Computational Statistics and Data Analysis 182 (2023) 107701

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Explaining classifiers with measures of statistical association

Emanuele Borgonovo [a,*], Valentina Ghidini [a], Roman Hahn [a], Elmar Plischke [b]

[a] *Department of Decision Sciences, Bocconi University, Milan, 20136, Italy*
[b] *Faculty of Energy Research and Economics, Clausthal University of Technology, Clausthal-Zellerfeld, 38678, Germany*

ARTICLE INFO

Article history:
Received 28 November 2021
Received in revised form 13 January 2023
Accepted 13 January 2023
Available online 1 February 2023

Keywords:
Explainability
Sensitivity measures
Measures of statistical association

ABSTRACT

A new class of probabilistic sensitivity measures that quantifies the degree of association between covariates and generic targets used in classification is proposed, and it is shown that such class possesses the zero-independence property. Corresponding estimators are introduced, asymptotic consistency is proven and bootstrap is used to quantify uncertainty in the estimates. The use of the new dependence measures as explanations in a statistical machine learning context is illustrated. The resulting approach, called Xi-method, is demonstrated through applications involving different data formats: tabular, visual and textual.

1. Introduction

The growing size and complexity of data structures and the simultaneous need of accurate predictions force analysts to employ black-box rather than transparent machine learning (ML) models in an increasing number of applications. While the success of such models extends the range of statistical ML applications, it also increases the need of methods that aid explainability (Dunson, 2018; Rudin, 2019; Murdoch et al., 2019).

Determining feature importance is essential for model simplification, dimensionality reduction, and for understanding whether predictions are at risk of unfair discrimination (Fisher et al., 2019). As underlined in Murdoch et al. (2019), there is a variety of techniques for calculating feature importance. Methods such as the Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016), the Layerwise Relevance Propagation (LRP) (Binder et al., 2016), and the SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017) yield feature importance measures at the individual prediction level. Techniques such as permutation or removal (Breiman, 2001), Shapley values (Lundberg and Lee, 2017; Lundberg et al., 2019), knock-offs (Barber and Candés, 2015; Candès et al., 2018) and Sure Independence Screening (Fan and Lv, 2008) provide an indication of importance at the dataset level.

Measures of statistical association are an alternative way for assessing feature importance: they provide dataset scores and are model agnostic (Murdoch et al., 2019). While they have been widely studied beginning with works such as Pearson (1905), Hotelling (1936) and Renyi (1959), the size and complexity of modern datasets have generated new interest in their construction, as the recent works of Chatterjee (2021), Pan et al. (2020) and Wiesel (2022) highlight. We note that several measures of statistical association rely on the assumption that the target is a real number (or vector). However, in certain ML applications, targets and/or features may be images or objects, for which a mathematical order relation may not be appropriate. To illustrate, consider an image recognition task in which the target consists of three different types of pictures

[*] Corresponding author.
E-mail address: emanuele.borgonovo@unibocconi.it (E. Borgonovo).

https://doi.org/10.1016/j.csda.2023.107701

representing, say, cats, tigers and rabbits. Then the alphabetical ordering "cat", "rabbit", "tiger" is a possible ordering, but it is as valid as the ordering "cat", "tiger", "rabbit" that is based on the number of the characters in each name, and none of them qualifies as a natural ranking. The issue is underlined in recent works (Da Veiga, 2021; Marrel and Charibon, 2021) and the definition of probabilistic sensitivity measures for non-ordered outputs is a topical research subject.

To bridge this gap we propose a family of measures of statistical association whose definition is well-posed also for non-ordered data. Our intuition is to rely on separation measurements between probability mass functions. Here, by separation measurement we mean any distance or divergence between probability mass functions that is positive, and that is null if and only if the probability mass functions coincide. Then, we show that the new class of sensitivity indices complies with Renyi's postulate D of measures of statistical dependence (Renyi, 1959). This postulate, called zero-independence property in the following, requires that a measure of association is null if and only if the two random variables are statistically independent. We address the estimation of this new class of indicators for generic samples, and discuss their asymptotic convergence. We then use these probabilistic sensitivity measures in the context of explainability. A relevant aspect related to measures of statistical association is that they can be computed directly on the original dataset without the need of actually fitting a machine learning model. Thus, not only are they model agnostic in explaining the behavior of a black box, but they also provide pre-hoc explanations. Our intuition is then to compare explanations provided by measures of statistical association first calculated on the original data (the pre-hoc explanations) and then on the forecasts of the machine learning model fitted to the data (post-hoc explanations). This comparison provides an indication on whether the ML model predictions capture the statistical dependence originally present in the data. We call the resulting approach Xi-method. We proceed as follows: first, we discuss the methodological framework, with a focus on the choice of the separation measurement. Then, we address estimation, with a focus on a partition-based approach that allows us to obtain the explanations from a given dataset. We perform experiments to investigate the asymptotic convergence of the estimates and highlight limitations of the approach with particular reference to the curse of dimensionality. We then test the approach by performing experiments on datasets of alternative types, comprising tabular, image and textual data. For the first set of experiments, in which the ML model is a random forest, we compare the pre-hoc and post-hoc explanations provided by measures of statistical dependence with the post-hoc explanations delivered by variable importance measures based on Split and Count (Breiman et al., 1984) and permutation (Breiman, 2001). The rationale of this comparison is to test the agreement between results provided by measures of statistical association with one representative of variable importance measures that are post-hoc and model-dependent (Split and Count) and one representative of measures that are post-hoc but model agnostic (permutation feature importance). Measures of statistical dependence produce additional and complementary insights with respect to alternatives currently in use, with the advantage of being model agnostic and computationally convenient.

The remainder of the work is organized as follows: Section 2 sets up the relevant framework and reviews the literature. Section 3 introduces the new dependence measures for classification. Section 4 presents the Xi-method. Section 5 illustrates results for alternative datasets. Section 6 offers conclusions.

2. Feature importance and measures of association

Let $(\Omega, \mathcal{B}(\Omega), \mathbb{P})$ be a reference probability space (where Ω is called sample space, $\mathcal{B}(\Omega)$ is a Borel σ-algebra, and $\mathbb{P} : \mathcal{B}(\Omega) \to [0,1]$ is the reference probability measure) and let $\mathbf{X} = (X_1, X_2, \dots, X_{n_{\mathbf{X}}})$ and L be random variables on this reference space with supports $\mathcal{X}$ and $\mathcal{L}$, respectively. Let also $\mathbf{X}$ have the meaning of vector of features and let L represent targets (labels). For example, in image classification, we might be dealing with a set of images of a given resolution, say n_{pixels}. Frequently, pixels themselves are selected as features, yielding $\mathbf{X} = (X_1, X_2, \dots, X_{n_{\text{pixels}}})$. The set of labels is then the list of objects depicted in the corresponding images. The support of each pixel $(\mathcal{X}_i, i = 1, \dots, n_{\text{pixels}})$ is its range of values (for instance $\mathcal{X}_i = (0,1)$ in case of grayscale images), and the overall support $\mathcal{X}$ is the Cartesian product of the ranges of all pixels. The support of the target $\mathcal{L}$ is the list of all the possible labels associated to each image in the dataset, that is the set of objects represented in the data collection (e.g., cats, rabbits, tigers as we were mentioning in the introduction).

We are interested in associating realizations of $\mathbf{X}$ to realizations of L through the input-output mapping $g : \mathcal{X} \times \Theta \to \mathcal{L}$ (Hastie et al., 2009; Zhao and Hastie, 2021)

$$\Lambda = g(\mathbf{X}, \boldsymbol{\theta}), \tag{1}$$

where Λ denotes a model forecast and $\boldsymbol{\theta} \in \Theta$ is a vector of parameters (or rules). In supervised learning, a dataset $(\mathbf{x}^{(n)}, L^{(n)})$, $n = 1, 2, \dots, N$, of realizations of $\mathbf{X}$ and L is usually available (henceforth, N denotes the sample size). Splitting the sample into training and testing subsamples of sizes N^{Tr} and N^{Te} respectively (on the meaning and rationale for training and testing in supervised learning see classical references such as Hastie et al. (2009)), the parameters of the input-output mapping are determined via the solution of an optimization problem of the form

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left\{ \frac{1}{N^{Tr}} \sum_{n=1}^{N^{Tr}} \mathcal{C}\left(L^{(n)}; g(\mathbf{x}^{(n)}, \boldsymbol{\theta})\right) \right\}, \tag{2}$$

where $\mathcal{C} : \mathcal{L} \times \mathcal{L} \to \mathbb{R}$ is a suitably defined objective (loss) function. In the remainder, we shall use the shorter $\widehat{g}(\mathbf{X})$ for $g(\mathbf{X}, \boldsymbol{\theta}^*)$. Then, let $\mathbf{X}_{\pi(i)}$ denote the design matrix in which we have randomly permuted the realizations of the i^{th} feature

X_i. Let also $\mathcal{C}(\boldsymbol{X};\boldsymbol{\theta}^*) = \frac{1}{N^{Tr}}\sum_{n=1}^{N^{Tr}} \mathcal{C}(L^{(n)}; g(\mathbf{x}^{(n)}, \boldsymbol{\theta}^*))$ denote the value of the optimized loss function and $\mathcal{C}(\mathbf{X}_{\pi(i)};\boldsymbol{\theta}^*) = \frac{1}{N^{Tr}}\sum_{n=1}^{N^{Tr}} \mathcal{C}(L^{(n)}, g(\mathbf{x}_{\pi(i)}^{(n)}, \boldsymbol{\theta}^*))\}$ the expected loss registered when feature i is permuted, both computed on the training dataset. Then, the permutation importance of feature X_i is given by

$$\mathrm{PI}_i = \mathcal{C}(\mathbf{X}_{\pi(i)};\boldsymbol{\theta}^*)/\mathcal{C}(\boldsymbol{X};\boldsymbol{\theta}^*). \tag{3}$$

This indicator measures the deterioration in the ML model performance caused by removing the dependence between Y and X_i.

Measures of statistical dependence instead quantify importance from a different perspective, evaluating the degree of association between the target and one or more features. The problem of measuring statistical association roots back to works such as Pearson (1895, 1905) and Hotelling (1936) and in the statistical literature we find several measures of association and tests for independence. Recently, renewed interest has been generated by the type and size of modern datasets (see Chatterjee (2021) for a review). In particular, among measures of statistical dependence recently studied, we find the Hilbert–Schmidt Independence Criterion (Gretton et al., 2005; Da Veiga, 2015), distance correlation (Székely et al., 2007; Székely and Rizzo, 2009; Chaudhuri and Hu, 2019) as well as a new correlation coefficient (Chatterjee, 2021).

To provide the background needed for the remainder of our investigation, we recall the following definition.

Definition 1 *(Separation Measurement (Glick, 1975)).* Let $\mathcal{P}$ be the set of all probability measures on $(\Omega, \mathcal{B}(\Omega))$. A separation measurement between $\mathbb{P}, \mathbb{Q} \in \mathcal{P}$ is given by a function $\zeta : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ such that a) $\zeta(\mathbb{P}, \mathbb{Q}) \geq 0$, and b) $\zeta(\mathbb{P}, \mathbb{Q}) = 0$ if and only if $\mathbb{P} = \mathbb{Q}$.

Let now Y and $X \subseteq \mathbf{X}$ (i.e. X is a subset of one or more covariates of interest contained in the original design matrix $\mathbf{X}$) be random variables on $(\Omega, \mathcal{B}(\Omega), \mathbb{P})$, and denote by $\mathbb{P}_Y$ and $\mathbb{P}_X$ their marginal laws, and by $\mathbb{P}_{Y|X}$ the conditional law of Y given X. Without loss of generality, we will assume X to be a univariate random variable (if not specified otherwise).

Definition 2 *(Probabilistic Sensitivity Measure).* We define

$$\xi_X = \mathbb{E}_X[\zeta(\mathbb{P}_Y, \mathbb{P}_{Y|X})] \tag{4}$$

as the probabilistic sensitivity measure of X with respect to Y based on the separation measurement $\zeta(\cdot,\cdot)$.

In Equation (4), the expectation is taken with respect to the law of X, that is,

$$\xi_X = \int_{\mathcal{X}} \zeta(\mathbb{P}_Y, \mathbb{P}_{Y|X=x}) dF_X(x), \tag{5}$$

where $F_X(x)$ is the cumulative distribution function of X and the integral is interpreted in a Riemann-Stieltjes sense. To illustrate, if X is continuous with density $f_X(x)$ then Equation (5) becomes

$$\xi_X = \int_{\mathcal{X}} \zeta(\mathbb{P}_Y, \mathbb{P}_{Y|X=x}) f_X(x) dx. \tag{6}$$

If X is discrete with realizations $x_1, x_2, \ldots, x_n$, then Equation (5) becomes

$$\xi_X = \frac{1}{n}\sum_{i=1}^{n} \zeta(\mathbb{P}_Y, \mathbb{P}_{Y|X=x_i}) p(X = x_i). \tag{7}$$

Also notice that if Y and X are independent, then $\mathbb{P}_Y = \mathbb{P}_{Y|X}$, implying $\zeta(\mathbb{P}_Y, \mathbb{P}_{Y|X}) = 0$ and consequently $\xi_X = 0$. This means that the index ξ_X complies with the zero-independence property (Renyi, 1959).

The choice of the separation measurement in (4) determines the properties of the dependence measure ξ_X. For instance, if $\zeta(\cdot,\cdot)$ is invariant for monotonic transformation of Y then ξ_X is also monotonic transformation invariant (Borgonovo et al., 2014). If Y is absolutely continuous then selecting the Kullback-Leibler divergence as separation measurement has a corresponding probabilistic sensitivity measure equal to the mutual information between X and Y (Soofi, 1994). Recently, Taverniers et al. (2021) propose this importance measure in the context of neural network interpretability. They develop a deep neural network to emulate the behavior of a complex system in a forecasting task, and then the mutual information is used to quantify feature importance with respect to the neural network predictions. Alternatively, if one selects the separation measurement as the Cramér-von Mises distance as in Gamboa et al. (2018), one obtains

$$\xi_X^{\mathrm{CvM}} = \mathbb{E}_X\left[\int_{\mathbb{R}} \left(F_Y(y) - F_{Y|X}(y)\right)^2 dF_Y(y)\right]. \tag{8}$$

Table 1
Three possible separation measurements based on the 1-norm, 2-norm and Kuiper distance between probability mass functions.

1-norm	2-norm	Kuiper distance
$\zeta^1(\mathbf{p},\mathbf{q})=\sum_{l=1}^{n_L}\lvert p_l-q_l\rvert$	$\zeta^2(\mathbf{p},\mathbf{q})=\sum_{l=1}^{n_L}(p_l-q_l)^2$	$\zeta^{\mathrm{KU}}(\mathbf{p},\mathbf{q})=\mathrm{range}(\mathbf{p}-\mathbf{q})$

Notice that ξ_X^{CvM} is the limiting value of Chatterjee's new correlation coefficient (Chatterjee, 2021). Both the mutual information and ξ_X^{CvM} are then transformation invariant and possess the zero-independence property.

The estimation of probabilistic sensitivity measures has been extensively studied, with computational breakthroughs obtained in works such as Chan et al. (2000); Saltelli (2002); Strong et al. (2012), and Gamboa et al. (2016), among others. Relevant to our work is a given-data estimation approach that enables direct estimation from a dataset. The key-intuition dates back to Pearson (1905), and is extensively studied in works such as Strong et al. (2012), Strong and Oakley (2013) for the calculation of variance-based sensitivity measures, and Plischke et al. (2013) for distribution-based sensitivity measures. Recently, Gamboa et al. (2020) show that the newly introduced Chatterjee's rank-based correlation coefficient (Chatterjee, 2021) can be used as a given-data estimator for ξ_X^{CvM}. The connection leads to an elegant and advantageous approach for the calculation of a global sensitivity measure in the form of Equation (4) from a given dataset. However, the advantage is lost in a classification setting, because the targets are not necessarily elements of an ordered space and cannot be ranked univocally. This then opens the question of defining probabilistic sensitivity measures for classification tasks. We address this issue next.

3. Probabilistic sensitivity measures for supervised classification

Let $\mathcal{L}=\{\ell_1,\ell_2,\ldots,\ell_{n_L}\}$ denote the support of L, i.e. the set of labels in the target variable of interest. Without loss of generality, we assume that for all $\ell\in\mathcal{L}$, $\mathbb{P}(\{L=\ell\})>0$. $\mathcal{L}$ represents in general a list of objects and consequently the realizations of L may not be orderable. Let $\mathbf{p}_L$ be the probability mass function (pmf) of L and $\mathbf{p}_{L|X}$ the conditional pmf given a feature (or a feature group) X. Without loss of generality, in the remainder we will consider X to be a single feature, if not explicited otherwise. We recall that $\mathbf{p}_L=\{\mathbb{P}(\{L=\ell_1\}),\ldots,\mathbb{P}(\{L=\ell_{n_L}\})\}=\{p_1,p_2,\ldots,p_{n_L}\}$ is a probability mass function if $p_l\geq 0$ for all $l=1,\ldots,n_L$ and $\sum_{l=1}^{n_L}p_l=1$. Let $\mathcal{P}^{\mathrm{mf}}$ denote the set of all probability mass functions on $(\mathcal{L},\mathcal{P}(\mathcal{L}))$ and let $\zeta(\cdot,\cdot)$ denote a separation measurement between probability mass functions, $\zeta:\mathcal{P}^{\mathrm{mf}}\times\mathcal{P}^{\mathrm{mf}}\to\mathbb{R}$. By Definition 1, ζ is such that given $\mathbf{p},\mathbf{q}\in\mathcal{P}$ it holds $\zeta(\mathbf{p},\mathbf{q})\geq 0$ and equality holds if and only if $\mathbf{p}=\mathbf{q}$. Then, we propose the following definition.

Definition 3. We call

$$\xi_X^L=\mathbb{E}_X\left[\zeta(\mathbf{p}_L,\mathbf{p}_{L|X})\right] \tag{9}$$

the probabilistic sensitivity measure of X with respect to L based on $\zeta(\cdot,\cdot)$.

Formally, ξ_X^L in Equation (9) is a particular case of ξ_X in Equation (4). However, Equation (9) explicitly considers the marginal and conditional probability mass functions of the labels. The rationale is that probability mass functions are defined without ambiguity also when labels are non-ordered data. In fact, the corresponding cumulative distribution function would require an additional convention, that is, we need to order the labels and then stick to such lexicographic order. If an alternative order is used, we get an alternative cumulative distribution function. Relying on the probability mass functions avoids the additional step of introducing an order relation. In Table 1, we report three potential choices for the separation measurement between two probability mass functions $\mathbf{p}=\{p_1,\ldots,p_{n_L}\}$ and $\mathbf{q}=\{q_1,\ldots,q_{n_L}\}$ defined on the same sample space. Specifically, $\zeta^1(\mathbf{p},\mathbf{q})$ is a separation measurement based on the 1-norm (absolute value of the differences), $\zeta^2(\mathbf{p},\mathbf{q})$ is a separation based on the 2-norm (square value of the differences) and $\zeta^{\mathrm{KU}}(\mathbf{p},\mathbf{q})$ is based on an extension to the discrete case of the Kuiper distance (Kuiper, 1960) (which is, in turn, an extension of the Kolmogorov-Smirnov distance). In $\zeta^{\mathrm{KU}}(\mathbf{p},\mathbf{q})$ we have $\mathrm{range}(\mathbf{p}-\mathbf{q})=\max_{l=1}^{n_L}(p_l-q_l)-\min_{l=1}^{n_L}(p_l-q_l)$.

Example 4. To illustrate the separation measurements in Table 1, let L be a categorical variable with support $\mathcal{L}=\{A,B,C\}$ and consider the probability mass functions given by $\mathbf{p}=[\frac{1}{3},\frac{1}{3},\frac{1}{3}]$ and $\mathbf{q}=[\frac{1}{4},\frac{1}{2},\frac{1}{4}]$. Then, we have:

$$\begin{aligned}
\zeta^1(\mathbf{p},\mathbf{q})&=\left|\frac{1}{3}-\frac{1}{4}\right|+\left|\frac{1}{3}-\frac{1}{2}\right|+\left|\frac{1}{3}-\frac{1}{4}\right|=\frac{1}{3},\\
\zeta^2(\mathbf{p},\mathbf{q})&=\left(\frac{1}{3}-\frac{1}{4}\right)^2+\left(\frac{1}{3}-\frac{1}{2}\right)^2+\left(\frac{1}{3}-\frac{1}{4}\right)^2=\frac{1}{24},\\
\zeta^{\mathrm{KU}}(\mathbf{p},\mathbf{q})&=\mathrm{range}\left(\frac{1}{3}-\frac{1}{4},\frac{1}{3}-\frac{1}{2},\frac{1}{3}-\frac{1}{4}\right)=\frac{1}{12}-\left(-\frac{1}{6}\right)=\frac{1}{4}.
\end{aligned}$$

Analysts can select separation measurements between probability mass functions that go beyond the ones listed in Table 1, such as the Kullback-Leibler divergence, or the Hellinger distance, or any other particular choice that best suites her/his needs for the application at hand.

The next results state that as long as the separation measurement follows the requirements in Definition 1, members of the ξ_X^L family in Equation (9) possess the zero-independence property. Indeed, notice that for all the measurements $\zeta(\cdot,\cdot)$ complying with Definition 1, $\zeta(\mathbf{p}_L,\mathbf{p}_{L|X})=0$ when $\mathbf{p}_L=\mathbf{p}_{L|X}$, that includes the case when L is independent of X.

Proposition 5. *Given the above setup, if $\zeta(\cdot,\cdot)$ is a separation measurement between probability mass functions then $\xi_X^L\geq 0$, and $\xi_X^L=0$ if and only if L is independent of X.*

Let us now turn to the estimation of the probabilistic sensitivity measures of $X_i\in\mathbf{X}$ with respect to L, for $i\in\{1,\ldots,n_X\}$. Let $\mathcal{X}_i$ denote the support of X_i. Let also $\mathcal{K}_i=\{\mathcal{X}_i^1,\mathcal{X}_i^2,\ldots,\mathcal{X}_i^K\}$ denote a partition of $\mathcal{X}_i$, i.e., a finite or countable collection of K subsets of $\mathcal{X}_i$ such that $\mathcal{X}_i=\bigcup_{k=1}^K\mathcal{X}_i^k$ and $\mathcal{X}_i^m\cap\mathcal{X}_i^l=\emptyset$, for all $m\neq l$. A given-data estimator of $\xi_i^L:=\xi_{X_i}^L$ in (9) is given by:

$$\xi_i(\mathcal{K}_i)=\sum_{k=1}^{K}p\big(X_i\in\mathcal{X}_i^k\big)\zeta\big(\mathbf{p}_L,\mathbf{p}_{L|X_i\in\mathcal{X}_i^k}\big),\tag{10}$$

where $\mathbf{p}_{L|X_i\in\mathcal{X}_i^k}=\big\{p_r^L(\mathcal{X}_i^k)=p(L=\ell_r|X_i\in\mathcal{X}_i^k);\ r=1,2,\ldots,n_L\big\}$ denotes the conditional distribution of L, for $\mathcal{X}_i^k\in\mathcal{K}_i$. Now, consider a fixed partition $\mathcal{K}_i=\{\mathcal{X}_i^1,\mathcal{X}_i^2,\ldots,\mathcal{X}_i^K\}$ with cardinality K and a dataset of features and target realizations. Let N be the sample size, and, for $r=1,\ldots,n_L$, let N_r the number of the observations labeled with ℓ_r. For $\mathcal{X}_i^k\in\mathcal{K}_i$, let $N(\mathcal{X}_i^k)$ denote the number of input observations in $\mathcal{X}_i^k$, and $N_r^L(\mathcal{X}_i^k)$ the corresponding number of target observations with label ℓ_r. Then, using the plug-in principle, we obtain an estimate of $\xi_i(\mathcal{K}_i)$ in (10) from

$$\widehat{\xi}_i(K,N)=\sum_{k=1}^{K}\widehat{p}\big(\mathcal{X}_i^k\big)\zeta\big(\widehat{\mathbf{p}}_L,\widehat{\mathbf{p}}_{L|X_i\in\mathcal{X}_i^k}\big),\tag{11}$$

where $\widehat{p}(\mathcal{X}_i^k)$, $\widehat{\mathbf{p}}_L$ and $\widehat{\mathbf{p}}_{L|X_i\in\mathcal{X}_i^k}$ are plug-in estimates of, respectively, $p(X_i\in\mathcal{X}_i^k),\mathbf{p}_L$ and $\mathbf{p}_{L|X_i\in\mathcal{X}_i^k}$. Observing that, for $i=1,2,\ldots,n_X$ and $r=1,2,\ldots,n_L$, $\widehat{p}(X_i\in\mathcal{X}_i^k)=N(\mathcal{X}_i^k)/N$, $\widehat{p}(L=\ell_r)=N_r/N$ and $\widehat{\mathbf{p}}_L=\{\widehat{p}(L=\ell_1),\ldots,\widehat{p}(L=\ell_{n_L})\}$, as well as $\widehat{p}_r^L(\mathcal{X}_i^k)=\widehat{p}(\mathcal{X}_i^k)\,N_r^L(\mathcal{X}_i^k)/N(\mathcal{X}_i^k)=N_r^L(\mathcal{X}_i^k)/N$ and $\widehat{\mathbf{p}}_{L|X_i\in\mathcal{X}_i^k}=\{\widehat{p}_1^L(\mathcal{X}_i^k),\ldots,\widehat{p}_{n_L}^L(\mathcal{X}_i^k)\}$, we can rewrite (11) as

$$\widehat{\xi}_i(K,N)=\sum_{k=1}^{K}\frac{N(\mathcal{X}_i^k)}{N}\zeta\big(\widehat{\mathbf{p}}_L,\widehat{\mathbf{p}}_{L|X_i\in\mathcal{X}_i^k}\big).\tag{12}$$

Proposition 6. *Let $\zeta(\cdot,\cdot)$ be a separation measure between probability mass functions, and let $\zeta(\cdot,\cdot)$ be bounded almost everywhere as X_i varies in $\mathcal{X}_i$. Let $\widehat{\xi}_i(K,N)$ be defined by (12). Then, if X_i is discrete,*

$$\lim_{N\to\infty}\widehat{\xi}_i(K,N)=\xi_i^L.$$

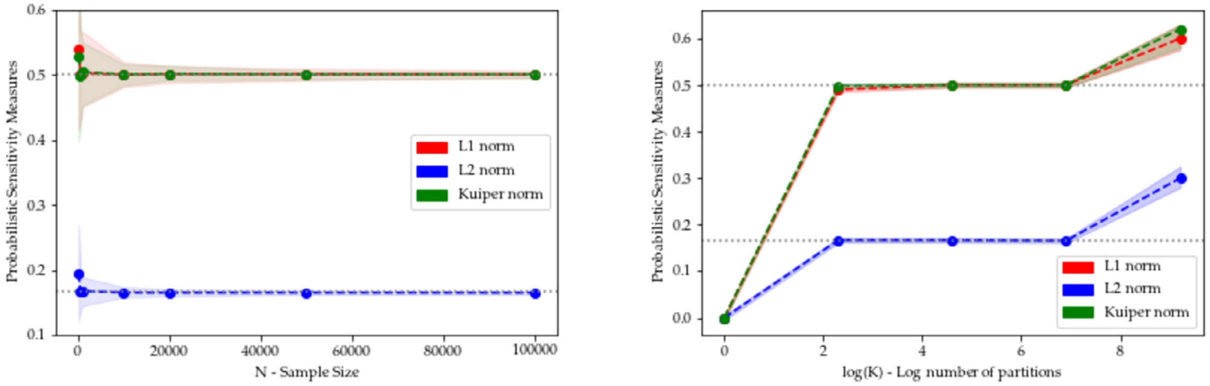
If X_i is continuous,

$$\lim_{K\to\infty}\lim_{N\to\infty}\widehat{\xi}_i(K,N)=\xi_i^L.\tag{13}$$

Proposition 6 reassures us of the consistency of the estimator of ξ_i^L. From an implementation viewpoint, we need to distinguish the case in which X_i is discrete or categorical from the case in which it is continuous. If X_i is discrete then the partition $\mathcal{K}_i$ is fixed and immediately given by the support of X_i. If X_i is continuous then the two limits with respect to the sample size and the partition cardinality are nested. Theoretically, first one lets the sample size N tend to infinity and then one refines the partitions sending K to infinity — as evidenced already in Pearson (1905). In practice, care must be taken in selecting the partition size at any finite sample N. Proposition 6 also implies that for bounded metrics $\zeta(\cdot,\cdot)$, the variance of $\zeta(\widehat{\mathbf{p}}_L,\widehat{\mathbf{p}}_{L|X_i\in\mathcal{X}_i^k})$ is finite for every $k=1,2,\ldots,K$. As an immediate consequence, the following holds for the separation measurements in Table 1 the following holds.

Corollary 7. *The estimators of $\zeta^1(\mathbf{p}_L,\mathbf{p}_{L|X_i\in\mathcal{X}_i^k})$, $\zeta^2(\mathbf{p}_L,\mathbf{p}_{L|X_i\in\mathcal{X}_i^k})$, and $\zeta^{KU}(\mathbf{p}_L,\mathbf{p}_{L|X_i\in\mathcal{X}_i^k})$ are asymptotically consistent.*

Consistency of the estimators has an immediate advantage also with regard to uncertainty quantification. In particular, a natural way of quantifying uncertainty in a context as the one we are dealing with is to make use of the bootstrap method.



(a) Bootstrap results as the sample size increases from $N = 10^2$ to $N = 10^5$.

(b) Bootstrap results as the partition size increases from $K = 1$ to $K = N$.

Fig. 1. Bootstrap results varying N and K for $\widehat{\xi}_1^{\mathrm{KU}}(K,N)$, $\widehat{\xi}_1^1(K,N)$ and $\widehat{\xi}_1^2(K,N)$. The shaded areas represent 95%-bootstrap confidence intervals. In both Figs. 1a and 1b, the y-axis represents the boostrap mean values of $\widehat{\xi}_1^L(K,N)$ based on the 1-norm (red), 2-norm (blue) and Kuiper norm (blue). In Fig. 1a the x-axis reports the values of the sample size N, in Fig. 1b the x-axis reports the values of the partition cardinality K on a logarithmic scale. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

We consider the bootstrap bias-reducing estimator in the version of Efron and Gong (1983), $\widehat{\widehat{\xi}}_i(K,N)$. More precisely, consider a sample of size N and a partition $\mathcal{K}_i$ of K elements. Then, we have

$$\widehat{\widehat{\xi}}_i(K,N) = 2\bar{\xi}_i(K,N) - \widehat{\xi}_i(K,N), \tag{14}$$

where $\bar{\xi}_i(K,N)$ is the estimate of $\xi_i(K,N)$ produced by taking the mean over the bootstrap replicates with fixed partition $\mathcal{K}_i$ and $\widehat{\xi}_i(K,N)$ is the corresponding point estimate. By the theory of the bootstrap method, the asymptotic consistency of $\widehat{\xi}_i(K,N)$ implies the asymptotic consistency of $\widehat{\widehat{\xi}}_i(K,N)$. In our case, the consistency of $\widehat{\xi}_i(K,N)$ is ensured by Proposition 6.

The selection of the partition cardinality K is a crucial step in implementing given-data estimators. The partition cardinality is, indeed, the sole hyperparameter of the design. It is well-known that this choice is associated with a bias-variance trade-off. On the one hand, the higher the value of K, the fewer the available realizations in each partition. We then have higher bias and lower variance for $\widehat{\mathbf{p}}_L$ and $\widehat{\mathbf{p}}_{L|X_i\in\mathcal{X}_i^k}$ for $k = 1, \dots, K$. On the other hand, the smaller K is, the lower the bias but the higher the variance of the estimators. The choice of K in relationship to the sample size has been studied in depth in Strong and Oakley (2013) — see also Borgonovo et al. (2016). The analysis in Strong and Oakley (2013) evidences a plateau effect: for large enough samples, the choice of K in a certain range does not impact the value of $\widehat{\xi}_i$. To illustrate, consider Figure 1 in Strong and Oakley (2013, p. 759), that shows results of experiments at alternative values of K: fixing a sample of size $N = 10'000$, choosing K in a range between $K = 10$ and $K = 1'000$ yields similar values of $\widehat{\xi}_i(K,N)$. For larger (smaller) values of K (at the same sample of size $N = 10'000$), Strong and Oakley (2013) showcase an upward (downward) bias. The plateau effect can then be used to obtain guidance on the choice of K: Strong and Oakley (2013) recommend a value of K in the range that causes the plateau as a natural choice. To illustrate this aspect in our case, we report results of numerical experiments involving an analytical test case.

Example 8. Consider the following fictitious binary classification problem, with a target random variable L with a binary support, i.e. $\mathcal{L} = \{0, 1\}$. We simulate L using the following data generating process:

$$L = \begin{cases} 0 & \text{if } 0 \le Y < 1, \\ 1 & \text{if } 1 \le Y \le 2, \end{cases}$$

where $Y = X_1 + X_2$, and X_1, X_2 are two independent and uniformly distributed random variables in [0,1], i.e. $X_1, X_2 \sim$ Unif$(0,1)$, here playing the role of covariates. Computing explicitly the true probability mass function for L on its support $\mathcal{L} = \{0, 1\}$, we obtain $p(L = 1) = 1 - p(L = 0) = \frac{1}{2}$. Applying the definition in Equation (9), the analytical values of the probabilistic sensitivity measures based on the 1-norm, 2-norm and the Kuiper distance are respectively obtained computing the integrals $\xi_i^1 = \int_0^1 |1 - 2x_i| dx_i = \frac{1}{2}$, $\xi_i^2 = \int_0^1 (2x_i^2 - 2x_i + \frac{1}{2}) dx_i = \frac{1}{6}$, and $\xi_i^{\mathrm{KU}} = \int_0^1 |1 - 2x_i| dx_i = \frac{1}{2}$, for $i = 1, 2$.

In Fig. 1a, we report results at increasing sample sizes. Using a Sobol' Quasi-Random sequence generator we produce a sequence of datasets with realizations of (X_1, X_2) and L, for the sample size ranging from $N = 100$ to $N = 100'000$. The shadows represent bootstrap confidence intervals. The estimates converge towards the analytical values as N increases, and the width of the bootstrap confidence intervals shrinks, thus confirming the asymptotic consistency of the estimates. Fig. 1b presents results for a fixed sample size ($N = 10'000$) but selecting alternative partitions of increasing cardinality. In particular, we vary K from $K = 1$ to $K = 10'000$. Fig. 1b shows that we register a plateau in the values of the estimates

$\widehat{\xi}_i(K, N)$ for $K \in [20, 900]$. For values of K smaller than 20, we have a downward bias. Conversely, for values of K exceeding 900, we register an upward bias.

The results in Fig. 1b are in agreement with the findings in Strong and Oakley (2013). Moreover, as discussed also in Borgonovo et al. (2016), the downward bias is empirically explained by the fact that as K gets smaller the conditional and unconditional distributions tend to coincide. In fact, in the limiting case $K = 1$ the conditional and unconditional distributions are the same and the value of any measures of statistical association is null. Conversely, choosing $K = N$, we obtain exactly one point per partition. Then, the numerical calculation is between the marginal distribution of Y and a Dirac-δ mass centered at the sole realization in the partition. Such distance is maximal or infinite for several metrics and therefore we register an upward bias.

Finally, any given-data estimator in Equation (11) is exposed to the curse of dimensionality when $X_i \in \mathbf{X}$ is multidimensional, that is, we are determining the joint importance of two or more features. In general, then, $X_i = (X_1, X_2 \ldots, X_s)$, $s \leq n_X$ requires us to condition on s features. We are then dealing with s-dimensional partition sets (to fix ideas, in one dimension we are dealing with intervals, in two dimension with rectangles, in s dimension with hyper-rectangles). These partition sets contain a number of realizations (data) that decreases exponentially with s. To illustrate, start with $s = 1$ and consider that we have available N realizations and a partition with cardinality K. Then, we can count on $J = N/K$ realizations per partition set (assuming equipopulated partition sets). For instance if $N = 10,000$ and $K = 25$, we have 400 data points per partition set. If we consider bi-dimensional partitions to find the joint importance of, say, features X_i and X_j $(i \neq j)$, then we could count on $J = N/(K_i K_j)$, where K_i and K_j are, respectively, the cardinalities of the partitions of the supports of X_i and X_j. At $N = 10,000$ selecting $K_i = K_j = 25$, we would have 16 data points per partition set. Compared to the previously available 400 points per partition set, this number of realizations is drastically lower and may prohibit an accurate estimation of the requested statistical quantities. However, this sparsity effect becomes noticeable when we increase the number of features (s) in the group, and is not related to the overall number of features n_X.

4. Application: the Xi-method

In this section, we introduce a framework for obtaining pre-hoc and post-hoc explanations for ML models through measures of statistical dependence. The idea is to understand how close the values of explanations computed from the data are to those computed from the forecasts. Performing this analysis at the overall dataset level as well as at the level of each individual class leads to several indications about the features that are statistically important in the problem at hand. Starting with the data collection $T_L = \left\{(\mathbf{x}^{(n)}, L^{(n)});\ n = 1, 2, \ldots, N\right\}$, *dataset explanations* are defined as the collection of probabilistic sensitivity measures estimates:

$$\widehat{\xi}_X^L = \left\{\widehat{\xi}_i^L,\ i = 1, 2, \ldots, n_X\right\}, \tag{15}$$

estimated from T_L by applying (11) (the superscript L denotes that the estimates are obtained using the data, i.e., $\widehat{\xi}_i^L$ is an estimate of $\xi_i^L = \mathbb{E}_X[\zeta(\mathbf{p}_L, \mathbf{p}_{L|X_i})]$). Consider now the dataset $T_\Lambda = \{(\mathbf{x}^{(n)}, \Lambda^{(n)});\ n = 1, 2, \ldots, N\}$, with $\Lambda^{(n)} = \widehat{g}(\mathbf{x}^{(n)})$, obtained from T_L by replacing the true labels with the corresponding ML model predictions. *Prediction explanations* are defined as the collection of probabilistic sensitivity measures

$$\widehat{\xi}_X^\Lambda = \left\{\widehat{\xi}_i^\Lambda,\ i = 1, 2, \ldots, n_X\right\}, \tag{16}$$

with $\widehat{\xi}_i^\Lambda$ defined in Equation (9) and estimated by applying Equation (11) to T_Λ.

Because they look at allthe target classes simultaneously, $\widehat{\xi}_X^L$ and $\widehat{\xi}_X^\Lambda$ yield an understanding of the most important covariates for the overall dataset or prediction task. However, we can make the analysis more granular by analyzing what are the statistically most important features to predict a given label. To do this, we resort to the so-called one-hot encoding technique for the response variable (Hastie et al., 2009). In particular, for n_L different labels $\ell_1, \ldots, \ell_{n_L}$, we codify the data by $L_1 = \mathbb{1}_{\{L=\ell_1\}}, \ldots, L_{n_L} = \mathbb{1}_{\{L=\ell_{n_L}\}}$. Then, we call the ℓ_r*-dataset explanations* the quantities $\widehat{\xi}_X^{L_r} = \left\{\widehat{\xi}_i^{L_r},\ i = 1, 2, \ldots, n_X\right\}$, where $\widehat{\xi}_X^{L_r}$ is the estimator of $\xi_i^{L_r} = \mathbb{E}\left[\zeta(\mathbf{p}_{L_r}, \mathbf{p}_{L_r|X_i})\right], r = 1, \ldots, n_L$. Similarly, to answer the question of *what are the statistically important features for the ML model when predicting label* ℓ_r, we define the ℓ_r*-prediction explanations* as $\widehat{\xi}_X^{\Lambda_r} = \left\{\widehat{\xi}_i^{\Lambda_r},\ i = 1, 2, \ldots, n_X\right\}$ with $\xi_i^{\Lambda_r} = \mathbb{E}\left[\zeta(\mathbf{p}_{\Lambda_r}, \mathbf{p}_{\Lambda_r|X_i})\right]$ estimated by $\widehat{\xi}_X^{\Lambda_r}$, and Λ_r defined in a similar fashion as L_r.

The next step is to compare $\widehat{\xi}_X^\Lambda$ and $\widehat{\xi}_X^L$ (or their counterparts $\widehat{\xi}_i^{L_r}, \widehat{\xi}_i^{\Lambda_r}$). This comparison can be done by a simple graphical visualization of the values of $\widehat{\xi}_X^\Lambda$ and $\widehat{\xi}_X^L$ or of the ranking they induce. For this task, one can use any discrepancy measure between vectors of real numbers. A possible choice is the Minkowski distance:

$$D^p\left(\widehat{\xi}_X^\Lambda, \widehat{\xi}_X^L\right) = \left\|\widehat{\xi}_X^\Lambda - \widehat{\xi}_X^L\right\|_p = \sqrt[p]{\frac{1}{n_L}\sum_{i=1}^{n_L}\left|\widehat{\xi}_X^\Lambda - \widehat{\xi}_X^L\right|^p}. \tag{17}$$

For $p = 1$, $D^p(\widehat{\xi}_X^\Lambda, \widehat{\xi}_X^L)$ is the Mean Absolute Deviation (MAD), while for $p = 2$ the Mean Squared Error (MSE). Information delivered by $D^p\left(\widehat{\xi}_X^\Lambda, \widehat{\xi}_X^L\right)$ can be used in alternative ways. For instance, the analyst may consider two set of explanations $\widehat{\xi}_X^\Lambda$ and $\widehat{\xi}_X^L$ far apart if $D^p\left(\widehat{\xi}_X^\Lambda, \widehat{\xi}_X^L\right) > \delta$, for some threshold $\delta > 0$. If that is the case, and the model is fitting well, then the model is making good predictions but it is not picking up the same statistical relationships present in the data. Conversely, if $D^p\left(\widehat{\xi}_X^\Lambda, \widehat{\xi}_X^L\right) < \delta$, the model is predicting accurately and its forecasts recreate well the statistical dependence of the original dataset. In this respect, it is interesting to conduct the comparison for the training as well as the test datasets, to see if differences emerge between the two subsamples. Regarding the value of the threshold δ, the choice depends on the application of interest and on the separation measure $\zeta(\cdot,\cdot)$; we therefore refrain from providing a general rule. Also, often the interest is on the ordinal ranking induced by $\widehat{\xi}_X^\Lambda$ and $\widehat{\xi}_X^L$. In this case, a quantitative comparison is carried out using well-known statistical techniques such as the Spearman rank correlation coefficient (ρ^{Spear}, henceforth) (Spearman, 1904). This very same procedure can be employed in a very similar fashion at the individual class level, i.e., when the sensitivity measures are calculated for each target class.

5. Test cases: tabular, visual and text data

In this section, we illustrate the method and apply it to well-known datasets in different formats: in Subsection 5.1 we analyze a tabular dataset, in Subsection 5.2 an image dataset and in Subsection 5.3 a text dataset. In all the experiments, we train a ML model on a subset of the available data (the training set), and then we compute the explanations ξ_X^Λ, ξ_X^L on the test set, using the estimators defined in Equation (10) and fixing the number of partitions at $K = 10$. To fit the models, we employ the well-known `scikit-learn` package in Python, using the default loss functions. We report the ML model test accuracy as the percentage of correctly classified instances.

5.1. Tabular data: the Wine dataset

In our first application, we use the well-known Wine dataset publicly available at https://archive.ics.uci.edu/ml. This dataset has been widely used in association with the task of predicting a wine quality based on its chemical properties (Dua and Graff, 2017). The dataset collects quantitative data on eleven features (ranging from alcohol content to the pH of the wine), and the output is the corresponding wine quality measured on a scale from 1 to 10. The sample size is $N = 4898$, split into 60% for training and 40% for testing. The only preprocessing has been to replace the 39 missing entries in the design matrix $\mathbf{X}$ with either the mean or the median of the corresponding variable (we also performed our analysis omitting these entries, with unchanged results). The classes in the final dataset are unevenly populated, with 1457, 2198 and 880 entries, respectively, for Qualities 5, 6 and 7, with 163 and 175 entries for Qualities 4 and 8, with only 20 entries for Quality 3, and zero entries for Quality 1 and 2. We trained alternative ML models and registered a very similar accuracy for all. In the remainder, we shall focus on results obtained using a Random Forest (RF), which registers a test accuracy of 66% (1-off accuracy $\sim 90\%$). Using the test set and the trained model, for all the variables $i = 1, \ldots, n_X$, we estimate the feature importance measures provided by Permutation Importance (PI_i) and Split and Count (SC_i), as well as the probabilistic sensitivity measures $\widehat{\xi}_i^L$, $\widehat{\xi}_i^\Lambda$ based on the 1-norm separation (for the sake of space we shall focus only on this norm in the remainder). Results are displayed in Fig. 2 Panels (a) and (b) in Fig. 2 compare the values of the probabilistic sensitivity measures on the real data versus the ones the model forecasts (both computed on the test set). A visual inspection suggests that the ranking induced by $\widehat{\xi}_X^L$ and $\widehat{\xi}_X^\Lambda$ are similar: alcohol stands out as the most important variable, the group volatile acidity, density and chlorides follows, and is followed in turn by a group comprising citric acid, total sulfur dioxide, free sulfur dioxide and sulphates, while the group pH, fixed acidity and residual sugar contains the three least relevant features.

Table 2 reports results of the quantitative comparisons between $\widehat{\xi}_X^L, \widehat{\xi}_X^\Lambda$ with MAD and MSE in the second and third columns, respectively, and ρ^{Spear} in the last column. The small values of MAD and MSE and the simultaneously high value of ρ^{Spear} confirm the visual impression of Fig. 2 concerning the overall agreement. Thus, the ML model forecasts actually reproduce well the original statistical dependence in the data and the covariates that are statistically important for the true data generating process are also important for the ML model predictions. Panels (c) and (d) in Fig. 2 compare these results with indications provided by the Split and Count measure (SC) and the Permutation Importance measure (PI). The values of ρ^{Spear} between the ranking induced by the alternative importance measure amount at around 0.7, indicating a lower ranking agreement than in the previous case (Table 3). Indeed, while all importance measures agree on alcohol as the most important feature, the values of SC indicate a rather homogeneous influence of the remaining features (the values of SC are similar), while $\widehat{\xi}_X^\Lambda$ and $\widehat{\xi}_X^L$ display greater variability in their values, with alcohol, density and volatile acidity being statistically more important than pH, fixed acidity and residual sugar. One also notes that the ranking of the three most important variables is consistent for $\widehat{\xi}_X^L$, $\widehat{\xi}_X^\Lambda$ and SC, while PI assigns density a much lower importance. All importance measures suggest alcohol as the most important variable, so that this feature is not only the feature on which the target depends most strongly, but also the one that drives the ML model performance the most.

Let us now examine results at the individual class level. We apply one-hot encoding to the model forecasts for the test set, with $N^{Te} = 2573$. Estimates of $\xi_i^{\Lambda_r}$ for $i = 1, \ldots, n_X$ and $r = 1, \ldots, n_L$ based on $\zeta^1(\cdot,\cdot)$ are reported in Fig. 3. Each group of bars displays estimates of $\xi_i^{\Lambda_r}$ for a given class, from Quality 4 to Quality 8, as the model never predicts

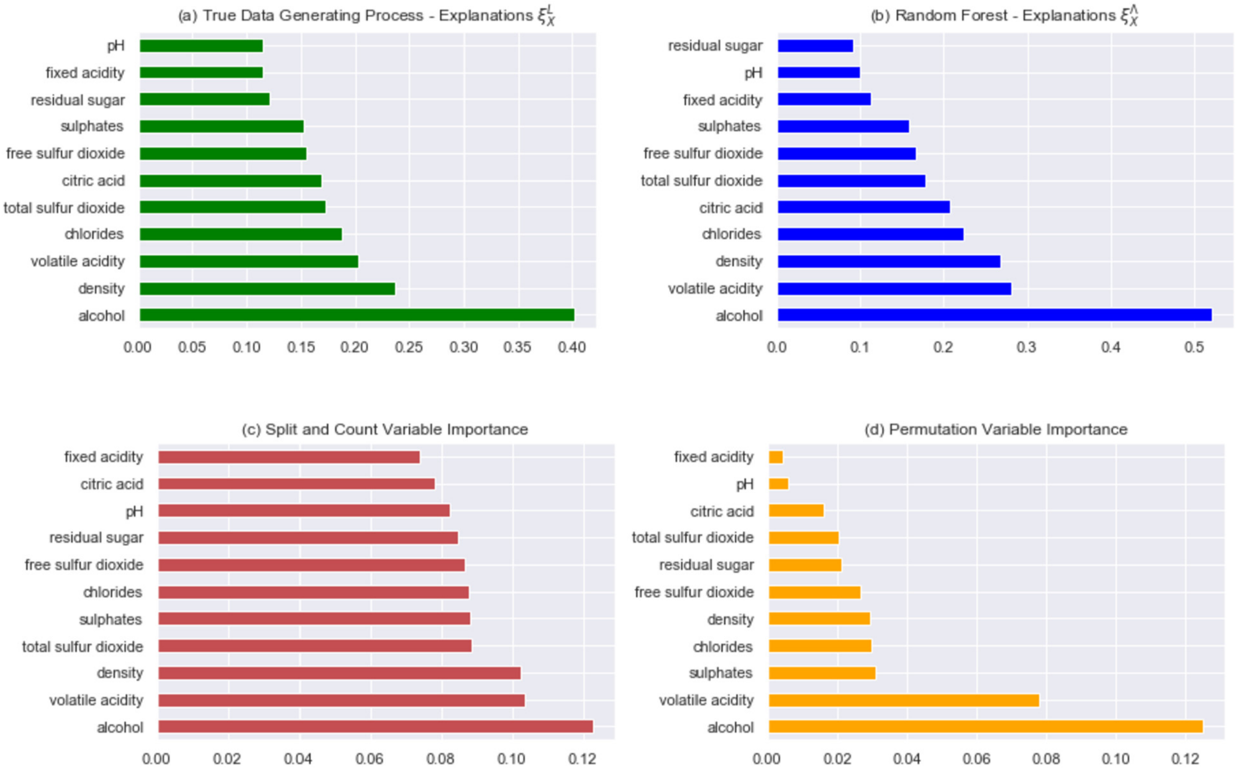


Fig. 2. Panel 2a: dataset explanations $\widehat{\xi}_i^L$ based on the 1-norm separation. Panel 2b: prediction explanations $\widehat{\xi}_i^\Lambda$ based on the 1-norm separation. Panel 2c: Split and Count measure $\widehat{\mathrm{SC}}_i$, Panel 2d: Permutation Importance measure $\widehat{\mathrm{PI}}_i$, for $i = 1, \dots, n_X$— tabular data.

Table 2
Quantitative similarity between dataset explanations $\widehat{\xi}_X^L$ and prediction explanations $\widehat{\xi}_X^\Lambda$ based on the 1-norm separation — tabular data.

	MAD	MSE	ρ^{Spear}
1-norm	0.034	0.048	0.955
2-norm	0.010	0.017	0.986
Kuiper	0.040	0.055	0.936

Table 3
Spearman correlation coefficient comparing the ranking of dataset explanations $\widehat{\xi}_X^L$, prediction explanations $\widehat{\xi}_X^\Lambda$ (both based on the 1-norm separation), Split and Count (SC) measure, and Permutation Importance (PI) — tabular data.

Spearman Correlation	$\rho^{\text{Spear}}(\mathrm{SC}, \widehat{\xi}_X^L)$	$\rho^{\text{Spear}}(\mathrm{SC}, \widehat{\xi}_X^\Lambda)$	$\rho^{\text{Spear}}(\mathrm{PI}, \widehat{\xi}_X^L)$	$\rho^{\text{Spear}}(\mathrm{PI}, \widehat{\xi}_X^\Lambda)$
1-norm	0.809	0.709	0.791	0.709
2-norm	0.745	0.736	0.718	0.736
Kuiper	0.745	0.709	0.727	0.709

Quality 3 or Quality 9 and thus the bars would be all null. Similarly, we register extremely low values of $\xi_i^{\Lambda_r}$ for Quality 4 and Quality 8. This is a consequence of the few predictions of the corresponding labels, with Quality 4 and 8 predicted, respectively, only 12 and 82 times on over 2'500 entries. As a result, the one-hot encoding vectors is, effectively, a vector of zeros independently of the values of the features. This effect is then captured by the values of the probabilistic sensitivity measures that are close to zero. The remaining bars indicate that alcohol is, statistically, the most important feature for the model when making a prediction on classes Quality 5, Quality 6 and Quality 7, followed by sulphates; fixed acidity raises in importance when the target is Quality 5 or Quality 6, while total sulfur dioxide becomes the third most important variable for predicting Quality 7.

In the previous discussion, we have referred to point estimates. However, we also performed corresponding uncertainty quantification to understand whether such observations are robust to variability in the estimates. We report a first set of

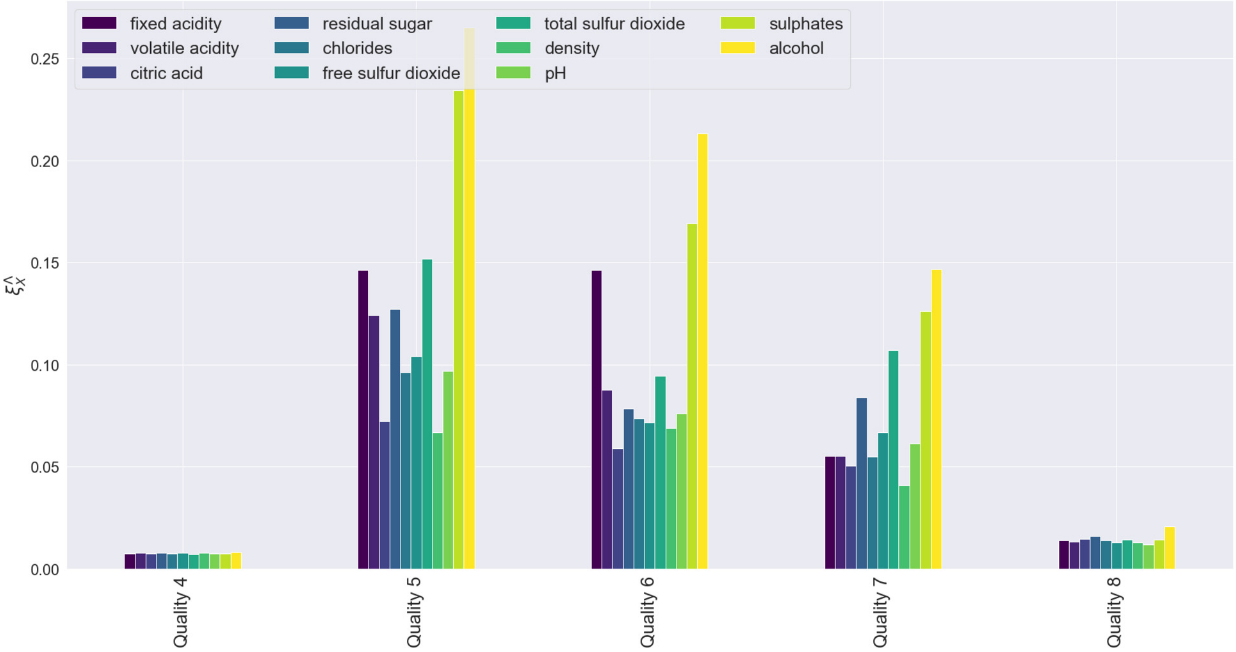


Fig. 3. ℓ_r-prediction explanations $\widehat{\xi}_X^{\Lambda_r}$ based on the 1-norm separation, $r = 4, \ldots 8$ — tabular data.

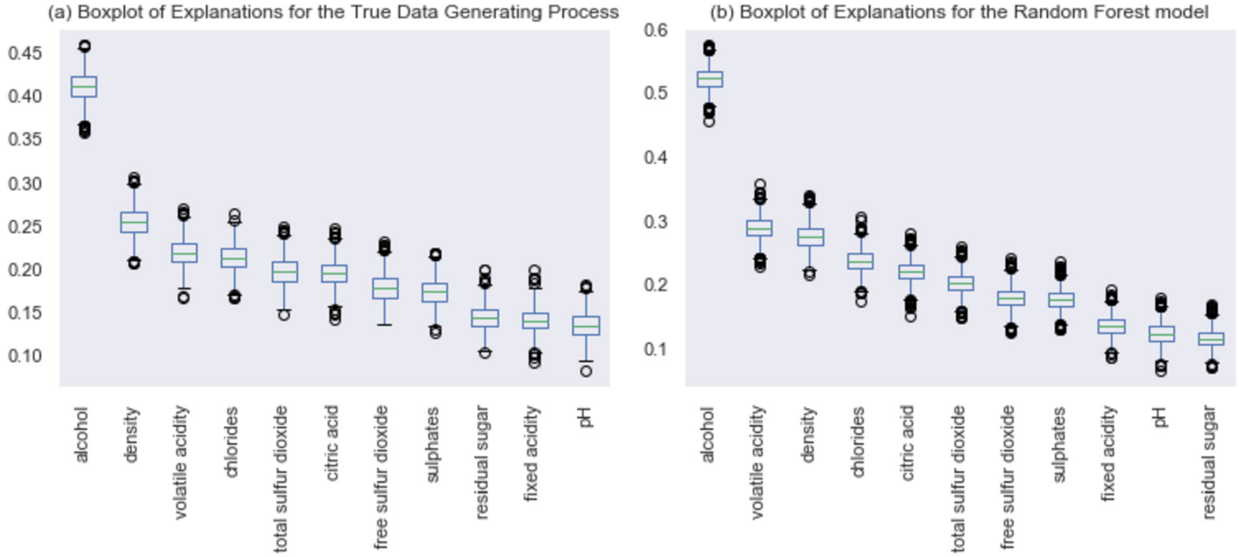


Fig. 4. Panel 2a: bootstrap distributions of dataset explanations $\widehat{\xi}_X^L$. Panel 2b: bootstrap distributions of prediction explanations $\widehat{\xi}_X^\Lambda$ (both based on the 1-norm separation) — tabular data.

results obtained by 2’500 bootstrap replicates of the test data and predictions, and for each sample we recompute $\widehat{\xi}_X^L$ as well as $\widehat{\xi}_X^\Lambda$. Fig. 4 reports the corresponding bootstrap distributions as boxplots. The bootstrap confidence intervals in Fig. 2a and 2b are narrow enough around the point estimates to let us state that the results obtained with point estimates are actually reliable, with a few (and negligible) outliers in the boxplot. Thus, in this case, uncertainty quantification does not alter the previous considerations.

5.2. Image data: Fashion MNIST

In this section, we report results of experiments conducted on the well-known Fashion MNIST dataset (Xiao et al., 2017), a database by Zalando research containing images of clothing articles, with N=70’000, of which 60’000 images are used for training and 10’000 for testing. The images are made of 28 × 28 grayscale pixels; each instance is associated with a

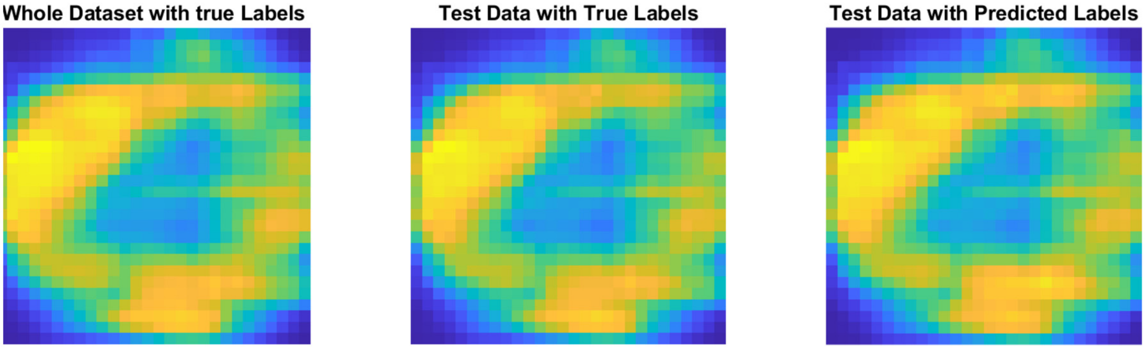


Fig. 5. From left to right: dataset explanations $\widehat{\xi}_X^L$ (computed respectively using the whole dataset and the test set) and prediction explanations $\widehat{\xi}_X^\Lambda$, both based on the 1-norm separation separation — image data.

Table 4
Quantitative similarity between the dataset explanations $\widehat{\xi}_X^L$ and the prediction explanations $\widehat{\xi}_X^\Lambda$ based on the 1-norm separation — image data.

	MAD	MSE	ρ^{Spear}
1-norm	0.024	0.030	0.997
2-norm	0.004	0.005	0.994
Kuiper	0.009	0.011	0.993

label from ten classes: T-shirt/Top, Trouser, Pullover, Dress, Coat, Sandal, Shirt, Sneaker, Bag, Ankle boot. The ML model in this application is a pre-trained convolutional neural network with the LeNet architecture (LeCun et al., 1998). LeNet is a 7-layer neural network consisting of the input layer, 2 convolutional layers each followed by a pooling layer and another convolutional layer followed by the output layer. The overall accuracy is 94%.

In this experiment, we focus on the explanations provided by the Xi-method using the 1-norm separation measure. Fig. 5 reports results of the investigation at the all-classes level. The first heatmap displays values of $\widehat{\xi}_X^L$, with lighter pixels being more important: to estimate the explanations, we use the entire dataset of 70’000 images. The second heatmap refers to dataset explanations $\widehat{\xi}_X^L$ on the test set, and the third one displays prediction explanations $\widehat{\xi}_X^\Lambda$ computed on the test set. A visual inspection shows a great similarity between the regions that are statistically important for the LeNet model and for the true data generating process. Interestingly, the exact center of the image does not contain the most important pixels, which are the ones surrounding the object (images are centred).

Table 4 reports results for the quantitative comparison between $\widehat{\xi}_X^L$ and $\widehat{\xi}_X^\Lambda$: MAD and MSE are close to 0, and ρ^{Spear} is almost 1, indicating a high agreement between $\widehat{\xi}_X^L$ and $\widehat{\xi}_X^\Lambda$. Thus, the statistical dependence in the true data generating process is maintained in the forecasts.

Consider now the analysis at the individual class level. Fig. 6 reports the heatmaps generated by $\widehat{\xi}_X^{\Lambda_r}$ for each of the image classes $r = 1, \dots, n_L$ using the test data (the heatmaps of $\widehat{\xi}_X^{L_r}$ are similar and we do not display them for sake of space). The explanations at the individual-class level for the LeNet model show interesting insights: for example, the most important pixels for predicting the class T-shirt/Top highlight the lack of sleeve in such items; the most important ones for predicting trousers are the pixels that signal the empty space between the legs; for classes Dress or Shirt the focus is on the overall clothing piece (light blue pixels), with specific parts that are more important than others (yellow pixels). Also for this case study we performed an uncertainty quantification via the bootstrap. Results show stability in the estimates (details are not reported for the sake of space).

5.3. *Text data: Asian Religious Texts*

The last application concerns the textual database available at https://archive.ics.uci.edu/ml/datasets/A+study+of++Asian+Religious+and+Biblical+Texts. In this case, the task is to predict the book of origin of an excerpt among eight sacred texts (Sah and Fokoué, 2019): Book Of Ecclesiasticus, Book Of Ecclesiastes, Book Of Proverb, Book Of Wisdom (Christians sacred books), Buddhism, Tao Te Ching (sacred text for Taoism), Upanishad and Yoga Sutra (sacred books for Hinduism), related to different religions in Asia (Hinduism, Buddhism, Taoism, Christianity). Features are the individual words in the Document Term Matrix (DTM) used to train the model, describing the frequency of terms that occur in the collection of documents. The DTM is provided by the data source. The classifier is a simple Naive Bayes, trained on 70% of the observations. The ML model test accuracy is 65%. As in the previous experiments, we estimate the probabilistic sensitivity measures based on the 1-norm using the test dataset. Fig. 7 reports the ten most important words according to dataset explanations $\widehat{\xi}_X^L$

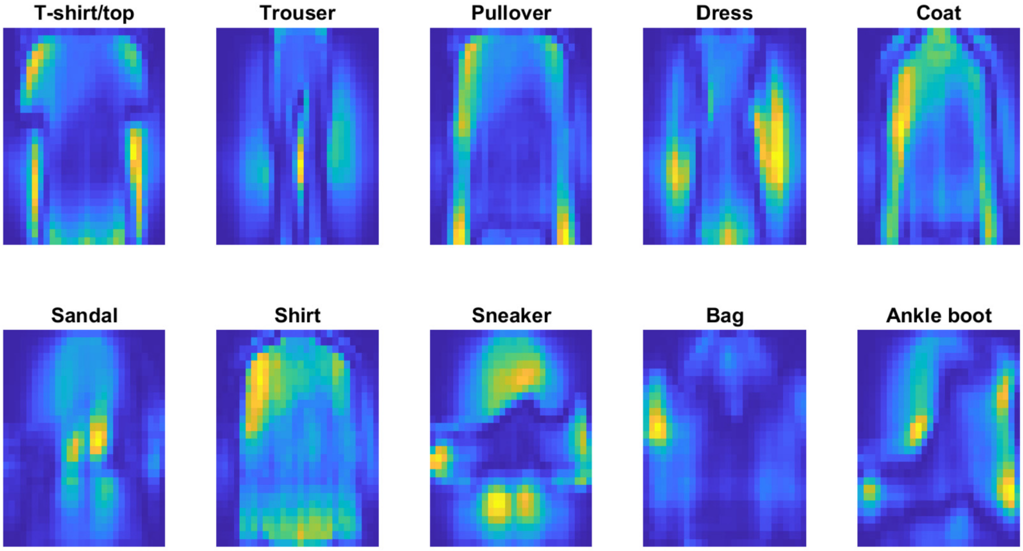
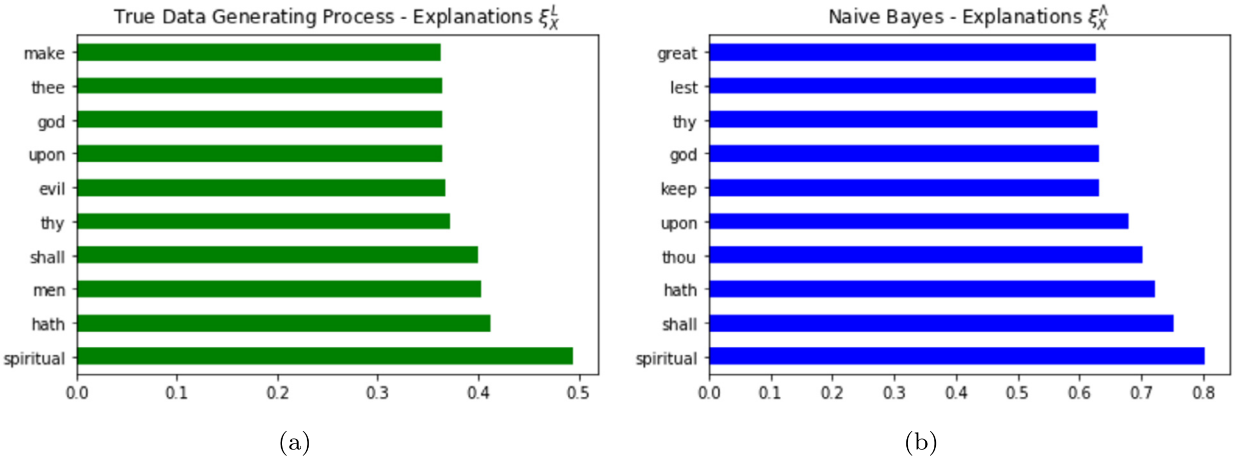


Fig. 6. ℓ_r-prediction explanations $\widehat{\xi}_X^{\Lambda_r}$ based on the 1-norm separation, $r = 1, \ldots, n_L$ — image data.



(a)

(b)

Fig. 7. Panel 7a: dataset explanations $\widehat{\xi}_X^L$ based on the 1-norm separation, Panel 7b: prediction explanations $\widehat{\xi}_X^\Lambda$ based on the 1-norm separation (ten most important features displayed) — text data.

Table 5
Quantitative similarity between the dataset explanations $\widehat{\xi}_X^L$ and the prediction explanations $\widehat{\xi}_X^\Lambda$ based on the 1-norm separation — text data.

	MAD	MSE	ρ^{Spear}
1-norm	0.030	0.049	0.923
2-norm	0.006	0.010	0.908
Kuiper	0.009	0.017	0.886

and prediction explanations $\widehat{\xi}_X^\Lambda$. It shows that while there is some difference in the actual values of the explanations, the rankings provided by both $\widehat{\xi}_X^L$ and $\widehat{\xi}_X^\Lambda$ are similar.

The quantitative comparison in Table 5 shows a value of ρ^{Spear} at about 90%; thus, we register a high overall ranking agreement, although the induced ranking is not perfectly coincident; at the same time values of the MAD and MSE between $\widehat{\xi}_X^L, \widehat{\xi}_X^\Lambda$ are close to 0. This last result shows that the values of the probabilistic sensitivity measures calculated on the data and on the forecasts are close on average; thus, small differences may result in different ranking, in spite of the features having a similar influence.

At the individual class level, results are presented in Table 6. The columns in Table 6 list the ten most important words

Table 6
Ten most important words provided by the ℓ_r-prediction explanations $\xi_i^{\Lambda_r}$ based on the 1-norm separation, $r = 1, \dots, n_L$ − text data.

Book Of Ecclestasticus	Book Of Ecclesiastes	Book Of Proverbs	Book Of Wisdom	Buddhism	Tao Te Ching	Upanishad	Yoga Sutra
shall	mortal	shall	therefore	hath	tao	brahman	spiritual
s	maketh	hath	consciousness	thy	young	called	great
lord	souls	like	form	right	never	whole	wise
heart	flesh	life	psychic	qualities	together	last	aged
god	treasure	without	though	therefore	person	devas	forgotten
great	silver	loveth	life	bring	appeared	knows	declareth
thee	spirit	soul	perception	made	looked	enters	approved
thou	born	glory	vision	come	perhaps	definite	wherewith
truth	know	poor	another	make	knows	teaching	number
knowledge	tongue	keepeth	know	nature	root	things	rites

for each target class according to $\widehat{\xi}_X^{\Lambda_r}$, $r = 1, \dots, n_L$. We note that the words that matter for the classification into a Hinduism sacred text (Upanishad, YogaSutra) mostly pertain to spiritualism and knowledge (spiritual, wise, knows, teaching) as well as with some specific terms of Hindu philosophy (Brahman, devas). On the other hand, terms more related to nature appear in Buddhism. The most impactful word to predict the Tao Te Ching sacred text is, unsurprisingly, Tao. Finally, the predictions regarding the Christian sacred texts are more influenced by terms dealing with the contraposition of mortal (flesh, mortal, born) and eternal (souls, spirit, form) life as well as hints to god and lord. We performed an uncertainty quantification via bootstrap. Estimates remain stable; we do not report the results for the sake of space. For this test case, we also computed Breiman's Permutation Variable Importance (PI), however obtaining a null value for all the features. We believe that this result can be explained by two main reasons: first, the high dimensionality of the DTM, which contains roughly 8′300 columns. Permuting one feature out of so many does not result in a significant loss. Second, the DTM is highly sparse, with feature entries consisting mostly of 0's. Then, the permutation leads to a basically identical vector of realizations, and does not yield any particular change in the loss functions.

6. Final remarks

This work has introduced probabilistic sensitivity measures that are well-posed on non-ordered data and has applied them to ML classification tasks. The proposed measures of association are based on the separation between probability mass functions and can be directly estimated from a given dataset, without the need of fitting a ML model. They possess the zero-independence property. Also, we have proven that, provided that the indices are based on a bounded and continuous separation measurement between probability mass functions, the corresponding estimators are asymptotically consistent.

We have then proposed a framework to explain relationships of statistical dependence in a classification context. A key part of the method is the comparison of measures of statistical association calculated first on the original data and then on the ML model forecasts. The first set of estimates uncovers the target-features dependence of the data generating process, the second the target-feature dependence that emerges when the ML model forecasts replace the original targets. The framework, here called Xi-method, offers several advantages: the importance measures are not computationally expensive, the explanations can be obtained for any kind of data (images, texts, tabular) and they come with theoretical guarantees. Also, when images are concerned, the sensitivity measures avoid data manipulations such as obscuring or removing pixels. This is a major advantage because it is well-known that results may change according to the color used to mask portions of the image. Moreover, even if this work has focused on classification, the Xi-method is applicable in a regression framework as well. In this case, measures of statistical association such as the one in Equation (8) (i.e., the Chatterjee correlation coefficient) and several others are available to be selected as statistical indicators (the measures introduced in this work are aimed for classification problems). A further advantage of the approach is that it allows a straightforward and computationally cheap uncertainty quantification, in line with the recommendation of, among others, Dunson (2018). As for any method based on measures of statistical association, a first limitation is that if two features X_i and X_j are highly correlated, measures of pairwise dependence between target and features will assign similar importance to X_i and X_j, even if Y is a function of X_i only. In principle, this issue can be overcome by conditioning on one or more of the correlated features and estimating conditional measures of pairwise dependence. The estimation procedure may also suffer from the curse of dimensionality or of lack of data in the case of small sample sizes. Investigations aimed at studying these probabilistic sensitivity measures in a conditional setting as well as the extension of the method to multivariate forecasting problems are future research avenues.

7. Proofs

Proof of Proposition 6. 1) Discrete input case. Consider X_i a discrete variable with support $\mathcal{X}_i = \left\{x_i^1, x_i^2, \dots, x_i^K\right\}$. In this case, the natural partition choice is $\mathcal{K}_i = \left\{\mathcal{X}_i^1, \mathcal{X}_i^2, \dots, \mathcal{X}_i^K\right\}$, with $\mathcal{X}_i^k = \{X_i : X_i = x_i^k\}$, for $k = 1, 2, \dots, K$. Note that $\xi_i^L = \xi_i(\mathcal{X}_i) = \xi_i(\mathcal{K}_i)$ becomes

$$\xi_i^L = \xi_i(\mathcal{K}_i) = \sum_{k=1}^{K} p(X_i = x_i^k)\zeta(\mathbf{p}_L, \mathbf{p}_{L|X_i=x_i^k}). \tag{18}$$

Then, a given-data estimator for $\xi_i(\mathcal{K}_i)$ in Equation (18) can be written as

$$\widehat{\xi}_i(K,N) = \sum_{k=1}^{K} \widehat{p}\big(X_i = x_i^k\big)\zeta\big(\widehat{\mathbf{p}}_L, \widehat{\mathbf{p}}_{L|X_i=x_i^k}\big). \tag{19}$$

Then, consider a dataset $(\mathbf{X}, L)$. Let N_i^k denote the number of realizations of X_i such that $X_i \in \mathcal{X}_i^k$ – i.e. $X_i = x_i^k$, for $k = 1, 2, \dots, K$. Then, $\widehat{p}(X_i = x_i^k) = N^{-1}N_k$ is a consistent estimator of $p(X_i = x_i^k)$, by the law of large numbers. Similarly, letting N_r^L be the number of labels in category ℓ_r, $r = 1, 2, \dots, n_L$, $\widehat{p}(L = \ell_r) = N^{-1}N_r^L$ is a consistent estimator of $p(L = \ell_r)$. Analogously, let $\widehat{p}(L = \ell_r|X_i = x_i^k) = N^{-1}N_r^L(\mathcal{X}_i^k)$, where $N_r^L(\mathcal{X}_i^k)$ counts the realizations of L equal to ℓ_r, when $X_i \in \mathcal{X}_i^k$. Then, this estimator is also consistent by the law of large numbers. Therefore $\widehat{\mathbf{p}}_L \to \mathbf{p}_L$ and $\widehat{\mathbf{p}}_{L|X_i=x_i^k} \to \mathbf{p}_{L|X_i=x_i^k}$ as $N \to \infty$. Now $\zeta(\widehat{\mathbf{p}}_L, \widehat{\mathbf{p}}_{L|X_i=x_i^k})$ is continuous, and, therefore, we have

$$\zeta\big(\widehat{\mathbf{p}}_L, \widehat{\mathbf{p}}_{L|X_i=x_i^k}\big) \underset{N\to\infty}{\to} \zeta(\mathbf{p}_L, \mathbf{p}_{L|X_i=x_i^k}),$$

so that

$$\sum_{k=1}^{K} \widehat{p}\big(X_i = x_i^k\big)\zeta\big(\widehat{\mathbf{p}}_L, \widehat{\mathbf{p}}_{L|X_i=x_i^k}\big) \underset{N\to\infty}{\to} \sum_{k=1}^{K} p\big(X_i = x_i^k\big)\zeta(\mathbf{p}_L, \mathbf{p}_{L|X_i=x_i^k}), \tag{20}$$

which implies $\widehat{\xi}_i(K,N) \to \xi_i^L$ as $N \to \infty$.

2) Absolutely continuous input case. If X_i is absolutely continuous, let $f_{X_i}(x_i)$ be the density of X_i. Consider now $\xi_i^L = \xi_i(\mathcal{X}_i)$ written as

$$\xi_i(\mathcal{X}_i) = \int_{\mathcal{X}_i} \zeta\big(\mathbf{p}_L, \mathbf{p}_{L|X_i=x_i}\big) f_{X_i}(x_i)dx_i.$$

First, we note that the integral above is the limit, if it exists, of the following Riemann-Stieltjes sum:

$$\xi_i(\mathcal{K}_i^\delta) = \lim_{\delta\to 0} \sum_{k=1}^{K(\delta)} \zeta\big(\mathbf{p}_L, \mathbf{p}_{L|X_i\in\mathcal{X}_i^k(\delta)}\big)p\big(X_i \in \mathcal{X}_i^k(\delta)\big),$$

where the set $\mathcal{X}_i^k$ is a member of a partition $\mathcal{K}_i^\delta$ of $\mathcal{X}_i$, and where K and δ denote the cardinality and norm of the partition, respectively. Consider now a dataset $(\mathbf{X}, L)$. Fixing a partition $\mathcal{K}_i(K) = \{\mathcal{X}_i^1, \dots, \mathcal{X}_i^K\}$, the given data estimator $\widehat{\xi}_i(K,N)$ is written as

$$\widehat{\xi}_i(K,N) = \sum_{k=1}^{K} \zeta\big(\widehat{\mathbf{p}}_L, \widehat{\mathbf{p}}_{L|X_i\in\mathcal{X}_i^k}\big)\widehat{p}\big(X_i \in \mathcal{X}_i^k\big).$$

Now, let $N \to \infty$. If $\zeta(\cdot,\cdot)$ is continuous, then

$$\begin{aligned}\lim_{N\to\infty} \widehat{\xi}_i(K,N) &= \lim_{N\to\infty} \sum_{k=1}^{K} \zeta\big(\widehat{\mathbf{p}}_L, \widehat{\mathbf{p}}_{L|X_i\in\mathcal{X}_i^k}\big)\widehat{p}\big(X_i \in \mathcal{X}_i^k\big) \\ &= \sum_{k=1}^{K} \zeta\big(\mathbf{p}_L, \mathbf{p}_{L|X_i\in\mathcal{X}_i^k}\big)p\big(X_i \in \mathcal{X}_i^k\big) = \xi_i(K)\end{aligned}$$

by the same argument holding for the consistency of the discrete case. Then, we consider a sequence of refining partitions of $\mathcal{X}_i$ such that $\mathcal{K}_i(K+1)$ is finer than $\mathcal{K}_i(K)$ and such that $\lim_{K\to\infty} \mathcal{K}_i(K) = \mathcal{X}_i$. Then, by Rohlin's disintegration theorem, we have that $\widehat{\mathbf{p}}_{L|X_i\in\mathcal{X}_i^k} \to \widehat{\mathbf{p}}_{L|X_i=x_i^k}$ for almost every $x_i^k \in \mathcal{X}_i$. Then, by the continuity and boundedness of $\zeta(\cdot,\cdot)$ and the definition of Riemann-Stieltjes integral, we have that

$$\lim_{K\to\infty} \xi_i(K) = \xi_i^L. \quad \square$$

Proof of Corollary 7. To prove the thesis, we first need to show that the three metrics are bounded. We start with the Kuiper metric. We have that

$$\zeta^{\mathrm{KU}}\big(\mathbf{p}_L, \mathbf{p}_{L|X_i \in \mathcal{X}_i^k}\big) = \sup_{r=1,2,\ldots,n_L} \Big\{ p(L=\ell_r) - p(L=\ell_r|X_i \in \mathcal{X}_i^k) \Big\} + \sup_{r=1,2,\ldots,n_L} \Big\{ p(L=\ell_r|X_i \in \mathcal{X}_i^k) - p(L=\ell_r) \Big\}.$$

Noting that the maximum value that $|p(L=\ell_r) - p(L=\ell_r|X_i \in \mathcal{X}_i^k)|$ can assume is 1, we have $\zeta^{\mathrm{KU}}(\mathbf{p}_L, \mathbf{p}_{L|X_i \in \mathcal{X}_i^k}) \leq 2$.

Similarly, for the 1-norm, we have

$$\zeta^{1}\big(\mathbf{p}_L, \mathbf{p}_{L|X_i \in \mathcal{X}_i^k}\big) = \sum_{r=1}^{n_L} |p(L=\ell_r) - p(L=\ell_r|X_i \in \mathcal{X}_i^k)| \leq \sum_{r=1}^{n_L} p(L=\ell_r) + \sum_{r=1}^{n_L} p(L=\ell_r|X_i \in \mathcal{X}_i^k) = 2.$$

For the 2-norm, we have

$$\zeta^{2}\big(\mathbf{p}_L, \mathbf{p}_{L|X_i \in \mathcal{X}_i^k}\big) = \sum_{r=1}^{n_L} \big[p(L=\ell_r) - \mathbf{p}(L=\ell_r|X_i \in \mathcal{X}_i^k)\big]^2 = \sum_{r=1}^{n_L} p(L=\ell_r)^2 - 2\sum_{r=1}^{n_L} p(L=\ell_r)p(L=\ell_r|X_i \in \mathcal{X}_i^k) + \sum_{r=1}^{n_L} p(L=\ell_r|X_i \in \mathcal{X}_i^k)^2.$$

Hence, because $p(L=\ell_r)^2 \leq p(L=\ell_r)$ and $p(L=\ell_r|X_i \in \mathcal{X}_i^k)^2 \leq p(L=\ell_r|X_i \in \mathcal{X}_i^k)$, then $\sum_{r=1}^{n_L} p(L=\ell_r)^2 \leq \sum_{r=1}^{n_L} p(L=\ell_r) \leq 1$, and similarly $\sum_{r=1}^{n_L} p(L=\ell_r|X_i \in \mathcal{X}_i^k)^2 \leq 1$. This then leads to

$$\zeta^{2}\big(\mathbf{p}_L, \mathbf{p}_{L|X_i \in \mathcal{X}_i^k}\big) \leq 2 - 2\sum_{r=1}^{n_L} p(L=\ell_r)p(L=\ell_r|X_i \in \mathcal{X}_i^k).$$

Because $\sum_{s=1}^{n_L} p(L=\ell_r)p(L=\ell_s|X_i = x_i)$ is positive, we have $\zeta^{2}\big(\mathbf{p}_L, \mathbf{p}_{L|X_i \in \mathcal{X}_i^k}\big) \leq 2$. Consistency of the estimators then follows from Proposition 6. □

References

Barber, R.F., Candés, E.J., 2015. Controlling the false discovery rate via knockoffs. Ann. Stat. 43, 2055–2085.

Binder, A., Bach, S., Montavon, G., Müller, K.R., Samek, W., 2016. Layer-wise relevance propagation for deep neural network architectures. Lect. Notes Electr. Eng. 376, 913–922.

Borgonovo, E., Hazen, G., Plischke, E., 2016. A common rationale for global sensitivity measures and their estimation. Risk Anal. 36, 1871–1895.

Borgonovo, E., Tarantola, S., Plischke, E., Morris, M.D., 2014. Transformations and invariance in the sensitivity analysis of computer experiments. J. R. Stat. Soc., Ser. B 76, 925–947.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C., 1984. Classification and Regression Trees. Chapman&Hall.

Candès, E., Fan, Y., Janson, L., Lv, J., 2018. Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. J. R. Stat. Soc., Ser. B, Stat. Methodol. 80, 551–577.

Chan, K., Saltelli, A., Tarantola, S., 2000. Winding stairs: a sampling tool to compute sensitivity indices. Stat. Comput. 10, 187–196.

Chatterjee, S., 2021. A new coefficient of correlation. J. Am. Stat. Assoc. 116, 2009–2022.

Chaudhuri, A., Hu, W., 2019. A fast algorithm for computing distance correlation. Comput. Stat. Data Anal. 135, 15–24.

Da Veiga, S., 2015. Global sensitivity analysis with dependence measures. J. Stat. Comput. Simul. 85, 1283–1305.

Da Veiga, S., 2021. Kernel-based ANOVA decomposition and Shapley effects: application to global sensitivity analysis. arXiv:2101.05487. 1–44.

Dua, D., Graff, C., 2017. UCI machine learning repository. URL: http://archive.ics.uci.edu/ml.

Dunson, D.B., 2018. Statistics in the Big Data era: failures of the machine. Stat. Probab. Lett. 136, 4–9.

Efron, B., Gong, G., 1983. A leisurely look at the bootstrap, the jackknife, and cross-validation. Am. Stat. 37, 36–48.

Fan, J., Lv, J., 2008. Sure independence screening for ultrahigh dimensional feature space. J. R. Stat. Soc., Ser. B, Stat. Methodol. 70, 849–911.

Fisher, A., Rudin, C., Dominici, F., 2019. All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. J. Mach. Learn. Res. 20, 1–81.

Gamboa, F., Gremaud, P., Klein, T., Lagnoux, A., 2020. Global sensitivity analysis: a new generation of mighty estimators based on rank statistics. arXiv: 2003.01772. 1–27.

Gamboa, F., Janon, A., Klein, T., Lagnoux, A., Prieur, C., 2016. Statistical inference for Sobol pick-freeze Monte Carlo method. Statistics 50, 881–902.

Gamboa, F., Klein, T., Lagnoux, A., 2018. Sensitivity analysis based on Cramér von Mises distance. SIAM/ASA J. Uncertain. Quantificat. 6, 522–548.

Glick, N., 1975. Measurements of separation among probability densities or random variables. Can. J. Stat. 3, 267–276.

Gretton, A., Bousquet, O., Smola, A., Schölkopf, B., 2005. Measuring statistical dependence with Hilbert-Schmidt norms. Algorithmic learning theory. In: 16th International Conference, ALT 2005, pp. 63–77.

Hastie, T.J., Tibshirani, R., Friedman, J.H., 2009. The Elements of Statistical Learning, second ed. Springer-Verlag.

Hotelling, H., 1936. Relations between two sets of variates. Biometrika 28, 321–377.

Kuiper, N.H., 1960. Tests concerning random points on a circle. Proc. K. Ned. Akad. Wet., Ser. A 63, 38–47.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86, 2278–2324.

Lundberg, S., Erion, G., Lee, S.I., 2019. Consistent individualized feature attribution for tree ensembles. arXiv:1802.03888. 1–9.

Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 2017, 4766–4775.

Marrel, A., Charibon, V., 2021. Statistical developments for target and conditional sensitivity analysis: application on safety studies for nuclear reactor. Reliab. Eng. Syst. Saf. 2014, 107711.

Murdoch, W.J., Signh, C., Kumbier, K., Abbasi-Asl, R., Yu, B., 2019. Definitions, methods and applications in interpretabile Machine Learning. Proc. Natl. Acad. Sci. 116, 22071–22080.

Pan, W., Wang, X., Zhang, H., Zhu, H., Zhu, J., 2020. Ball Covariance: a generic measure of dependence in Banach space. J. Am. Stat. Assoc. 115, 307–317.

Pearson, K., 1895. Notes on regression and inheritance in the case of two parents. Proc. R. Soc. Lond. 58, 240–242.

Pearson, K., 1905. On the general theory of skew correlation and non-linear regression. In: Mathematical Contributions to the Theory of Evolution. London.

Plischke, E., Borgonovo, E., Smith, C., 2013. Global sensitivity measures from given data. Eur. J. Oper. Res. 226, 536–550.

Renyi, A., 1959. On measures of statistical dependence. Acta Math. Acad. Sci. Hung. 10, 441–451.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. “Why should I trust you?” Explaining the predictions of any classifier. In: Proceedings of the International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144.

Rudin, C., 2019. Stop explaining black-box Machine Learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. 1, 206–215.

Sah, P., Fokoué, E., 2019. What do Asian religions have in common? An unsupervised text analytics exploration. arXiv:1912.10847. 1–34.

Saltelli, A., 2002. Making best use of model valuations to compute sensitivity indices. Comput. Phys. Commun. 145, 280–297.

Soofi, E.S., 1994. Capturing the intangible concept of information. J. Am. Stat. Assoc. 89, 1243–1254.

Spearman, C., 1904. The proof and measurement of the association between two things. Am. J. Psychol. 15, 72–101.

Strong, M., Oakley, J., 2013. An efficient method for computing partial expected value of perfect information for correlated inputs. Med. Decis. Mak. 33, 755–766.

Strong, M., Oakley, J.E., Chilcott, J., 2012. Managing structural uncertainty in health economic decision models: a discrepancy approach. J. R. Stat. Soc., Ser. C 61, 25–45.

Székely, G.J., Rizzo, M.L., 2009. Brownian distance covariance. Ann. Appl. Stat. 3, 1236–1265.

Székely, G.J., Rizzo, M.L., Bakirov, N.K., 2007. Measuring and testing dependence by correlation of distances. Ann. Stat. 35, 2769–2794.

Taverniers, S., Hall, E.J., Katsoulakis, M.A., Tartakovsky, D.M., 2021. Mutual information for explainable deep learning of multiscale systems. J. Comput. Phys. 444, 110551.

Wiesel, J.C., 2022. Measuring association with Wasserstein distances. Bernoulli 28, 2816–2832.

Xiao, H., Rasul, K., Vollgraf, R., 2017. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. arXiv:1708.07747. 1–20.

Zhao, Q., Hastie, T., 2021. Causal interpretations of Black-Box models. J. Bus. Econ. Stat. 39, 272–281.