

# Feature Selection via Mutual Information: New Theoretical Insights

Mario Beraha<sup>\*†</sup>, Alberto Maria Metelli<sup>†</sup>, Matteo Papini<sup>†</sup>, Andrea Tirinzoni<sup>†</sup>, Marcello Restelli<sup>†</sup>

<sup>\*</sup> Università degli Studi di Bologna, Bologna, Italy

<sup>†</sup> DEIB, Politecnico di Milano, Milan, Italy

Email:{mario.beraha, albertomaria.metelli, matteo.papini, andrea.tirinzoni, marcello.restelli}@polimi.it

**Abstract**—Mutual information has been successfully adopted in filter feature-selection methods to assess both the relevancy of a subset of features in predicting the target variable and the redundancy with respect to other variables. However, existing algorithms are mostly heuristic and do not offer any guarantee on the proposed solution. In this paper, we provide novel theoretical results showing that conditional mutual information naturally arises when bounding the ideal regression/classification errors achieved by different subsets of features. Leveraging on these insights, we propose a novel stopping condition for backward and forward greedy methods which ensures that the ideal prediction error using the selected feature subset remains bounded by a user-specified threshold. We provide numerical simulations to support our theoretical claims and compare to common heuristic methods.

**Index Terms**—feature selection, mutual information, regression, classification, supervised learning, machine learning

## I. INTRODUCTION

The abundance of massive datasets composed of thousands of attributes and the widespread use of learning models able of large representational power pose a significant challenge to machine learning algorithms. Feature selection allows to effectively address some of these challenges with a potential benefit in terms of computational costs, generalization capabilities and interpretability. A large variety of approaches has been proposed by the machine learning community [1]. A simple dimension for classifying the feature selection methods is whether they are aware of the underlying learning model. A first group of methods take advantage of this knowledge and try to identify the best subset of features for the specific model class. This group can be further split into *wrapper* and *embedded* methods. Wrappers [2] employ the learning process as a subroutine of the feature selection process, using the validation error of the trained model as a score to decide whether to keep or discard a feature. Clearly, this potentially leads to good generalization capabilities at the cost of iterating the learning process multiple times, which might become impractical for high-dimensional datasets. Embedded methods [3], still assume the knowledge of the model class, but the feature selection and the learning process are carried out together (a remarkable example is [4] in which a generalization bound on the SVM is optimized for both learning the features and the model). Although less demanding than wrappers from a computational standpoint, embedded methods heavily rely on the peculiar properties of the model class. A second group

of methods do not incorporate knowledge of the model class. These approaches are known as *filters*. Filters [5] perform the feature selection using scores that are independent of the underlying learning model. For this reason, they tend not to overfit but they might result less effective than wrappers and embedded methods as they are general across all the possible model classes. From a computational perspective, filters are the most efficient feature selection methods.

Filter methods have been deeply studied in the supervised learning field [6]. A relevant amount of literature focused on using the *mutual information* (MI) as a score for identifying a suitable subset of features [7]. The MI [8] is an index of statistical dependence between random variables. Intuitively, the MI measures how much knowing the value of one variable reduces the uncertainty on the other. Differently from other indexes, like the Pearson correlation coefficient, the MI is able to capture also non-linear dependences and is invariant under invertible and differentiable transformations of the random variables [8]. Thanks to these properties, the MI has been employed extensively as a score for filter methods [9]–[14]. Nonetheless, all these techniques are rather empirical as they try to encode with MI the intuition that “a feature can be discarded if it is useless for predicting the target or it is predictable from the other features”. This notion can be made more formal by introducing the notion of *relevance*, *redundancy* and *complementarity* [7].

To the best of our knowledge, the only work that draws a connection among the several approaches based on the MI is [15]. The authors claim that selecting features using as a score the *conditional mutual information* (CMI) is equivalent to maximizing the conditional likelihood between the target and the features. This observation provides a justification to the well-known iterative backward and forward algorithms in which the features are considered one-by-one for insertion in or removal from the feature set, like in the Markov Blanket approach [16]. Although this work offers a wide perspective on the feature selection methods based on the MI, it does not investigate the relation between the mutual information of a feature set and the prediction error, which, of course, will depend on the specific choice of the model class.

In this paper, we address the problem of controlling the prediction (regression and classification) error when performing the feature selection process via CMI. We claim that selecting features using CMI has the effect on controlling the

ideal error, i.e., the error attained by the Bayes classifier for classification and the minimum MSE (Mean Squared Error) model for regression. We start in Section II by revising some fundamental concepts of information theory. In Section III, we introduce our main theoretical contribution. We derive a pair of inequalities, one for regression (Section III-A) and one for classification (Section III-C), that upper bound the increment of the ideal error obtained by removing a set of features. Such increment is expressed in terms of the CMI between the target and the removed features, given the remaining features. These results support the intuition that a set of features can be safely removed if it does not increase significantly the “information” about the target, assuming we observed the remaining features. Since the result holds for the *ideal* error, we assert that a filter method based on CMI selects the features assuming that the model employed for solving the regression/classification problem has “infinite capacity”. We show that, when considering linear models for regression, the bound does not hold and we propose an adaptation for this specific case (Section III-B). These results can be effectively employed to derive a novel and principled stopping condition for the feature selection process (Section IV). Differently from the typical stopping conditions, such as a fixed number of features or the increment of the score, our approach allows to explicitly control the ideal error introduced in the feature selection process. After contextualizing our work in the feature selection literature (Section V), we evaluate our approach in comparison with several different stopping criteria on both synthetic and real datasets (Section VI).

## II. PRELIMINARIES

We indicate with  $\mathcal{X} \subseteq \mathbb{R}^d$  the feature space and with  $\mathcal{Y}$  the target space. In case of classification  $\mathcal{Y} = \{y_1, y_2, \dots, y_m\}$  is a finite set of classes, whereas in case of regression  $\mathcal{Y} \subseteq \mathbb{R}$  is a subset of the real numbers. We consider a distribution  $p(\mathbf{X}, Y)$  over  $\mathcal{X} \times \mathcal{Y}$  from which a finite dataset  $\mathcal{D} = \{(\mathbf{x}_i, y_i) | i \in \{1, \dots, N\}\}$  of  $N$  i.i.d. instances is drawn, i.e.,  $(\mathbf{x}_i, y_i) \sim p(\mathbf{X}, Y)$  for all  $i$ . For regression problems, we assume there exists  $B \in \mathbb{R}$  such that  $|Y| \leq B$  almost surely. A key term for a regression/classification problem is the conditional distribution  $p(Y|\mathbf{x})$ , which allows to predict the target associated with any given  $\mathbf{x} \in \mathcal{X}$ .

### A. Notation

Given a (random) vector  $\mathbf{X} \in \mathcal{X}$  and a set of indices  $A \subseteq \{1, 2, \dots, d\}$ , we denote by  $\mathbf{X}_A$  the vector of components of  $\mathbf{X}$  whose indices are in  $A$ . Notice that the vectors  $\mathbf{X}_A$  and  $\mathbf{X}_{\bar{A}}$ , for  $\bar{A} = \{1, 2, \dots, d\} \setminus A$ , form a partition of  $\mathbf{X}$ .

For a  $d$ -dimensional random vector  $\mathbf{X}$  we indicate with  $\mathbb{E}_{\mathbf{X}}[\mathbf{X}]$  the  $d$ -dimensional vector of the expectations of each component. Given two random vectors  $\mathbf{X}$  and  $\mathbf{Y}$ , we indicate with  $\text{Cov}_{\mathbf{X}, \mathbf{Y}}[\mathbf{X}, \mathbf{Y}] = \mathbb{E}_{\mathbf{X}, \mathbf{Y}}[(\mathbf{X} - \mathbb{E}_{\mathbf{X}}[\mathbf{X}]) (\mathbf{Y} - \mathbb{E}_{\mathbf{Y}}[\mathbf{Y}])^T]$  the covariance matrix between the two. We indicate with  $\text{Cov}_{\mathbf{X}}[\mathbf{X}] = \text{Cov}_{\mathbf{X}}[\mathbf{X}, \mathbf{X}]$  the covariance matrix of  $\mathbf{X}$ . We denote with  $\text{Var}_{\mathbf{X}}[\mathbf{X}] = \text{tr}(\text{Cov}_{\mathbf{X}}[\mathbf{X}, \mathbf{X}])$  the trace of the covariance matrix of  $\mathbf{X}$ .

Whenever clear by the context we will remove the subscripts from  $\mathbb{E}$ ,  $\text{Var}$  and  $\text{Cov}$ . Given two random (scalar) random variables  $X$  and  $Y$  we denote with  $\rho(X, Y) = \frac{\text{Cov}[X, Y]}{\sqrt{\text{Var}[X] \text{Var}[Y]}}$  the Pearson correlation coefficient between  $X$  and  $Y$ .

### B. Entropy and Mutual Information

We now introduce the basic concepts from information theory that we employ in the remaining of this paper. For simplicity, we provide the definitions for continuous random variables, although all these concepts straightforwardly generalize to discrete variables [8].

The *entropy*  $H(X)$  of a random variable  $X$ , having  $p$  as probability density function, is a common measure of uncertainty:

$$H(X) := \mathbb{E}_X[p(X)] = - \int p(x) \log p(x) dx. \quad (1)$$

Given two distributions  $p$  and  $q$ , we define the Kullback-Leibler (KL) divergence as:

$$D_{\text{KL}}(p||q) := \mathbb{E}_X \left[ \frac{p(X)}{q(X)} \right] = \int p(x) \log \frac{p(x)}{q(x)} dx.$$

The *mutual information* (MI) between two random variables  $X$  and  $Y$  is defined as:

$$\begin{aligned} I(X; Y) &:= H(Y) - H(Y|X) \\ &= \mathbb{E}_X [D_{\text{KL}}(p(Y|X)||p(Y))] \\ &= \int \int p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dx dy, \end{aligned}$$

Intuitively, the MI between  $X$  and  $Y$  represents the reduction in the uncertainty of  $Y$  after observing  $X$  (and viceversa). Notice that the MI is symmetric, i.e.,  $I(X; Y) = I(Y; X)$ . This definition can be straightforwardly extended by conditioning on a third random variable  $Z$ , obtaining the *conditional mutual information* (CMI) between  $X$  and  $Y$  given  $Z$ :

$$\begin{aligned} I(X; Y|Z) &:= \mathbb{E}_Z [I(X|Z; Y|Z)] \\ &= \mathbb{E}_Z [\mathbb{E}_X [D_{\text{KL}}(p(Y|X, Z)||p(Y|Z))]] \\ &= \int p(z) \int \int p(x, y|z) \log \frac{p(x, y|z)}{p(x|z)p(y|z)} dx dy dz. \end{aligned}$$

The CMI fulfills the useful chain rule:

$$I(X; Y, Z) = I(X; Z) + I(X; Y|Z). \quad (2)$$

As we shall see later, the CMI can be used to define a score of both relevancy and redundancy for our feature selection problem, which arises naturally when bounding the ideal regression/classification error. Given a set of indices  $A$ , we denote the CMI between  $Y$  and  $\mathbf{X}_A$  given  $\mathbf{X}_{\bar{A}}$  as:

$$\nu(A) := I(Y; \mathbf{X}_A | \mathbf{X}_{\bar{A}}).$$

This quantity intuitively represents the importance of the feature subset  $\mathbf{X}_A$  in predicting the target  $Y$  given that we are also using  $\mathbf{X}_{\bar{A}}$ .

## III. FEATURE SELECTION VIA MUTUAL INFORMATION

In this section, we introduce our novel theoretical results that shed light on the relationship between CMI and the ideal prediction error. Then, in the next section, we employ these results to propose a new stopping condition that ensures

bounded error. We discuss relationships to existing bounds in Section V.

#### A. Bounding the Regression Error

We start by analyzing an ideal regression problem under the mean square error (MSE) criterion. Consider the subspace  $\mathcal{X}_{\bar{A}}$  of  $\mathcal{X}$  which includes only the features with indices in  $\bar{A}$  and define  $\mathcal{G}_{\bar{A}} = \{g : \mathcal{X}_{\bar{A}} \rightarrow \mathcal{Y}\}$  as the space of all functions mapping  $\mathcal{X}_{\bar{A}}$  to  $\mathcal{Y}$ . The ideal regression problem consists of finding the function  $g^* \in \mathcal{G}_{\bar{A}}$  minimizing the expected MSE,

$$\inf_{g \in \mathcal{G}_{\bar{A}}} \mathbb{E}_{\mathbf{X}, Y} \left[ (Y - g(\mathbf{X}_{\bar{A}}))^2 \right], \quad (3)$$

where the expectation is taken under the full distribution  $p(\mathbf{X}, Y)$ , i.e., under *all* features and the target. The following result relates the ideal error to the expected CMI  $\nu(A)$ .

**Theorem 1.** *Let  $\sigma^2 = \mathbb{E}_{\mathbf{X}, Y} \left[ (Y - \mathbb{E}[Y|\mathbf{X}])^2 \right]$  be the irreducible error and  $A$  be a set of indices, then the regression error obtained by removing features  $\mathbf{X}_A$  can be bounded as:*

$$\inf_{g \in \mathcal{G}_{\bar{A}}} \mathbb{E}_{\mathbf{X}, Y} \left[ (Y - g(\mathbf{X}_{\bar{A}}))^2 \right] \leq \sigma^2 + 2B^2\nu(A). \quad (4)$$

*Proof.* The infimum  $\inf_{g \in \mathcal{G}_{\bar{A}}} \mathbb{E}_{\mathbf{X}, Y} \left[ (Y - g(\mathbf{X}_{\bar{A}}))^2 \right]$  is attained by the minimum MSE regression function  $g(\mathbf{x}_{\bar{A}}) = \mathbb{E}_Y[Y|\mathbf{x}_{\bar{A}}]$ . Therefore, we have

$$\begin{aligned} \inf_{g \in \mathcal{G}_{\bar{A}}} \mathbb{E}_{\mathbf{X}, Y} \left[ (Y - g(\mathbf{X}_{\bar{A}}))^2 \right] &= \mathbb{E}_{\mathbf{X}, Y} \left[ (Y - \mathbb{E}[Y|\mathbf{X}_{\bar{A}}])^2 \right] \\ &= \int p(\mathbf{x}) \int p(y|\mathbf{x}) (y - \mathbb{E}[Y|\mathbf{x}_{\bar{A}}] \pm \mathbb{E}[Y|\mathbf{x}])^2 dy d\mathbf{x} \\ &= \sigma^2 + \int p(\mathbf{x}) (\mathbb{E}[Y|\mathbf{x}] - \mathbb{E}[Y|\mathbf{x}_{\bar{A}}])^2 d\mathbf{x} \\ &= \sigma^2 + \int p(\mathbf{x}) \left( \int y (p(y|\mathbf{x}) - p(y|\mathbf{x}_{\bar{A}})) dy \right)^2 d\mathbf{x} \\ &\leq \sigma^2 + B^2 \int p(\mathbf{x}) \left( \int |p(y|\mathbf{x}) - p(y|\mathbf{x}_{\bar{A}})| dy \right)^2 d\mathbf{x} \\ &\leq \sigma^2 + 2B^2 \int p(\mathbf{x}) D_{\text{KL}}(p(\cdot|\mathbf{x}) \| p(\cdot|\mathbf{x}_{\bar{A}})) d\mathbf{x} \\ &= \sigma^2 + 2B^2\nu(A). \end{aligned}$$

The second inequality follows from Pinsker's inequality [17]–[19] by noting that  $\int |p(y|\mathbf{x}) - p(y|\mathbf{x}_{\bar{A}})| dy = 2D_{\text{TV}}(p(\cdot|\mathbf{x}) \| p(\cdot|\mathbf{x}_{\bar{A}}))$  is twice the total variation distance between  $p(\cdot|\mathbf{x})$  and  $p(\cdot|\mathbf{x}_{\bar{A}})$ .  $\square$

Theorem 1 tells us that the minimum possible MSE that we can achieve by predicting  $Y$  only with the feature subset  $\bar{A}$  can be bounded by the CMI between  $Y$  and  $\mathbf{X}_A$ , conditioned on  $\mathbf{X}_{\bar{A}}$ . This result formalizes the intuitive belief that whenever a subset of features  $A$  has low relevancy or high redundancy (i.e.,  $\nu(A)$  is small), such features can be safely removed without affecting the resulting prediction error too much. In fact, when  $\nu(A) = 0$ , Theorem 1 proves that it is possible to achieve the irreducible MSE  $\sigma^2$  without using any of the features in  $A$ .

Interestingly, this score accounts for both the relevancy of  $\mathbf{X}_A$  in the prediction of  $Y$  and its redundancy with respect to the other features  $\mathbf{X}_{\bar{A}}$ . To better verify this fact, we can rewrite  $\nu(A)$  as:

$$\int p(\mathbf{x}_{\bar{A}}) \int p(y, \mathbf{x}_A | \mathbf{x}_{\bar{A}}) \log \frac{p(y, \mathbf{x}_A | \mathbf{x}_{\bar{A}})}{p(y|\mathbf{x}_{\bar{A}})p(\mathbf{x}_A|\mathbf{x}_{\bar{A}})} d\mathbf{x}_A dy, \quad (5)$$

and notice that the inner integral is zero whenever: i)  $\mathbf{x}_{\bar{A}}$  perfectly predicts  $y$ , i.e.,  $\mathbf{x}_A$  is *irrelevant*, or ii)  $\mathbf{x}_{\bar{A}}$  perfectly predicts  $\mathbf{x}_A$ , i.e.,  $\mathbf{x}_A$  is *redundant*. In both cases we have  $p(y, \mathbf{x}_A | \mathbf{x}_{\bar{A}}) = p(y|\mathbf{x}_{\bar{A}})p(\mathbf{x}_A|\mathbf{x}_{\bar{A}})$  and, thus,  $\nu(A) = 0$ .

#### B. Regression Error in Linear Models

As previously mentioned the *actual* error introduced by removing a set of features depends on the choice of the model class. We remark that Theorem 1 bounds the *ideal* prediction error, i.e., the error achieved by a model of infinite capacity. Unfortunately, in practical applications the chosen model has often very limited capacity (e.g., linear). In such cases, our bound, and all CMI-based methods, might be over-optimistic. Indeed, there are situations in which CMI leads to discarding an apparently redundant feature which would reveal itself to be useful when considering the finite capacity of the chosen model. Let us consider the following example.

**Example 1.** *Consider a regression problem with two features,  $X_1$  and  $X_2$ , and target  $Y = aX_1 + bX_2$ , for two scalars  $a$  and  $b$ . Furthermore, assume that  $X_1 = Z$  and  $X_2 = e^Z$ , for  $Z \sim \mathcal{N}(0, \sigma^2)$ , with  $\sigma^2 \gg 0$ . It is clear that  $\nu(\{x_1\}) \simeq 0$  and  $\nu(\{x_2\}) \simeq 0$ , since the two features can be perfectly recovered from one another. However, if the chosen model is linear, both features are fundamental for predicting  $Y$ . In fact, the squared Pearson correlation coefficients  $\rho^2(X_1, Y)$  and  $\rho^2(X_2, Y)$  are high, while  $\rho^2(X_1, X_2)$  is small.*

We show now that, when linear models are involved, the correlation between the features and between a feature and the target can be used to bound the regression error.

**Theorem 2.** *Let  $\sigma_{\mathbf{X} \rightarrow Y}^2 = \min_{\mathbf{w}, b} \mathbb{E}_{\mathbf{X}, Y} \left[ (Y - \mathbf{w}^T \mathbf{X} - b)^2 \right]$  be the minimum MSE of the linear model that predicts  $Y$  with all the features and  $(\mathbf{w}^*, b^*)$  be the optimal weights and bias. Let  $A$  be a set of indices and  $\sigma_{\mathbf{X}_{\bar{A}} \rightarrow X_i}^2 = \min_{\mathbf{w}_{i, \bar{A}}, b_{i, \bar{A}}} \mathbb{E}_{X_i, \mathbf{X}_{\bar{A}}} \left[ (X_i - \mathbf{w}_{i, \bar{A}}^T \mathbf{X}_{\bar{A}} - b_{i, \bar{A}})^2 \right]$  be the minimum MSE of the linear model that predicts  $X_i$  from the features  $\mathbf{X}_{\bar{A}}$ . Then the minimum MSE of the linear model that predicts  $Y$  from the features  $\mathbf{X}_{\bar{A}}$  can be bounded as:*

$$\begin{aligned} \min_{\mathbf{w}_{\bar{A}}, b_{\bar{A}}} \mathbb{E}_{\mathbf{X}, Y} \left[ (Y - \mathbf{w}_{\bar{A}}^T \mathbf{X}_{\bar{A}} - b_{\bar{A}})^2 \right]^{\frac{1}{2}} \\ \leq \sigma_{\mathbf{X} \rightarrow Y} + \sqrt{|A|} \sum_{i \in A} w_i^* \sigma_{\mathbf{X}_{\bar{A}} \rightarrow X_i}. \end{aligned}$$

Furthermore, let  $\sigma_Y^2 = \text{Var}[Y]$ . If  $\rho(X_i, X_j) = 0$  for all  $i, j \in A$  and  $i \neq j$  and  $\rho(X_i, X_j) = 0$  for all  $i, j \in \bar{A}$  and  $i \neq j$ ,<sup>1</sup> then it holds that:

$$\begin{aligned} \min_{\mathbf{w}_{\bar{A}}, b_{\bar{A}}} \mathbb{E}_{\mathbf{X}, Y} \left[ (Y - \mathbf{w}_{\bar{A}}^T \mathbf{X}_{\bar{A}} - b_{\bar{A}})^2 \right]^{\frac{1}{2}} &\leq \sigma_{\mathbf{X} \rightarrow Y} \\ &+ \sqrt{|A|} \sigma_Y \sum_{i \in A} \rho(Y, X_i) \left( 1 - \sum_{j \in \bar{A}} \rho(X_i, X_j)^2 \right)^{\frac{1}{2}}. \end{aligned}$$

<sup>1</sup>We are requiring that all features in  $\mathbf{X}_A$  are uncorrelated and that all features in  $\mathbf{X}_{\bar{A}}$  are uncorrelated; but, of course, there might exist  $i \in A$  and  $j \in \bar{A}$  such that  $\rho(X_i, X_j) \neq 0$ .

*Proof.* Consider the linear regression problem for predicting  $Y$  with all the features,  $\min_{\mathbf{w}, b} \mathbb{E}_{Y, \mathbf{X}} \left[ (Y - \mathbf{w}^T \mathbf{X} - b)^2 \right]$ , having  $(\mathbf{w}^*, b^*)$  as the optimal solution. The expression of the optimal weights and the minimum MSE  $\sigma^2$  are given by:

$$\mathbf{w}^* = \text{Cov}[\mathbf{X}]^{-1} \text{Cov}[\mathbf{X}, Y],$$

$$\sigma_{\mathbf{X} \rightarrow Y}^2 = \text{Var}[Y] - \text{Cov}[Y, \mathbf{X}] \text{Cov}[\mathbf{X}]^{-1} \text{Cov}[\mathbf{X}, Y].$$

Consider now a partition of  $\mathbf{X}$  into  $\mathbf{X}_{\bar{A}}$  and  $\mathbf{X}_A$  and the linear regression problem to predict  $\mathbf{X}_A$  from  $\mathbf{X}_{\bar{A}}$ , i.e.,  $\min_{\mathbf{W}_{A, \bar{A}}, b_{A, \bar{A}}} \mathbb{E}_{\mathbf{X}_A, \mathbf{X}_{\bar{A}}} \left[ (\mathbf{X}_A - \mathbf{W}_{A, \bar{A}} \mathbf{X}_{\bar{A}} - b_{A, \bar{A}})^2 \right]$ . We can express the optimal weights and the minimum MSE as:

$$\mathbf{W}_{A, \bar{A}}^* = \text{Cov}[\mathbf{X}_A, \mathbf{X}_{\bar{A}}] \text{Cov}[\mathbf{X}_{\bar{A}}]^{-1}.$$

$$\sigma_{\mathbf{X}_{\bar{A}} \rightarrow \mathbf{X}_A}^2 = \text{Var}[\mathbf{X}_A] - \text{Cov}[\mathbf{X}_A, \mathbf{X}_{\bar{A}}] \text{Cov}[\mathbf{X}_{\bar{A}}]^{-1} \text{Cov}[\mathbf{X}_{\bar{A}}, \mathbf{X}_A]$$

Let us now consider the linear regression problem for predicting  $Y$  from the features  $\mathbf{X}_{\bar{A}}$ .

$$\begin{aligned} \min_{\mathbf{w}_{\bar{A}}, b_{\bar{A}}} \mathbb{E}_{\mathbf{X}, Y} \left[ (Y - \mathbf{w}_{\bar{A}}^T \mathbf{X}_{\bar{A}} - b_{\bar{A}})^2 \right] \\ \leq \mathbb{E}_{\mathbf{X}, Y} \left[ \left( Y - \mathbf{w}_{\bar{A}}^{*T} \mathbf{X}_{\bar{A}} - b^* \pm \mathbf{w}_{\bar{A}}^{*T} (\mathbf{W}_{A, \bar{A}}^* \mathbf{X}_{\bar{A}} - b_{A, \bar{A}}^*) \right)^2 \right]^{\frac{1}{2}} \\ \leq \mathbb{E}_{\mathbf{X}, Y} \left[ \left( Y - \mathbf{w}_{\bar{A}}^{*T} \mathbf{X}_{\bar{A}} - \mathbf{w}_A^{*T} \mathbf{X}_A - b^* \right)^2 \right]^{\frac{1}{2}} \\ + \mathbb{E}_{\mathbf{X}, Y} \left[ \left( \mathbf{w}_A^{*T} (\mathbf{X}_A - \mathbf{W}_{A, \bar{A}}^* \mathbf{X}_{\bar{A}} - b_{A, \bar{A}}^*) \right)^2 \right]^{\frac{1}{2}} \end{aligned} \quad (6)$$

$$\leq \sigma_{\mathbf{X} \rightarrow Y} + \sqrt{|\bar{A}|} \mathbb{E}_{\mathbf{X}} \left[ \sum_{i \in \bar{A}} w_i^{*2} (X_i - \mathbf{w}_{i, \bar{A}}^{*T} \mathbf{X}_{\bar{A}} - b_{i, \bar{A}}^*)^2 \right]^{\frac{1}{2}} \quad (7)$$

$$\begin{aligned} &= \sigma_{\mathbf{X} \rightarrow Y} + \sqrt{|\bar{A}|} \left( \sum_{i \in \bar{A}} w_i^{*2} \mathbb{E}_{\mathbf{X}} \left[ (X_i - \mathbf{w}_{i, \bar{A}}^{*T} \mathbf{X}_{\bar{A}} - b_{i, \bar{A}}^*)^2 \right] \right)^{\frac{1}{2}} \\ &\leq \sigma_{\mathbf{X} \rightarrow Y} + \sqrt{|\bar{A}|} \sum_{i \in \bar{A}} w_i^* \mathbb{E}_{\mathbf{X}} \left[ (X_i - \mathbf{w}_{i, \bar{A}}^{*T} \mathbf{X}_{\bar{A}} - b_{i, \bar{A}}^*)^2 \right]^{\frac{1}{2}} \end{aligned} \quad (8)$$

$$= \sigma_{\mathbf{X} \rightarrow Y} + \sqrt{|\bar{A}|} \sum_{i \in \bar{A}} w_i^* \sigma_{\mathbf{X}_{\bar{A}} \rightarrow X_i},$$

where (6) derives from Minkowski inequality after having summed and subtracted  $\mathbf{w}_A^{*T} \mathbf{X}_A$ , (7) is obtained from Cauchy-Schwarz inequality (for  $d$  dimensional vectors we have  $(\mathbf{a}^T \mathbf{b})^2 \leq d \sum_{i=1}^d a_i^2 b_i^2$ ) and (8) derives from subadditivity of the square root.

By recalling that  $w_i^* = \sum_{j \in \bar{A}} \text{Cov}[X_i, X_j]^{-1} \text{Cov}[X_j, Y]$ , for uncorrelated features we get:

$$w_i^* = \text{Var}[X_i]^{-1} \text{Cov}[X_i, Y] = \left( \frac{\text{Var}[Y]}{\text{Var}[X_i]} \right)^{\frac{1}{2}} \rho(Y, X_i).$$

If the features in  $\mathbf{X}_{\bar{A}}$  are uncorrelated as well, we have that  $\text{Cov}[\mathbf{X}_{\bar{A}}]$  is diagonal. Therefore, we have:

$$\begin{aligned} \sigma_{\mathbf{X}_{\bar{A}} \rightarrow \mathbf{X}_A}^2 &= \text{Var}[\mathbf{X}_A] - \sum_{j \in \bar{A}} \text{Cov}[X_i, X_j]^2 \text{Var}[X_j]^{-1} \\ &= \text{Var}[X_i] \left( 1 - \sum_{j \in \bar{A}} \rho(X_i, X_j)^2 \right), \end{aligned}$$

from which the result follows directly.  $\square$

This result allows highlighting two relevant points. First, when considering linear models what matters is the correlation among the features and the correlation between the features and the class. Most importantly, the Pearson correlation coefficient is a weaker index of dependency between random

variables compared to the MI as it identifies linear dependency only. As suggested by Example 1, using MI for discarding features when the model used for prediction is too weak might be dangerous. Second, Theorem 2 highlights once again two relevant properties of the features. In the linear case, a feature  $X_i$  is relevant if it is highly correlated with the target  $Y$ , i.e.,  $\rho^2(Y, X_i) \gg 0$ , and a feature is redundant if it is highly correlated with the others, i.e.,  $\rho^2(X_i, X_j) \gg 0$ . Both these contributions appear clearly in Theorem 2.

### C. Bounding the Classification Error

A similar result to Theorem 1 can be obtained for an ideal classification problem. Here the goal is to find the function  $g^* \in \mathcal{G}_{\bar{A}}$  minimizing the ideal prediction loss,

$$\inf_{g \in \mathcal{G}_{\bar{A}}} \mathbb{E}_{\mathbf{X}, Y} [\mathbb{1}_{\{Y \neq g(\mathbf{X}_{\bar{A}})\}}], \quad (9)$$

where  $\mathbb{1}_E$  denotes the indicator function over an event  $E$ .

**Theorem 3.** Let  $\epsilon = \mathbb{E}_{\mathbf{X}, Y} [\mathbb{1}_{\{Y \neq \arg \max_{y \in \mathcal{Y}} p(y|\mathbf{X})\}}]$  be the Bayes error and  $A$  be a set of indices, then the classification error obtained by removing features  $\mathbf{X}_A$  can be bounded as:

$$\inf_{g \in \mathcal{G}_{\bar{A}}} \mathbb{E}_{\mathbf{X}, Y} [\mathbb{1}_{\{Y \neq g(\mathbf{X}_{\bar{A}})\}}] \leq \epsilon + \sqrt{2\nu(A)}. \quad (10)$$

*Proof.* Let us denote by  $y^* = \arg \max_{y \in \mathcal{Y}} p(y|\mathbf{x})$  the optimal prediction given  $\mathbf{x}$  and by  $y_{\bar{A}}^* = \arg \max_{y \in \mathcal{Y}} p(y|\mathbf{x}_{\bar{A}})$  the optimal prediction given the subset of features in  $\bar{A}$ . We have:

$$\begin{aligned} \inf_{g \in \mathcal{G}_{\bar{A}}} \mathbb{E}_{\mathbf{X}, Y} [\mathbb{1}_{\{Y \neq g(\mathbf{X}_{\bar{A}})\}}] &= \mathbb{E}_{\mathbf{X}, Y} [\mathbb{1}_{\{Y \neq \arg \max_{y \in \mathcal{Y}} p(y|\mathbf{X}_{\bar{A}})\}}] \\ &= \mathbb{E}_{\mathbf{X}, Y} [\mathbb{1}_{\{Y \neq \arg \max_{y \in \mathcal{Y}} p(y|\mathbf{X}_{\bar{A}})\}} \pm \mathbb{1}_{\{Y \neq \arg \max_{y \in \mathcal{Y}} p(y|\mathbf{X})\}}] \\ &= \epsilon + \int p(\mathbf{x}) (p(y^*|\mathbf{x}) - p(y_{\bar{A}}^*|\mathbf{x})) d\mathbf{x} \\ &= \epsilon + \int p(\mathbf{x}) (p(y^*|\mathbf{x}) \pm p(y_{\bar{A}}^*|\mathbf{x}_{\bar{A}}) - p(y_{\bar{A}}^*|\mathbf{x})) d\mathbf{x}. \end{aligned}$$

Let us now bound the term inside the expectation point-wisely. For the term  $p(y^*|\mathbf{x}) - p(y_{\bar{A}}^*|\mathbf{x}_{\bar{A}})$ , we have:

$$\begin{aligned} p(y^*|\mathbf{x}) - p(y_{\bar{A}}^*|\mathbf{x}_{\bar{A}}) &= \max_{y \in \mathcal{Y}} p(y|\mathbf{x}) - \max_{y \in \mathcal{Y}} p(y|\mathbf{x}_{\bar{A}}) \\ &\leq \max_{y \in \mathcal{Y}} |p(y|\mathbf{x}) - p(y|\mathbf{x}_{\bar{A}})| \\ &\leq D_{\text{TV}}(p(\cdot|\mathbf{x}) \| p(\cdot|\mathbf{x}_{\bar{A}})). \end{aligned}$$

Following a similar argument for the term  $p(y_{\bar{A}}^*|\mathbf{x}_{\bar{A}}) - p(y_{\bar{A}}^*|\mathbf{x})$ , we find that the inner term is always less or equal than the total variation distance between  $p(\cdot|\mathbf{x})$  and  $p(\cdot|\mathbf{x}_{\bar{A}})$ . Then, by applying Pinsker's inequality:

$$\begin{aligned} \inf_{g \in \mathcal{G}_{\bar{A}}} \mathbb{E}_{\mathbf{X}, Y} [\mathbb{1}_{\{Y \neq g(\mathbf{X}_{\bar{A}})\}}] &\leq \epsilon + 2 \int p(\mathbf{x}) D_{\text{TV}}(p(\cdot|\mathbf{x}) \| p(\cdot|\mathbf{x}_{\bar{A}})) d\mathbf{x} \\ &\leq \epsilon + \int p(\mathbf{x}) \sqrt{2D_{\text{KL}}(p(\cdot|\mathbf{x}) \| p(\cdot|\mathbf{x}_{\bar{A}}))} d\mathbf{x} \\ &\leq \epsilon + \sqrt{2 \int p(\mathbf{x}) D_{\text{KL}}(p(\cdot|\mathbf{x}) \| p(\cdot|\mathbf{x}_{\bar{A}})) d\mathbf{x}} \\ &= \epsilon + \sqrt{2\nu(A)}. \end{aligned}$$

Here the last inequality follows from Jensen's inequality and the concavity of the square root.  $\square$

Similarly to the result for regression problems, Theorem 3 bounds the minimum ideal classification error achievable by a model which uses only the subset of features in  $\bar{A}$  by the score  $\nu(A)$ . The astute reader might have noticed a slightly better dependence on  $\nu(A)$  with respect to the

regression case (square root versus linear). This is due to the fact that minimizing the MSE gives rise to a squared total variation distance between the conditional distributions  $p(\cdot|\mathbf{x})$  and  $p(\cdot|\mathbf{x}_{\bar{A}})$ , which leads to a linear dependence on  $\nu(A)$ .

#### IV. ALGORITHMS

In this section, we rephrase the forward and backward feature selection algorithms based on the findings of Section III. Furthermore, we propose a novel stopping condition that allows to bound the error introduced by removing a set of features, assuming the predictor will make the best possible use of the remaining features. Actively searching for the optimal subset of features is combinatorial in the number of features and, thus, unfeasible [20]. Instead, we can start from the complete feature set and remove one feature at a time, greedily minimizing the score. In this spirit, we propose the following iterative procedure.

**Algorithm 1** (Backward Elimination). *Given a dataset  $\mathbf{X}, Y$ , select a threshold  $\delta \geq 0$ , the maximum error that the filter is allowed to introduce. Then:*

- Start with the full feature set, i.e.,  $A_1 = \emptyset$ , where  $A_t$  denotes the index set of features removed prior to step  $t$ .
- For each step  $t = 1, 2, \dots$ , remove the feature that minimizes the conditional mutual information between itself and the target  $Y$  given the remaining features, i.e.:  

$$i_t = \arg \min_i I(Y; X_i | \mathbf{X}_{\bar{A}_t} \setminus X_i), \quad (11)$$

$$I_t = I(Y; X_{i_t} | \mathbf{X}_{\bar{A}_t} \setminus X_{i_t}), \quad (12)$$

$$A_{t+1} = A_t \cup \{i_t\} \quad (13)$$
- Stop as soon as  $\sum_{h=1}^t I_h \geq \frac{\delta}{2B^2}$  for regression and  $\sum_{h=1}^t I_h \geq \frac{\delta^2}{2}$  for classification. The selected features are the remaining ones, indexed by  $\bar{A}_T$ , where  $T$  is the last step.

This algorithm, apart from the stopping condition, is described by Brown et al. [15] as *Backward Elimination with Mutual Information*. The same authors show that this procedure greedily maximizes the conditional likelihood of the selected features given the target, as long as  $I_k$  is always zero. This would correspond to selecting  $\delta = 0$  as a threshold in our algorithm. The same backward elimination step is used as a subroutine in the IAMB algorithm [16]. Our stopping condition allows selecting the maximum error that the feature selection procedure is allowed to add to the ideal error, i.e., the unavoidable error that even a perfect predictor using all the features would commit. The fact that the threshold will be actually observed is guaranteed by the following result.

**Theorem 4.** *Algorithm 1 achieves an error of  $\sigma^2 + \delta$  for regression, where  $\sigma^2$  is the irreducible error and  $\epsilon + \delta$  for classification, where  $\epsilon$  is the Bayes error.*

*Proof.* We prove the result for regression using Theorem 1. The proof for classification is analogous, but based on Theorem 3. We have:

$$\inf_{g \in \mathcal{G}_{\bar{A}}} \mathbb{E}_{\mathbf{X}, Y} \left[ (Y - g(\mathbf{X}_{\bar{A}_t}))^2 \right] \leq \sigma^2 + 2B^2 \nu(A_t), \quad (14)$$

where  $t$  is any iteration of the algorithm. By repeatedly applying the chain rule of CMI (2), we can rewrite the score as:

$$\begin{aligned} \nu(A_{t+1}) &= I(Y; \mathbf{X}_{A_{t+1}} | \mathbf{X}_{\bar{A}_{t+1}}) \\ &= I(Y; \mathbf{X}) - I(Y; \mathbf{X}_{\bar{A}_{t+1}}) \\ &= I(Y; \mathbf{X}_{A_t}, \mathbf{X}_{\bar{A}_t}) - I(Y; \mathbf{X}_{\bar{A}_{t+1}}) \\ &= I(Y; \mathbf{X}_{A_t} | \mathbf{X}_{\bar{A}_t}) + I(Y; \mathbf{X}_{\bar{A}_t}) - I(Y; \mathbf{X}_{\bar{A}_{t+1}}) \\ &= \nu(A_t) + I(Y; X_{i_t}, \mathbf{X}_{\bar{A}_{t+1}}) - I(Y; \mathbf{X}_{\bar{A}_{t+1}}) \\ &= \nu(A_t) + I(Y; X_{i_t} | \mathbf{X}_{\bar{A}_{t+1}}) \pm I(Y; \mathbf{X}_{\bar{A}_{t+1}}) \\ &= \nu(A_t) + I(Y; X_{i_t} | \mathbf{X}_{\bar{A}_t} \setminus X_{i_t}) \\ &= \nu(A_t) + I_t. \end{aligned} \quad (15)$$

Noting that  $\nu(A_1) = I(Y; \emptyset | \mathbf{X}) = 0$ , we can unroll this recursive equation, obtaining:

$$\nu(A_T) = \sum_{t=1}^{T-1} I_t \leq \frac{\delta}{2B^2}, \quad (16)$$

where the inequality is due to the stopping condition. Plugging (16) into (14), we get the thesis.  $\square$

Our Theorems 1 and 3 suggest that a backward elimination procedure allows keeping the error controlled. In the following, we argue that we can resort also to forward selection methods and still have a guarantee on the error. Using the chain rule of the CMI we can express our score  $\nu(A)$  as:

$$\nu(A) = I(Y; \mathbf{X}) - I(Y; \mathbf{X}_{\bar{A}}),$$

where  $\mathbf{X}_{\bar{A}}$  is the set of features that have not been eliminated yet. If we plug this equation into the bounds of Theorems 1 and 3 we get:

$$\begin{aligned} \inf_{g \in \mathcal{G}_{\bar{A}}} \mathbb{E}_{\mathbf{X}, Y} \left[ (Y - g(\mathbf{X}_{\bar{A}}))^2 \right] &\leq \sigma^2 + 2B^2 [I(Y; \mathbf{X}) - I(Y; \mathbf{X}_{\bar{A}})], \\ \inf_{g \in \mathcal{G}_{\bar{A}}} \mathbb{E}_{\mathbf{X}, Y} \left[ \mathbb{1}_{\{Y \neq g(\mathbf{X}_{\bar{A}})\}} \right] &\leq \epsilon + \sqrt{2 [I(Y; \mathbf{X}) - I(Y; \mathbf{X}_{\bar{A}})]}, \end{aligned}$$

for the regression and classification cases respectively. Since  $I(Y; \mathbf{X})$  does not depend on the selected features  $\mathbf{X}_{\bar{A}}$ , in order to minimize the bound we need to maximize the term  $I(Y; \mathbf{X}_{\bar{A}})$ . This matches the intuition that we should select the features that provide the maximum information on the class. Using this result, we can easily provide a forward feature selection algorithm.

**Algorithm 2** (Forward Selection). *Given a dataset  $\mathbf{X}, Y$ , select a threshold  $\delta \geq 0$ , the maximum error that the filter is allowed to introduce. Then:*

- Start with the empty feature set, i.e.,  $A_1 = \emptyset$ , where  $A_t$  denotes the index set of features selected prior to step  $t$ .
- For each step  $t = 1, 2, \dots$ , add the feature that maximizes the conditional mutual information between itself and the target  $Y$  given the remaining features, i.e.:  

$$i_t = \arg \max_i I(Y; X_i | \mathbf{X}_{A_t}), \quad (17)$$

$$I_t = I(Y; X_{i_t} | \mathbf{X}_{A_t}), \quad (18)$$

$$A_{t+1} = A_t \cup \{i_t\} \quad (19)$$
- Stop as soon as  $\sum_{h=1}^t I_h \geq \frac{\delta}{2B^2}$  for regression and  $\sum_{h=1}^t I_h \geq \frac{\delta^2}{2}$  for classification. The selected features are those indexed by  $A_T$ , where  $T$  is the last step.

Apart from the stopping condition, this algorithm was also presented in Brown et al. [15] and named *Forward Selection with Mutual Information*. Like for the backward case, we are able to provide a guarantee on the final error.

**Theorem 5.** *Algorithm 2 achieves an error of  $\sigma^2 - \delta + 2B^2I(Y; \mathbf{X})$  for regression, where  $\sigma^2$  is the irreducible error and  $\epsilon - \delta + \sqrt{2I(Y; \mathbf{X})}$  for classification, where  $\epsilon$  is the Bayes error.*

*Proof.* We prove the result just for the regression case, as the derivation for classification is analogous. Using the chain rule (2), we have the following recursion:

$$\begin{aligned} I(Y; \mathbf{X}_{A_{t+1}}) &= I(Y; \mathbf{X}_{A_t}, X_{i_t}) \\ &= I(Y; X_{i_t} | \mathbf{X}_{A_t}) + I(Y; \mathbf{X}_{A_t}) \\ &= I_t + I(Y; \mathbf{X}_{A_t}). \end{aligned}$$

By observing that  $I(Y; \mathbf{X}_{A_1}) = I(Y; \emptyset) = 0$ , we unroll the recursion and we get

$$I(Y; \mathbf{X}_{A_T}) = \sum_{t=1}^T I_t \geq \frac{\delta}{2B^2},$$

from which the result follows.  $\square$

#### A. Estimation of the Conditional Mutual Information

So far, we have assumed to be able to compute the CMI  $I(Y; X_i | \mathbf{X}_{\bar{A}_t} \setminus X_i)$  and  $I(Y; X_i | \mathbf{X}_{A_t})$  exactly. In practice, they need to be estimated from data. Estimating the MI can be reduced to the estimation of several entropies [21]; numerous methods have been employed in feature selection, either based on nearest neighbors approaches [22] or on histograms [15]. The main challenge arise in classification where we need to estimate CMI between a discrete variable (the class) and possibly continuous features. For this reason, we resort to the recent nearest neighbor estimator proposed by [23], which collapses to the more traditional KSG estimator [24] when both  $X$  and  $Y$  have a continuous density. These estimators are proved to be consistent when the number of samples and the number of neighbors grows to infinity [23].

#### V. RELATED WORKS

A related theoretical study of feature selection via MI has been recently proposed by Brown et al. [15]. The authors show that the problem of finding the minimal feature subset such that the conditional likelihood of the targets is maximized is equivalent to minimizing the CMI. Based on this result, common heuristics for information-theoretic feature selection can be seen as iteratively maximizing the conditional likelihood. Similarly, we show a connection between the CMI and the optimal prediction error. Differently from [15], we additionally propose a novel stopping condition that is well motivated by our theoretical findings.

In the information theory literature, [25] also analyzes the connection between CMI and minimum mean square error, deriving a similar result to our Theorem 1. However, classification problems (i.e., minimum zero-one loss) are not considered and the focus is not on feature selection.

The authors of [22] propose a nearest neighbor estimator for the CMI and show how it can be used in a classic forward

feature selection algorithm. One of the authors' questions is how to devise a suitable stopping condition for such methods. Here we propose a possible answer: our stopping criterion (Section IV) is intuitive, applicable to both forward and backward algorithms, and theoretically well-grounded.

Several existing approaches use linear correlation measures to score the different features [26]–[30]. Such algorithms are mostly based on the heuristic intuition that a good feature should be highly correlated with the class and lowly correlated with the other features. Instead, we provide a more theoretical justification for this claim (Section III), showing a connection between these two properties and the minimum MSE.

#### VI. EXPERIMENTS

We evaluate the performance of our stopping condition on synthetic and real-world datasets, by comparing different stopping criteria, employing a backward feature selection approach:

- *error (ER)*: stop when the bound on the prediction error, as in Theorem 4, is greater than a fixed threshold  $\delta$ ;
- *feature score (FS)*: stop when a feature with a CMI score greater than a fixed threshold  $\delta$  is encountered;
- *delta feature score ( $\Delta FS$ )*: stop when the difference between the score of two consecutive features is greater than a threshold  $\delta$  (as in knee-elbow analysis);
- *number of features (#F)*: stop with exactly  $k$  features are selected.

For all the experiments, we use Python's scikit-learn implementation of SVM with default parameters (RBF kernel and  $C = 1$ ).

##### A. Synthetic Data

The synthetic data consist in several binary classification problems. Each dataset is composed of 500 samples. The datasets are generated, similarly to [31], as follows: fix the number of useful features  $k$  (i.e., the number of features that are actually needed to predict the class); given  $Y = 1$ ,  $X_1, \dots, X_k$  are  $\mathcal{N}(0, 1)$  conditioned on  $\sum_{i=1}^k X_i > 3(k-2)$ , while  $X_{k+1}, \dots, X_{15} \sim \mathcal{N}(0, 1)$ . The choice of  $k$  will be specified for each experiment.

*Stopping Condition Comparison.* The first experiment is meant to compare the stopping conditions presented above across datasets for classification with different a number of useful features. We generate 6 independent problems with 30 features. Among the 30 features only  $k \in \{9, 12, 15, 18, 21, 24\}$  are useful to predict the target. In Figure 1, we show the accuracy of SVM for the different datasets and different stopping conditions. We can see that our stopping condition (ER) performs better than choosing a fixed number of features (#F) in most cases. More notably, the feature selection algorithm shows a greater robustness w.r.t the stopping condition's hyper parameter with our error-based criterion, as one would expect. Furthermore, in Figure 1, we notice that the delta feature score ( $\Delta FS$ ) (which is similar to the knee-elbow analysis) is highly inefficient (as the outputs are almost identical for both choices of the threshold) and is clearly the

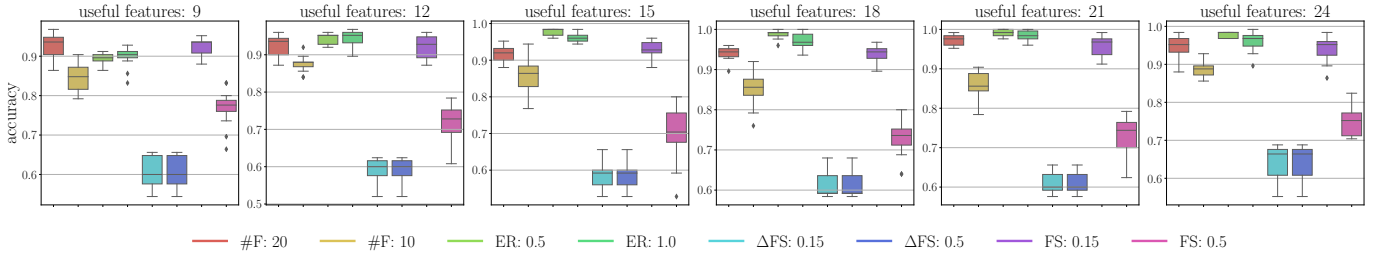


Fig. 1. SVM test accuracy for different choices of the number of features that generate the problem and different stopping criteria.

worst performer. The feature score (FS) stopping criterion is highly sensitive to the threshold, achieving good performance with a low threshold and a significantly worse performance when the threshold is increased. The choice of the threshold in both  $\Delta FS$  and FS poses a significant problem as it has no relation to the prediction error and its optimal value is highly problem-specific. On the contrary, for  $\#F$  and ER criteria the hyper parameter has a precise meaning and thus it can be selected more easily.

**Robustness.** To have a better grasp of the proposed stopping criterion we generate 50 binary classification problems, with the only difference of choosing  $k$  accordingly to  $k \sim \text{Uniform}(3, 15)$  and having only 15 total features. In Figure 2 we show the accuracy of a SVM classifier on a test set, after the feature selection has been performed, as a function of the error threshold  $\delta$ . Moreover, we overlay the fraction of selected features over the original 15. We can notice two interesting facts. i) Even with a threshold close to zero<sup>2</sup> a great number of features is discarded. ii) The classification accuracy is rather constant despite a high error threshold while the number of selected features decreases significantly. We can conclude that our method was effectively able to identify irrelevant features and discard them.

**CMI Estimation.** To see how the estimation of the (conditional) mutual information impacts on the performance of the stopping condition, we consider one last problem, generated as before with  $N = 30$  features and fixed  $k = 10$ . In Figure 3, we look at the performance of an SVM classifier on the same test set for increasing sizes (number of samples) of the training set. We select the number of neighbors in the mutual information estimation as a fixed fraction of the training set size. Notice how, when the data points are too few, the estimated mutual information “overfits”, and actually very little to no features are discarded in the feature selection step. As a consequence, also the SVM classifier overfits the training set and leads to poor performance on the test set. On the other hand, as the number of samples increases, the estimation of the mutual information becomes more precise and the appropriate set of features is selected, resulting in a great increase in the classification accuracy on the test set. Moreover, for a small number of data points, the number of neighbors used in the MI estimation is not too relevant, while it is evident that for a

<sup>2</sup>Since the CMI is estimated from data as well, we cannot set the threshold to exactly 0, thus, we used 0.05 in the experiments.

TABLE I  
REAL DATA RESULTS.

Dataset	$\delta = 0.05$	$\delta = 0.1$	$\delta = 0.25$	$\delta = 0.5$	$\delta = 1.0$
ORL	0.8	0.75	0.7	0.7375	0.7125
warpAR10P	0.97	0.98	0.98	0.98	0.98
glass*	0.99	0.99	0.99	0.99	0.99
wine	0.96	0.96	0.96	0.95	0.83
ALLAML	1.0	1.0	1.0	0.92	0.78

\*: no feature removed

large enough sample size, it is better to increase the number of neighbors.

## B. Real-World Data

We further tested the proposed feature selection algorithm on several popular real world datasets, publicly available on the ASU feature selection website and the UCI repository [32]. In Table I, we report the classification accuracy on a test set after the feature selection procedure for different values of the threshold  $\delta$ . We notice how the upper bound on the error is stricter in some examples and larger in others. In particular, the actual classification accuracy follows the theoretical error bound in cases where the dataset has a bigger number of samples and a number of features that is not too big, for example ORL. Conversely, if the number of features is too big in comparison to the number of samples, the error bound tends to be pessimistic and the actual accuracy is much bigger than the expected one (warpAR10P, ALLAML). Interestingly enough, the number of classes does not play a significant role.

## VII. DISCUSSION AND CONCLUSION

Conditional Mutual Information is an effective statistical tool to perform feature selection via filter methods. In this paper, we proposed a novel theoretical analysis showing that using CMI allows to control the ideal prediction error, assuming that the trained model has infinite capacity. This is a rather new insight, as filter methods are typically employed when no assumptions are made on the underlying trained model. We proved that, when using linear models, the correlation coefficient becomes a suitable criterion for ranking and selecting features. On the bases of our findings, we proposed a new stopping condition, that can be applied to both forward and backward feature selection, with theoretical guarantees on the prediction error. The experimental evaluation showed that, compared against classical filter methods and stopping



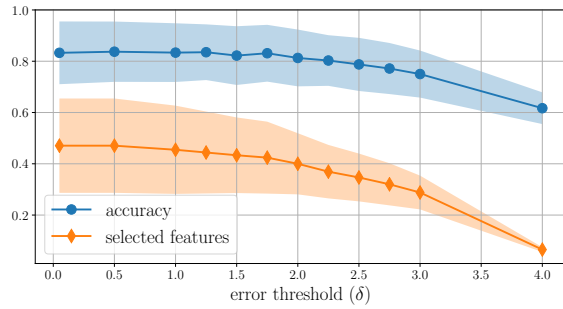


Fig. 2. Classification accuracy and fraction of selected features as a function of the error threshold  $\delta$ . Estimates are reported as mean values  $\pm$  standard deviation.

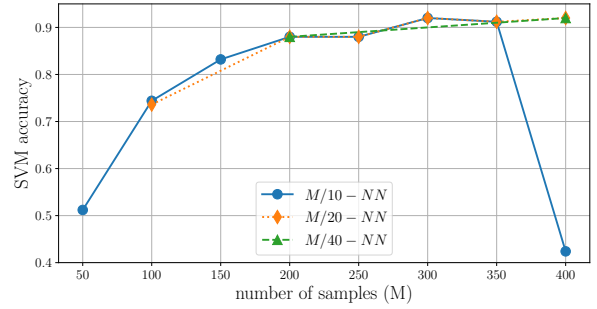


Fig. 3. Classification accuracy on a binary classification dataset, generated with 30 features and  $k = 10$ , for different values of number of samples and number of neighbors used for estimating the MI.

criteria, our approach, besides the theoretical foundation, is less sensitive to the choice of the threshold hyper-parameter and allows reaching state-of-the-art results.

## REFERENCES

- [1] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.
- [2] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.
- [3] T. N. Lal, O. Chapelle, J. Weston, and A. Elisseeff, "Embedded methods," in *Feature extraction*. Springer, 2006, pp. 137–165.
- [4] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik, "Feature selection for svms," in *Advances in neural information processing systems*, 2001, pp. 668–674.
- [5] W. Duch, T. Winiarski, J. Biesiada, and A. Kachel, "Feature selection and ranking filters," in *International conference on artificial neural networks (ICANN) and International conference on neural information processing (ICONIP)*, vol. 251. Citeseer, 2003, p. 254.
- [6] W. Duch, "Filter methods," in *Feature Extraction*. Springer, 2006, pp. 89–117.
- [7] J. R. Vergara and P. A. Estévez, "A review of feature selection methods based on mutual information," *Neural computing and applications*, vol. 24, no. 1, pp. 175–186, 2014.
- [8] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.
- [9] D. D. Lewis, "Feature selection and feature extraction for text categorization," in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 212–217.
- [10] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Transactions on neural networks*, vol. 5, no. 4, pp. 537–550, 1994.
- [11] H. H. Yang and J. Moody, "Data visualization and feature selection: New algorithms for nongaussian data," in *Advances in Neural Information Processing Systems*, 2000, pp. 687–693.
- [12] F. Fleuret, "Fast binary feature selection with conditional mutual information," *Journal of Machine Learning Research*, vol. 5, no. Nov, pp. 1531–1555, 2004.
- [13] D. Lin and X. Tang, "Conditional infomax learning: an integrated framework for feature extraction and fusion," in *European Conference on Computer Vision*. Springer, 2006, pp. 68–82.
- [14] g. Cheng, Z. Qin, C. Feng, Y. Wang, and F. Li, "Conditional mutual information-based feature selection analyzing for synergy and redundancy," *Etri Journal*, vol. 33, no. 2, pp. 210–218, 2011.
- [15] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján, "Conditional likelihood maximisation: a unifying framework for information theoretic feature selection," *Journal of machine learning research*, vol. 13, no. Jan, pp. 27–66, 2012.
- [16] I. Tsamardinos, C. F. Aliferis, A. R. Statnikov, and E. Statnikov, "Algorithms for large scale markov blanket discovery," in *FLAIRS conference*, vol. 2, 2003, pp. 376–380.
- [17] M. S. Pinsker, "Information and information stability of random variables and processes," 1960.
- [18] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observation," *studia scientiarum Mathematicarum Hungarica*, vol. 2, pp. 229–318, 1967.
- [19] S. Kullback, "A lower bound for discrimination information in terms of variation (corresp.)," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 126–127, 1967.
- [20] G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant features and the subset selection problem," in *Machine Learning Proceedings 1994*. Elsevier, 1994, pp. 121–129.
- [21] L. Paninski, "Estimation of entropy and mutual information," *Neural computation*, vol. 15, no. 6, pp. 1191–1253, 2003.
- [22] A. Tsimpiris, I. Vlachos, and D. Kugiumtzis, "Nearest neighbor estimate of conditional mutual information in feature selection," *Expert Systems with Applications*, vol. 39, no. 16, pp. 12 697–12 708, 2012.
- [23] W. Gao, S. Kannan, S. Oh, and P. Viswanath, "Estimating mutual information for discrete-continuous mixtures," in *Advances in Neural Information Processing Systems*, 2017, pp. 5986–5997.
- [24] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E*, vol. 69, p. 066138, Jun 2004. [Online]. Available: <https://link.aps.org/doi/10.1103/PhysRevE.69.066138>
- [25] Y. Wu and S. Verdú, "Functional properties of minimum mean-square error and mutual information," *IEEE Transactions on Information Theory*, vol. 58, no. 3, pp. 1289–1301, 2012.
- [26] S. K. Das, "Feature selection with a linear dependence measure," *IEEE transactions on Computers*, vol. 100, no. 9, pp. 1106–1109, 1971.
- [27] M. A. Hall, "Correlation-based feature selection for machine learning," 1999.
- [28] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *Proceedings of the 20th international conference on machine learning (ICML-03)*, 2003, pp. 856–863.
- [29] J. Biesiada and W. Duch, "Feature selection for high-dimensional data: a pearson redundancy based filter," in *Computer recognition systems 2*. Springer, 2007, pp. 242–249.
- [30] H. F. Eid, A. E. Hassanien, T.-h. Kim, and S. Banerjee, "Linear correlation-based feature selection for network intrusion detection model," in *Advances in Security of Information and Communication Networks*. Springer, 2013, pp. 240–248.
- [31] J. Chen, M. Stern, M. J. Wainwright, and M. I. Jordan, "Kernel feature selection via conditional covariance minimization," in *Advances in Neural Information Processing Systems*, 2017, pp. 6946–6955.
- [32] D. Dheeru and E. Karra Taniskidou, "UCI machine learning repository," 2017. [Online]. Available: <http://archive.ics.uci.edu/ml>