TEST
https://doi.org/10.1007/s11749-024-00958-2

ORIGINAL PAPER

Distribution-free tests for lossless feature selection in classification and regression

László Györfi[1] · Tamás Linder[2] · Harro Walk[3]

Received: 6 April 2024 / Accepted: 13 November 2024

Abstract

We study the problem of lossless feature selection for a d-dimensional feature vector $X = (X^{(1)}, \ldots, X^{(d)})$ and label Y for binary classification as well as nonparametric regression. For an index set $S \subset \{1, \ldots, d\}$, consider the selected $|S|$-dimensional feature subvector $X_S = (X^{(i)}, i \in S)$. If L^* and $L^*(S)$ stand for the minimum risk based on X and X_S, respectively, then X_S is called lossless if $L^* = L^*(S)$. For classification, the minimum risk is the Bayes error probability, while in regression, the minimum risk is the residual variance. We introduce nearest-neighbor-based test statistics to test the hypothesis that X_S is lossless. This test statistic is an estimate of the excess risk $L^*(S) - L^*$. Surprisingly, estimating this excess risk turns out to be a functional estimation problem that does not suffer from the curse of dimensionality in the sense that the convergence rate does not depend on the dimension d. For the threshold $a_n = \log n/\sqrt{n}$, the corresponding tests are proved to be consistent under conditions on the distribution of (X, Y) that are significantly milder than in previous work. Also, our threshold is universal (dimension independent), in contrast to earlier methods where for large d the threshold becomes too large to be useful in practice.

Keywords Classification · Nonparametric regression · Lossless feature selection · Nearest-neighbor estimate · Consistent test

L. Györfi.: The research of László Györfi was supported by the National Research, Development and Innovation Fund of Hungary under the 2023-1.1.1-PIACI-FÓKUSZ-2024-00051 funding scheme. T. Linder: Tamás Linder's research was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

✉ Tamás Linder
tamas.linder@queensu.ca

László Györfi
gyorfi@cs.bme.hu

Harro Walk
harro.walk@mathematik.uni-stuttgart.de

1 Department of Computer Science and Information Theory, Budapest University of Technology and Economics, Budapest, Hungary

2 Department of Mathematics and Statistics, Queen's University, Kingston, ON, Canada

3 Institut für Stochastik und Anwendungen, Universität Stuttgart, Stuttgart, Germany

Mathematics Subject Classification 62G05 · 62G10 · 62G08

1 Introduction

In this paper we study the problem of lossless feature selection for classification and nonparametric regression.

Binary classification deals with the problem of deciding on a ± 1-valued random label Y based on a random feature vector X taking values in $\mathbb{R}^d$, so that the risk, measured by the decision error probability, is as small as possible. If the joint distribution of X and Y is known, then the optimal decision with minimum risk (error probability), called the Bayes decision, can be derived. In the standard setup of classification, the joint distribution is unknown, but instead an observed random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of n independent copies of (X, Y) is available from which an estimate of the Bayes decision is to be constructed. Although estimates (classification rules) exist whose error probabilities converge to the optimum as $n \to \infty$ without any condition on the distribution of X and Y, the convergence rate of the error probability of any such classification rule to the Bayes error is very sensitive to the dimension of the feature vector. This suggests that dimension reduction, also called feature selection, is crucial before constructing a classification rule.

For nonparametric regression, Y is a real-valued random variable with $\mathbb{E}[Y^2] < \infty$, the risk is the mean squared error, and the minimum risk is the residual variance.

In this paper we are interested in feature selection that is lossless, i.e., does not incur any loss of information. For lossless feature selection, the minimum risk based on the feature vector obtained by leaving out some components of X and that based on the original feature vector X, are equal.

To make this notion precise, let S denote a proper subset of $\{1, \ldots, d\}$, and for an $\mathbb{R}^d$-valued feature vector

$$X = (X^{(1)}, \ldots, X^{(d)}),$$

consider the $\mathbb{R}^{|S|}$-valued subvector picked out by S, given by

$$X_S = (X^{(i)}, i \in S). \tag{1}$$

Let L^* and $L^*(S)$ denote the minimum risk based on X and X_S, respectively. Then X_S is called lossless if $L^* = L^*(S)$. One goal in this paper is to construct a nonparametric (distribution-free) test for the null hypothesis $\mathcal{H}_0 : L^* = L^*(S)$, i.e., for the null hypothesis that the minimum risks based on X and X_S are equal. In our setup, the alternative hypothesis $\mathcal{H}_1$ is that $L^* < L^*(S)$, i.e., $\mathcal{H}_1$ is the complement of the null hypothesis, and therefore there is no separation gap between the hypotheses. It is not at all obvious that consistency is possible without any separation gap. Dembo and Peres (1994) and Nobel (2006) characterized hypotheses pairs that admit strongly consistent tests, i.e., tests that, with probability one, only make finitely many Type I and II errors. This property is called discernibility. As an illustration of the intricate nature of the discernibility concept, Dembo and Peres (1994) demonstrated an exotic example, where the null hypothesis is that the mean of a random variable is rational, while the alternative hypothesis is that this mean minus $\sqrt{2}$ is rational. (See also Cover (1973) and Kulkarni and Zeitouni (1991).) The discernibility property shows up in Biau and Györfi (2005) (testing homogeneity), Devroye and Lugosi (2002) (classification of densities), Gretton and Györfi (2010) (testing independence), Györfi and Walk (2012) and Györfi et al. (2023) (testing conditional independence), Morvai and Weiss (2021) and Nobel (2006) (classification of stationary processes), among others.

Consistent tests can be constructed by estimating the functionals L^* and $L^*(S)$ using distribution-free consistent nonparametric estimates, i.e., estimates that (in some sense) converge to the target functional as $n \to \infty$. For example, Györfi and Walk (2017) consider the Bayes error probability, Devroye et al. (2018) the residual variance, Berrett et al. (2019), Delattre and Fournier (2017), and Kozachenko and Leonenko (1987) the differential entropy, Gretton and Györfi (2010), Silva and Narayanan (2010), and Wang et al. (2005) the mutual information, and Beirlant et al. (2001) the total variation.

Most of these estimators are based on consistent nonparametric estimators of the corresponding function (e.g., for the residual variance, the estimator of the regression function). However, the rate of convergence of such functional estimators is determined by the rate of convergence of the corresponding function estimator, which can be slow. In general, estimating the function itself is a harder problem than estimating the corresponding functional because the variance of the functional estimate can be of order $O(1/n)$, as is the case in this work, see for example inequalities (29), (42) and (43) in our analysis. In fact, it surprisingly turns out that functional estimators with good rate of convergence are based on non-consistent function estimates, e.g., the 1-NN-based estimators of residual variance in Devroye et al. (2018) and differential entropy in Kozachenko and Leonenko (1987).

In this paper, we introduce a distribution-free test which uses an estimate of the difference of the minimum risks $L^* - L^*(S)$, called the excess risk. This is in contrast to estimating L^* and $L^*(S)$ separately and then taking the difference of these two estimates. This is an important step since all separate estimates of L^* and $L^*(S)$ suffer from the curse of dimensionality. We propose a nearest-neighbor-based test statistic T_n and the threshold $a_n = \log n/\sqrt{n}$. The null hypothesis $L^* = L^*(S)$ is accepted if $T_n \leq a_n$, and otherwise rejected. Our main results show that the resulting tests (for classification and regression, respectively) are consistent in the sense that the Type I and II errors converge to zero as the sample size n tends to infinity. Surprisingly, estimating the excess risk turns out to be a functional estimation problem where there is no curse of dimensionality in the sense that the convergence rate (under the null hypothesis) depends on d only through a multiplicative factor (see (36) at the end of the proof of Theorem 1.)

For the alternative hypothesis, the analysis is relatively straightforward because the consistency of the test naturally follows from the consistency of the minimum risk estimates. However, under the null hypothesis, the problem is more challenging. On the one hand, the estimation error $T_n - \mathbb{E}[\, T_n \,]$ is of order $1/\sqrt{n}$. On the other hand, the absolute value of the bias $\mathbb{E}[\, T_n \,]$ can be of order $n^{-1/d}$ (d is the dimension of the feature vector X), which can be much larger than $1/\sqrt{n}$. The resulting threshold is about $a_n =$

$\log n/n^{1/d}$, which becomes impractically large for high-dimensional X, because the test rarely rejects the null hypothesis, see, e.g., Györfi and Walk (2017) (classification) and Devroye et al. (2018), Section 3 (regression). To address this problem, we develop statistics T_n such that the bias of the estimator of $L^*(S)$ is much larger than the bias of the estimator of L^*, resulting in a negative bias $\mathbb{E}[\, T_n \,]$. This approach allows us to set a universal, dimension independent threshold a_n, yielding smaller and more practically useful threshold values.

We note that if for a threshold sequence a_n', $T_n < a_n'$ is a consistent test, then so is $f_n(T_n) < f_n(a_n')$ for any sequence of increasing scaling functions f_n, so by replacing the test statistic T_n by $f_n(T_n)$, the threshold $a_n = f_n(a_n')$ can easily be made dimension independent. However, in this case $f_n(T_n)$ would not be a consistent estimate of the excess risk $L^* - L^*(S)$, which would make a test based on $f_n(T_n)$ much less attractive.

Under mild smoothness conditions, the main results in this paper show the consistency of the constructed tests for the lossless feature selection property for classification and nonparametric regression. However, one may in addition be interested in the power of these tests. Without any separation gap between the hypotheses, for any test the rate of convergence of the Type II error (power) can be arbitrarily slow. In future work one may consider a formulation of separation that makes it possible to derive sharper bounds on the power on these tests. For example, with the null hypothesis $\mathcal{H}_0$ as above, one may consider the alternative hypothesis $\mathcal{H}_1$ to be $L^* + \delta < L^*(S)$ for some fixed but unknown $\delta > 0$ and investigate the power of the test in this setup as a function of n and d.

The paper is organized as follows. In Sect. 2 we introduce a novel k-nearest-neighbor-based test statistic and threshold for lossless feature selection in classification and state the consistency of the corresponding test under a mild Lipschitz-type condition (Theorem 1). We also state a lemma (Lemma 1) that gives sharp upper and lower bounds on the first absolute moment of the average of n independent and identically distributed ± 1-valued random variables, which plays an important role in proving the test's consistency under the null hypothesis. In Sect. 3 we introduce a similar test for lossless feature selection for nonparametric regression and state its consistency (Theorem 2). Finally, the proofs are presented in Sect. 4.

2 Lossless feature selection for classification

The task of binary classification is to decide on the ± 1-valued random variable Y given an $\mathbb{R}^d$-valued random vector X by finding a decision function g, defined on the range of X, such that $g(X) = Y$ with large probability. If $g : \mathbb{R}^d \to \{-1, 1\}$ is an arbitrary measurable decision function, then its error probability is denoted by

$$L(g) = \mathbb{P}\{g(X) \neq Y\}.$$

Let

$$m(x) = \mathbb{E}[\, Y \mid X = x].$$

It is well-known that the so-called Bayes decision g^*, the decision function that minimizes the error probability $L(g)$, is given by

$$g^*(x) = \operatorname{sgn} m(x),$$

where $\operatorname{sgn}(z) = 1$ if $z \geq 0$ and $\operatorname{sgn}(z) = -1$ if $z < 0$ for any $z \in \mathbb{R}$.

The minimum error probability, also called the Bayes error probability, is given by

$$L^* = \mathbb{P}\{g^*(X) \neq Y\} = \min_g L(g),$$

which can be considered as the minimum risk in classification.

Constructing the Bayes decision requires the knowledge of the distribution of (X, Y). Typically, this distribution is unknown and instead one observes the training samples

$$\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\},$$

consisting of independent and identically distributed (i.i.d.) copies of (X, Y), which arrive in a stream with increasing sample size n. The monograph by Devroye et al. (1996) provides constructions of classification algorithms, based on the data $\mathcal{D}_n$, that are universally consistent in the sense that the error probability of these algorithms tends to the Bayes error probability for all distributions of (X, Y) as $n \to \infty$. However, the rate of convergence of the error probabilities heavily depends on regularity (smoothness) properties of the function m and on the dimension d of the feature vector X. Detecting a subset of ineffective features that, in the presence of the other features, has no influence on L^* allows lossless reduction of the dimension.

This section deals with testing the hypothesis of ineffectiveness of specific features. The test uses an estimate of the difference of the Bayes error probabilities with and without these features.

As before, for $S \subset \{1, \ldots, d\}$ let $L^*(S)$ denote the Bayes error probability when Y is estimated using the subvector $X_S = (X_i, i \in S)$. Our aim is to construct a test for the hypothesis $L^* = L^*(S)$, i.e., the hypothesis that the subset of neglected features $(X^{(i)}, i \notin S)$ do not provide more information beyond what is contained in X_S. A set S with this property provides lossless feature selection. Note, however, that the neglected features $(X^{(i)}, i \notin S)$ may still be informative, e.g., it is possible that $L^* = L^*(S) = L^*(S^c)$.

Most dimension reduction algorithms for classification are modified versions of principal component analysis (PCA), where one looks for a linear transformation of the feature vector into a lower dimensional subspace. In the resulting optimization problem the Bayes error probability is replaced by a smooth error proxy, see, e.g., Siblini et al. (2021) and Tang et al. (2014).

A particular way of dimension reduction is feature selection. Similarly to the PCA, in feature selection algorithms, instead of the Bayes error probability, one usually considers other, more treatable criteria, see Guyon and Elisseeff (2003). Usually, such feature selection algorithms are looking for individual relevant features. Another, more direct goal is to find a good feature set S that has small size $|S|$ and error probability $L^*(S)$ that is close to L^*. If one fixes $|S| = d'$ for some integer $1 \le d' < d$, the problem is to find a d'-element selection S with minimal Bayes error $L^*(S)$. In this respect, the examples of Cover and van Campenhout (1977) and Toussaint (1971) show that the d' features that are individually the best do not necessarily constitute the best d'-dimensional vector, and therefore, every algorithm has to search exhaustively through all d'-element subsets S, see Chapter 32 in Devroye et al. (1996), This procedure is prohibitively complex and instead one may search for the smallest feature set S that provides lossless feature selection, a task that may be easier. In this paper we are concentrating on the problem of testing whether or not a given S provides lossless feature selection, but we do not deal with the problem of searching for such a smallest feature set S. Finding effective algorithms for identifying such S may be the subject of future work.

When searching for such an S based on the training samples $\mathcal{D}_n$, the classification null-hypothesis $\mathcal{H}_0$, defined as

$$L^*(S) = L^* \tag{2}$$

must be tested. The null-hypothesis (2) means that the subset of neglected features $(X^{(i)}, i \notin S)$ of the vector X carries no additional information, i.e., it has no additional predictive power.

An obvious approach to this problem is to estimate $L^*(S)$ and L^* from the data and to accept the hypothesis (2) if the difference of the estimates is small. Unfortunately, it seems that no such estimates with fast rate of convergence are available in the literature. Antos et al. (1999) proved that without any regularity conditions, the rate of convergence for any estimate of L^* can be arbitrarily slow. In view of this, our approach will be to estimate directly the difference $L^*(S) - L^*$. This task will prove easier in the sense that we can construct an estimate T_n of $L^*(S) - L^*$, for which, without any condition,

$$T_n \to L^*(S) - L^* \tag{3}$$

in L_1, and if $L^*(S) - L^* = 0$, then

$$\mathbb{E}[\,|T_n|\,] \to 0$$

with a nontrivial rate of convergence. More specifically, we introduce an estimate T_n and a threshold $a_n \to 0$ such that (3) holds and if $L^*(S) - L^* = 0$, then

$$\lim_{n\to\infty} \mathbb{P}\{T_n > a_n\} = 0.$$

In order to detect ineffective features, Györfi and Walk (2017) proposed nearest neighbor and partitioning-based statistics and, for the threshold $a_n \approx 1/\sqrt{n} + n^{-2/(2+d)}$, proved the consistency of the resulting tests under fairly restrictive conditions on the distribution of the pair (X, Y). For the case of large d, which is the focus in feature selection, the corresponding a_n is too large to be useful in practice. In this section we weaken the restrictive conditions and significantly decrease the threshold value.

Let $S \subset \{1, \ldots, d\}$ be fixed and introduce the notation

$$\widehat{X} = X_S,$$
$$\widehat{m}(\widehat{X}) = \mathbb{E}[\, Y \mid \widehat{X}] = \mathbb{E}[\, m(X) \mid \widehat{X}]$$

and

$$\widehat{L}^* = L^*(S) = \mathbb{P}\{\operatorname{sgn} \widehat{m}(\widehat{X}) \neq Y\}.$$

For any measurable $g : \mathbb{R}^d \to \{-1, 1\}$ we have

$$L(g) - L^* = \mathbb{E}\left[\mathbb{I}_{\{g(X) \neq g^*(X)\}} |m(X)|\,\right], \tag{4}$$

where $\mathbb{I}$ denotes the indicator function (see Devroye et al. 1996, Theorem 2.2). Letting $g = -g^*$ in (4), we obtain $(1 - L^*) - L^* = \mathbb{E}[\,|m(X)|\,]$. Therefore

$$L^* = \frac{1}{2}\big(1 - \mathbb{E}\big[|m(X)|\big]\big),$$

and similarly

$$\widehat{L}^* = \frac{1}{2}\big(1 - \mathbb{E}\big[|\widehat{m}(\widehat{X})|\big]\big),$$

implying that

$$\widehat{L}^* - L^* = \frac{1}{2}\left(\mathbb{E}\big[|m(X)|\big] - \mathbb{E}\big[|\widehat{m}(\widehat{X})|\big]\right), \tag{5}$$

Therefore the classification null hypothesis (2) is equivalent to

$$\mathbb{E}\big[|m(X)|\big] - \mathbb{E}\big[|\widehat{m}(\widehat{X})|\big] = 0. \tag{6}$$

To test the null hypothesis, we propose a nearest-neighbor-based test statistic. Fix $x \in \mathbb{R}^d$ and reorder the data $(X_1, Y_1), \ldots, (X_n, Y_n)$ according to increasing values of $\|X_i - x\|$, where $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^d$. The reordered data sequence is denoted by

$$(X_{(n,1)}(x), Y_{(n,1)}(x)), \ldots, (X_{(n,n)}(x), Y_{(n,n)}(x)), \tag{7}$$

so that $X_{(n,k)}(x)$ is the kth nearest-neighbor of x. In case of a tie, i.e., if X_i and X_j are equidistant from x, then X_i is declared "closer" if $i < j$. In this paper we assume that ties occur with probability 0. This assumption can be enforced by endowing X with a $(d+1)$th component Z that is independent of (X, Y) and is uniformly distributed on $[0, 1]$, see Section 11.2 in Devroye et al. (1996). In this situation, the training samples are similarly modified. We note that this procedure does not change either L^* or $\widehat{L}^*$ and m and $\widehat{m}$ remain unchanged as well.

Choose an integer k_n less than n. The k-nearest-neighbor (k-NN) regression estimate of m is

$$m_n(x) = \frac{1}{k_n} \sum_{j=1}^{k_n} Y_{(n,j)}(x). \tag{8}$$

Analogously to (8), from the training subsamples

$$\widehat{\mathcal{D}}_n = \{(\widehat{X}_1, Y_1), \ldots, (\widehat{X}_n, Y_n)\}$$

(where $\widehat{X}_i$ is the subvector picked out from X_i by the selected features S), we introduce the k-NN estimate $\widehat{m}_n$ of $\widehat{m}$:

$$\widehat{m}_n(\widehat{x}) = \frac{1}{k_n} \sum_{j=1}^{k_n} \widehat{Y}_{(n,j)}(\widehat{x}). \tag{9}$$

The k-nearest-neighbor classification rule is defined by the plug-in estimator

$$g_n(x) = \mathrm{sgn}(m_n(x)).$$

Assuming ties occur with probability 0, if k_n is chosen so that $k_n \to \infty$ and $k_n/n \to 0$, then without any other condition on the distribution of (X, Y), the k-NN classification rule is strongly consistent, i.e.,

$$L(g_n) \to L^* \quad \text{a.s.},$$

where $L(g_n) = \mathbb{P}\{g_n(X) \neq Y \mid \mathcal{D}_n\}$, see Sections 11.1 and 11.2 in Devroye et al. (1996).

To construct the test we assume for convenience that $2n$ i.i.d. training samples are available, so that in addition to $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, the i.i.d. copies

$$\mathcal{D}'_n = \{(X'_1, Y'_1), \ldots, (X'_n, Y'_n)\}$$

of (X, Y) are also available ((X, Y), $\mathcal{D}_n$, and $\mathcal{D}'_n$ are independent). Recalling (5), we estimate the difference

$$\mathbb{E}\Big[|m(X)| - |\widehat{m}(\widehat{X})|\Big]$$

by means of the nearest-neighbor-based test statistic

$$T_n = \frac{1}{n} \sum_{i=1}^{n} \left(Y'_i \,\mathrm{sgn}(m_n(X'_i)) - |\widehat{m}_n(\widehat{X}'_i)|\right)$$

with k_n chosen as

$$k_n = \lfloor \sqrt{\log n} \rfloor.$$

One accepts the classification null hypothesis if

$$T_n \le a_n,$$

where

$$a_n = \frac{\log n}{\sqrt{n}}. \tag{10}$$

To prove the consistency of this test, we need the following modified Lipschitz condition. It is a combined smoothness and tail condition that weakens two rather restrictive conditions that are used in the literature: the Lipschitz continuity of m and the condition that X is bounded.

Definition 1 *(* Chaudhuri and Dasgupta (2014), Döring et al. (2018)*)* If μ stands for the distribution of X, then m satisfies the modified Lipschitz condition if there exists $C^* > 0$ such that for any $x, z \in \mathbb{R}^d$,

$$|m(x) - m(z)| \le C^* \mu(S_{x,\|x-z\|})^{1/d},$$

where $S_{x,r} = \{y \in \mathbb{R}^d : \|y - x\| \le r\}$ denotes the closed Euclidean ball centered at x having radius r.

The following theorem, which is one of the main results in this paper, states the consistency of our test.

Theorem 1 *Let $d \ge 2$ and assume that ties occur with probability* 0*. Then*

(a) Under the classification alternative hypothesis, one has

$$\lim_{n\to\infty} \mathbb{P}\{T_n \le a_n\} = 0.$$

(b) If $\widehat{m}$ satisfies the modified Lipschitz condition and the residual variance satisfies $\mathbb{E}\big[(Y - \widehat{m}(\widehat{X}))^2\big] = 1 - \mathbb{E}[\widehat{m}(\widehat{X})^2] > 0$, then under the classification null hypothesis, one has

$$\lim_{n\to\infty} \mathbb{P}\{T_n > a_n\} = 0.$$

Remarks:

(i) The proof of the theorem also works if instead of the modified Lipschitz condition one assumes the standard Lipschitz condition and the boundedness of X.

(ii) In the proof we make use of the Efron-Stein inequality (Boucheron et al. 2013, Theorem 3.1). At the price of a larger threshold $(\log n)^2/\sqrt{n}$, the Efron-Stein inequality can be replaced by an exponential concentration inequality, such as McDiarmid's inequality (Boucheron et al. 2013, Theorem 6.2), to obtain strong consistency. (A test is called strongly consistent if, almost surely, with increasing sample size it makes finitely many errors.)

In the analysis we show that, under the null hypothesis, the expectation of the test statistic is negative, and in the proof we need to lower bound the expected L_1 norm of the corresponding regression estimate. In contrast to the L_2 setup, where the difference of the second moment and the squared expectation is equal to the variance, here the difference of the first absolute moment and the absolute value of the expectation is much less than the standard deviation if the expectation is nonzero. In this respect the lower bound in the following lemma plays a crucial role. The proof of the lemma is given in Sect. 4.2.

Lemma 1 *Let $Z_1, \dots, Z_n$ be ± 1-valued i.i.d. random variables with mean a and variance $\sigma^2 = 1 - a^2$. Then,*

$$|a| + \frac{\sqrt{2}}{n^{3/2}}\sigma^n \le \mathbb{E}\left|\frac{1}{n}\sum_{i=1}^{n} Z_i\right| \le |a| + \frac{c^*}{n^{3/2}}\sigma^n \tag{11}$$

with $c^* = c^*(a) < \infty$ if $a \neq 0$.

We note that for $a = 0$ the upper bound in (11) is not valid, but the lower bound is. In fact, for $a = 0$ the Berry-Esseen theorem implies the asymptotics

$$\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^{n} Z_i\right| = \sqrt{\frac{2}{\pi n}} + O\left(\frac{1}{n}\right).$$

Lemma 1 can be considered as a binomial distribution analogue of the second part of Lemma 5.8 in Devroye and Györfi (1985), which deals with the normal distribution. However, an application of that result, together with the normal approximation to the binomial distribution using the Berry-Esseen theorem according to the first half of this Lemma 5.8, does not give the bound of Lemma 1 because the normal approximation is too rough for our setting.

3 Lossless feature selection for nonparametric regression

An analogous problem can be posed in the context of nonparametric regression, where Y is a real-valued random variable with $\mathbb{E}[Y^2] < \infty$. Here the functional

$$L^* = \mathbb{E}\left[(Y - m(X))^2\right].$$

with $m(x) = \mathbb{E}[Y \mid X = x]$, plays the central role. It is known that for any measurable function $g : \mathbb{R}^d \to \mathbb{R}$,

$$\mathbb{E}\left[(Y - g(X))^2\right] = L^* + \mathbb{E}\left[(m(X) - g(X))^2\right]$$

and therefore

$$L^* = \min_g \mathbb{E}\left[(Y - g(X))^2\right],$$

where the minimum is taken over all measurable functions $g : \mathbb{R}^d \to \mathbb{R}$. The functional L^* is often referred to as the residual variance; it is the minimum mean squared error in predicting Y based on the observation X.

For $S \subset \{1, \ldots, d\}$, the predictive power of a subvector $X_S = (X^{(i)}, i \in S)$ of X is measured by the residual variance

$$L^*(S) = \mathbb{E}\left[(Y - \mathbb{E}[Y \mid X_S])^2\right]$$

that can be achieved using the features as explanatory variables. For possible dimensionality reduction, one needs, in general, to test the regression null hypothesis $\mathcal{H}_0$ that the two residual variances are equal:

$$L^* = L^*(S). \tag{12}$$

Again, a set S with this property provides lossless feature selection.

Identifying a set of features S with the property (12) is equivalent to finding a set of features $\bar{S} = S^c$ that are irrelevant in inference. Lei and Wasserman (2014) introduced the Leave-One-Covariate-Out (LOCO) value

$$\mathrm{LOCO}(S^c) = L^*(S) - L^*,$$

for characterizing the importance of a feature or a subset of features, see also Gan et al. (2023), Verdinelli and Wasserman (2023), and Williamson et al. (2021). These papers mostly deal with finding a single element set S^c that minimizes $\mathrm{LOCO}(S^c)$. Testing the null hypothesis (12) is equivalent to testing $\mathrm{LOCO}(S^c) = 0$ a task of which Williamson et al. (2021) says "Developing valid inference under this particular null hypothesis appears very difficult."

As in De Brabanter et al. (2014), a natural way of approaching this testing problem is by estimating both residual variances L^* and $L^*(S)$, and accept the regression null hypothesis if the two estimates are close to each other.

For fixed $S \subset \{1, \ldots, d\}$, let $\widehat{X} = X_S$ and $\widehat{m}(\widehat{x}) = \mathbb{E}[Y \mid \widehat{X} = \widehat{x}]$. Due to the identities

$$\begin{aligned} L^* - L^*(S) &= \left(E[Y^2] - E\left[m(X)^2\right]\right) - \left(E[Y^2] - E\left[\widehat{m}(\widehat{X})^2\right]\right) \\ &= \mathbb{E}\left[(m(X) - \widehat{m}(\widehat{X}))^2\right], \end{aligned}$$

the regression null-hypothesis $\mathcal{H}_0$ defined by (12) is equivalent to both

$$\mathbb{E}\big[m(X)^2\big] = \mathbb{E}\big[\widehat{m}(\widehat{X})^2\big]. \tag{13}$$

and

$$\mathbb{P}\big\{m(X) = \widehat{m}(\widehat{X})\big\} = 1. \tag{14}$$

As an estimate of $\mathbb{E}\big[m(X)^2\big] - \mathbb{E}\big[\widehat{m}(\widehat{X})^2\big]$, (Devroye et al. 2018, Section 3) introduced the following 1-NN based test statistic

$$\tilde{T}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i'(Y_{(n,1)}(X_i') - \widehat{Y}_{(n,1)}(\widehat{X}_i')). \tag{15}$$

and accepted the null-hypothesis $\mathcal{H}_0$ if

$$\tilde{T}_n \le \tilde{a}_n = \log n\left(n^{-1/2} + n^{-1/d}\right).$$

For bounded Y and X, X with a density, and m satisfying the ordinary Lipschitz condition, Devroye et al. (2018) showed that this test is strongly consistent. For large d, the threshold above is too large to be of practical use.

We slightly modify the test statistic $\tilde{T}_n$ in a way that results in a negative bias. Define a nearest-neighbor-based test statistic

$$T_n = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i' m_n(X_i') - \widehat{m}_n(\widehat{X}_i')^2\right),$$

where m_n and $\widehat{m}_n$ are the k-NN regression estimators defined in (8) and (9), respectively. Again, here we set

$$k = k_n = \lfloor \log n \rfloor,$$

and with $a_n = \log n/\sqrt{n}$ as in (10), we accept the regression null hypothesis if

$$T_n \le a_n.$$

The following theorem is the main result of this section:

Theorem 2 *Let $d \ge 2$ and assume that ties occur with probability* 0.

(a) Under the regression alternative hypothesis

$$\lim_{n\to\infty} \mathbb{P}\{T_n \le a_n\} = 0.$$

(b) *If $\widehat{m}$ satisfies the modified Lipschitz condition, $\mathbb{E}\left[Y^4 \mid X\right] \le C$ a.s. for some finite $C > 0$, and the residual variance satisfies $\mathbb{E}\left[(Y - \widehat{m}(\widehat{X}))^2\right] > 0$, then under the regression null hypothesis*

$$\lim_{n\to\infty} \mathbb{P}\{T_n > a_n\} = 0.$$

Similarly to Theorem 1, if in the proof of this theorem we assume that Y is bounded and the Efron-Stein inequality is replaced by an exponential concentration inequality, then with the larger threshold $(\log n)^2/\sqrt{n}$ we obtain strong consistency.

4 Proofs

4.1 Some moment inequalities

A cone with angle θ centered at the origin is the collection of all points $y \in \mathbb{R}^d$ such that angle$(y, z) \le \theta$ for some given $z \in \mathbb{R}^d$. The following inequalities will be needed in the proofs of Theorems 1 and 2.

Lemma 2 *Let γ_d be the minimum number of cones centered at the origin and having angle $\pi/6$ whose union covers $\mathbb{R}^d$. Set*

$$m_n^*(x) = \frac{1}{k_n}\sum_{i=2}^{n} Y_i \mathbb{I}_{\{X_i \text{ is among the } k_n \text{ NNs of } x \text{ in } \{X_2,\dots,X_n\}\}}.$$

(a) *Under the assumptions of Theorem 1, one has*

$$\mathbb{E}\left[\left(\int |m_n(x)|\,\mu(dx) - \int |m_n^*(x)|\,\mu(dx)\right)^2\right] \le \frac{16\gamma_d^2}{n^2}.$$

(b) *Under the assumptions of Theorem 2, one has*

$$\mathbb{E}\left[\left(\int m(x)m_n(x)\,\mu(dx) - \int m(x)m_n^*(x)\,\mu(dx)\right)^2\right] \le \frac{16C\gamma_d^2}{n^2}$$

and

$$\mathbb{E}\left[\left(\int m_n(x)^2\mu(dx) - \int m_n^*(x)^2\mu(dx)\right)^2\right] \le \frac{64C\gamma_d^2}{n^2}.$$

Proof We only prove the second half of Lemma 2(b); the proofs of the other two inequalities are similar, but easier. We have

$$\mathbb{E}\left[\left(\int m_n(x)^2\mu(dx) - \int m_n^*(x)^2\mu(dx)\right)^2\right]$$

$$\leq \mathbb{E}\left[\left(\int |m_n(x) + m_n^*(x)| \cdot |m_n(x) - m_n^*(x)|\mu(dx)\right)^2\right]$$
$$= \mathbb{E}\Big[\big|m_n(X_{n+1}) + m_n^*(X_{n+1})\big| \cdot \big|m_n(X_{n+1}) - m_n^*(X_{n+1})\big|$$
$$\cdot \big|m_n(X_{n+2}) + m_n^*(X_{n+2})\big| \cdot \big|m_n(X_{n+2}) - m_n^*(X_{n+2})\big|\Big],$$

where X_{n+1} and X_{n+2} are independent of the training samples $\mathcal{D}_n$ and have the common distribution of the X_i. Therefore, using the notation in (7),

$$\mathbb{E}\left[\left(\int m_n(x)^2\mu(dx) - \int m_n^*(x)^2\mu(dx)\right)^2\right]$$
$$\leq \mathbb{E}\Big[\mathbb{E}\Big[(|m_n(X_{n+1})| + |m_n^*(X_{n+1})|)\frac{|Y_1| + |Y_{(n,k_n+1)}(X_{n+1})|}{k_n}$$
$$\cdot (|m_n(X_{n+2})| + |m_n^*(X_{n+2})|)\frac{|Y_1| + |Y_{(n,k_n+1)}(X_{n+2})|}{k_n}$$
$$\cdot \mathbb{I}_{\{X_1 \text{is among the} k_n \text{NNs of } X_{n+1} \text{ and} X_{n+2} \text{ in} \{X_1,\dots,X_n\}\}}$$
$$\Big| X_1, \dots, X_n, Y_1, \dots, Y_n\Big]\Big]$$
$$\leq \frac{16Ck_n^2}{k_n^4}$$
$$\cdot \mathbb{E}\Big[\mathbb{E}\Big[\mathbb{I}_{\{X_1 \text{ is among the } k_n \text{ NNs of} X_{n+1} \text{ and } X_{n+2} \text{ in } \{X_1,\dots,X_n\}\}}$$
$$\Big|X_1, \dots, X_{n+2}\Big]\Big] \leq \frac{16C}{k_n^2}$$
$$\cdot \mathbb{P}\{X_1 \text{ is among the } k_n + 2 \text{ NNs of } X_{n+1} \text{ and } X_{n+2} \text{ in } \{X_1, \dots, X_{n+2}\}\}.$$

Thus,

$$\mathbb{E}\left[\left(\int m_n(x)^2\mu(dx) - \int m_n^*(x)^2\mu(dx)\right)^2\right]$$
$$\leq \frac{16C}{k_n^2(n+1)n}\mathbb{E}\Big[\Big(\sum_{j=2}^{n+2} \mathbb{I}_{\{X_1 \text{ is among the } k_n+2 \text{ NN's of } X_j \text{ in } \{X_1,\dots,X_{n+2}\}\}}\Big)^2\Big]$$
$$= \frac{16C}{k_n^2(n+1)n}\mathbb{E}\Big[\Big(\sum_{j=2}^{n+2}$$
$$\mathbb{I}_{\{X_1 \text{ is among the } k_n+1 \text{ NN's of } X_j \text{ in } \{X_1,\dots,X_{j-1},X_{j+1},\dots,X_{n+2}\}\}}\Big)^2\Big]$$
$$\leq \frac{16C}{k_n^2 n^2}((k_n+1)\gamma_d)^2,$$

where for the last step we refer to in Corollary 6.1, Györfi et al. (2002). □

Lemma 3 (Extended Efron-Stein inequality, Boucheron et al. (2004), Theorem 6) *Let A be a measurable set and $Z = (Z_1, \dots, Z_n)$ be an i.i.d. n-tuple of A-valued random variables. Set $Z^{(1)} = (Z_2, \dots, Z_n)$. Let f and g be real-valued measurable functions on A^n and A^{n-1}, respectively, such that $f(Z)$ is integrable. Then,*

$$\mathbb{V}ar(f(Z)) \leq n\mathbb{E}\big[\,(f(Z) - g(Z^{(1)}))^2\,\big].$$

4.2 Proof of Theorem 1

Proof of Theorem 1(a) Assume the classification alternative hypothesis and define c^* as

$$c^* = \mathbb{E}\left[(|m(X)| - |\widehat{m}(\widehat{X})|)\right] > 0. \tag{16}$$

For n sufficiently large,

$$\mathbb{P}\{T_n \leq a_n\} \leq \mathbb{P}\{T_n \leq c^*/2\} \leq \mathbb{P}\{|T_n - c^*| \geq c^*/2\} \leq \frac{2\mathbb{E}[\,|T_n - c^*|\,]}{c^*}.$$

Therefore, it suffices to show that

$$\lim_{n\to\infty} \mathbb{E}[\,|T_n - c^*|\,] = 0. \tag{17}$$

We use the decomposition

$$T_n = (T_n - \mathbb{E}[\,T_n \mid \mathcal{D}_n\,]) + \mathbb{E}[\,T_n \mid \mathcal{D}_n\,]. \tag{18}$$

One obtains

$$\begin{aligned}\mathbb{E}\big[(T_n - \mathbb{E}[\,T_n \mid \mathcal{D}_n\,])^2 \mid \mathcal{D}_n\big] &\leq \frac{2\mathbb{E}[Y^2]}{n} + \frac{2\mathbb{E}[m_n(X_1')^2 \mid \mathcal{D}_n]}{n} \\ &\leq \frac{4}{n} \quad \text{a.s.,}\end{aligned} \tag{19}$$

and thus

$$\mathbb{E}\big[\,(T_n - \mathbb{E}[\,T_n \mid \mathcal{D}_n\,])^2\big] \to 0. \tag{20}$$

Furthermore, one has

$$\begin{aligned}\mathbb{E}[\,T_n \mid \mathcal{D}_n\,] &= \mathbb{E}\big[\,Y_1' \operatorname{sgn}(m_n(X_1')) \mid \mathcal{D}_n\,\big] - \mathbb{E}\big[\,|\widehat{m}_n(\widehat{X}_1')| \,\big|\, \mathcal{D}_n\,\big] \\ &= \int m(x) \operatorname{sgn}(m_n(x))\, \mu(dx) - \int |\widehat{m}_n(\widehat{x})|\, \widehat{\mu}(d\widehat{x}).\end{aligned} \tag{21}$$

We notice

$$|m(x)\,\mathrm{sgn}(m_n(x)) - |m(x)|\,| \le 2|m_n(x) - m(x)|$$

and obtain

$$\begin{aligned}
&\mathbb{E}\left[\,\left|\int m(x)\,\mathrm{sgn}(m_n(x))\,\mu(dx) - \int |m(x)|\,\mu(dx)\right|\,\right]\\
&\quad\le 2\mathbb{E}\left[\int |m_n(x) - m(x)|\,\mu(dx)\right]\\
&\quad\to 0, \qquad (22)
\end{aligned}$$

where the last step follows from (Györfi et al. 2002, Theorem 6.1). This theorem also yields

$$\begin{aligned}
&\mathbb{E}\left[\,\left|\int |\widehat{m}_n(\widehat{x})|\,\widehat{\mu}(d\widehat{x}) - \int |\widehat{m}(\widehat{x})|\,\widehat{\mu}(d\widehat{x})\right|\,\right]\\
&\quad\le \mathbb{E}\left[\,\int |\widehat{m}_n(\widehat{x}) - \widehat{m}(\widehat{x})|\,\widehat{\mu}(d\widehat{x})\,\right] \to 0. \qquad (23)
\end{aligned}$$

From (21), (22), (23), and the definition of c^* in (16) we get

$$\mathbb{E}\left[\,\left|\mathbb{E}[\,T_n \mid \mathcal{D}_n\,] - c^*\right|\,\right] \to 0. \qquad (24)$$

Now (18), (20), and (24) yield (17) and part (a) of Theorem 1 is proved. □

Proof of Lemma 1 With the notation $p = (a+1)/2$, one has

$$\sum_{i=1}^{n} Z_i = 2B(n,p) - n,$$

where $B(n,p)$ is a binomial random variable with parameters n and $p \in [0,1]$. Without loss of generality we assume that $\sigma^2 = 1 - a^2 = 4p(1-p) > 0$; otherwise (11) holds with equality. Therefore, $p \in (0,1)$. Due to the chain of equalities

$$\begin{aligned}
\mathbb{E}\big|2B(n,p) - n\big| &= \mathbb{E}\big|2(n - B(n,1-p)) - n\big|\\
&= \mathbb{E}\big| - 2B(n,1-p)) + n\big|\\
&= \mathbb{E}\big|2B(n,1-p)) - n\big|,
\end{aligned}$$

it suffices to consider the case of $p \le 1/2$, i.e., $a \le 0$. We have

$$\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^{n} Z_i\right| = \mathbb{E}\left|\frac{2B(n,p) - 2np}{n} + a\right|$$

$$= \mathbb{E}\left(\frac{2B(n,p)-2np}{n}+a\right)^+ + \mathbb{E}\left(\frac{-2B(n,p)+2np}{n}-a\right)^+$$
$$= 2\mathbb{E}\left(\frac{2B(n,p)-2np}{n}+a\right)^+ + \mathbb{E}\left[\frac{-2B(n,p)+2np}{n}-a\right]$$
$$= 2\mathbb{E}\left[\frac{2B(n,p)-n)^+}{n}\right] + |a| \tag{25}$$
$$\geq 2\mathbb{P}\{B(n,p)/n > 1/2\}/n + |a|. \tag{26}$$

Stirling's formula implies that

$$\mathbb{P}\{B(n,p)/n > 1/2\} \geq \frac{1}{\sqrt{2n}} e^{-nD(1/2\|p)},$$

where

$$D(\epsilon\|p) = \epsilon \ln \frac{\epsilon}{p} + (1-\epsilon)\ln\frac{1-\epsilon}{1-p},$$

see (4.7.2) on p. 115 in Ash (1990). Since $D\,(1/2\|p) = -\ln\sigma$, we obtain

$$\mathbb{P}\{B(n,p)/n > 1/2\} \geq \frac{1}{\sqrt{2n}}\sigma^n,$$

which, in view of (26), proves the lower bound in the lemma.

For the sake of simplicity, in the proof of the upper bound assume that n is even. As in the proof of the upper bound assuming that $0 < p < 1/2$, we have

$$\begin{aligned}
\mathbb{E}\big[(B(n,p)-n/2)^+\big] &= \sum_{n/2<j\leq n} (j-n/2)\binom{n}{j} p^j (1-p)^{n-j} \\
&= \sum_{0<j\leq n/2} j\binom{n}{n/2+j} p^{n/2+j}(1-p)^{n/2-j} \\
&= \sum_{0<j\leq n/2} j\binom{n}{n/2+j} 2^{-n}\left(\frac{p}{1-p}\right)^j 2^n\sqrt{p(1-p)}^n \\
&= \sum_{0<j\leq n/2} j\binom{n}{n/2+j} 2^{-n}\left(\frac{p}{1-p}\right)^j \sigma^n.
\end{aligned}$$

Therefore, in view of (25), we have to prove that

$$\sum_{0<j\leq n/2} j\binom{n}{n/2+j} 2^{-n}\left(\frac{p}{1-p}\right)^j = O\left(1/\sqrt{n}\right).$$

Stirling's formula implies that

$$\sqrt{\frac{n}{8j(n-j)}} \le \binom{n}{j} 2^{-nh(j/n)} \le \sqrt{\frac{n}{2\pi j(n-j)}} \quad \text{if } \ 1 \le j \le n-1$$

and

$$\binom{n}{j} 2^{-nh(j/n)} \le 1 \quad \text{if } \ 0 \le j \le n,$$

where, for $\epsilon \in (0,1)$,

$$h(\epsilon) = -\epsilon \log_2 \epsilon - (1-\epsilon)\log_2(1-\epsilon)$$

is the binary entropy function, see p. 530 in Gallager (1968). This implies

$$\begin{aligned}
&\sum_{0<j\le n/2} j \binom{n}{n/2+j} 2^{-n} \left(\frac{p}{1-p}\right)^j \\
&\le \sum_{0<j\le n/4} j \sqrt{\frac{n}{2\pi(n^2/4-j^2)}} \left(\frac{p}{1-p}\right)^j \\
&+ \sum_{n/4<j\le n/2} j 2^{-n(1-h((n/2+j)/n))} \left(\frac{p}{1-p}\right)^j \\
&\le \sqrt{\frac{n}{2\pi(n^2/4-n^2/16)}} \sum_{0<j\le n/4} j \left(\frac{p}{1-p}\right)^j + 2^{-n(1-h(3/4))} n^2 \\
&= O\left(1/\sqrt{n}\right) + O\left(2^{-n(1-h(3/4))} n^2\right) \\
&= O\left(1/\sqrt{n}\right),
\end{aligned}$$

since $p < 1/2$. □

Proof of Theorem 1(b) Assume the classification null hypothesis. We use the decomposition of T_n in (18). The upper bound in (19) implies

$$\mathbb{P}\left\{T_n - \mathbb{E}[\, T_n \mid \mathcal{D}_n \,] \ge a_n/2\right\} \le \frac{8}{a_n^2 n}. \tag{27}$$

On the other hand, (21) and the classification null hypothesis (see (6)) yield that

$$\begin{aligned}
\mathbb{E}[\, T_n \mid \mathcal{D}_n \,] &\le \int (|m(x)| - |\widehat{m}_n(\widehat{x})|)\, \mu(dx) \\
&= \int (|\widehat{m}(\widehat{x})| - |\widehat{m}_n(\widehat{x})|)\, \widehat{\mu}(d\widehat{x})
\end{aligned}$$

$$= -\left(\int |\widehat{m}_n(\widehat{x})|\,\widehat{\mu}(d\widehat{x}) - \mathbb{E}\left[\int |\widehat{m}_n(\widehat{x})|\,\widehat{\mu}(d\widehat{x})\right]\right)$$
$$+ \left(\int |\widehat{m}(\widehat{x})|\,\widehat{\mu}(d\widehat{x}) - \mathbb{E}\left[\int |\widehat{m}_n(\widehat{x})|\,\widehat{\mu}(d\widehat{x})\right]\right). \tag{28}$$

Using first Lemma 3 and the corresponding definition of $\widehat{m}_n^*$, and then part (a) of Lemma 2, one obtains

$$\begin{aligned} \mathbb{V}ar\left(\int |\widehat{m}_n(\widehat{x})|\,\widehat{\mu}(d\widehat{x})\right) &\leq n\mathbb{E}\left[\left(\int |\widehat{m}_n(\widehat{x})|\,\widehat{\mu}(d\widehat{x}) - \int |\widehat{m}_n^*(\widehat{x})|\widehat{\mu}(d\widehat{x})\right)^2\right] \\ &\leq \frac{16\gamma_{d'}^2}{n} \\ &= \frac{C_d}{n}, \end{aligned} \tag{29}$$

where $C_d > 0$ is finite. We will prove that

$$\int |\widehat{m}(\widehat{x})|\,\widehat{\mu}(d\widehat{x}) - \mathbb{E}\left[\int |\widehat{m}_n(\widehat{x})|\,\widehat{\mu}(d\widehat{x})\right] < 0 \tag{30}$$

for all n large enough. We use the decomposition

$$\int |\widehat{m}(\widehat{x})|\,\widehat{\mu}(d\widehat{x}) - \mathbb{E}\left[\int |\widehat{m}_n(\widehat{x})|\,\widehat{\mu}(d\widehat{x})\right]$$
$$= \int |\widehat{m}(\widehat{x})|\,\widehat{\mu}(d\widehat{x}) - \int \left|\mathbb{E}[\,\widehat{m}_n(\widehat{x})\,]\right|\,\widehat{\mu}(d\widehat{x}) \tag{31}$$
$$+ \int \left|\mathbb{E}[\,\widehat{m}_n(\widehat{x})\,]\right|\,\widehat{\mu}(d\widehat{x}) - \mathbb{E}\left[\int |\widehat{m}_n(\widehat{x})|\,\widehat{\mu}(d\widehat{x})\right]. \tag{32}$$

For (31), under the modified Lipschitz condition the proof of Theorem 6 in Döring et al. (2018) implies that

$$\begin{aligned} \int |\widehat{m}(\widehat{x})|\,\widehat{\mu}(d\widehat{x}) - \int \left|\mathbb{E}[\,\widehat{m}_n(\widehat{x})\,]\right|\,\widehat{\mu}(d\widehat{x}) &\leq \int \left|\widehat{m}(\widehat{x}) - \mathbb{E}[\,\widehat{m}_n(\widehat{x})\,]\right|\,\widehat{\mu}(d\widehat{x}) \\ &= O\left(\left(\frac{k_n}{n}\right)^{1/d'}\right). \end{aligned} \tag{33}$$

Let $R_{n,k_n}(\widehat{x}) = \|\widehat{x} - \widehat{X}_{(n,k_n)}(\widehat{x})\|$ be the k_n-NN distance and recall that ties occur with probability 0 by assumption. Since $\widehat{m}_n(\widehat{x})$, defined in (9), can be written in the form

$$\widehat{m}_n(\widehat{x}) = \frac{1}{k_n}\sum_{j=1}^{n} Y_j\, \mathbb{I}_{\{\|\widehat{x}-\widehat{X}_j\| \leq R_{n,k_n}(\widehat{x})\}},$$

one can see that given $R_{n,k_n}(\widehat{x})$, the estimate $\widehat{m}_n(\widehat{x})$ is conditionally distributed as the average of k_n i.i.d. ±1-valued random variables. Conditioned on $R_{n,k_n}(\widehat{x})$, these random variables have common variance

$$k_n \cdot \mathbb{V}ar\big(\widehat{m}_n(\widehat{x}) \mid R_{n,k_n}(\widehat{x})\big) = 1 - \mathbb{E}\big[\widehat{m}_n(\widehat{x}) \mid R_{n,k_n}(\widehat{x})\big]^2. \tag{34}$$

Then Jensen's inequality and Lemma 1 imply

$$\begin{aligned}
&\int \mathbb{E}\big[\, |\widehat{m}_n(\widehat{x})| \,\big] \widehat{\mu}(d\widehat{x}) - \int \big|\mathbb{E}[\, \widehat{m}_n(\widehat{x}) \,]\big| \widehat{\mu}(d\widehat{x}) \\
&\geq \int \mathbb{E}\left[\mathbb{E}\big[\, |\widehat{m}_n(\widehat{x})| \,\big|\, R_{n,k_n}(\widehat{x})\big]\right] \widehat{\mu}(d\widehat{x}) \\
&- \int \mathbb{E}\Big[\Big|\mathbb{E}[\, \widehat{m}_n(\widehat{x}) \mid R_{n,k_n}(\widehat{x}) \,]\Big|\Big] \widehat{\mu}(d\widehat{x}) \\
&\geq \frac{\sqrt{2}}{k_n^{3/2}} \int \mathbb{E}\left[\left(1 - \mathbb{E}[\, \widehat{m}_n(\widehat{x}) \mid R_{n,k_n}(\widehat{x}) \,]^2\right)^{k_n/2}\right] \widehat{\mu}(d\widehat{x}).
\end{aligned}$$

Again, apply Jensen's inequality twice:

$$\begin{aligned}
&\int \mathbb{E}\big[\, |\widehat{m}_n(\widehat{x})| \,\big] \widehat{\mu}(d\widehat{x}) - \int \big|\mathbb{E}[\, \widehat{m}_n(\widehat{x}) \,]\big| \widehat{\mu}(d\widehat{x}) \\
&\geq \frac{\sqrt{2}}{k_n^{3/2}} \left(\int \mathbb{E}\Big[1 - \mathbb{E}[\, \widehat{m}_n(\widehat{x}) \mid R_{n,k_n}(\widehat{x}) \,]^2\Big] \widehat{\mu}(d\widehat{x})\right)^{k_n/2} \\
&\geq \frac{\sqrt{2}}{k_n^{3/2}} \left(\int \mathbb{E}\Big[1 - \mathbb{E}[\, \widehat{m}_n(\widehat{x})^2 \mid R_{n,k_n}(\widehat{x}) \,]\Big] \widehat{\mu}(d\widehat{x})\right)^{k_n/2} \\
&= \frac{\sqrt{2}}{k_n^{3/2}} \left(1 - \int \mathbb{E}\big[\, \widehat{m}_n(\widehat{x})^2 \big] \widehat{\mu}(d\widehat{x})\right)^{k_n/2} \\
&\geq \frac{\sqrt{2}}{k_n^{3/2}} \left(1 - \int \widehat{m}(\widehat{x})^2 \widehat{\mu}(d\widehat{x}) + o(1)\right)^{k_n/2},
\end{aligned}$$

where the last step holds because $\int \mathbb{E}\{\widehat{m}_n(\widehat{x})^2\}\widehat{\mu}(d\widehat{x}) \to \int \widehat{m}(\widehat{x})^2\widehat{\mu}(d\widehat{x})$ as $n \to \infty$ by (Györfi et al. 2002, Theorem 6.1). By the condition $1 - \int \widehat{m}(\widehat{x})^2\widehat{\mu}(d\widehat{x}) = 1 - \mathbb{E}[\widehat{m}(\widehat{X})^2] > 0$, we therefore obtain

$$\int \big|\mathbb{E}[\, \widehat{m}_n(\widehat{x}) \,]\big| \widehat{\mu}(d\widehat{x}) - \mathbb{E}\left[\int |\widehat{m}_n(\widehat{x})| \widehat{\mu}(d\widehat{x})\right] \leq -\frac{e^{-c_1 k_n}}{k_n^{3/2}} \tag{35}$$

for some $c_1 > 0$. Since $k_n = \lfloor \log n \rfloor$, (33) and (35) yield

$$\int |\widehat{m}(\widehat{x})| \widehat{\mu}(d\widehat{x}) - \mathbb{E}\left[\int |\widehat{m}_n(\widehat{x})| \widehat{\mu}(d\widehat{x})\right]$$

$$\leq O\left(\left(\frac{\sqrt{\log n}}{n}\right)^{1/d'}\right) - \frac{1}{n^{c_1/\sqrt{\log n}}(\log n)^{3/4}}$$
$$< 0$$

if n is large enough, and so (30) is verified. From (18), (27), (28), (29) and (30), we now get for all n large enough,

$$\begin{aligned}
\mathbb{P}\{T_n > a_n\} \leq \mathbb{P}\{T_n - \mathbb{E}[\,T_n \mid \mathcal{D}_n\,] \geq a_n/2\} \\
&+ \mathbb{P}\left\{-\left(\int |\widehat{m}_n(\widehat{x})|\widehat{\mu}(d\widehat{x}) - \mathbb{E}\left[\int |\widehat{m}_n(\widehat{x})|\widehat{\mu}(d\widehat{x})\right]\right) > a_n/2\right\} \\
&+ \mathbb{I}_{\{\int |\widehat{m}(\widehat{x})|\widehat{\mu}(d\widehat{x}) - \mathbb{E}\{\int |\widehat{m}_n(\widehat{x})|\widehat{\mu}(d\widehat{x})\}>0\}} \\
&\leq \frac{8}{a_n^2 n} + \frac{4\mathbb{V}ar\left(\int |\widehat{m}_n(\widehat{x})|\widehat{\mu}(d\widehat{x})\right)}{a_n^2} \\
&\leq \frac{8}{a_n^2 n} + \frac{4C_d}{na_n^2} \qquad (36)\\
&\to 0,
\end{aligned}$$

which yields part (b) of Theorem 1. □

4.3 Proof of Theorem 2

Proof of Theorem 2(a) Assume the regression alternative hypothesis. We will prove that

$$T_n \to \int m(x)^2 \mu(dx) - \int \widehat{m}(\widehat{x})^2\, \widehat{\mu}(d\widehat{x}) > 0$$

in probability. We have

$$T_n = L_n - \widehat{L}_n,$$

where

$$L_n = \frac{1}{n}\sum_{i=1}^{n} Y_i' m_n(X_i')$$

and

$$\widehat{L}_n = \frac{1}{n}\sum_{i=1}^{n} \widehat{m}_n(\widehat{X}_i')^2.$$

Again, one can show that

$$\begin{aligned}\mathbb{E}\big[\,(T_n - \mathbb{E}[\,T_n \mid \mathcal{D}_n\,])^2\big] &\leq \frac{2\mathbb{E}[\,Y_1'^2 m_n(X_1')^2\,]}{n} + \frac{2\mathbb{E}[\,\widehat{m}_n(\widehat{X}_1')^4\,]}{n} \\ &= \frac{4\mathbb{E}[\,Y^4\,](1+o(1))}{n},\end{aligned} \tag{37}$$

which yields

$$\lim_{n\to\infty}(T_n - \mathbb{E}[\,T_n \mid \mathcal{D}_n\,]) = 0 \tag{38}$$

in probability. One has

$$\mathbb{E}[\,L_n \mid \mathcal{D}_n\,] = \int m(x)m_n(x)\,\mu(dx) \to \int m(x)^2\mu(dx)$$

and

$$\mathbb{E}[\,\widehat{L}_n \mid \mathcal{D}_n\,] = \int \widehat{m}_n(\widehat{x})^2\,\widehat{\mu}(d\widehat{x}) \to \int \widehat{m}(\widehat{x})^2\,\widehat{\mu}(d\widehat{x})$$

in probability, because by (Györfi et al. 2002, Theorem 6.1),

$$\int (m_n(x) - m(x))^2\,\mu(dx) \to 0$$

and

$$\int (\widehat{m}_n(\widehat{x}) - \widehat{m}(\widehat{x}))^2\,\widehat{\mu}(d\widehat{x}) \to 0$$

in probability. Thus we obtain that

$$\lim_{n\to\infty}\mathbb{E}[\,T_n \mid \mathcal{D}_n\,] = 0 \tag{39}$$

in probability. Therefore, by (38) and (39), under the alternative hypothesis

$$T_n \to \int m(x)^2\mu(dx) - \int \widehat{m}(\widehat{x})^2\,\widehat{\mu}(d\widehat{x}) > 0$$

in probability, which proves part (a) of Theorem 2. □

Proof of Theorem 2(a) Assume the regression null hypothesis. We have

$$\begin{aligned}\mathbb{P}\{T_n > a_n\} &\leq \mathbb{P}\big\{T_n - \mathbb{E}[\,T_n \mid \mathcal{D}_n\,] \geq a_n/2\big\} \\ &\quad + \mathbb{P}\big\{\mathbb{E}[\,T_n \mid \mathcal{D}_n\,] - \mathbb{E}[\,T_n\,] \geq a_n/2\big\} \\ &\quad + \mathbb{I}_{\{\mathbb{E}[\,T_n\,]>0\}}\end{aligned} \tag{40}$$

The bound (37) implies that

$$\mathbb{P}\left\{T_n - \mathbb{E}[\, T_n \mid \mathcal{D}_n \,] \geq a_n/2\right\} \leq \frac{8\mathbb{E}[\, Y^4 \,]}{a_n^2 n} \to 0. \tag{41}$$

Similarly to the proof of Theorem 1, using first Lemma 3 and the corresponding definition of $\widehat{m}_n^*$, and then part (b) of Lemma 2, one obtains

$$\begin{aligned}
&\mathbb{V}ar\left(\int m(x)m_n(x)\,\mu(dx)\right)\\
&\leq n\mathbb{E}\left[\left(\int m(x)m_n(x)\,\mu(dx) - \int m(x)m_n^*(x)\,\mu(dx)\right)^2\right]\\
&\leq \frac{16C\gamma_d^2}{n}
\end{aligned} \tag{42}$$

and

$$\begin{aligned}
\mathbb{V}ar\left(\int \widehat{m}_n(\widehat{x})^2\,\widehat{\mu}(d\widehat{x})\right) &\leq n\mathbb{E}\left[\left(\int \widehat{m}_n(\widehat{x})^2\,\widehat{\mu}(d\widehat{x}) - \int \widehat{m}_n^*(\widehat{x})^2\,\widehat{\mu}(d\widehat{x})\right)^2\right]\\
&\leq \frac{64C\gamma_{d'}^2}{n}.
\end{aligned} \tag{43}$$

Thus,

$$\begin{aligned}
&\mathbb{V}ar\,(\mathbb{E}[\, T_n \mid \mathcal{D}_n \,])\\
&= \mathbb{V}ar\left(\int m(x)m_n(x)\mu(dx) - \int \widehat{m}_n(\widehat{x})^2\widehat{\mu}(d\widehat{x})\right)\\
&\leq 2\mathbb{V}ar\left(\int m(x)m_n(x)\mu(dx)\right) + 2\mathbb{V}ar\left(\int \widehat{m}_n(\widehat{x})^2\widehat{\mu}(d\widehat{x})\right)\\
&\leq \frac{c_d}{n}
\end{aligned}$$

with finite $c_d > 0$. Therefore

$$\begin{aligned}
\mathbb{P}\left\{\mathbb{E}[\, T_n \mid \mathcal{D}_n \,] - \mathbb{E}[\, T_n \,] \geq a_n/2\right\} &\leq \frac{4\mathbb{V}ar\,(\mathbb{E}[\, T_n \mid \mathcal{D}_n \,])}{a_n^2}\\
&\leq \frac{4c_d}{na_n^2} \to 0.
\end{aligned} \tag{44}$$

In view of (40), (41), and (44), it remains to prove that under the regression null hypothesis

$$\mathbb{E}[\, T_n \,] < 0 \tag{45}$$

if n is large enough. Under the null hypothesis (see (14)) one can use the decomposition

$$\begin{aligned}\mathbb{E}[\,T_n\,] &= \int m(x)\mathbb{E}[\,m_n(x)\,]\,\mu(dx) - \int \mathbb{E}[\,\widehat{m}_n(\widehat{x})^2]\,\widehat{\mu}(d\widehat{x})\\ &= \int \widehat{m}(\widehat{x})\mathbb{E}[\,m_n(x)\,]\,\mu(dx) - \int \widehat{m}(\widehat{x})^2\,\widehat{\mu}(d\widehat{x})\\ &\quad + \int \widehat{m}(\widehat{x})^2\,\widehat{\mu}(d\widehat{x}) - \int \mathbb{E}[\,\widehat{m}_n(\widehat{x})\,]^2\,\widehat{\mu}(d\widehat{x})\\ &\quad - \int \mathbb{V}ar\,(\widehat{m}_n(\widehat{x}))\,\widehat{\mu}(d\widehat{x}).\end{aligned} \tag{46}$$

Again, under the modified Lipschitz condition the proof of Theorem 6 in Döring et al. (2018) implies that

$$\begin{aligned}&\int \widehat{m}(\widehat{x})\mathbb{E}[\,m_n(x)]\,\mu(dx) - \int \widehat{m}(\widehat{x})^2\,\widehat{\mu}(d\widehat{x})\\ &\le \sqrt{\int \widehat{m}(\widehat{x})^2\,\widehat{\mu}(d\widehat{x})}\sqrt{\int \left(\mathbb{E}[\,m_n(x)] - \widehat{m}(\widehat{x})\right)^2\mu(dx)}\\ &= \sqrt{\int \widehat{m}(\widehat{x})^2\,\widehat{\mu}(d\widehat{x})}\sqrt{\int \left(\mathbb{E}[\,m_n(x)] - m(x)\right)^2\mu(dx)}\\ &= O\left(\left(\frac{k_n}{n}\right)^{1/d}\right)\end{aligned} \tag{47}$$

and

$$\begin{aligned}&\int \widehat{m}(\widehat{x})^2\,\widehat{\mu}(d\widehat{x}) - \int \mathbb{E}[\,\widehat{m}_n(\widehat{x})]^2\,\widehat{\mu}(d\widehat{x})\\ &\le \sqrt{\int \left(|\widehat{m}(\widehat{x})| + |\mathbb{E}[\,\widehat{m}_n(\widehat{x})]|\right)^2\widehat{\mu}(d\widehat{x})}\sqrt{\int \left(\mathbb{E}[\,\widehat{m}_n(\widehat{x})] - \widehat{m}(\widehat{x})\right)^2\widehat{\mu}(d\widehat{x})}\\ &= O\left(\left(\frac{k_n}{n}\right)^{1/d'}\right).\end{aligned} \tag{48}$$

Analogously to (34) and by Jensen's inequality we obtain that

$$\begin{aligned}&k_n\cdot\int \mathbb{V}ar\,(\widehat{m}_n(\widehat{x}))\;\widehat{\mu}(d\widehat{x})\\ &\ge k_n\cdot\int \mathbb{E}\left[\mathbb{V}ar\,\left(\widehat{m}_n(\widehat{x})\mid \widehat{X}_1,\dots,\widehat{X}_n\right)\right]\widehat{\mu}(d\widehat{x})\\ &= \int \mathbb{E}\left[\frac{1}{k_n}\sum_{j=1}^{k_n}\mathbb{V}ar\,\left(Y_{n,j}(\widehat{x})\mid \widehat{X}_1,\dots,\widehat{X}_n\right)\right]\widehat{\mu}(d\widehat{x})\end{aligned}$$

$$= \int \frac{1}{k_n} \sum_{j=1}^{k_n} \mathbb{E}\left[\left(Y_{n,j}(\widehat{x}) - \widehat{m}(\widehat{X}_{n,j}(\widehat{x}))\right)^2\right] \widehat{\mu}(d\widehat{x})$$
$$\to \mathbb{E}\big[(Y - \widehat{m}(\widehat{X}))^2\big],$$

where for the limit relation we refer to Theorem 6.1 in Györfi et al. (2002). Therefore, under the condition $\mathbb{E}\big[(Y - \widehat{m}(\widehat{X}))^2\big] > 0$, one obtains

$$- \int \mathbb{V}ar\ (\widehat{m}_n(\widehat{x}))\, \widehat{\mu}(d\widehat{x}) \leq -\frac{c_2}{k_n} \tag{49}$$

with $c_2 > 0$ for n large enough. Thus, (46), (47), (48) and (49) yield

$$\mathbb{E}\left[T_n\right] \leq O\left(\left(\frac{1}{n}\right)^{1/d}\right) + O\left(\left(\frac{k_n}{n}\right)^{1/d'}\right) - \frac{c_2}{k_n} < 0$$

for n sufficiently large (since $k_n = \lfloor \log n \rfloor$) and so (45) is verified. This completes the proof of part (b) of Theorem 2. □

Declarations

Conflict of interest The authors declare no conflict of interest.

References

Antos A, Devroye L, Györfi L (1999) Lower bounds for Bayes error estimates. IEEE Trans Pattern Anal Mach Intell 21:643–645

Ash RB (1990) Information theory. Dover Publications, New York

Beirlant J, Devroye L, Györfi L, Vajda I (2001) Large deviations of divergence measures on partitions. J Stat Plan Inference 93:1–16

Berrett TB, Samworth RJ, Yuan M (2019) Efficient multivariate entropy estimation via k-nearest-neighbour distances. Ann Stat 47:288–318

Biau G, Györfi L (2005) On the asymptotic properties of a nonparametric L_1-test statistic of homogeneity. IEEE Trans Inf Theory 51:3965–3973

Boucheron S, Lugosi G, Bousquet O (2004) Concentration inequalities. In: Lectures in Advanced Machine Learning. ML 2003, O. Bousquet, U. von Luxburg and G. Rätsch, (eds) 208–240, vol 3176. Lecture Notes in Computer Science. Springer, Berlin

Boucheron S, Lugosi G, Massart P (2013) Concentration inequalities: a nonasymptotic theory of independence. Oxford University Press, Oxford

Chaudhuri K, Dasgupta S (2014) Rates of convergence for nearest neighbor classification. In: Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ (eds) Advances in neural information processing systems, vol 27, pp 3437–3445

Cover T (1973) On determining the irrationality of the mean of a random variable. Ann Stat 1:862–871

Cover T, van Campenhout J (1977) On the possible orderings in the measurement selection problem. IEEE Trans Syst Man Cybern 7:657–661

De Brabanter K, Ferrario PG, Györfi L (2014) Detecting ineffective features for nonparametric regression. In: Suykens JAK, Signoretto M, Argyriou A (eds) Regularization, Optimization, Kernels, and Support Vector Machines. Machine Learning and Pattern Recognition Series, Chapman & Hall/CRC, pp 177–194

Delattre S, Fournier N (2017) On the Kozachenko-Leonenko entropy estimator. J. Stat Plan Inference 185:69–93
Dembo A, Peres Y (1994) A topological criterion for hypothesis testing. Ann Stat 22:106–117
Devroye L, Györfi L (1985) Nonparametric density estimation: the L_1 view. Wiley, New York
Devroye L, Györfi L, Lugosi G (1996) A probabilistic theory of pattern recognition. Springer, New York
Devroye L, Györfi L, Lugosi G, Walk H (2018) A nearest neighbor estimate of the residual variance. Electron J Stat 12:1752–1778
Devroye L, Lugosi G (2002) Almost sure classification of densities. J Nonparametric Stat 14:675–698
Döring M, Györfi L, Walk H (2018) Exact rate of convergence of k-nearest-neighbor classification rule. J Mach Learn Res 18:1–16
Gallager RG (1968) Information theory and reliable communication. Wiley, New York
Gan L, Zheng L, Allen GI (2023) Model-agnostic confidence intervals for feature importance: a fast and powerful approach using minipatch ensembles. arXiv preprint arXiv:2206.02088v2
Gretton A, Györfi L (2010) Consistent nonparametric tests of independence. J Mach Learn Res 11:1391–1423
Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182
Györfi L, Kohler M, Krzyżak A, Walk H (2002) A distribution-free theory of nonparametric regression. Springer, New York
Györfi L, Linder T, Walk H (2023) Lossless transformations and excess risk bounds in statistical inference. Entropy 25(10):1394
Györfi L, Walk H (2017) Detecting ineffective features for pattern recognition. Oberwolfach Preprint, http://publications.mfo.de/handle/mfo/1314
Györfi L, Walk H (2012) Strongly consistent nonparametric tests of conditional independence. Stat Probab Lett 82:1145–1150
Han Y, Jiao J, Weissman T, Wu Y (2020) Optimal rates of entropy estimation over Lipschitz balls. Ann Stat 48:3228–3250
Kozachenko LF, Leonenko NN (1987) Sample estimate of entropy of a random vector. Probl Inf Transm 23:95–101
Kulkarni SR, Zeitouni O (1991) Can one decide the type of the mean from the empirical distribution? Stat Probab Lett 12:323–327
Lei J, Wasserman L (2014) Distribution-free prediction bands for non-parametric regression. J R Stat Soc Ser B Stat Methodol 76:71–96
Morvai G, Weiss B (2021) On universal algorithms for classifying and predicting stationary processes. Probab Surv 18:77–131
Nobel AB (2006) Hypothesis testing for families of ergodic processes. Bernoulli 12:251–269
Siblini W, Kuntz P, Meyer F (2021) A review on dimensionality reduction for multi-label classification. IEEE Trans Knowl Data Eng 33:839–857
Silva J, Narayanan S (2010) Nonproduct data-dependent partitions for mutual information estimation: strong consistency and applications. IEEE Trans Signal Process 58:3497–3511
Tang J, Alelyani S, Liu H (2014) Feature selection for classification: a review. In: Aggarwal CC (ed) Data classification: algorithms and applications. Chapman and Hall/CRC, Boston, pp 37–64
Toussaint G (1971) Note on the optimal selection of independent binary-valued features for pattern recognition. IEEE Trans Inf Theory 17:618
Verdinelli I, Wasserman L (2023) Feature importance: a closer look at Shapley values and LOCO. arXiv:2303.05981
Wang Q, Kulkarni SR, Verdu S (2005) Divergence estimation of continuous distributions based on data-dependent partitions. IEEE Trans Inf Theory 51:3064–3074
Williamson BD, Gilbert PB, Carone M, Simon N (2021) Nonparametric variable importance assessment using machine learning techniques. Biometrics 77:9–22