Feature Selection for Trustworthy Regression Using Higher Moments

Fabian Hinder(✉), Johannes Brinkrolf, and Barbara Hammer

Cognitive Interaction Technology (CITEC), Bielefeld University,
Inspiration 1, 33619 Bielefeld, Germany
{fhinder,jbrinkro,bhammer}@techfak.uni-bielefeld.de

Abstract. Feature Selection is one of the most relevant preprocessing techniques in machine learning. Yet, it is usually only considered in the context of classification tasks. Although many methods designed for classification can be carried over to regression tasks, they usually lack some of the theoretical guarantees, that are provided for classification. In particular, reject-option and certainty measures or, more generally, operations which depend on the posterior distribution rather than its expectation only, are not supported. As machine learning is increasingly used in all areas of the daily life including high risk areas like medicine, such tools are essential. In this work, we focus on the problem how to extend feature selection techniques, such that certainty measures are taken into consideration during the selection process. We show that every method which is applicable in multi-value regression can be extended to take into account the complete distribution by making use of higher moments. We prove that the resulting method can be applied to preserve various certainty measures for regression tasks, including variance and confidence intervals, and we demonstrate this in example applications.

Keywords: Feature selection · Feature relevance · Trustworthy regression · Higher moments · Non-parametric methods

1 Introduction

As machine learning systems become more and more relevant in every day life, including critical infrastructure, medical applications, autonomous driving, etc., the demand for trustworthy AI becomes increasingly relevant [5,12,15]. Many existing approaches including feature selection technologies mainly focus on improving model precision or efficiency and often ignore the model confidence [11]. This is particularly problematic if critical decisions are based on possibly insufficient predictions e.g. due to statistical fluctuations in the (training) data or a small sample size. Popular approaches that tackle such restrictions per design include, e.g., reject-options [5,14,15], background classification [12],

Funding in the frame of the BMWi project KI-Marktplatz, 01MK20007E, is gratefully acknowledged.

E. Pimenidis et al. (Eds.): ICANN 2022, LNCS 13532, pp. 76–87, 2022.
https://doi.org/10.1007/978-3-031-15937-4_7

or confidence intervals [15]. These methods rely on quantities which are derived from the conditional label distribution or posterior rather than the expected prediction only.

Issues regarding trustworthiness or certainty can be amplified if relevant information is removed during preprossessing steps or ignored by analysis tools. Feature selection is a common preprocessing technique that has a high risk for such mistakes, by removing features that are not relevant to model accuracy but could be critical to trustworthiness. A simple example to show this point is given by the regression task $y = x_1 + x_2\varepsilon$ where ε is independent Gaussian noise. In this case a simple feature selection that only pays attention to the mean value would consider x_2 to be irrelevant, although it is of high relevance for the certainty of the prediction. Currently, many common feature selection methods suffer from this problem.

In this contribution, we provide a theoretical framework that allows us to understand the discrepancy between accuracy-based feature selection methods and those that take the entire posterior, including model certainty, into account. We relate this challenge to feature relevance theory. Thereon, we derive a simple extension to standard feature selection methods and analysis techniques that extends the selection objective to consider the entire posterior, thus including all certainty related quantities. The model can efficiently be implemented based on the higher moments of the label variable.

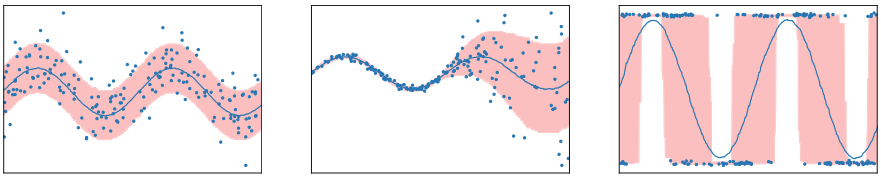
This paper is organized as follows: First, we recall common certainty measures for regression and discuss their strength and weaknesses (Sect. 2). Then, we recall the definition of classical feature relevance (Sect. 3.1) and provide a new definition that focuses on performance (Sect. 3.2). We recall several feature selection methods from the literature and set them into the context of the theoretical framework provided before (Sect. 4) and compare the relevance notions on a theoretical level (Sect. 5) and derive new approach to deal with them (Sect. 6). We conclude our work by a numerical evaluation of our criterion (Sect. 7) and summarize our findings (Sect. 8).

2 Trustworthy Regression

As it is a common assumption in classification that the classes are rather well separated, a very common assumption in regression tasks is that the noise is independent of the data (see Fig. 1a). Both assumptions can be violated in practical applications, leading to the need for confidence or certainty measures to avoid misleading over-precise results.

In classification, one commonly used type are reject options: If a probabilistic classification model predicts very comparable likelihoods for more than one most likely class then the model refuses to perform the prediction [14]. This idea has been extended by adding a "background class" that has a low uniform density everywhere, which allows the model to identify regions where there is insufficient data to make a trustworthy prediction [12].

Under the common assumption that the conditional label $y \mid X$ follows an unimodal distributed, e.g. normal distribution, we can extend the idea of reject



(a) Common regression example with independent Gaussian noise.

(b) Regression example with dependent Gaussian noise.

(c) Regression on bimodal distribution with dependent mean.

Fig. 1. Regression examples with comparable MSE but different levels of trustworthyness. Graphic shows data points (blue), regression line (blue), and 10%–90%-quantiles (red). (Color figure online)

options to regression tasks by estimating not only the mean value but also a confidence interval. Similar to reject options for classification, the model can reject a sample if the confidence interval is too large [15] or present the interval to the user to allow them to make an informed decision.

We illustrate the idea of confidence bounds in Fig. 1. Although, the MSE is comparable in all three cases, only in the case of Fig. 1a it is also a valid measure for the model certainty. In case of Fig. 1b, the variance drastically increases to the right, so that the model certainty heavily depends on the considered point. In case of Fig. 1c, the estimation of the mean value is very precise, however, as the distribution is not unimodal, the conditional mean itself is misleading. In all three cases the 10%–90%-quantile area provides a good insight into the specific certainty issue.

(Conditional) Variance (and Higher Moments): A common way to quantify the certainty of a measurement is offered by its (conditional) variance. In the case of normal distributed noise, the variance fully represents the uncertainty. However, in cases where the uncertainty depends on the input value, it might be more reasonable to consider conditional variance $\mathrm{Var}(y \mid X) = \mathbb{E}[(y - \mathbb{E}[y \mid X])^2 \mid X]$. It has been successfully used in the context of regression with reject option [15], for example. Notice that conditional variance can easily be estimated using a second model to estimate $\mathbb{E}[y^2 \mid X]$ based on a mean squared error loss.

Albeit the variance offers a suitable certainty measure under the assumption of normal distributed noise, higher moments like skewness or kurtosis can help to understand the peculiarities of more general distributions. For example, if overestimating the true value is less problematic than underestimating it, skewness can offer important information. As another example if an estimation error is critical in extreme cases only, it might be sufficient to take care of the tail distribution – such information is provided by the kurtosis of the distribution.

Quantiles: One drawback of moment-based confidence measures is that they are sometimes hard to interpret in a specific setup. Quantiles provide an alternative

in such cases: since they are the level points of the cumulative distribution function (cdf), their interpretation is straight forward. As in the case of moments, several estimation methods exist. However, as those are based on a different loss function [9] not every model can be used in a straight forward fashion. Another drawback of quantiles in comparison to moments is that they provide information of a single point of the distribution only. In particular, they do not take information regarding the tail distribution into account.

3 Feature Relevance

We will now recall the notion of feature relevance. We consider classification and regression over $\mathbb{R}^d$ to $\{c_1, \ldots, c_m\}$ and $\mathbb{R}$, respectively, with pairs of random variables $(\mathbf{X}, Y)$, corresponding to data and label. We refer to the i-th feature of $\mathbf{X}$ as X_i. For a set $R = \{r_1, \ldots, r_n\} \subseteq \{1, \ldots, d\}$, we denote the sub-vector containing all features in R as $X_R = (X_{r_1}, \ldots, X_{r_n})$. We also make use of the shorthand notation $C_i = \{i\}^C$ and X_{C_i} for the subset and sub-vector of all features except i. In the next two section we will first recall the notion of feature relevance from the literature and then extend it as needed.

3.1 Feature Relevance for Classification

We recall the notion of relevance of a feature to the label variable Y as given by [7]. Roughly speaking, a feature X_i is relevant, if it provides information regarding Y. More formally, feature relevance is defined as follows:

Definition 1. *A feature X_i is* relevant *to Y if and only if there exists a set $R \subseteq C_i$ such that X_i and Y are not independent given X_R, i.e.*

$$Y \not\perp\!\!\!\perp X_i \mid X_R.$$

A relevant feature is called strongly relevant*, if and only if we may choose $R = C_i$, otherwise it is called* weakly relevant*. A feature that is not relevant is called* irrelevant*.*

Some authors prefer a slightly different but equivalent definition of strong and weak relevance:

Corollary 1. *A feature X_i is strong relevant if and only if Y is not conditionally independent of X_i given the remaining features, i.e. $Y \not\perp\!\!\!\perp X_i \mid X_{C_i}$. It is weakly relevant if and only if it can be made relevant by restricting the feature set, i.e. there exists a proper subset of features $R \subsetneq C_i$ for which X_i is strong relevant.*

As pointed out in [4], the distinction between strong and weak relevance is inspired by the observation that some features may carry redundant information regarding Y. As an example, consider the case where two features are identical copies of each other, i.e. $\mathbf{X} = (X_1, X_2, X_2)$. Supposing that Y can be predicted using $X_1 + X_2$, then the first feature is clearly relevant. The other features carry

relevant information but they are redundant and only one of those is required. In the framework of Definition 1, the first feature is strongly relevant, while features two and three are weakly relevant.

In the context of feature relevance, two problems are of particular interest: the minimal-optimal problem and the all-relevant problem.

The *minimal-optimal* problem refers to the problem of finding a smallest set of features, that are relevant to the Bayesian classifier and contain all important information. Hence adding further features does not improve the prediction accuracy for Y. It can be shown that there exists exactly one such minimal set for strictly positive distributions [10], which is equivalent to the set of strongly relevant features. When it comes to feature selection, one is usually interested in a minimal feature set.

The *all-relevant* problem refers to the problem of identifying all features relevant to Y. It was shown that this problem needs exhaustive search for general distributions [10]. Under assumptions on the distribution exact but computational expensive algorithms exist. Due to this restriction, other model based approaches were designed to find approximate solutions [4,8,13].

3.2 Feature Relevance for (MSE-)Regression

Although the definition of relevance given above can be applied to regression tasks, and is often referred to in the literature [3,8,13], it is usually approximated by a simpler form in practical applications, at least for regression tasks. The term of statistical independence is empirically tested only by considering a "decrease of the loss of the model", i.e. a feature X_i is relevant if it contains information that may help to predict Y. For universal MSE-regression models, this can be formalized as follows:

Definition 2. *A feature X_i is* $\mathbb{E}$-relevant *to Y if and only if there exists a set $R \subseteq C_i$, such that the conditioning of Y on X_i and X_i, X_R differ, i.e.*

$$\mathbb{E}[Y \mid X_i, X_R] \neq \mathbb{E}[Y \mid X_R].$$

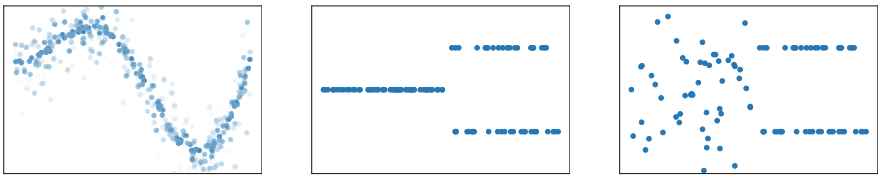
A $\mathbb{E}$-relevant feature is called strongly $\mathbb{E}$-relevant *if and only if we may choose $R = C_i$, otherwise it is called* weakly $\mathbb{E}$-relevant. *A feature that is not $\mathbb{E}$-relevant is called* $\mathbb{E}$-irrelevant.

Notice that the first formulation of weak and strong relevance (Corollary 1) carries over to weak and strong $\mathbb{E}$-relevance.

A key observation in this context is that (conditional) variance and bias are decreasing if the set of considered features is increased. As a consequence we observe that for the optimal model $f^*_X(x) = \mathbb{E}[Y \mid X = x]$ it holds

$$\mathrm{MSE}(f^*_{X_i, X_R}) \leq \mathrm{MSE}(f^*_{X_R})$$

and equality holds for all R if and only if X_i is $\mathbb{E}$-irrelevant. However, one can easily construct an example of a strongly-relevant but $\mathbb{E}$-irrelevant feature:



(a) Sample drawn from distribution of Example 1.

(b) Sample drawn from distribution of Example 2

(c) Sample drawn from distribution of Example 3

Fig. 2. Example distributions. Axis are x_1 and y, α-channel is x_2 (if present).

Example 1. Let $\mathbf{X} = (X_1, X_2)$ be a $\mathbb{R}^2$-valued random variable, $\varepsilon \sim \mathcal{N}(0, 1)$ an independent standard normal distributed random variable, $f : \mathbb{R} \to \mathbb{R}$ be some function. If we set $Y = f(X_1) + X_2\varepsilon$, then it holds $\mathbb{E}[Y \mid \mathbf{X}] = f(X_1) + X_2\mathbb{E}[\varepsilon] = f(X_1)$. So X_2 is $\mathbb{E}$-irrelevant, but since $\mathrm{Var}(Y \mid \mathbf{X}) = X_2^2$, X_2 is strong relevant.

We illustrate the distribution in this example in Fig. 2a. Notice that this example shows that $\mathbb{E}$-relevance is not sufficient for trustworthy regression, as only the first feature is relevant for the prediction, the second feature, however, is very important to estimate the certainty of the prediction. This gives rise to the question under which circumstances we lose information regarding the certainty of a model if we apply feature selection based on the notion of $\mathbb{E}$-relevance. As it turns out, this is a problem which is specific to regression tasks.

Furthermore, the question occurs, how the properties of relevance and $\mathbb{E}$-relevance are related in general. We will consider this question in Sect. 5.

4 Feature Selection Methods

In the following, we will discuss some classical feature selection methods from the literature and consider them in the context of the definition of relevance and $\mathbb{E}$-relevance and all-relevant and minimal-optimal problem, respectively. Notice that we will concentrate on wrapper methods which derive a relevance measure from an underlying model, as our approach is based on a transformation of the labels for the prediction task at hand, to select the features.

Recursive Feature Elimination (RFE) [3] is based on models that assign importances to features. This can be the feature weight, as in the case of linear models, or an implicit quantity, such as feature importance in case of decision trees or random forests. The algorithm proceeds in a recursive fashion: starting with all features, the model is trained and the feature with the smallest importance value is remove. This procedure is repeated until a certain number of features is obtained. Thereby, the desired number of features can either be predefined by the user or determined by cross-validation.

As RFE does not directly consider the features but only importances assigned by the model its relation to relevance or $\mathbb{E}$-relevance is unclear. However, if it

is used together with models that rely on an optimization of the MSE, only $\mathbb{E}$-relevance is considered. This model dependence also makes it hard to determine precisely the set of features RFE aims for, as not all models can process all information equivalently well. Furthermore, it also depends on whether the model is sparse, in which case we will obtain a model dependent analog to the minimal-optimal set, or not, in which case we are closer to the all-relevant set.

Sequential Feature Selection (SFS) [2] works similar to RFE in the sense that it either removes ("backward") or adds ("forward") features in a recursive fashion. However, instead of relying on the model to obtain the feature importance, cross-validation is used. It can be shown that "backward" and "forward" procedure do not lead to the same results, in general.

As in the case of RFE, it is neither clear whether relevant or $\mathbb{E}$-relevant features are found. However, in case of "backward" direction, features that are not necessarily relevant for the model are assigned a small value. Hence one can consider SFS as a minimal-optimal search strategy.

Boruta [8] uses the feature importance assigned by a learning model, most commonly a random forest. However, in contrast to RFE it adds a randomly permuted versions of the features, dubbed as "shadow-features". These are irrelevant for the prediction of the label by design. As a consequence, any feature that is ranked less important than a shadow feature cannot be relevant. A relevance ranking is then obtained by repeating the steps several time and then performing a statistical analysis.

Boruta is presented as an all-relevant search by the authors. Since it relies on the model error, it aims for $\mathbb{E}$-relevant features in the sense of Definition 2.

5 On the Relation of Relevance Notions

As already shown in Example 1, the notions of relevance and $\mathbb{E}$-relevance are not equivalent. Indeed, as suggested above the notion of relevance is stronger, in the sense that $\mathbb{E}$-relevance implies relevance. However, it is also true that under certain circumstances, e.g. in case of binary classification, both notions are equivalent. This justifies the usage of methods that aim for $\mathbb{E}$-relevance from a theoretical point of view in the case of classification problems. Furthermore, it also shows that in the case of classification, standard feature selection methods do not deteriorate trustworthiness of a model.

Theorem 1. *Let* $\mathbf{X}$ *and* Y *be random variables. If* X_i *is (weakly/strongly)* $\mathbb{E}$*-relevant, then it is (weakly/strongly) relevant. Conversely if* X_i *is irrelevant, then it is* $\mathbb{E}$*-irrelevant.*

Furthermore, if $Y \mid X_R$ *is Bernoulli distributed (or more general the set of all distributions can be parametrized by their mean value) for all* R*, then the notions of (weak/strong) relevance and (weak/strong)* $\mathbb{E}$*-relevant coincide.*

Proof. It suffices to show the first statement in the case where X_i is strongly $\mathbb{E}$-relevant. In case of weak $\mathbb{E}$-relevance, we reduce to the set R for which i is strongly $\mathbb{E}$-relevant (Corollary 1), the converse statement for irrelevance follows directly from the definition. Let X_i be strongly $\mathbb{E}$-relevant and assume that it is not strongly relevant. By the rules of conditional expectation regarding independence it then holds $\mathbb{E}[Y \mid X_i, X_{C_i}] = \mathbb{E}[Y \mid X_{C_i}]$ which is a contradiction.

For the second statement, it again suffices to show that strong relevance implies strong $\mathbb{E}$-relevance. Let X_i be strongly relevant and assume that it is not strongly $\mathbb{E}$-relevant. Then $\mathbb{E}[Y \mid X_i, X_{C_i}] = \mathbb{E}[Y \mid X_{C_i}]$, but since expectation is (by assumption) the only parameter of $\mathbb{P}_{Y|X_R}$ for all R, this implies that $\mathbb{P}_{Y|X_i,X_{C_i}} = \mathbb{P}_{Y|X_{C_i}}$ which is equivalent to conditional independence and therefore a contradiction.

Considering Theorem 1, one might ask for which type of distribution the notion of relevance and $\mathbb{E}$-relevance coincide. However, as shown by Example 1 this does not hold even for very simple distributions. Indeed, equality basically only holds in the described case. In particular, for the simplest regression task the notion actually do coincide:

Corollary 2. *If there exists a function f such that $Y \mid \mathbf{X} = f(\mathbf{X}) + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma)$ is independent, normal distributed noise, the notions of (weakly/strong) relevance and (weakly/strong) $\mathbb{E}$-relevant coincide.*

However, in this case the certainty is independent of $\mathbf{X}$ and thus of no interest. For every set that contains more than two points there exists a distribution for which equality no longer holds:

Example 2. Let $\mathbf{X} = (X_1) \sim \mathcal{U}([-1, 1])$ a uniformly distributed random variable and σ be an independent Bernoulli distributed random variable with mean value $1/2$. Set $Y = 2\sigma - 1$ if $X_1 > 0$ and $Y = 0$ otherwise. Then Y is supported on three points only. Furthermore, since $\mathbb{E}[Y \mid \mathbf{X}] = 0 = \mathbb{E}[Y]$ it follows that X_1 is not $\mathbb{E}$-relevant, but the value of the conditional variance implies that X_1 is relevant.

We illustrate this example in Fig. 2b. As restricting the considered distributions yields only trivial solutions, we are looking for a way that allows us to relate relevance and $\mathbb{E}$-relevance. Considering Example 1 and 2, the size of the variance may provide a sufficient criterion. However, it is easy to construct an example where this is no longer the case:

Example 3. Consider the same setup as in Example 2 and let ε be an independent standard normal distributed random variable. Set $Y' = \varepsilon(1 - |Y|) + Y$. Clearly $\mathbb{E}[Y' \mid \mathbf{X}] = \mathbb{E}[Y'] = 0$, furthermore $\mathrm{Var}(Y' \mid \mathbf{X}) = 1$.

Instead of considering only the first two moments, i.e. mean and variance, we can also consider the higher moments which solves the problem:

Theorem 2. *Let Y be a real-valued random variable. Assume Y is compactly supported or fulfills Carleman's condition [1]. If X_i is (weakly/strongly) relevant to Y then there exists a k such that X_i is (weakly/strongly) $\mathbb{E}$-relevant to Y^k.*

Proof. It is again suffices to show the case where X_i is strongly relevant (see the proof of Theorem 1). As X_i is strongly relevant it holds $Y \not\perp\!\!\!\perp X_i \mid X_{C_i}$ which is equivalent to $\mathbb{P}_{Y|X_i,X_{C_i}} \neq \mathbb{P}_{Y|X_{C_i}}$, i.e. there exists a set C such that $\mathbb{P}_{Y|X_i,X_{C_i}}(C) \neq \mathbb{P}_{Y|X_{C_i}}(C)$ as L^1-functions. Denote the conditionals (with and without X_i) difference by $f(x_i, x_r) = \mathbb{P}_{Y|X_i=x_i,X_{C_i}=x_r}(C) - \mathbb{P}_{Y|X_{C_i}=x_r}(C)$. As $f \neq 0$, we may find a set $A \times B \subset \mathbb{R} \times \mathbb{R}^{d-1}$ on which f has positive expectation, i.e. $\mathbb{E}[f, A \times B] > 0$, using a monotonous class argument. As $\mathbb{E}[f \mid X_{C_i}] = 0$, it follows that $\mathbb{E}[f, A \times B] = -\mathbb{E}[f, A^C \times B]$ and thus $\mathbb{P}_{Y|X_i \in A, X_{C_i} \in B} \neq \mathbb{P}_{Y|X_i \in A^C, X_{C_i} \in B}$. If Y is compactly supported, the statement follows by applying Weierstrass's approximation theorem. To use Carleman's condition observe that either $\mathbb{P}_{Y|X_i \in A, X_{C_i} \in B^C} = \mathbb{P}_{Y|X_i \in A^C, X_{C_i} \in B^C}$, in which case we may replace B by $\mathbb{R}^{d-1}$, or it does not hold, so that we end up with two pairs of distributions, each with a global split along X_i. As Carleman's condition holds globally for Y, it follows by Jensen's inequality that it holds for at least one of the two/four distributions, which is then uniquely determined by its moments. Notice that this suffices to show the statement since the partner distributions either fulfills the condition, too, in which case the uniqueness applies, or it does not, in which case the moments have to be different.

Notice that Theorem 2 allows us to connect relevance and $\mathbb{E}$-relevance of a higher dimensional problem, using a transformation of the labels. We will derive an algorithmic solution from this observation in the next section.

6 Application: Moment Feature Relevance

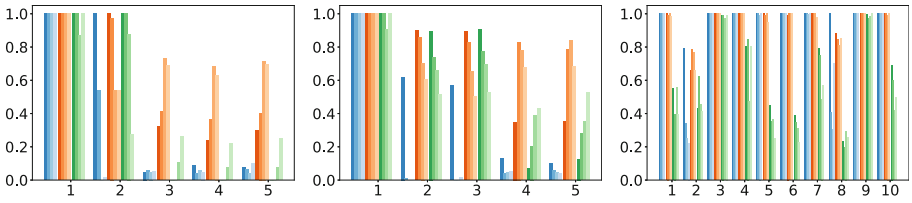
By applying Theorem 2 we connect the task to determine relevant and $\mathbb{E}$-relevant features. For this purpose, we perform relevance analysis for the regression task with respect to the label vector $(Y, Y^2, \ldots, Y^d)$ rather than Y [6]. One can also use a different transformation, e.g. based on the Legendre polynomials or a Fourier transformation, which can be beneficial in practice. Indeed, if Y is compactly supported, any function basis can be used. We refer to this method as *Moment Feature Relevance*.

Although we have to take $d \to \infty$ to monitor relevance, considering the first d moments is usually sufficient in practice due to noise and estimation errors. Indeed, we can quantify the error by following corollary of [6, Theorem 2] which shows that for large d only features with small impact on the label are missed:

Corollary 3. *Let Y be a $[-1, 1]$-valued random variable and assume that $\mathbb{E}[Y^k \mid \mathbf{X}] = \mathbb{E}[Y^k \mid X_{C_i}]$ for all $k \leq d$. Then it holds*

$$\int_0^1 \left| F^{-1}_{Y|X_{C_i}}(q) - F^{-1}_{Y|\mathbf{X}}(q) \right| \mathrm{d}q = \int_{-\infty}^{\infty} \left| F_{Y|X_{C_i}}(y) - F_{Y|\mathbf{X}}(y) \right| \mathrm{d}y \leq \sqrt{\frac{16}{2d-1}},$$

where $F_{Y|\mathbf{X}}$, $F^{-1}_{Y|\mathbf{X}}$, $F_{Y|X_{C_i}}$, and $F^{-1}_{Y|X_{C_i}}$ denote the cdf and quantile function (i.e. inverse cdf) of $\mathbb{P}_{Y|\mathbf{X}}$ and $\mathbb{P}_{Y|X_{C_i}}$, respectively.



(a) Line (b) SinSin (c) Diabetes

Fig. 3. Mean probability to select a feature (1–5/1–10) for used datatasets, selection methods, and transformations. Block order: Boruta (blue), REF (orange), SFS (green). In block order (dark to light): Fourier, Legendre, Moment, Raw. (Color figure online)

Notice that this result connects the two confidence measures considered in Sect. 2. In particular, when computing the feature relevances using the moment technique, we also take care of the quantiles. Furthermore, notice that though Corollary 3 formally require Y to take values in $[-1, 1]$, the statement can also be applied in the case where Y takes on any values in $\mathbb{R}$ using a suitable preprocessing function like tanh.

7 Empirical Evaluation

We empirically evaluate our method. The method uses a label transformation followed by a standard feature selection on the resulting (multi-)regression problem (see Sect. 6 for details). We use two theoretical datasets with ground truth, i.e. the relevant features are known. Further, we use one real world dataset. We use the following feature selection methods: SFS (backward search, predefined number of features), RFE (using cross-validation to determine number of features), and Boruta (default parameters). We use the following moment transformations: Fourier $(y \mapsto (\exp(k\pi \mathbf{i} y))_{k=-d}^{d})$, Legendre $(y \mapsto (L_k(y))_{k=1}^{d}$, where L_k denotes the Legendre polynomial of degree k), Moment: $(y \mapsto (y^k)_{k=1}^{d})$, and Raw $(y \mapsto y$, i.e. the base case without transformation). In all cases, we choose $d = 5$ as suggested in [6] and random forests as base model. Before the transformation, Y is normalized to the interval -1 and 1.

For the theoretical data, $\mathbf{X}$ follows a 5-dimensional uniform distribution. Y is distributed according to

$$\begin{aligned} \text{Line: } Y &= 8X_1 - (36X_2^2 - 36X_2 + 1)\varepsilon \\ \text{SinSin: } Y &= \sin(2\pi X_1 - \pi) + 3\sin(2\pi X_2 - \pi)\sin(2\pi X_3 - \pi)\varepsilon, \end{aligned}$$

whereby $\varepsilon \sim \mathcal{N}(0, 1)$ is an independent standard normal distribution. As can be seen X_1 and X_2 are relevant for Line, and X_1, X_2, and X_3 relevant for SinSin. In both cases only X_1 is $\mathbb{E}$-relevant. Notice that the all-relevant and minimal-optimal feature sets coincide. To evaluate the method we compare the feature set selected by the method and the set of relevant features. The results are presented in Fig. 3a, b, and in Table 1. The Fourier transformation works best over

Table 1. Mean results over 300 runs. Table shows how many features are selected by the method (all) and how many of them are relevant (rel., number in brackets indicates number of truly relevant features), and precision (prc.), recall (rec.), and F1-score (F1) comparing selected and relevant features.

Method		Line					SinSin				
	Trans.	rel. (2)	all	prc	rec	F1	rel. (3)	all	prc	rec	F1
Boruta	Fourier	2.0	2.2	0.9	**1.0**	**1.0**	2.2	2.4	0.9	**0.7**	**0.8**
	Legendre	1.5	1.7	0.9	0.8	0.8	1.0	1.1	1.0	0.3	0.5
	Moment	1.0	1.1	0.9	0.5	0.6	1.0	1.1	1.0	0.3	0.5
	Raw	1.0	1.2	0.9	0.5	0.6	1.0	1.1	1.0	0.3	0.5
RFE	Fourier	2.0	2.9	**0.8**	**1.0**	**0.9**	2.8	3.5	**0.8**	**0.9**	**0.9**
	Legendre	2.0	3.2	0.7	**1.0**	0.8	2.7	4.3	0.7	**0.9**	0.7
	Moment	1.5	3.7	0.5	0.8	0.5	2.4	4.0	0.6	0.8	0.7
	Raw	1.5	3.6	0.5	0.8	0.6	2.1	3.5	0.7	0.7	0.6
SFS	Fourier	2.0	2.0	**1.0**	**1.0**	**1.0**	2.8	3.0	**0.9**	**0.9**	**0.9**
	Legendre	2.0	2.0	**1.0**	**1.0**	**1.0**	2.5	3.0	0.8	0.8	0.8
	Moment	1.7	2.0	0.9	0.9	0.9	2.3	3.0	0.8	0.8	0.8
	Raw	1.3	2.0	0.6	0.6	0.6	2.0	3.0	0.7	0.7	0.7

all selection methods, followed by the Legendre transformation, which however seems to be less compatible with Boruta and RFE on SinSin. Simple moments seem to work with SFS, only. Furthermore, without moments (Raw) all methods are only capable of identifying the $\mathbb{E}$-relevant features, the probability of detecting the other relevant features is random.

We also apply the method to the UCI benchmark diabetes. We evaluate the method by splitting the dataset in two halves (50%), one for selecting the features and the remaining to score the selection using the test error (5-fold, 2/3–1/3 train test split) of a quantile regression model ($q = 0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95$) on the selected features only. The selection probability for Boruta and REF are very comparable (see Fig. 3c). SFS considers features 1, 2, 5, 6, and 8 less relevant. The scores of all methods are comparable and well inside the range of statistical fluctuations.

8 Conclusion

In this paper, we considered the problem of feature selection for trustworthy regression. We showed that feature relevance is a suitable relevance notion to take confidence intervals into account. We established a formal framework that allows us to connect feature selection via wrapper methods and feature relevance. Using this, we showed that commonly used wrapper methods are not sufficient for detecting all features that are needed if confidence is a target. We suggested and evaluated an extension via a label transformation to solve this problem.

References

1. Akhiezer, N.I.: The Classical Moment Problem and Some Related Questions in Analysis. SIAM, Philadelphia (2020)
2. Chandrashekar, G., Sahin, F.: A survey on feature selection methods. Comput. Electr. Eng. **40**(1), 16–28 (2014). https://doi.org/10.1007/s10462-021-10072-6
3. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Mach. Learn. **46**(1), 389–422 (2002). https://doi.org/10.1023/A:1012487302797
4. Göpfert, C., Pfannschmidt, L., Göpfert, J.P., Hammer, B.: Interpretation of linear classifiers by means of feature relevance bounds. Neurocomputing **298**, 69–79 (2018). https://doi.org/10.1016/j.neucom.2017.11.074
5. Hendrickx, K., Perini, L., Van der Plas, D., Meert, W., Davis, J.: Machine learning with a reject option: A survey. arXiv preprint arXiv:2107.11277 (2021)
6. Hinder, F., Vaquet, V., Brinkrolf, J., Hammer, B.: Fast non-parametric conditional density estimation using moment trees. In: 2021 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 1–7 (2021). https://doi.org/10.1109/SSCI50451.2021.9660031
7. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artif. Intell. **97**(1), 273– 324 (1997). https://doi.org/10.1016/S0004-3702(97)00043-X, relevance
8. Kursa, M.B., Rudnicki, W.R.: Feature selection with the Boruta package. J. Stat. Softw. **36**(11), 1–13 (2010). https://doi.org/10.18637/jss.v036.i11
9. Meinshausen, N., Ridgeway, G.: Quantile regression forests. J. Mach. Learn. Res. **7**(6), 983–999 (2006)
10. Nilsson, R., Peña, J., Björkegren, J., Tegner, J.: Consistent feature selection for pattern recognition in polynomial time. J. Mach. Learn. Res. **8**, 589–612 (2007)
11. Osborne, J.W., Waters, E.: Four assumptions of multiple regression that researchers should always test. Pract. Assess. Res. Eval. **8**(1), 2 (2002)
12. Perello-Nieto, M., Filho, T.D.M.E.S., Kull, M., Flach, P.: Background check: a general technique to build more reliable and versatile classifiers. In: 2016 IEEE 16th International Conference on Data Mining (ICDM), pp. 1143–1148 (2016). https://doi.org/10.1109/ICDM.2016.0150
13. Pfannschmidt, L., Hammer, B.: Sequential feature classification in the context of redundancies. CoRR **abs/2004.00658** (2020). https://arxiv.org/abs/2004.00658
14. Villmann, T., et al.: Self-adjusting reject options in prototype based classification. In: Merényi, E., Mendenhall, M.J., O'Driscoll, P. (eds.) Advances in Self-Organizing Maps and Learning Vector Quantization. AISC, vol. 428, pp. 269–279. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-28518-4_24
15. Zaoui, A., Denis, C., Hebiri, M.: Regression with reject option and application to KNN. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems, vol. 33, pp. 20073–20082. Curran Associates, Inc. (2020)