on "core" features (`CoreML`) show improved

stability, especially when there exists considerable correlation among the features.[7] This will be demonstrated in the next two sections. Second, our approach is not tied to any particular machine learning method; it is completely model-agnostic and can be integrated with any arbitrary algorithm: choose a specific classifier $\texttt{ML}_0$ and compute (3.3) and (3.4) to generate the associated infogram.

***Example 4*** *MONK's problems* (Thrun et al. 1991). It is a collection of three binary artificial classification problems (MONK-1, MONK-2 and MONK-3) with $p = 6$ attributes; available in the UCI Machine Learning Repository. As shown in Fig. 5, infogram selects $\{X_1, X_2, X_5\}$ for the MONK-1 data, and $\{X_2, X_5\}$ for the MONK-3 data as the core features. MONK-2 is an idiosyncratic case, where all six features turned out to be core! This indicates the possible complex nature of the classification rule for the MONK-2 problem.

#### 4.1.3 COREtree: high-dimensional microarray data analysis

How does one distill a compact (parsimonious) ML model by balancing accuracy, robustness, and interpretability to the best extent? To answer that, we introduce `COREtree`, whose construction is guided by infogram. The methodology is illustrated using two real datasets, namely Prostate cancer and SRBCT tumor data. The main findings are striking: it shows how one can systematically search and construct robust and interpretable shallow decision tree models (often with just two or three genes) for noisy high-dimensional microarray datasets that are as powerful as the most elaborate and complex machine learning methods.

***Example 5*** *Prostate cancer gene expression data*. The data consist of $p = 6033$ gene expression measurements on 50 control and 52 prostate cancer patients. It is available at https://web.stanford.edu/~hastie/CASI_files/DATA/prostate.html. Our analysis is summarized below.

`Step` 1. *Identifying CoreGenes*. GBM-selected top 50 genes are shown in Fig. 6. We generate the infogram[8] of these 50 variables (displayed on the top-right corner), which identifies five core-genes $\{1627, 2327, 77, 1511, 1322\}$.

`Step` 2. *Rank-transform: Robustness and Interpretability*. Instead of directly operating on the gene expression values, we transform them into their ranks. Let $\{x_{j1}, \ldots, x_{jn}\}$ be the measurements on $j$th gene with empirical cdf $\widetilde{F}_j$. Convert the raw $x_{ji}$ to $u_{ji}$ by

$$u_{ji} = \widetilde{F}_j(x_{ji}), \quad i = 1, \ldots, n \tag{3.5}$$

and work on the resulting $\mathbf{U}_{n\times p}$ matrix instead of the original $\mathbf{X}_{n\times p}$. We do this transformation for two reasons: first, to robustify, since it is known that gene expressions are inherently noisy. Second, to make it unit-free, since the raw gene expression values depend on the type of preprocessing, thus carries much less scientific meaning. On the other hand, percentiles are much more easily interpretable to convey "how overexpressed a gene is."

[7] Numerous studies have found that many current methods like partial dependence plots, LIME, and SHAP could be highly misleading, particularly when there is strong dependence among features.

[8] To reduce unnecessary clutter, we have displayed the infogram using top 50 features, since the rest of the genes will be cramped inside the nonessential L-zone anyway.

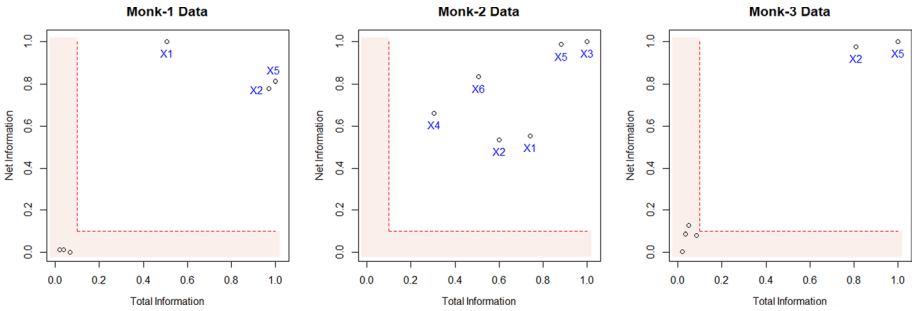



**Fig. 5** Infograms of Monk's problems. CoreSets are denoted in *blue* (Colour figure online)

`Step` 3. *Shallow Robust Tree*. We build a single decision tree using the infogram-selected coregenes. This is displayed in the bottom-right panel of Fig. 6. Interestingly, the `CoreTree` retained only two genes $\{1627, 2327\}$ whose scatter plot (in the rank-transform domain) is shown in the bottom-left corner of Fig. 6. A simple eyeball estimate of the discrimination surfaces are shown in bold (black and red) lines, which closely matches with the decision tree rule. It is quite remarkable that we have reduced the original 6033-dimensional problem to a simple bivariate two-sample one, just by wisely selecting the features based on the infogram.

`Step` 4. *Stability*. Note the tree that we build is based only on the infogram-selected core features. These features have less redundancy and high relevance, which provide an extraordinary stability (over different runs on the same dataset) to the decision-tree–a highly desirable characteristic.

`Step` 5. *Accuracy*. The accuracy of our *single* decision tree (on a randomly selected 20% test set, averaged over 100 times) is more than 95%. On the other hand, the full-data gbm (with $p = 6033$ genes) is only 75% accurate. Huge simplification of the model-architecture with significant gain in the predictive performance!

`Step` 6. *Gene Hunting: Beyond Marginal Screening*. We compute two-sample $t$-test statistic for all $p = 6033$ genes and rank them according to their absolute values (the gene with the largest absolute $t$-statistic gets ranked 1–the most differentially expressed gene). The $t$-scores for the `coregenes` along with their $p$-values and ranks are:

$$\begin{aligned} |t_{1627}| &= 0.15\,; p\text{-value} = \ 0.88; \quad \text{rank} = \ 5383. \\ |t_{2327}| &= 1.40\,; p\text{-value} = \ 0.17; \quad \text{rank} = \ 1228. \end{aligned}$$

Thus, it is hopeless to find `coregenes` by any marginal-screening method–they are *too weak marginally (in isolation), but jointly an extremely strong predictor*. The good news is that our approach can find those multivariate hidden gems in a completely nonparametric fashion.

`Step` 7. *Lasso Analysis and Results*. We have used the `glmnet` R-package. Lasso with $\lambda_{\min}$ (minimum cross-validation error) selects 70 genes, where as $\lambda_{1se}$ (the largest lambda such that error is within 1 standard error of the minimum) selects 60 genes. Main findings are:

(i) The `coregenes` $\{1627, 2327\}$ were never selected, probably because they are marginally very weak; and the significant interaction is not detectable by standard-lasso.

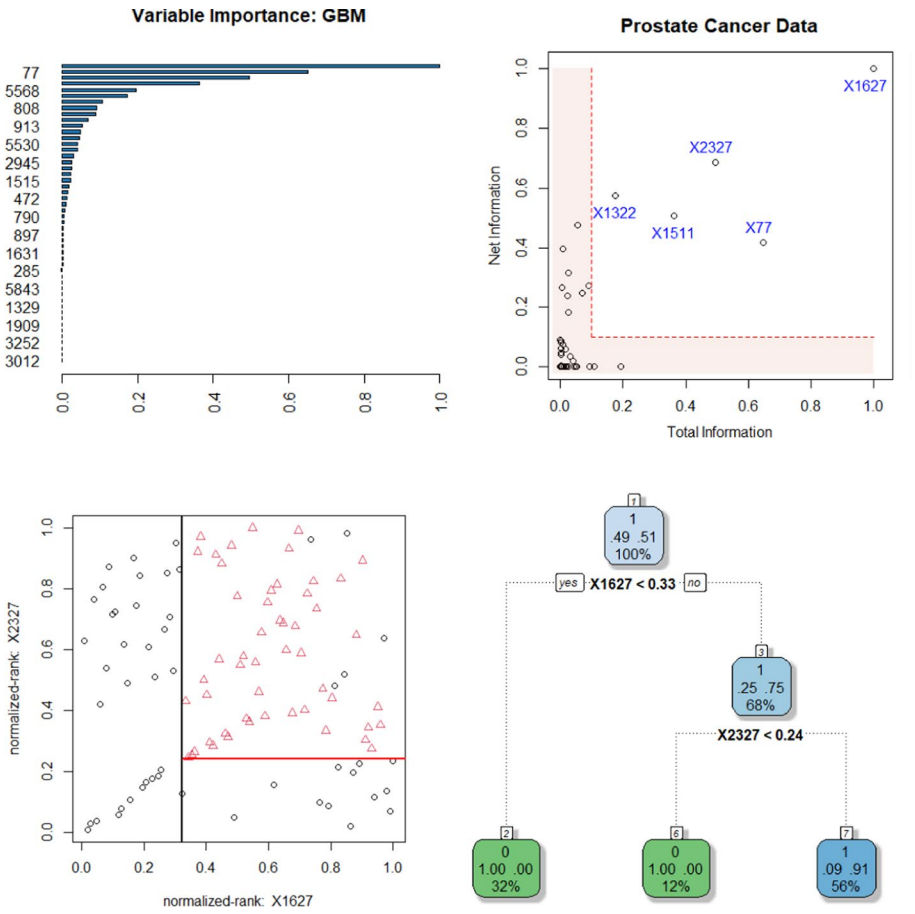



**Fig. 6** Prostate data analysis. *Top panel*: the gbm-feature importance graph, along with the infogram for the top 50 genes. *Bottom-left*: the scatter plot of Gene 1627 vs. 2327. For clarity, we have plotted them in the quantile domain $(u_i, v_i)$, where $u = \text{rank}(X[, 1627])/n$ and $v = \text{rank}(X[, 2327])/n$. The b*lack dots* denote control samples with $y = 0$ class and red triangles are prostate cancer samples with $y = 1$ class. *Bottom-right*: the estimated `CoreTree` with just two decision-nodes, which is good enough to be 95% accurate

(ii) Accuracy of Lasso with $\lambda_{\min}$ is around 78% (each time we have randomly selected 85% data for training; computed the $\lambda_{\text{cv}}$ for making prediction; averaged over 100 runs).

`Step` 8. *Explainability*. The final "two-gene model" is so simple and elegant that it can be easily communicated to doctors and medical practitioners: a patient with overexpressed gene 1627 and gene 2327 has a higher risk of getting prostate cancer. Biologists can use these two genes as robust prognostic markers for decision-making (or for recommending the proper drug). It is hard to imagine there could be a more accurate algorithm, one that is at least as compact as the "two-gene model." We should not forget that the success behind this dramatic model-reduction hinges on discovering multivariate `coregenes`, which: (i) help us to gain insights into biological mechanisms [clarifying 'who' and 'how'], and (ii) provide a simple explanation of the predictions [justifying 'why'].

***Example 6*** *SRBCT Gene Expression Data*. It is a microarray experiment of Small Round Blue Cell Tumors (SRBCT) taken from a childhood cancer study. It contain information on $p = 2,308$ genes on 63 training samples and 25 test samples. Among $n = 63$ tumor examples, 8 are Burkitt Lymphoma (BL), 23 are Ewing Sarcoma (EWS), 12 are neuroblastoma (NB), and 20 are rhabdomyosarcoma (RMS). The dataset is available in the `plsgenomics` R-package. The top-panel of Fig. 7 shows the infogram, which identifies five core genes $\{123, 742, 1954, 246, 2050\}$. The associated coretree with only three decision-nodes is shown in the bottom panel, which accurately classifies 95% of the test cases. In addition, it enjoys all the advantages that were ascribed to the prostate data—we don't repeat them again.

***Remark 8*** We end this section with a general remark: when applying machine learning algorithms in scientific applications, it is of the utmost importance to design models that can clearly explain the 'why and how' behind their decision-making process. We should not forget that scientists mainly use machine learning as a *tool* to gain a mechanistic understanding, so that they can judiciously intervene and control the system. Sticking with the old way of building inscrutable predictive black-box models will severely slow down the adoption of ML methods in scientific disciplines like medicine and healthcare.

#### 4.1.4 COREglm: breast cancer wisconsin data

***Example 7*** *Wisconsin Breast Cancer Data*. The Breast Cancer dataset is available in the UCI machine learning repository. It contains $n = 569$ malignant and benign tumor cell samples. The task is to build an admissible (interpretable and accurate) ML classifier based on $p = 31$ features extracted from cell nuclei images.

`Step 1.` Infogram Construction: Fig. 8 displays the infogram, which provides a quick understanding of the phenomena by revealing its 'core.' Noteworthy points: (i) there are three highly predictive inadmissible features (green bubbles in the plot: perimeter_worst, area_worst, and concave_points_worst), which have large overall predictive importance but almost zero net individual contributions. We have called these variables '`Imitators`' in Sec. 3.1.1. (ii) Three among the four 'core' admissible features (texture_worst, concave_points_mean, and texture_mean) are not among the top features based on usual predictive information, yet they contain a considerable amount of new exclusive information (net-predictive information) that is useful for separating malignant and benign tumor cells. In simple terms, infogram help us to track down where the 'core' discriminatory information is hidden.

`Step 2.` Core-Scatter plot. The right panel of Fig. 8 shows the scatter plot of the top two core features[9] and *how* they separate the malignant and benign tumor cells.

`Step 3.` Infogram-assisted `CoreGLM` model: The simplest possible model that one could build is a logistic regression based on those four admissible features. Interestingly, the Akaike information criterion (AIC) based model selection further drops the variable `texture_mean`, which is hardly surprising considering that it has the least net and total information among the four admissible core features. The final logistic regression model

[9] Based on distance from (1, 1).

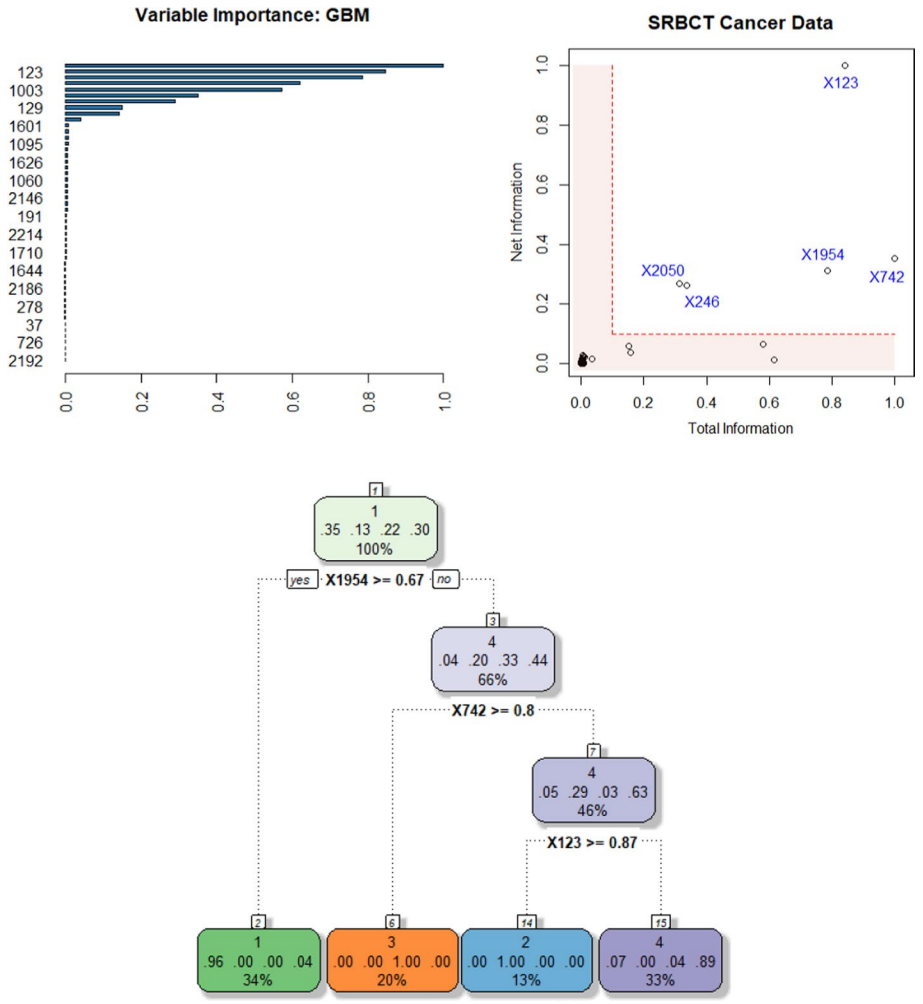



**Fig. 7** SRBCT data analysis. *Top-left*: GBM-feature importance plot; top 50 genes are shown. *Top-right*: The associated infogram. *Bottom panel*: The estimated coretree with just three decision nodes

with three core variables is displayed below (output of `glm` R-function) — a compact yet accurate classifier:

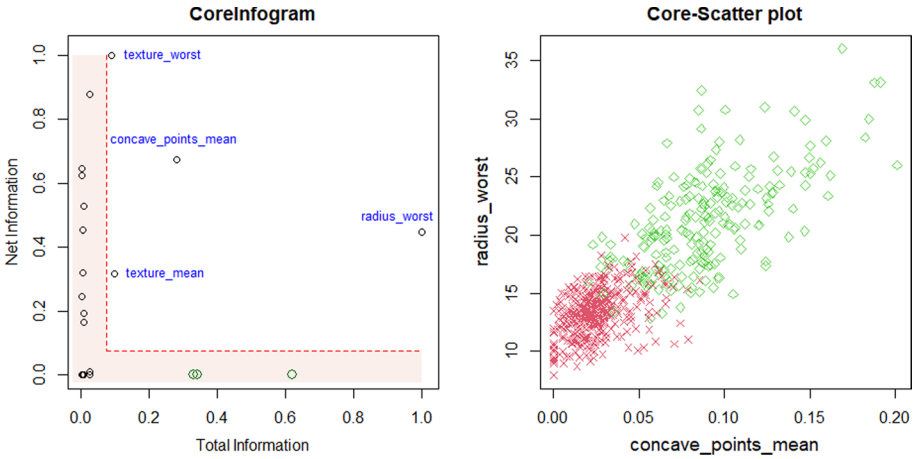



**Fig. 8** Breast Cancer Wisconsin Data. *Left*: Infogram reveals where the crux of the information is hidden. *Right*: scatter plot of top two core features where color green denotes the malignant samples

```
#COREglm Model: UCI breast cancer data
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -29.42361    3.85131  -7.640 2.17e-14 ***
concave_points_mean  96.48880   16.11261   5.988 2.12e-09 ***
radius_worst          0.99767    0.16792   5.941 2.83e-09 ***
texture_worst         0.30451    0.05302   5.744 9.27e-09 ***
```

This simple parametric model achieves a competitive accuracy of 96.50% (on a 15% test set; averaged over 50 trials). Compare this with full-fledged big ML models (like gbm, random forest, etc.) which attain accuracy in the range of 95−−97%. This example again shows how infogram can guide the design of a highly transparent and interpretable `CoreGLM` model with a few handful of variables—which is as powerful as complex black-box ML methods.

***Remark 9*** (Integrated statistical modeling culture) One should bear in mind that the process by which we arrived at simple admissible models actually utilizes the power of modern machine learning—needed to estimate the formula (3.4) of definition 3, as described by the theory laid out in Sect. 2. For more discussion on this topic, see Appendix A.6 and Mukhopadhyay and Wang (2020). In short, we have developed a process of constructing an admissible (explainable and efficient) ML procedure starting from a 'pure prediction' algorithm.

### 4.2 FINEml: algorithmic fairness

ML-systems are increasingly used for automated decision-making in various high-stakes domains such as credit scoring, employment screening, insurance eligibility, medical diagnosis, criminal justice sentencing, and other regulated areas. To ensure that we are making responsible decisions using such algorithms, we have to deploy admissible models that can balance Fairness, INterpretability, and Efficiency (FINE) to the best possible extent. This section discusses principles and tools for designing such FINE-algorithms.

#### 4.2.1 FINE-ML: approaches and limitations

Imagine that a machine learning algorithm is used by a bank to accurately predict whether to approve or deny a loan application based on the probability of default. This ML-based risk-assessing tool has access to the following historical data:

- $Y$: {0, 1} Loan status variable–1 whether the loan was approved and 0 if denied.
- $\mathbf{X}$: Feature matrix {income, loan amount, education, credit history, zip code}
- $\mathbf{S}$: Collection of protected attributes {gender, marital status, age, race}.

To automate the loan-eligibility decision-making process, the bank wants to develop an accurate classifier that will not discriminate among applicants on the basis of their protected features. Naturally, the question is: how to go about designing such ML-systems that are accurate and at the same time provide safeguards against algorithmic discrimination?

**Approach 1** *Non-constructive*: We can construct a myriad of ML models by changing and tuning different hyper-parameters, base learners, etc. One can keep building different models until one finds a perfect one that avoids adverse legal and regulatory issues. There are at least two problems with this 'try until you get it right' approach: first, it is non-constructive. The whole process gives zero guidance on how to rectify the algorithm to make it less-biased; also see Appendix A.9. Second, there is no single definition of fairness—more than twenty different definitions have been proposed over the last few years (Narayanan 2018). And the troubling part is that these different fairness measures are mutually incompatible[10] to each other and cannot be satisfied simultaneously (Kleinberg 2018); see Appendix A.4. Hence this laborious process could end up being a wild-goose chase, resulting in a huge waste of computation.

**Approach 2** *Constructive*: Here we seek to construct ML models that—by design—mitigate bias and discrimination. To execute this task successfully, we must first identify and remove proxy variables (e.g., zip code) from the learning set, which prevent a classification algorithm from achieving desired fairness. But how to define a proper mathematical criterion to detect those surrogate variables? Can we develop some easily interpretable graphical exploratory tools to systematically uncover those problematic variables? If we succeed in doing this, then ML developers can use it as a *data filtration* tool to quickly spot and remove the potential sources of biases in the pre-modeling (data-curation) stage, in order to mitigate fairness issues in the downstream analysis.

[10] Thus, cataloging a huge library of inherently contradictory model validation metrics is hardly going to help model developers to search for an admissible and deployable model. Instead of *searching in a dark*, we need some other practical and prudent strategies.

### 4.2.2 InfoGram and admissible feature selection

We offer a diagnostic tool for identification of admissible features that are predictive and safe. Before going any further, it is instructive to formally define what we mean by 'safe.'

**Definition 4** (Safety-index and Inadmissibility) Define the safety-index for variable $X_j$ as

$$F_j \;=\; \mathrm{MI}\big(Y, X_j \mid \{S_1, \ldots, S_q\}\big) \tag{3.6}$$

This quantifies how much extra information $X_j$ carries for $Y$ that is not acquired through the sensitive variables $\mathbf{S} = (S_1, \ldots, S_q)$. For interpretation purposes, we standardize $F_j$ between zero and one by dividing by the $\max_j F_j$. Variables with "small" $F$-values (F-stands for fairness) will be called inadmissible, as they possess little or no informational value beyond their use as a dummy for protected characteristics.

*Construction*. In the context of fairness, we construct the infogram by plotting $\{(R_j, F_j)\}_{j=1}^{p}$, where recall $R_j$ denotes the relevance score (3.3) for $X_j$. The goal of this graphical tool is to assist identification of admissible features which have little or no information-overlap with sensitive attributes $\mathbf{S}$, yet are reasonably predictive for $Y$.

*Interpretation*. Fig. 9 displays an infogram with six covariates. The L-shaped highlighted region contains variables that are either inadmissible (the horizontal slice of L) or inadequate (the vertical slice of L) for prediction. The complementary set $\mathrm{L}^c$ comprises of the desired admissible features. Focus on variables `A` and `B`: both have the same predictive power, but are gained through a completely different manner. The variable `B` gathered information for $Y$ entirely through the protected features (verify it from the graphical representation of `B`), and is thus inadmissible. On the other hand, the variable `A` carries direct informational value, having no connection with the prohibitive $\mathbf{S}$, and is thus totally admissible. Unfortunately, though, reality is usually more complex than this clear-cut black and white `A`-`B` situation. The fact of the matter is: admissibility (or fairness, per se) is not a yes/no concept, but a matter of *degree*[11], which is explained at the bottom two rows of Fig. 9 utilizing variables C to F.

***Remark 10*** The graphical exploratory nature of the `infogram` makes the whole learning process much more transparent, interactive, and human-centered.

*Legal doctrine*. Note that in our framework the protected variables are used only in the pre-deployment phase to determine what other (admissible) attributes to include in the algorithm to mitigate unforeseen downstream bias, which is completely legal (Hellman 2020). It is also advisable that once inadmissible variables are identified using an infogram, not to throw them (especially the highly predictive ones such as the feature B in Fig. 9) blindly from the analysis without consulting domain experts—including some of them may not necessarily imply violation of the law; ultimately, it is up to the policymakers and judiciary to determine their appropriateness (legal permissibility) based on the given context. Our job as statisticians is to discover those hidden inadmissible L-features (preferably in a fully data-driven and automated manner) and raise a red flag for further investigation.

[11] "Zero bias" is an illusion. All models are biased (to a different degree), but some are admissible. The real question is how to methodically construct those admissible ones from possibly biased data.

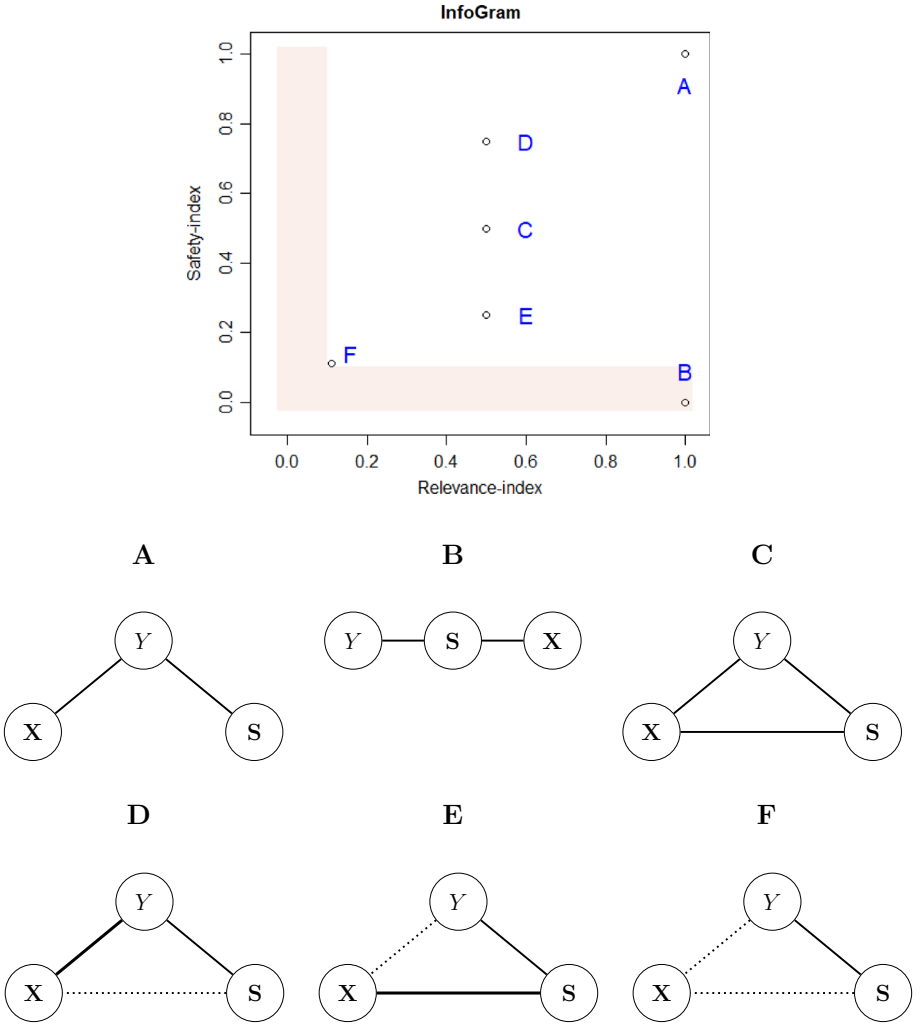



**Fig. 9** Infogram maps variables in a two dimensional (effectiveness vs. safety) diagram. It is a *pre-modeling* nonparametric exploratory tool for admissible feature selection. Infogram is interpreted based on graphical (conditional) independence structure. In real problems, all variables will have some degree of correlation with the protected attributes. Important part is to quantify the "degree," which is measured through eq. (3.6)—as indicated by varying thicknesses of the edges (bold to dotted) between **S** and **X**. Ultimately, the purpose of this graphical diagnostic tool is to provide the necessary guardrails to construct an appropriate learning algorithm that can retain as much of the predictive accuracy as possible, while defending against unforeseen biases—tool for risk-benefit analysis

### 4.2.3 FINEtree and ALFA-test: financial industry applications

***Example 8*** *The Census Income Data*. The dataset is extracted from 1994 United States Census Bureau database, available in UCI Machine Learning Repository. It is also known as the "Adult Income" dataset, which contains $n = 45,222$ records involving personal details such as yearly income (whether it exceeds \$50,000 or not), education level, age,

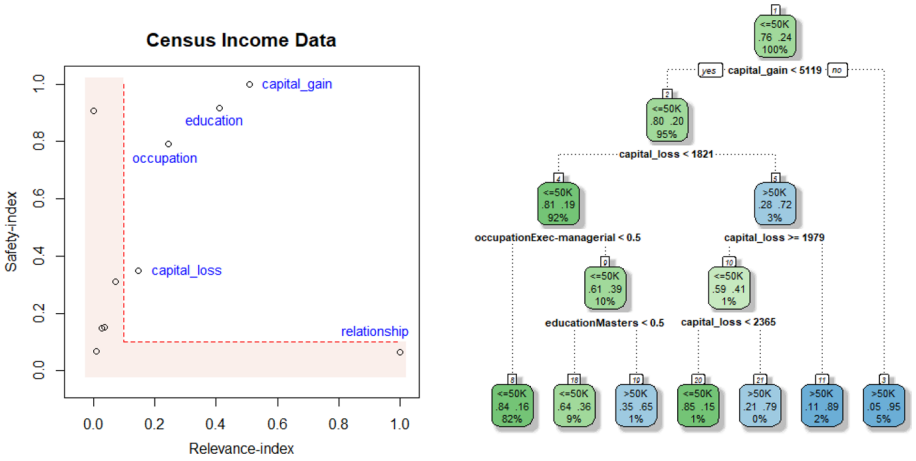



**Fig. 10** Census income data. The left plot shows the infogram. And FINEtree is displayed on the right

gender, marital-status, occupation, etc. The classification task is to determine whether a person makes $50k per year based on a set of 14 attributes, of which four are protective:

$$\mathbf{S} = \{\texttt{Age}, \texttt{Gender}, \texttt{Race}, \texttt{Marital\_Status}\}.$$

`Step 1`. Trust in data. Is there any evidence of built-in bias in the data? That is to say, whether a 'significant' portion of the decision-making ($Y$ is greater or less than 50k per year) was influenced by the sensitive attributes $\mathbf{S}$ beyond what is already captured by other covariates $\mathbf{X}$? One may be tempted to use $\text{MI}(Y, \mathbf{S} \mid \mathbf{X})$ as a measure for assessing fairness. But we need to be careful while interpreting the value of $\text{MI}(Y, \mathbf{S} \mid \mathbf{X})$. It can take a 'small' value for two reasons: First, a genuine case of fair decision-making where individuals with similar $\mathbf{x}$ received a similar outcome irrespective of their age, gender, and other protected characteristics; see Appendix A.4 for one such example. Second, there is a collusion between $\mathbf{X}$ and $\mathbf{S}$ in the sense that $\mathbf{X}$ contains some proxies of $\mathbf{S}$ which *reduce* its effect-size—leading one to falsely declare a decision-rule fair when it is not.

***Remark 11*** The presence of a highly-correlated surrogate variable in the conditional set drastically reduces the size of the CMI-statistics. We call this contraction phenomenon of effect-size in the presence of proxy feature the "shielding effect." To guard against this effect-distortion phenomenon we first have to identify the admissible features from the infogram.

`Step 2`. Infogram to identify inadmissible proxy features. The infogram, shown in the left panel of Fig. 10, finds four admissible features

$$\mathbf{X}_A = \{\texttt{Capital\_gain}, \texttt{Capital\_loss}, \texttt{Occupation}, \texttt{Education}\}.$$

They share very little information with $\mathbf{S}$ yet are highly predictive. In other words, they enjoy high relevance and high safety-index. Next, we also see that there is a feature that appears at the lower-right corner

$$\mathbf{X}_R = \{\texttt{Relationship}\}$$

which is the prime source of bias; the subscript 'R' stands for risky. The variable `relationship` represents the respondent's role in the family—i.e., whether the earning member is husband, wife, child, or other relative.

***Remark 12*** Since $\mathbf{X}_R$ is highly predictive, most *unguided* "pure prediction" ML algorithms will include it in their models, even though it is quite unsafe. Admissible ML models should avoid using variables like `relationship` to reduce unwanted bias.[12] A careful examination reveals that there could be some unintended association between `relationship` and other protected attributes due to social constructs. Without any formal method, it is a hopeless task (especially for practitioners and policymakers; see Lakkaraju and Bastani 2020, Sec. 5.2) to identify these innocent-looking proxy variables in a scalable and automated way.

`Step 3`. ALFA-test and encoded bias. We can construct an honest fairness assessment metric by conditioning CMI with $\mathbf{X}_A$ (instead of $\mathbf{X}$):

$$\widehat{\mathrm{MI}}(Y, \mathbf{S} \mid \mathbf{X}_A) = 0.13, \quad \text{with pvalue almost 0.} \tag{3.7}$$

This strongly suggests historical bias or discrimination is encoded in the data. Our approach not only quantifies but also allows ways to mitigate bias to create an admissible prediction rule; this will be discussed in Step 4. The preceding discussions necessitate the following, new general class of fairness metrics.

**Definition 5** (Admissible Fairness Criterion) To check whether an algorithmic decision is fair given the sensitive attributes and the set of admissible features (identified from infogram), define AdmissibLe FAirness criterion, in short the ALFA-test, as

$$\alpha_Y := \alpha(Y \mid \mathbf{S}, \mathbf{X}_A) = \mathrm{MI}(Y, \mathbf{S} \mid \mathbf{X}_A). \tag{3.8}$$

**Three Different Interpretations**. The ALFA-statistic (3.8) can be interpreted from three different angles.

- It quantifies the trade-off between fairness and model performance: how much net-predictive value is contained within $\mathbf{S}$ (and its close associates)? This is the price we pay in terms of accuracy to ensure a higher degree of fairness.
- A small $\alpha$-inadmissibility value ensures that individuals with similar 'admissible characteristics' receive a similar outcome. Note that our strategy of comparing individuals with respect to only (infogram-learned) "admissible" features allows us to avoid the (direct and indirect) influences of sensitive attributes on the decision making.
- Lastly, the $\alpha$-statistic can also be interpreted as "bias in response $Y$." For a given problem, if we have access to several "comparable" outcome variables[13] then we choose the

[12] or at least should be assessed by experts to determine their appropriateness.

[13] e.g, Obermeyer et al. (2019) showed that healthcare cost can be a poor proxy of health, especially for Black patients; similarly, Blattner and Nelson (2021) showed that credit scores could be a poor proxy for creditworthiness especially for low-income and minority groups.

one which minimizes the $\alpha$-inadmissibility measure. In this way, we can minimize the loss of predictive accuracy while mitigating the bias as best as we can.

***Remark 13*** (Generalizability) Note that, unlike traditional fairness measures, the proposed ALFA-statistic is valid for multi-class problems with a set of multivariate mixed protected attributes—which is, in itself, a significant step forward.

`Step 4`. FINEtree. The inherent historical footprints of bias (as noted in eq. 3.7) need to be deconstructed to build a less-discriminatory classification model for the income data. Fig. 10 shows FINEtree—a simple decision tree based on the four admissible features, which attains 83.5% accuracy.

***Remark 14*** FINEtree is an inherently explainable, fair, and highly competent (decent accuracy) model whose design was guided by the principles of admissible machine learning.

`Step 5`. Trust in algorithm through risk assessment and ALFA-ranking: The current standard for evaluating ML models is primarily based on predictive accuracy on a test set, which is narrow and inadequate. For an algorithm to be deployable it has to be *admissible*; an unguided ML carries the danger of inheriting bias from data. To see that, consider the following two models:

$\text{Model}_\text{A}$: Decision tree based on $\mathbf{X}_A$ (FINEtree)

$\text{Model}_\text{R}$: Decision tree based on $\mathbf{X}_A \cup \{\text{relationship}\}$.

Both models have comparable accuracy around 83.5%. Let $\widehat{Y}_\text{A}$ and $\widehat{Y}_\text{R}$ be the predicted labels based on these two models, respectively. Our goal is to compare and rank different models based on their risk of discrimination using ALFA-statistic:

$$\widehat{\alpha}_\text{A} = \widehat{\text{MI}}(\widehat{Y}_A, \mathbf{S} \mid \mathbf{X}_A) \;=\; 0.00042, \quad \text{with pvalue } 0.95 \tag{3.9}$$

$$\widehat{\alpha}_\text{R} = \widehat{\text{MI}}(\widehat{Y}_\text{R}, \mathbf{S} \mid \mathbf{X}_A) \;=\; 0.195, \quad \text{with pvalue almost } 0. \tag{3.10}$$

$\alpha$-inadmissibility statistic measures how much the final decision (prediction) was impacted by the protective features. A smaller value is better in the sense that it indicates improved fairness of the algorithm's decision. Eqs (3.9)–(3.10) immediately imply that $\text{Model}_\text{A}$ is better (less discriminatory without being inefficient) than $\text{Model}_\text{R}$, and can be safely put into production.

***Remark 15*** Infogram and ALFA-testing can be used (by oversight board or regulators) as a fully-automated exploratory auditing tool that can systematically monitor and discover signs of bias or other potential gaps in compliance[14]; see Appendix A.3.

***Example 9*** *Taiwanese Credit Card data*. This dataset was collected in October 2005, from a Taiwan-based bank (a cash and credit card issuer). It is available in the UCI Machine Learning Repository. We have records of $n = 30,000$ cardholders, and for each we have a

[14] Under the Algorithmic Accountability Act, large AI-driven corporations have to perform broader "admissibility" tests to keep a check on their algorithms' fairness and trustworthiness; see Appendix A.2.

response variable $Y$ denoting: default payment status (Yes = 1, No = 0), along with $p = 23$ predictor variables, including demographic factors, credit data, history of payment, etc. Among these 23 features we have two protected attributes: `gender` and `age`.

The infogram, shown in the left panel of Fig. 11, clearly selects the variable `Pay_0` and `Pay_2` as the key admissible factors that determine the likelihood of default. Once we know the admissible features, the next question is: 'how' `Pay_0` and `Pay_2` are impacting the credit risk? Can we extract an admissible decision rule? For that we construct the FINEtree: a decision tree model based on the infogram-selected admissible features; see Fig. 11. The resulting predictive model is extremely transparent (with shallow yet accurate decision trees[15]) and also mitigates unwanted bias by avoiding inadmissible variables. Lenders, regulators, and bank managers can use this model for automating credit decisions.

#### 4.2.4 Admissible criminal justice risk assessment

***Example 10*** *ProPublica's COMPAS Data.* COMPAS–an acronym for Correctional Offender Management Profiling for Alternative Sanctions–is a most widely used commercial algorithm within the criminal justice system for predicting recidivism risk (the likelihood of re-offending). The data[16]—complied by a team of journalists from ProPublica—constitute all criminal defendants who were subject to COMPAS screening in Broward County, Florida, during 2013 and 2014. For each defendant, $p = 14$ features were gathered, including demographic information, criminal history, and other administrative information. Besides, the dataset also contains information on whether the defendant did in fact actually recidivate (or not) within two years of the COMPAS administration date (i.e., through the end of March 2016); and 3 additional sensitive attributes (gender, race, and age) for each case.

The goal is to develop a accurate and fairer algorithm to predict whether a defendant will engage in violent crime or fail to appear in court if released. Figure 12 shows our results. Infogram selects `event` and `end` as the vital admissible features. The bottom row of Fig. 12 confirms their predictive power. Unfortunately, these two variables are not explicitly defined by ProPublica in the data repository. Based on Brennan et al. (2009), we feel that `event` indicates some kind of crime that resulted in a prison sentence during a past observation period (we suspect the assessments were conducted by local probation officers during some period between January 2001 and December 2004), and the variable `end` denotes the number of days under observation (first event or end of study, whichever occurred first). The associated FINEtree recidivism algorithm based on `event` and `end` reaches 93% accuracy with AUC 0.92 on a test set (consist of 20% of the data). Also see Appendix A.5.

[15] One can slightly improve accuracy by combining hundreds or thousands of trees (based on only the admissible features) using random forest or boosting. But the opacity of such models renders them unfit for deployment in financial and bank sectors (Fahner 2018).

[16] Data: https://github.com/propublica/compas-analysis/raw/master/compas-scores-two-years.csv

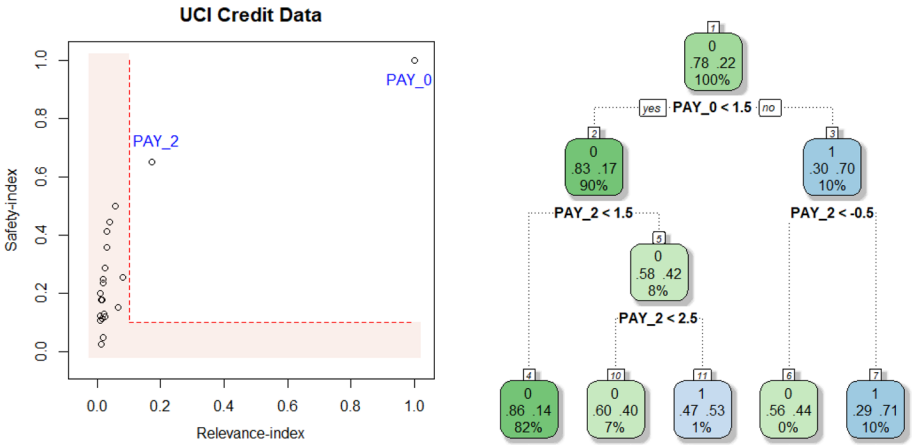



**Fig. 11** *Left*: Infogram of UCI credit card data. It selects two admissible features (i.e., those that are relevant and less-biased) that lie in the complementary of the "L"-shaped region. *Right*: The FINEtree (test data accuracy 82%)

### 4.2.5 FINEglm and application to marketing campaign

We are interested in the following question: how does one systematically build fairness-enhancing parametric statistical algorithms, such as a generalized linear model (GLM)?

***Example 11*** *Thera Bank Financial Marketing Campaign*. This is a case study about Thera Bank, the majority of whose customers are liability customers (depositors) with varying sizes of deposits—and among them, very few are borrowers (asset customers). The bank wants to expand its client network to bring more loan business and in the process, earn more through the interest on loans. To test the viability of this business idea they ran a small marketing campaign with $n = 5000$ customers where a 480 (= 9.6%) accepted the personal loan offer. Motivated by the healthy conversion rate, the marketing department wants to devise a much more targeted digital campaign to boost loan applications with a minimal budget.

*Data and the problem*. For each of 5000 customers, we have binary response $Y$: customer response to the last personal loan campaign, and 12 other features like customer's annual income, family size, education level, value of house mortgage if any, etc. Among these 12 variables, there are two protected features: `age` and `zip code`. We consider zip code as a sensitive attribute, since it often acts as a proxy for race.

Based on this data, we want to devise an AI-tool for automatic and *fair* digital marketing campaign that will maximize the targeting effectiveness of the advertising campaign while minimizing the discriminatory impact on protected classes to avoid legal landmines.

*Customer targeting using admissible machine learning*. Our approach is summarized below:

`Step 1`. Graphical tool for algorithmic risk management. Fig. 13 shows the infogram, which identifies two admissible features for loan decision: `Income` (annual income in $000), and `CCAvg` (Avg. spending on credit cards per month in $000).

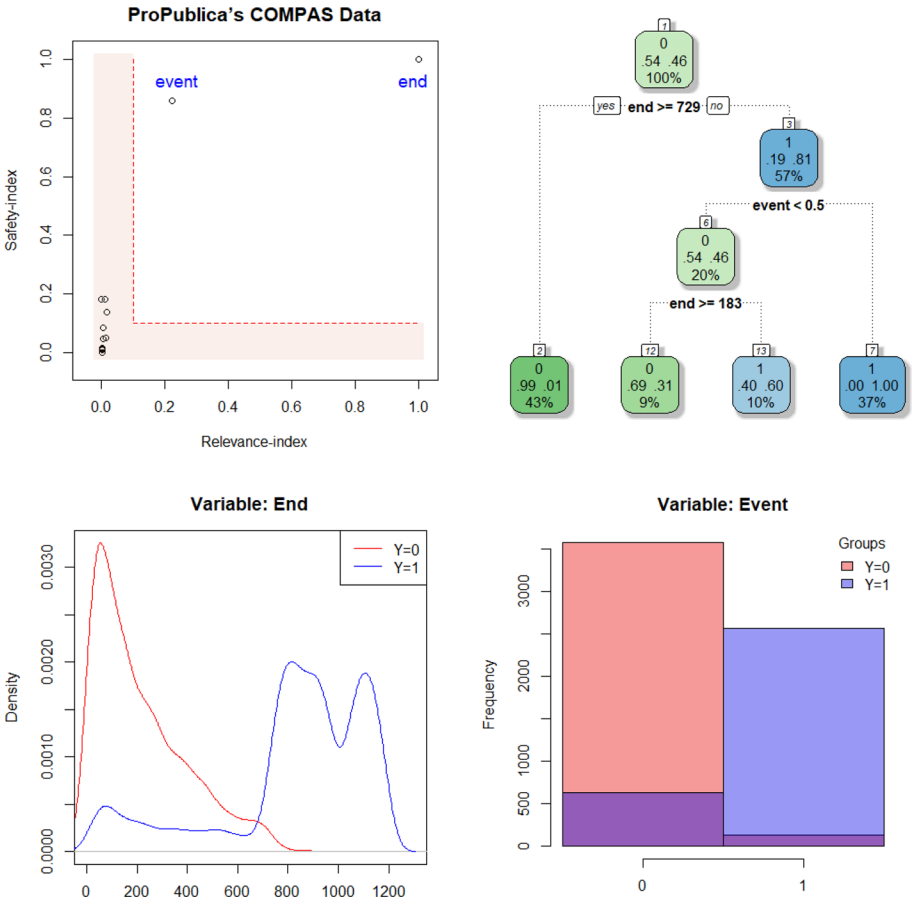



**Fig. 12** ProPublica's COMPAS Data: *Top row*: infogram and the estimated FINEtree. *Bottom row*: The two-sample distribution of the continuous variable `end` and binary `event` show their usefulness for predicting whether a defendant will recidivate or not

However, the two highly predictive variables `education` (education level: undergraduate, graduate, or advanced) and `family` (family size of the customer) turn out to be inadmissible, even though they look completely "reasonable" on the surface. Consequently, including these variables in a model can do more harm than good by discriminating against minority applicants.

***Remark 16*** Infogram provides an *explanatory* risk management interface that provides explanation and insights into *what* (are the key sources of bias) and *how* (to combat and mitigate unwanted bias)—leading to faster deployment of ML-models. Regulators can use infogram to quickly spot and remediate issues of historic discrimination; see Appendix A.3.

***Remark 17*** Infogram runs a 'combing operation' to distill down a large, complex problem to its *core* that holds the bulk of the "admissible information." In our problem, the useful

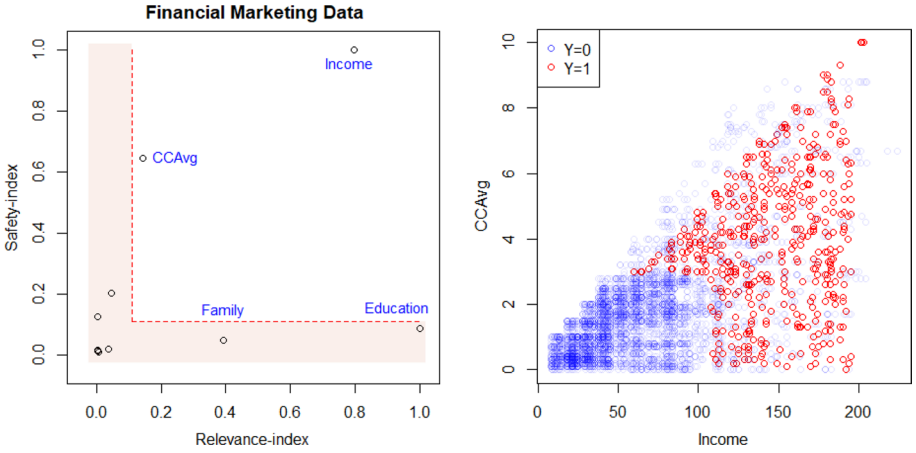



**Fig. 13** Thera Bank marketing campaign data. *Left*: infogram. *Right*: scatter plot based on the two admissible features; the color *blue* and *red* indicate two different classes

information is mostly concentrated into two variables—Income and CCAvg, as seen in the scatter diagram.

`Step 2`. FINE-Logistic model: We train a logistic regression model based on the two admissible features, leading to the following model:

$$\text{logit}\{\mu(x)\} \;=\; -6.13 \;+\; .04\,\texttt{Income} \;+\; .06\,\texttt{CCAvg}, \tag{3.11}$$

where $\mu(x) = \Pr(Y = 1|\mathbf{X} = \mathbf{x})$. This simple model achieves 91% accuracy. It provides a clear understanding of the 'core' factors that are driving the model's recommendation.

***Remark 18*** (Trustworthy algorithmic decision-making) FINEml models provide a transparent and self-explainable algorithmic decision-making system that comes with protection against unfair discrimination—which is essential for earning the trust and confidence of customers. The financial services industry can benefit from this tool.

`Step 3. FINElasso`. One natural question would be, How can we extend this idea to high-dimensional glm models? In particular, we are interested in the following question: Is there any way we can directly embed 'admissibility' into the lasso regression model? The key idea is as follows: use adaptive regularization by choosing the weights to be the inverse of safety-indices, as computed in formula (3.6) of definition 4. Estimate `FINElasso` model by solving the following adaptive version:

$$\widehat{\boldsymbol{\beta}}_{\text{FINE}} \;=\; \text{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} \Big[ -y_i(\mathbf{x}_i^T\boldsymbol{\beta}) + \log\left(1 + e^{\mathbf{x}_i^T\boldsymbol{\beta}}\right)\Big] \;-\; \lambda \sum_{j=1}^{p} w_j|\beta_j|, \tag{3.12}$$

where the weights are defined as

$$w_j^{-1} \;= \text{MI}\big(Y, X_j \mid \{S_1, \ldots, S_q\}\big). \tag{3.13}$$

The adaptive penalization in (3.12) acts as a bias-mitigation mechanism by dropping (that is, heavily penalizing) the variables with very low safety-indices. This whole procedure can be easily implemented using the `penalty.factor` argument of `glmnet` R-package (Friedman et al. 2010). No doubt a similar strategy can be adopted for other regularized methods such as ridge or elastic-net. For an excellent review on different kinds of regularization procedures, see Hastie (2020).

***Remark 19*** A full lasso on $\mathbf{X}$ selects the strong surrogates (variables `family` and `education`) as some of the top features due to their high predictive power, and hence carries enhanced risk of being discriminatory. On the other hand, an infogram-guided `FINElasso` provides an automatic defense mechanism for combating bias without significantly compromising accuracy.

***Remark 20*** (Towards A Systematic Recipe) This idea of data-adaptive 're-weighting' as a bias mitigation strategy, can be easily translated to other types of machine learning models. For example, to incorporate fairness into the traditional random forest method, choose splitting variables at each node by performing weighted random sampling. The selection probability is determined by

$$\Pr(\text{selecting variable } X_j) \;=\; \frac{F_j}{\sum_j F_j}, \tag{3.14}$$

where the F-values $F_j$ is defined in Eq. (3.6). This can be easily operationalized using the `mtry.select.prob` argument of the randomForest() function in `iRF` R-package. Following this line of thought, one can (re)design a variety of less-discriminatory ML techniques without changing a single architecture of the original algorithms.

## 5 Conclusion

> *Faced with the profound changes that AI technologies can produce, pressure for "more" and "tougher" regulation is probably inevitable.* (Stone et al. 2019).

Over the last 60 years or so—since the early 1960s—there's been an explosion of powerful ML algorithms with increasing predictive performance. However, the challenge for the next few decades will be to develop sound theoretical principles and computational mechanisms that *transform* those conventional ML methods into more safe, reliable, and trustworthy ones.

The fact of the matter is that doing machine learning in a 'responsible' way is much harder than developing another complex ML technique. A highly accurate algorithm that does not comply with regulations is (or will soon be) unfit for deployment, especially in safety-critical areas that directly affect human lives. For example, the Algorithmic Accountability Act[17] (see Appendix A.2) introduced in April 2019 requires large corporations (including tech companies, as well as banks, insurance, retailers, and many other

[17] Also see, EU's "Artificial Intelligence Act" released on April 21, 2021, whose key points are summarized in Appendix A.8.

consumer businesses) to be cognizant of the potential for biased decision-making due to algorithmic methods; otherwise, civil lawsuits can be filed against those firms. As a result, it is becoming necessary to develop tools and methods that can provide ways to *enhance* interpretability and efficiency of classical ML models while guarding against bias. With this goal in mind, this paper introduces a new kind of statistical learning technology and information-theoretic automated monitoring tools that can guide a modeler to *quickly* build "better" algorithms that are less-biased, more-interpretable, and sufficiently accurate.

One thing is clear: rather than being passive recipients of complex automated ML technologies, we need more general-purpose statistical *risk management tools* for algorithmic accountability and oversight. This is critical to the responsible adoption of regulatory-compliant AI-systems. This paper has taken some important steps towards this goal by introducing the concepts and principles of 'Admissible Machine Learning.'

## Appendix

### A.1 Proof of Theorem 1

The conditional entropy $H(Y \mid \mathbf{X}, \mathbf{S})$ can be expressed as

$$\begin{aligned} H(Y \mid \mathbf{X}, \mathbf{S}) &= \iint_{\mathbf{x},\mathbf{s}} H(Y \mid \mathbf{X} = \mathbf{x}, \mathbf{S} = \mathbf{s}) \, \mathrm{d}F_{\mathbf{x},\mathbf{s}} \\ &= \iint_{\mathbf{x},\mathbf{s}} \Big\{ - \int_y f_{Y|\mathbf{X},\mathbf{S}}(y, \mathbf{x}|\mathbf{s}) \log \big( f_{Y|\mathbf{X},\mathbf{S}}(y, \mathbf{x}|\mathbf{s}) \big) \, \mathrm{d}y \Big\} \, \mathrm{d}F_{\mathbf{x},\mathbf{s}} \\ &= - \iiint_{\mathbf{x},\mathbf{s},y} \log \big( f_{Y|\mathbf{X},\mathbf{S}}(y, \mathbf{x}|\mathbf{s}) \big) \, \mathrm{d}F_{\mathbf{x},\mathbf{s},y}. \end{aligned} \tag{5.1}$$

Similarly,

$$\begin{aligned} H(Y \mid \mathbf{S}) &= \int_{\mathbf{s}} H(Y \mid \mathbf{S} = \mathbf{s}) \, \mathrm{d}F_{\mathbf{s}} \\ &= \int_{\mathbf{s}} \Big\{ - \int_y f_{Y|\mathbf{S}}(y|\mathbf{s}) \log \big( f_{Y|\mathbf{S}}(y|\mathbf{s}) \big) \, \mathrm{d}y \Big\} \, \mathrm{d}F_{\mathbf{s}} \\ &= - \iiint_{\mathbf{x},\mathbf{s},y} \log \big( f_{Y|\mathbf{S}}(y|\mathbf{s}) \big) \, \mathrm{d}F_{\mathbf{x},\mathbf{s},y}. \end{aligned} \tag{5.2}$$

Take the difference $H(Y|\mathbf{S}) - H(Y|\mathbf{X}, \mathbf{S})$ by substituting (5.2) and (5.1) to complete the proof. □

### A.2 The algorithmic accountability act

This bill[18] was introduced by Senators Cory Booker (D-NJ) and Ron Wyden (D-OR) in the Senate and Rep. Yvette Clarke (D-N.Y.) in the House on April, 2019. It requires large companies to conduct automated decision system impact assessments of their algorithms.

[18] https://www.congress.gov/bill/116th-congress/house-bill/2231/all-info

Entities that develop, acquire, and/or utilize AI must be cognizant of the potential for biased decision-making and outcomes resulting from its use, otherwise civil lawsuits can be filed against those firms. Interestingly, on Jan. 13, 2020, the Office of Management and Budget released a draft memorandum[19] to make sure the federal government doesn't over-regulate industry's AI to the extent that it hampers innovation and development.

### A.3 Fair housing act's disparate impact standard

Detecting inadmissible (proxy) variables can be used as a first defense against algorithmic bias. Consider the Fair Housing Act's Disparate Impact Standard[20] (U.S. Aug. 19, 2019)–according to §100.500 (c)(2)(i) of the Act, a defendant can rebut a claim of discrimination by showing that "none of the factors used in the algorithm rely in any material part on factors which are substitutes or close proxies for protected classes under the Fair Housing Act." Therefore regulators, judges, and model developers can use `infogram` as a statistical diagnostic tool to keep a check on the algorithmic disparity of automated decision systems.

### A.4 Beware of The "Spurious Bias" problem

Using a real data example, here we alert practitioners some of the flaws of current fairness criteria and discuss their remedies. Consider the admission data shown in Table 1. We are interested to know: is there a gender bias in the admission process?

- Marginal analysis: the overall acceptance rate in two departments for female applicants is 37%, whereas for male applicants it is roughly 50%. The disparity can be quantified using the adverse impact ratio (AIR), also known as disparate impact:

$$\text{AIR}(Y, G) = \frac{\Pr(Y = 1 \mid G = \text{female})}{\Pr(Y = 1 \mid G = \text{male})} = \frac{.37}{.50} = 0.74 < 0.80 \tag{5.3}$$

  The conventional "80% rule"[21] indicates that the admission process is biased.
- The bias-reversal phenomena: admission chances within Department I: Male 63% (male), and female 68%; within Department II: Male 33%, and female F 35%. Thus, when we investigate the admissions by department, the discrimination against women vanishes; in fact, the bias gets reversed (in the favor of women)! • Department-specific "subgroup" analysis: Here we investigate the adverse impact ratio (AIR) within each department.

For Dept I (no bias):

[19] The draft memo is available at:whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf

[20] https://www.govinfo.gov/content/pkg/FR-2019-08-19/pdf/2019-17542.pdf

[21] The US Equal Employment Opportunity Commission states that fair employment should abide the 80% rule: the acceptance rate for any group should be no less than 80% of that of the highest-accepted group.

**Table 1** Admission data classified by gender and departments. This is actually a part of the 1973 UC Berkeley graduate admission data; here, for simplicity, we have taken the data of Departments B and D

| Dept. (D) | Gender (G) | Admitted ($y = 1$) | Rejected ($y = 0$) |
|---|---|---|---|
| I | Male | 353 | 207 |
| | Female | 17 | 8 |
| II | Male | 138 | 279 |
| | Female | 131 | 244 |

$$\text{AIR}(Y, G \mid D = \text{I}) \;=\; \frac{\Pr(\text{Y} = 1 \mid G = \text{male})}{\Pr(\text{Y} = 1 \mid G = \text{female})} \;=\; .63/.68 = 0.92 \;>\; 0.80. \tag{5.4}$$

For Dept II (no bias):

$$\text{AIR}(Y, G \mid D = \text{II}) \;=\; \frac{\Pr(\text{Y} = 1 \mid G = \text{male})}{\Pr(\text{Y} = 1 \mid G = \text{female})} \;=\; .33/.35 = 0.94 \;>\; 0.80. \tag{5.5}$$

Eqs. (5.3)-(5.5) present us with a paradoxical situation. What will be our final conclusion on the fairness of the admission process? How to resolve it in a principled way?

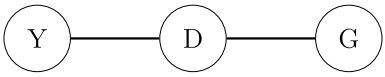
- A resolution: Compute a measure of overall (university-wide) discrimination by ALFA-statistic (see definition 5 for more details):

$$\alpha_Y \;:=\; \text{MI}(Y, G|D) \;=\; \sum_{d=0}^{1} \Pr(D = d)\text{MI}(Y, G|D = d), \tag{5.6}$$

  where $\alpha$-inadmissibility statistic measures the discrimination (how predictive the admission variable $Y$ is based on gender $G$) in a particular department's admission. Applying the formula (2.6) we get

$$\widehat{\alpha}_Y = \widehat{\text{MI}}(Y, G|D) = 0.000285, \;\; \text{with } p\text{-value: } 0.715.$$

This suggests $Y \perp\!\!\!\perp G \mid D$, i.e., the gender contains no additional predictive information for admission beyond what is already captured by the department variable. The apparent gender bias can be 'explained away' by the choice of the department. Graphically, this can be represented as a Markov chain:




Note that there is no direct link between the gender (G) and admission (Y). Conclusion: there is no evidence of any direct sex-discrimination in the admission process.

- Improved AIR measure: One can generalize the (marginal) adverse impact ratio (5.3) to the following conditional one (which is similar in spirit to eq. (5.6)):

$$\text{CAIR}(Y, G|D) \;=\; \int \text{AIR}(Y, G|D = d)\; \text{d}F_D, \tag{5.7}$$

  which, in this case, can be decomposed as

$$\text{CAIR}(Y, G|D) \;=\; \Pr(D = \text{I})\text{AIR}(Y, G|D = \text{I}) \;+\; \Pr(D = \text{II})\text{AIR}(Y, G|D = \text{II}). \quad (5.8)$$

Applying (5.8) for our Berkeley example data yields the following estimate:

$$\begin{aligned}\widehat{\text{CAIR}}(Y, G|D) =& 0.43 \times 0.92 + 0.57 \times 0.94\\ =& 0.93 \;>\; 0.80.\end{aligned}$$

This shows no evidence of sex bias in graduate admissions! The moral is: beware of spurious bias, and be aware of two types of errors that might occur due to an incorrect fairness-metric: falsely rejecting a fair algorithm as unfair (Type-I fairness error), and falsely accepting an unfair algorithm as fair (Type-II fairness error).

### A.5 Revisiting COMPAS data

There is another version of the COMPAS data[22] (binarized features) that researchers have used for evaluating the accuracy of their algorithms. This dataset contains a list of hand-picked $p = 22$ features over $n = 10,747$ criminal records. Goal is to build an interpretable and accurate recidivism prediction model. Infogram-selected COREtree is displayed below (Fig.14).

10-fold cross-validation shows $(72 \pm 1.50)\%$ classification accuracy of our model, which is close to the best known performance on this version of the COMPAS data.

### A.6 Two cultures of machine learning

Black-box ML culture: it builds large complex models, keeping solely the predictive accuracy in mind. White-box ML culture: it directly builds interpretable models, often by enforcing domain-knowledge-based constraints on traditional ML algorithms like decision tree or neural net. Orthodox 'black-or-white thinkers' of each camp have been at loggerheads for some time. This raises the question: is there any way to get the best of both worlds? If so, how?

*An Integrated Third Culture*: In this paper, we have taken the middle path between two extremes. We leverage (instead of boycotting) the power (scalability and flexibility) of modern machine learning methods by viewing them as a heavy-duty "toolkit" that can efficiently drill through big complex datasets to systematically search for the hidden admissible models.

### A.7 COREtree: Iris data

The dataset includes three kinds of iris flowers (setosa, versicolor, or virginica) with 50 samples from each class. The task is to develop a model (preferably a compact model based on only important features) to accurately classify iris flowers based on their sepals and petals' length and width ($p = 4$). Before we start our analysis, it is important to be aware of the highly-correlated nature of the 4-features; the estimated $4 \times 4$ correlation matrix is displayed below:

[22] https://raw.githubusercontent.com/Jimmy-Lin/TreeBenchmark/master/datasets/compas/data.csv

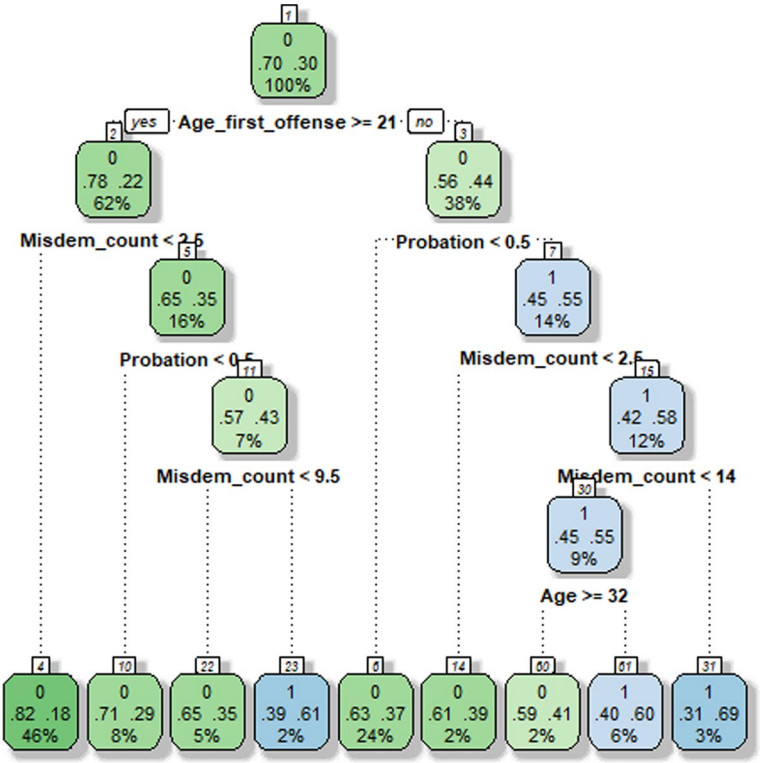



**Fig. 14** Infogram-selected COREtree

$$\hat{\Sigma}_{\rho} = \begin{bmatrix} 1.000 & -0.118 & \mathbf{0.872} & \mathbf{0.818} \\ -0.118 & 1.000 & -0.428 & -0.366 \\ \mathbf{0.872} & -0.428 & 1.000 & \mathbf{0.963} \\ \mathbf{0.818} & -0.366 & \mathbf{0.963} & 1.000 \end{bmatrix} \tag{5.9}$$

The infogram for the iris data, constructed using the recipe given in sect. 3.1, is shown at the top-left corner of Fig. 15, which clearly identifies `petal.length` and `petal.width` as the *core* relevant features. Since we have reduced the problem to a bivariate one (variables: `petal.length` and `petal.width`), we can now simply plot the data. This is done in the top-right of Fig. 15. We can even visually draw the linear decision surfaces to separate the three classes; see the red and blue lines in the scatter plot. Finally, we train a decision tree classifier based on the selected core features: `petal.length` and `petal.width`. The estimated `COREtree` is shown in the bottom panel, which gives a beautifully crisp (readily interpretable) decision rule for classifying iris flowers.

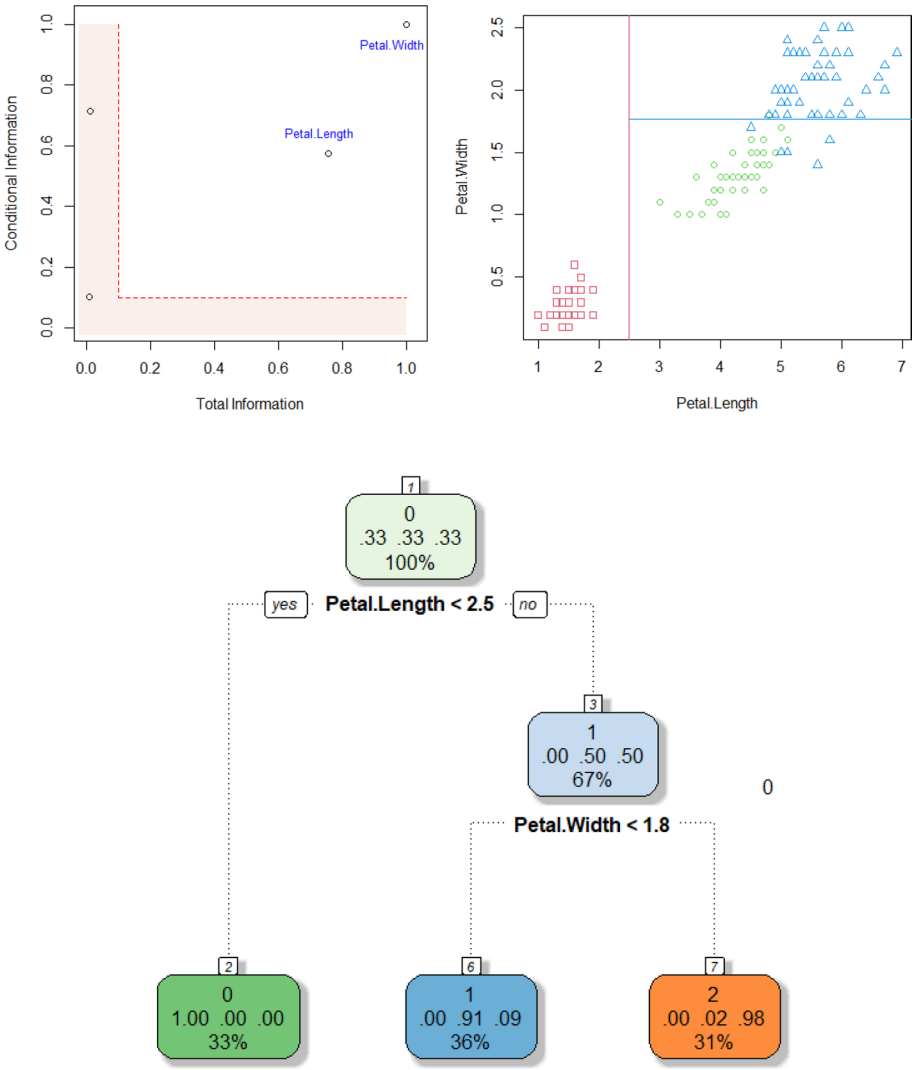



**Fig. 15** Iris data analysis. *Top left*: infogram; *top right*: the scatter plot of the data based on the selected core features; three different classes are indicated by red, green, and blue colors; *bottom*: the estimated decision tree classifier using the variables petal-length and petal-width (Colour figure online)

## A.8 EU's artificial intelligence act

On 21st April 2021, the European Union (EU) unveiled strict regulations[23] to govern high-risk AI systems, which provides one of the first formal and comprehensive regulatory

[23] The full report is available online at https://bit.ly/EUAI_act. Also see the New York Times article https://www.nytimes.com/2021/04/16/business/artificial-intelligence-regulation.html

frameworks on AI. Few key takeaways from the report:

- A risk management system shall be established, implemented, documented and maintained in relation to high-risk AI systems.
- In identifying the most appropriate risk management measures, the following shall be ensured: elimination or reduction of risks as far as possible through adequate design and development.
- Bias monitoring, detection, and correction mechanism should be at place for high-risk AI systems in the pre-as well as the post-deployment stages.
- High-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately.
- High-risk AI systems should equip with appropriate human-machine interface tools—which allows the system to be effectively overseen by decision-makers during the period in which the AI system is in use.
- High-risk AI technology providers shall ensure that their systems undergo regulatory compliant assessments. If the AI system is not in conformity with the requirements, they need to take the necessary corrective actions before putting them into service. Companies that fail to do so could face fines of up to 6% of their global sales.

### A.9 Existing bias mitigation strategies

Broadly speaking, existing bias mitigation staregies comes in three flavors: (i) *Pre-processing*: Re-weights or re-labels the original data to minimize the given fairness measure; (ii) *In-processing*: Optimizes hyperparameters of a blackbox ML by imposing the given fairness measure as constraint; and (iii) *Post-processing*: Controls the given (un)fairness metric by artificially changing the classification thresholds for each protected group.

Unfortunately, all three categories of unfairness mitigation strategies carry serious legal compliance risks. The reason being, these methods entertain either (i) data massaging/manipulation; or (ii) using protected attributes during model training or decision making, both of which are forbidden by law.

**Acknowledgements** This paper is dedicated to the memory of Leland Wilkinson (November 5, 1944–December 10, 2021) --- a wonderful person, friend, and one of the strong supporters of this research project. The author thanks the editor, associate editor, and four anonymous reviewers for their helpful suggestions. The author was benefited from many useful discussions with Erin LeDell, Michael Guerzhoy, Hany Farid, Julia Dressel, Beau Coker, and Hanchen Wang on demystifying some aspects of COMPASS data; Daniel Osei on the data pre-processing steps of Lending Club loan data.

**Author Contributions** Not applicable, since this is a single authored paper.

**Availability of data and material** The datasets used for analysis are available upon request to the author.

## Declarations

**Code availability** The R-code written for the analysis is available upon request to the corresponding author.

**Conflict of interest** The authors declare that they have no conflicts of interest/competing interests.

**Ethics approval** Not applicable.

**Consent to participate** Not applicable.

**Consent for publication** Not applicable.

## References


Allen, B., Agarwal, S., Coombs, L., Wald, C., & Dreyer, K. (2021). 2020 ACR data science institute artificial intelligence survey. *Journal of the American College of Radiology*

Berrett, T. B., Wang, Y., Barber, R. F., & Samworth, R. J. (2019). The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*

Blattner, L., & S. Nelson (2021). How costly is noise? Data and disparities in consumer credit. arXiv preprint: arXiv:2105.07554.

Breiman, L., et al. (2004). Population theory for boosting ensembles. *The Annals of Statistics, 32*(1), 1–11.

Brennan, T., Dieterich, W., & Ehret, B. (2009). Evaluating the predictive validity of the compas risk and needs assessment system. *Criminal Justice and Behavior, 36*(1), 21–40.

Candes, E., Fan, Y., Janson, L., & Lv, J. (2018). Panning for gold: 'model-x' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 80*(3), 551–577.

Chouldechova, A., & Roth, A. (2020). A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM, 63*(5), 82–89.

Fahner, G. (2018). Developing transparent credit risk scorecards more effectively: An explainable artificial intelligence approach. *Data Analysis, 2018,* 17.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software, 33*(1), 1.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics,* 1189–1232.

Hastie, T. (2020). Ridge regularizaton: an essential concept in data science. *Technometrics* (forthcoming).

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: The lasso and generalizations*. Boca Raton: CRC Press.

Hellman, D. (2020). Measuring algorithmic fairness. *Virginia Law Review, 106,* 811.

Kleinberg, J. (2018). Inherent trade-offs in algorithmic fairness. In *Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems*, pp. 40–40.

Lakkaraju, H., & O. Bastani (2020). "How do I fool you?" manipulating user trust via misleading black box explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 79–85.

Mukhopadhyay, S., & K. Wang (2020). Breiman's "Two Cultures" revisited and reconciled. arXiv:2005.13596, 1–51.

Narayanan, A. (2018). Translation tutorial: 21 fairness definitions and their politics. In *Proceedings of the Conference Fairness Accountability Transportation, New York, USA*, Vol. 1170.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453.

Reardon, S. (2019). Rise of robot radiologists. *Nature, 576*(7787), S54.

Rosenbaum, P. R. (1984). Conditional permutation tests and the propensity score in observational studies. *Journal of the American Statistical Association, 79*(387), 565–574.

Stone, P., R. Brooks, E. Brynjolfsson, et al. (2019). One hundred year study on artificial intelligence. *Stanford University;* https://ai100.stanford.edu.

Thrun, S. B., J. Bala, E. Bloedorn, I. Bratko, B. Cestnik, J. Cheng, K. De Jong, S. Dzeroski, S. E. Fahlman, D. Fisher, et al. (1991). The monk's problems a performance comparison of different learning algorithms.

Wall, L. D. (2018). Some financial regulatory implications of artificial intelligence. *Journal of Economics and Business, 100*, 55–63.

Wyner, A. D. (1978). A definition of conditional mutual information for arbitrary ensembles. *Information and Control, 38*(1), 51–59.

Yeh, I.-C., & Lien, C.-H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications, 36*(2), 2473–2480.

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Medicine, 15*(11), e1002683.
Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, pp. 56–85.